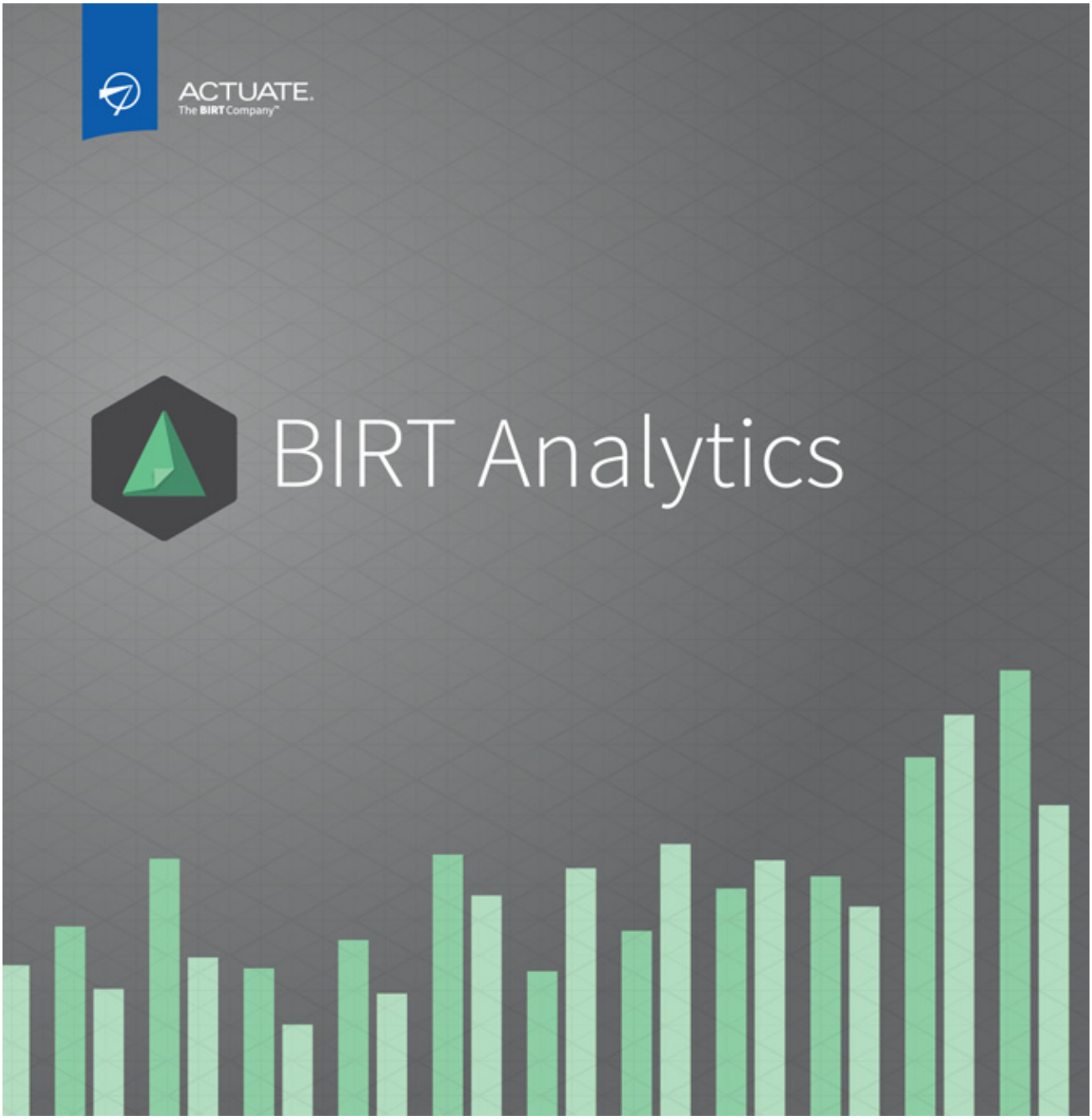


Using BIRT Analytics

Contains information proprietary to:
Actuate Corporation, 951 Mariners Island Boulevard, San Mateo, CA 94404

www.actuate.com

Document No. 150331-2-580331 April 20, 2015

Contents

About *Using BIRT Analytics*

BIRT Analytics is a web application, data repository, and set of web services that support analysis and forecasting of big data. In today's business climate, business users must be able to address the problems presented by big data, drawn from both structured and unstructured sources. BIRT Analytics supports selecting, grouping, analyzing, and presenting big data in a way that makes it actionable. BIRT Analytics enables a business user to process massive amounts of data, predict business outcomes, and make informed decisions. By making better decisions faster, business strategists can deliver vibrant and informative visual analysis of inherent trends in big data.

BIRT Analytics consists of three key components:

- The main user interface described in this manual, *Using BIRT Analytics.*
- BIRT Analytics Loader, a tool described in *Using BIRT Analytics Loader.* BIRT Analytics Loader extracts, transforms, and loads records from multiple external data sources to FastDB, the BIRT Analytics data repository.
- Admin, a tool-set described in *Administering BIRT Analytics*. Admin technology supports administering access and privileges for each module and component of BIRT Analytics.

Using BIRT Analytics describes how to use Actuate BIRT Analytics technology to carry out dynamic analyses. *Using BIRT Analytics* includes the following chapters:

- *About Using BIRT Analytics.* This chapter provides an overview of this guide.
- *Chapter 1. Understanding BIRT Analytics.* This chapter introduces Actuate BIRT Analytics and provides information about the application's home page.
- *Chapter 2. Understanding BIRT Analytics work areas.* This chapter describes the BIRT Analytics work areas: Data Explorer, Data Tree, and Scratchpad.
- *Chapter 3. Working with your data.* This chapter describes how to select your data for analysis using BIRT Analytics fundamental tools.
- *Chapter 4. Analyzing your data.* This chapter describes how to analyze data.
- *Chapter 5. Visualizing your data.*This chapter describes how to create appealing data analysis visualizations.
- *Chapter 6. Identifying and predicting data trends.* This chapter describes how to use BIRT Analytics to mine data.
- *Chapter 7. Managing campaigns.*This chapter describes how set up and run a business campaign using BIRT Analytics.
- *Chapter 8. Scheduling tasks.* This chapter describes how to automate tasks and events using BIRT Analytics.
- *Glossary.* This chapter provides definitions of terms used in the BIRT Analytics product and documentation.

Chapter

1

Understanding BIRT Analytics

This chapter contains:

- About BIRT Analytics main interface
- Understanding the home page
- Understanding the sample data model

About BIRT Analytics main interface

The BIRT Analytics main interface provides the following tools:

- Data exploration
- Purely analytical tools: crosstabs, Venn diagrams, bubble diagrams, evolutions, profiles, maps, and Pareto analyses
- Advanced analytical tools for data mining: clustering and forecasting
- Segmentation tool: selections
- Data engineering and enrichment tools: aggregates, decodes, expressions, numeric ranges, quantile ranges, parameters, and rankings
- Data export and import tools
- Configuration tools for user preferences: password, language, and theme

Understanding the home page

The home page appears in the browser when you open BIRT Analytics.

A "**Recent analysis**" panel, on the right-hand side of the page, lists recent analyses that have been accessed. A "garbage bin" icon at the bottom lets you clear this list.

A "**My folders"** panel to the left of the Recent analysis panel gives you access to your own saved analyses. Also, clicking on the "*Folder*" tab, in the upper left hand corner displays your folders in the "**Data Tree"** panel below it. (Go to the section "Understanding Data Tree" in Chapter 2, "Understanding BIRT Analytics work areas" for more information on the Data Tree.) Both the "Recent analysis"and the "My folders" panels are shown below in Figure 1-1.

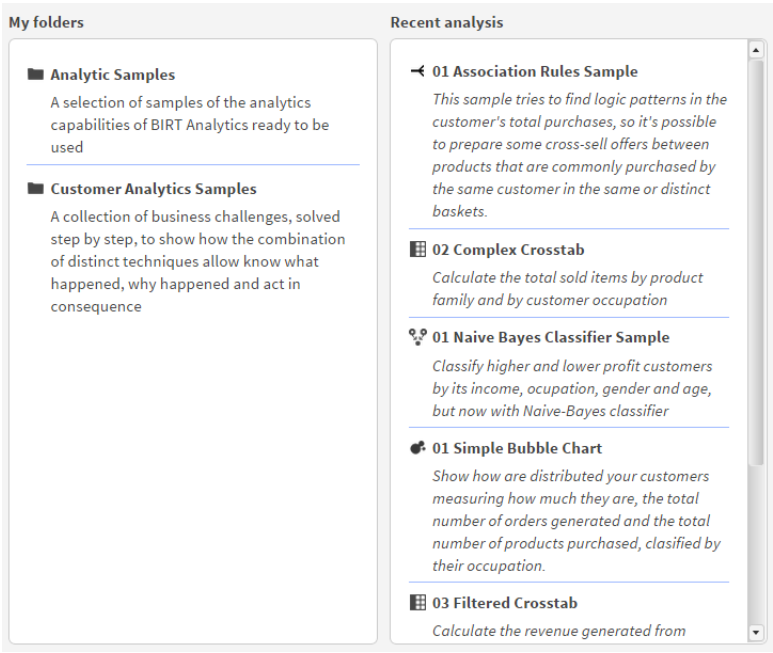


Figure 1-1 My folders and Recent analysis panels in the home page

This page provides access through icons to all of the features of the product: Table 1-1 lists the icons and describes the functionality provided.

Table 1-1 Icons for BIRT Analytics features

Icon	Label	Purpose of the BIRT Analytics feature
	Start	The initial view
	Explore	Viewing and interacting with data in a segment or table
	Enrichment	Creating more informative data by defining new fields in existing tables or using the data outside BIRT Analytics
	Enrichment—Engineering	Enriching analyses by creating new fields in existing tables
	Enrichment—Import-Export	Inputting and outputting data using various formats
	Analytics	Performing in-depth examinations of data and drawing conclusions from the patterns
	Analytics—Analysis	Examining data in depth using statistical tools
	Analytics—Advanced	Mining data to produce information from operations such as grouping and prediction
	Analytics—Selections	Segmenting the data by identifying groups of items that meet certain criteria
	Gallery	Displaying data in visual form, providing a quick way to see trends and patterns
	Campaign Workflow	Planning, configuring and managing campaigns.
	Tasks and Events	Scheduling tasks based on Time or on Events

Laying out the feature tabs

A tools menu provides customization options for each set of tabs in the user interface, as shown in Figure 1-2. These options support placing the tabs in different positions and displaying text, an icon, or both on each tab. To access these options, right-click a tool's tab.

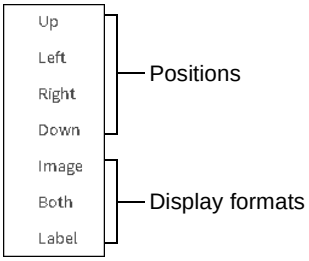


Figure 1-2 Options for tool tab placement and text

Setting preferences

Use Preferences to change the password, language, and theme for BIRT Analytics, as shown in Figure 1-3. To save one change and prepare to make further changes, choose Change. To discard unsaved changes, choose Cancel. To save changes and return to BIRT Analytics, choose Accept. Changes take effect after exiting BIRT Analytics and reentering.

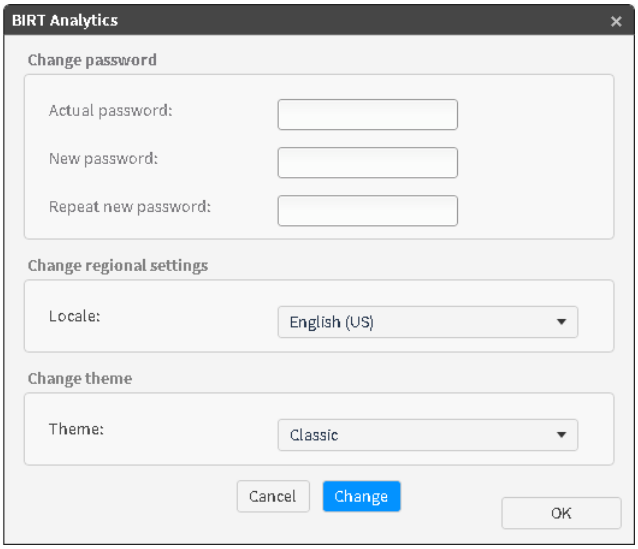


Figure 1-3 Available user preferences

Logging Out

The application does not log out automatically, even when there has been no use for a prolonged period of time. However, it does lock the screen after prolonged inactivity. Simply enter your password to unlock it.

You can lock your screen at any time via the User icon at the top right-hand side of the screen, as shown in Figure 1-4.

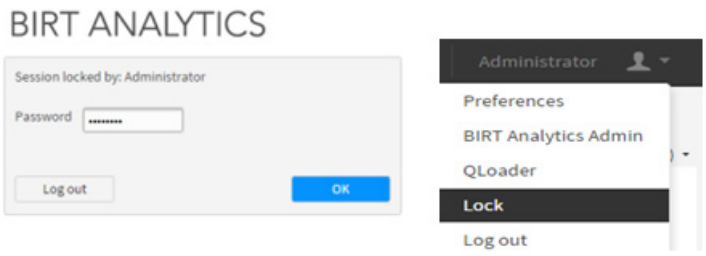


Figure 1-4 Using the Lock command

Changing a password

The new password must comply with the password policy associated with the user account and must contain at least one of each of the following character types:

- Lowercase letters
- Uppercase letters
- The following special characters _, -, !, ·, $, %, (,), =, |, @, #, €, *, ~, ",", ., " "
- Numeric characters

Changing regional settings for language/locale

You can change your language/locale settings by choosing the desired setting from the dropdown list accessed from the language field in the preferences window. (See Figure 1-5 below.)

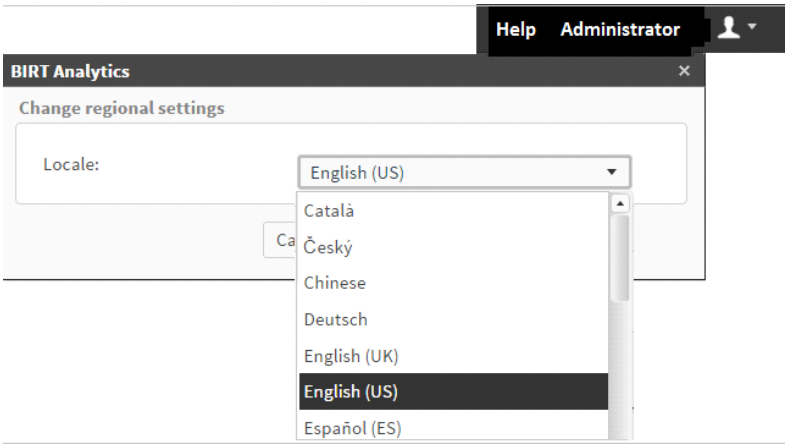


Figure 1-5 Changing regional settings

How changing regional settings affects date and date/time display

Setting your locale also sets your date and date/time values so that they will be displayed in your local format - appearing correctly in the Discrete Values grid, in Crosstab row dimensions, in Bubbles, Pareto and Profiles.

Subsequent editing of date and date/time values is done using a Calendar form.

Changing the theme

You can change the appearance of BIRT Analytics using predefined themes. BIRT Analytics provides two themes: Augusta (by default) and Classic (original BIRT Analytics theme, as shown in Figure 1-6.

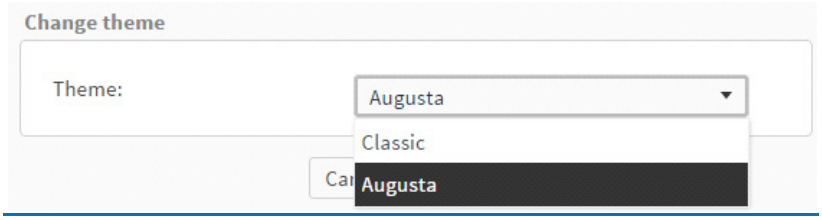


Figure 1-6 Changing the theme

Identifying hidden buttons and tabs

When you minimize the size of a window, an icon appears next to the Window's drop-down list on the top right of the screen. It gives access to a list of any buttons or tabs that are no longer visible on your screen. In Figure 1-7 below, clicking on the this icon in a minimized window in the Advanced Tab of the Analysis toolset shows that both the Logistic regression and the Linear regression tabs are hidden from view. This feature is available in all windows and tabs except Start and Explore.

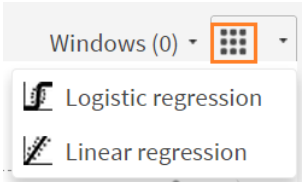


Figure 1-7 Identifying hidden tabs

Access to other resources

The "Resources" panel found at the bottom of the home page gives users access to the following online resources:

- BIRT Analytics website
- BIRT Analytics Documentation page
- Webcasts and Training Videos
- Actuate Analytics Blog
- Professional Services for BIRT Analytics

This panel also gives access to information concerning the installed version of the software.

Understanding the sample data model

Figure 1-8 shows an outline of the demo database analytical model. Not all fields from each table appear in this example.

- Household is the parent table.
- Customer is child table 1.
- Order is child table 2.
- Order Detail is child table 3.

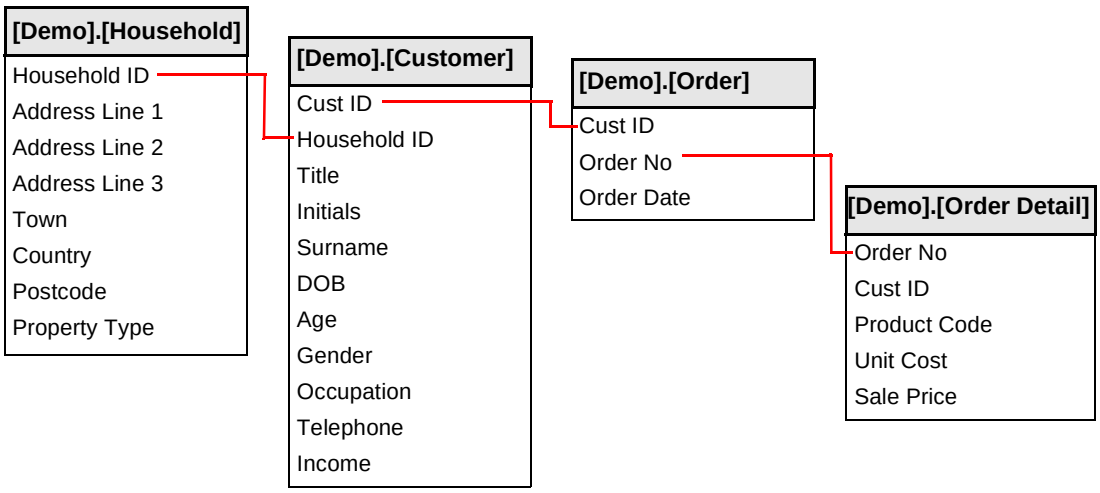


Figure 1-8 Demo database tables, fields, and associations

When the administrator loads the demo database into BIRT Analytics, the loader sets up the associations among the tables. Use these implied associations in BIRT Analytics to change the resolution on many types of analyses:

- A household can have from 1 to N associated customers.
- A customer can have from 1 to N associated orders. A customer must have an associated household.
- An order can have from 1 to N associated order detail lines. An order must have an associated customer.
- An order detail line must have an associated order.

Chapter

2

Understanding BIRT Analytics work areas

This chapter contains:

- About BIRT Analytics work areas
- Understanding Data Tree
- Understanding Scratchpad
- Understanding Data Explorer
- Understanding table resolution

About BIRT Analytics work areas

The BIRT Analytics user interface consists of three main areas:

- Data Tree
- Scratchpad
- Data Explorer

Understanding Data Tree

Data Tree provides the following three views to navigate through your data and saved analysis files: My Data, Discrete Values, and My Folders, as shown in Figure 2-1.

Figure 2-1 Data Tree showing the tables in the Demo database

My Data works with the Discrete Values viewer, which shows the content of explored fields, their values, and the number of stored entries with each value. Values are sometimes referred to as categories and entries as records.

Using My Data

My Data is available from Data Tree. Use My Data to display and navigate through databases, tables, and fields. To view the tables in a database, select the triangle icon beside the database name. To view the fields in a table, select the triangle next to the table, as shown in Figure 2-2. In this figure, the Demo database and the Customer table in the database are expanded and the fields in the Customer table are visible. Selecting the triangle next to an expanded database or table collapses the view of the items.

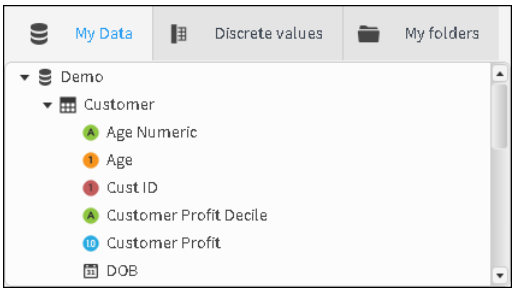


Figure 2-2 My Data showing fields in the Demo Customer table

Data items appear with a different identifying icon for each item type, as shown in Table 2-1. Unindexed fields appear without color but are otherwise the same as the associated physical field type.

Table 2-1 Icons for database, table, and field data types

Icon	Data structure or type
	Database
	Table
	Full numeric field
	Real numeric field
	Text field
	Date field
	Time field
	Date/time field
	Calculated field
	Unindexed field
	Unicode field
	Long integer field

In My Data, right-click an item (Database, Table or column) to enable the following field operation options:

- **Chart:** Display the Chart view of the item.
- **Delete:** Permanently remove the Database, table or column. Delete is available only for users who have the Delete privilege granted by an administrator.
- **Discrete values:** Display the discrete values for the item.
- **Edit:** Edit the segment in the Selections window that opens, making necessary changes or corrections.
- **Explore:** Display all relevant table columns of the selected database in the Data Explorer panel, filtering with the chosen segment.
- **Frequency:** Provides frequency information about the item
- **Make Permanent:** Write a calculated field as an entry in the database.
- **Rename:** Rename the item.
- **Selection:** Select a range of values stored in a column.
- **Statistics:** Provides statistical information about the item
- **Summary:** Displays a summary view of relevant item information
- **Index/Unindex:** Indexes or unindexes columns.
- **View definition:** View definition of a domain column

As you can see below, not all options are available to every Data Tree item:

- **Database options:** *Summary, Rename and Delete*
- **Table Options:** *Summary, Explore, Rename and Delete*
- **Column Options:** *Chart, Discrete values, Explore, Summary, Rename, Delete and Index/Unindex*
  - **Numeric Column Options:** *Chart, Frequency, Discrete values, Explore, Statistics, Summary, Remove, Delete, Select and Index/Unindex*
  - **Derived Column Options:** *Chart, Frequency, Discrete values, Explore, Statistics, Summary, Rename, Edit, Make permanent, Delete, Select and Index/Unindex*
  - **Derived Domain Column Options:** *Chart, Discrete values, Explore, Summary, Rename, Delete, Index/Unindex and View definitions*

Making a calculated field permanent

A calculated field can be stored as a permanent field in the associated table for use by others. Identify a calculated field by the small, gray cog in the lower left corner of the calculated field icon. Unlike a calculated field, a permanent field can be used to sort columns in a report. Making a calculated field permanent replaces the calculated field. The database does not retain the original calculated field definition.

How to make a calculated field permanent

1. Right-click a calculated field appearing in My Data.
2. Choose Make permanent.

Using Discrete Values

Discrete Values is a view that shows the contents of individual database fields. Each unique value, or category, is represented by a total of all its occurrences or records. Discrete Values appears when you double-click a database field in My Data. For example, double clicking the Age Numeric field in the Customers table displays a list of age ranges and the number of customer records matching each age range, as shown in Figure 2-3.

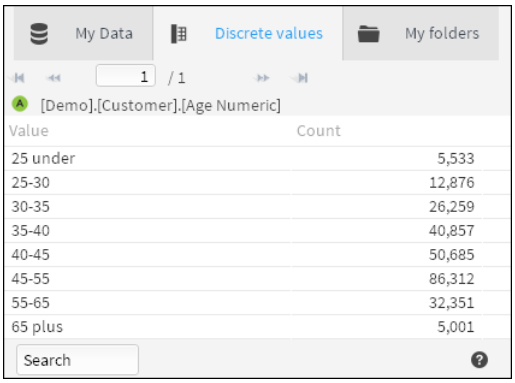


Figure 2-3 Discrete values showing the values in the Age Numeric field

Discrete Values also supports dragging fields and dropping them in places where selections or segments are used. For example, drag one or more specific categories to Scratchpad to examine the values in detail or store them for future use.

When a field has many categories, the viewer displays them on multiple pages. Typically, the list of discrete values is paginated, because most fields have a large number of categories. By default, the viewer shows 100 categories per page. For example, [Household].[Town] is a field that contains many discrete values. [Order].[Order No] contains more discrete values as there is a unique category for every order.

Categories can be sorted by name and the number of records they contain. The default sort order is by number of records. To sort categories in alphabetical order, click Value. A search engine supports finding particular categories by Name. Select multiple discrete values by holding CTRL while clicking the desired categories.

Discrete Value Search

The Discrete Values searching process is instantaneous due to its ability to take multiple keystrokes into account rather than individual keystrokes.

Using My Folders

My Folders displays the files and folders available to a user, as shown in Figure 2-4.

Figure 2-4 My Folders showing a personal folder, Income

Files and folders in the My Folders section in My Folders are available only to you. Files in the Shared section are visible to all users. Files can contain analyses, selections, and exports. There is no fixed limit to the number of folders you can create. To share a file with other users, move the file under the Shared heading.

Create new folders directly or while saving a file:

- Right-click the red folder labeled My folders. Choose New. Type the name and description of the new folder, as shown in Figure 2-5.

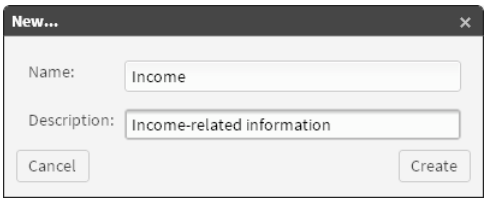


Figure 2-5 Creating a new folder named Income

- When saving an analysis, choose Save As to create a folder in which to save the analysis.

To save a complete data structure into a file with a .ddw extension, right-click a file and choose Export. The file can be downloaded to your local system. To load previously stored structures, right-click a folder and choose Import.

Understanding Scratchpad

Scratchpad is a work area located below Data Tree, as shown in Figure 2-6. Segments can be placed in Scratchpad for use in an analysis or for examining them in detail. Items in Scratchpad are not saved between sessions.

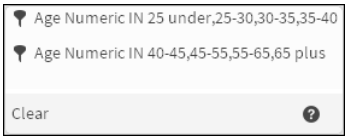


Figure 2-6 Scratchpad showing two segments

Use Scratchpad in the following ways:

- Drag segments directly from Data Tree to use as filters in analyses.
- Drag results of different analyses to Scratchpad for detailed analysis.
- Manipulate segments from Data Tree or from an analysis to create intersections, invert content, and so on.

All changes applied to a segment in Scratchpad affect only the current session. To retain these segments for later use, save them in My Folders, as described in "Using My Folders."

Scratchpad consists of four elements:

- Segments window. Drag segments from other windows in BIRT Analytics and drop them in this main area of Scratchpad. Available segments are tables from a database, discrete values, and the results of some analyses.
- Status bar. This information section appears below the Scratchpad segments window and displays the number of records in the selected segment.
- Segment handling menu. When you drag one segment over another in Scratchpad, the join functions in Table 2-2 appear in a menu.

Table 2-2 Functions used to join segment contents

Icon	Function	Description
	AND	Joins the two categories with the resulting segment containing those records that meet both conditions simultaneously. For example, you could select Customers who are Sales Assistants and under the age of 25.
	OR	Joins the two categories with the resulting segment containing those records that meet one or both of the conditions. For example, you could select customers who are either Sales Assistants or under the age of 25, which would include Sales Assistants under the age of 25.
	Exclude	Joins the two categories, with the resulting segment containing those records that meet the conditions in the first category but not the second one. For example, you could select customers who are Sales Assistants but not under the age of 25, which would exclude Sales Assistants under the age of 25.
	NAND	Inversion of AND.
	NOR	Inversion of OR.

- Secondary menu. When you right-click a segment, the options in Table 2-3 are available.

Table 2-3 Segment options

Option	Description
Rename	Prompts for a new name for the segment. Use this option if you have changed the results of the segment by joining it with another one and the default name no longer describes the segment accurately.
Delete	Removes the segment from the scratchpad.
Edit	Edits the segment in the Selections window that opens, making necessary changes or corrections.
Explore	Display all relevant table columns of the selected database in the Data Explorer panel, filtering with the chosen segment.

Table 2-3 Segment options

Option	Description
Invert	Change the segment to return results that exclude the selected segment. For example, to select all records from households that are not in London, select London and invert the selection. [Household].[Town] contains 1335 different towns. Selecting all towns, except London, one by one would be very time-consuming. Invert takes the category of London and returns all of the records that do not belong to this category.
Make permanent	Save the segment as a new field. The field appears in the table currently being used for resolution. Save a segment to make it available in a subsequent BIRT Analytics session. After refreshing the screen view, the new field appears in the table in which it belongs.
Change resolution level	You can resolve results at different levels within the database. These levels correspond to the tables that make up the segment. For example, resolve a category from the Customer table at the Household level to display the households of the customers. To interpret the results of changing the resolution, keep in mind the direction of the resolution change, either many (N) to one or one to N.
First discrete	Choose a record for each of the selection's attributes based on a field in the database.
View definition	Examine the definition of the segment. The definition contains all of the operations applied to the segment, indicating the type of operation, the segment's total records, the number of records after performing the operation, and the query that obtains this segment. After you make a change in resolution, the operations performed are displayed in groups.
Sort	Uses a field from the database to sort the segment. The result of the sort appears in Scratchpad and not in the data explorer.
Select sample	Returns a data sample with the size determined by the selection criteria: Top, Bottom, Middle, 1 in N, Random.
Refresh	Updates the count of records in the segment.

Understanding Data Explorer

Data Explorer is a tool that displays detailed and summary information about databases, tables, and fields. To access Data Explorer, choose Explore. Then, drag a segment from Data Tree or Scratchpad, and drop it in Record View.

Depending on the selection you make, different views are available, as shown in Table 2-4.

Table 2-4 Data Explorer views for data selection types

Selection type	Available Data Explorer view
Database	Summary
Table	Record, Summary
Unindexed field	Record, Summary
Indexed textual field	Record, Summary, Discrete Values, Chart
Indexed numeric field	Record, Summary, Discrete Values, Chart, Statistics, Frequency

Data Explorer allows you to export data as .csv files which can be opened in Microsoft Excel. To remove data from Data Explorer, choose Clear.

About Record View

Record View displays the records in the table you are exploring or the table that contains the field you are exploring. For example, when you explore the Customers table you see 50 of the 259,874 records. You can browse the pages into which the records are divided by using the arrows on the pagination bar. You can navigate to a specific page by typing the page number and pressing Enter.

Select the specific columns/fields you want to view exclusively in the Data Explorer main window by clicking on the small "Column Selector" icon in the upper right-hand corner of the window.

Double-clicking on a column/field in the Data Tree moves it into the main Data Explorer window on the right.

You can remove columns from the main Data Explorer panel by simply dragging them out.

To remove all data from the Data Explorer, click on "Clear".

About Summary

Summary displays information about the database, table, or field you are exploring.

For a database, Summary displays the name of the database; the tables in the database; and the number of rows, columns, and cells in each table.

For a table, Summary displays the name of the table; the number of rows, columns, and cells it contains; other database tables to which it is joined; and information about each field in the table, such as data type, number of discrete values, and whether or not the column is indexed.

For a field, Summary displays the name field; the table containing it; and other information, such as data type, number of discrete values, and whether the column is indexed.

To export the contents of the Summary tab to an PDF file, choose the export icon in the upper-right corner.

About Discrete Values

Discrete Values shows the categories contained in a column, the number of occurrences or records for each category, and the percentage for each category. To export the contents of Discrete Values to a CSV file, choose Export. CSV format is compatible with Microsoft Excel and text editors, such as Ultraedit and Notepad++.

About Chart

Chart displays a graphical representation of the records of categories in a database column. The values are ordered from most frequently occurring to least frequently occurring. The following chart types are supported:

- Columns 3D (default)
- Areas
- Bars
- Columns
- Doughnut
- Doughnut 3D
- Lines
- Pie
- Pie 3D

To change the chart type, use the “chart” icon in the upper right corner.

To export a chart as an image file, choose the “image” icon.

The export icon displayed in the Chart view exports the full analysis (data and graphic) in PDF format.

Sorting Charts in Explorer

By default, and for performance reasons, the Chart view displays records sorted by Value (ascending). This default sort is determined by how the values are sorted in the Discrete Values view. To display your Chart sorted by Count (descending) instead of by Value (ascending),” you need to go to the Discrete Values view and click on the “Count” column label. Your Chart will then be displayed by Count (descending). (See Figure 2-7 below).

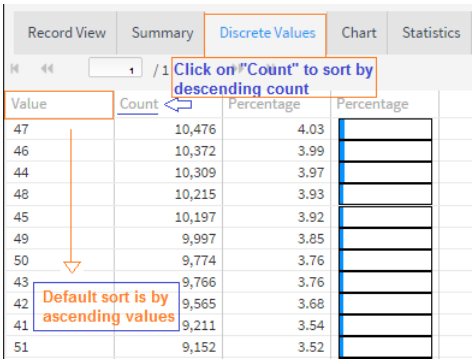


Figure 2-7 Sorting Charts in Explorer

About Statistics

Statistics displays the following information in a tabular format for numeric fields.

Table 2-5 Information provided for a numeric field

Column name	Value displayed
Count	Total number of non-discrete values in the field.
Kurtosis	A measure of the distribution of values, or peakedness. The value varies between -0.5 and 0.5. The Kurtosis coefficient indicates how sharp a distribution is, relative to a standard normal distribution.
Maximum	Largest value in the field.
Mean	Arithmetic average of all non-discrete values in the field.
Median	Value for which there are an equal number of larger values and smaller values in the field.
Minimum	Lowest value in the field.
Mode	The most frequently occurring value in the field.
Skewness	A measure of symmetry that varies between -0.5 and 0.5. If skewness is negative, the distribution is skewed to the left. If skewness is positive, the distribution is skewed to the right.
Standard deviation	A measure of how much variation or dispersion there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean. A high standard deviation indicates that the data points are spread over a large range of values.

(continues)

Table 2-5 Information provided for a numeric field (continued)

Column name	Value displayed
Sum	The sum of all the values in the field.
Sum of squares	The sum of all the squared values.

To export the contents of the Statistics tab to a PDF file, choose the export icon in the upper-right corner.

About Frequency

Frequency displays a graphical representation of the number of occurrences of values in a database column. The values are in ascending order. The following chart types are supported:

- Areas (default)
- Bars
- Columns
- Columns 3D
- Lines

To change the chart type, use the chart icon in the upper-right corner. To export a chart to a JPG file, choose the export icon.

Exploring views of a database

Data Explorer supports viewing different levels of a database, depending on the level of data you choose to view. For example, you can choose a database to view summary information for that database. You can choose a column in a database table to view records, discrete values, a chart of values, or summary information about the column.

Note: *When you minimize the size of a window, an icon appears that lists any buttons or tabs that are no longer visible on the screen.*

How to view a database summary

1. Drag a database from My Data and drop it in Data Explorer.
2. Examine the information appearing on Summary, as shown in Figure 2-8.

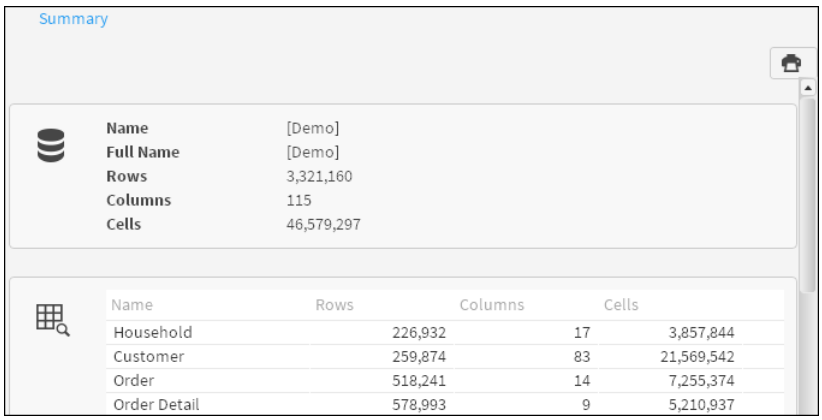


Figure 2-8 Viewing a database summary in Data Explorer

How to view a database table

1. From a database appearing in My Data, drag a table and drop it in Data Explorer.

2 Examine the information on Record View and Summary, as shown in Figure 2-9.

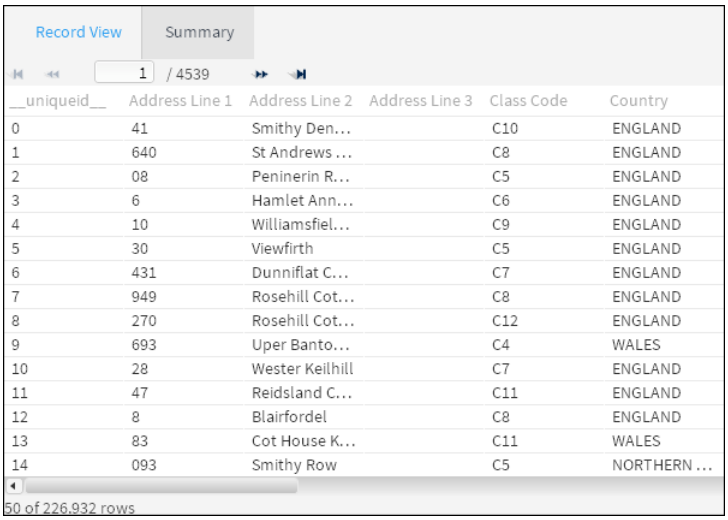


Figure 2-9 Viewing records from a table in Data Explorer

How to view a database column

1 From a database table appearing in My Data, drag a column and drop it in Data Explorer.

2 Examine the information in Record View, Summary, Discrete Values, and Chart, as shown in Figure 2-10.

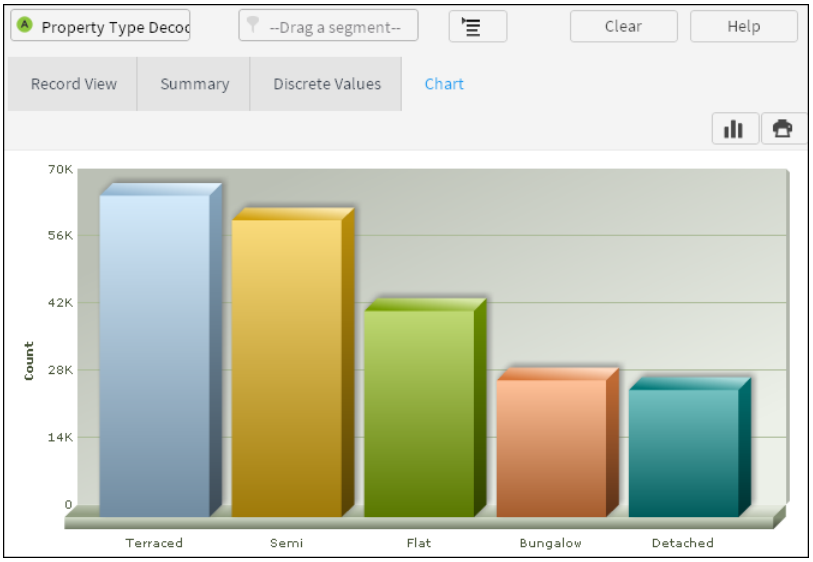


Figure 2-10 Viewing a database column using a chart in Data Explorer

Filtering views of a database

You can filter the view of a table or column appearing in Data Explorer using segments. Adding a segment to the view of a table or column limits the number of records shown to only those records having a value that matches the segment.

How to filter a database view using a segment

1 Drag a column from My Data and drop it in Data Explorer.

2 Select Discrete Values.

3 Drag a row in Discrete Values and drop it in Scratchpad. For example, from the Age Numeric column, drag Age Numeric EQ 25 under. A segment appears in Scratchpad.

4 In Data Explorer, select Record View.

5 Drag one or multiple segments from Scratchpad and drop the segments in Drag a segment. A view of data, filtered by the selected segment, appears in Data Explorer.

Understanding table resolution

BIRT Analytics supports saving your selections, queries and analyses for reuse. To retrieve records from a different part of your database, replace the table from which you retrieved records in a previous selection, query, or analysis, then recalculate the results. To demonstrate how the concept of table resolution works in practice, this section presents examples of viewing different tables in the BIRT Analytics demo database using Data Explorer.

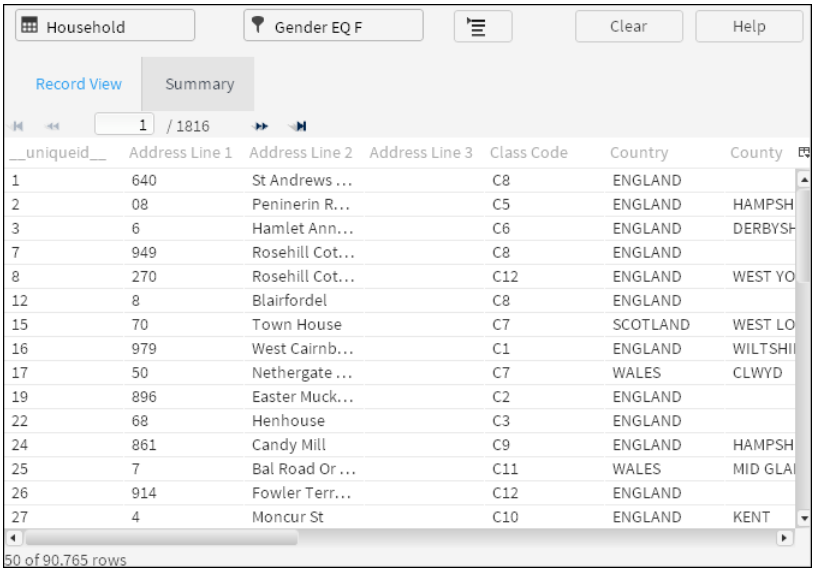
Viewing results of simple queries

Examining results returned by a simple query from one database table shows the discrete values in that table. For example, using My Data, expand the Household table, then double-click Property Types Decode. In Discrete Values, you see 28,514 records for households having the type Bungalow.

For a similar example, expand Customers, then double-click Gender. In Discrete Values, you see 102,042 F and 157,832 M records, which represent 102,042 female and 157,832 male customers in the database. To resolve questions about customers in each household, you can change the resolution or perspective of your query.

Changing table resolution

To demonstrate changing table resolution, modify the second example from the preceding section in the following way. Drag the Female value from Discrete Values and drop it in Record View of Data Explorer. You see complete records for 102,042 female customers. Choose My Data, then replace Customer with Household. To do this task, drag Household from My Data and drop it on Customer in Data Explorer. 90,765 records that represent households having one or more female customers appear in Record View, as shown in Figure 2-11.



__uniqueid__	Address Line 1	Address Line 2	Address Line 3	Class Code	Country	County
1	640	St Andrews ...		C8	ENGLAND	
2	08	Peninerin R...		C5	ENGLAND	HAMPSH
3	6	Hamlet Ann...		C6	ENGLAND	DERBYSH
7	949	Rosehill Cot...		C8	ENGLAND	
8	270	Rosehill Cot...		C12	ENGLAND	WEST YO
12	8	Blairfordel		C8	ENGLAND	
15	70	Town House		C7	SCOTLAND	WEST LO
16	979	West Cairnb...		C1	ENGLAND	WILTSHI
17	50	Nethergate ...		C7	WALES	CLWYD
19	896	Easter Muck...		C2	ENGLAND	
22	68	Henhouse		C3	ENGLAND	
24	861	Candy Mill		C9	ENGLAND	HAMPSH
25	7	Bal Road Or ...		C11	WALES	MID GLA
26	914	Fowler Terr...		C12	ENGLAND	
27	4	Moncur St		C10	ENGLAND	KENT

Figure 2-11 Changing table resolution in Data Explorer

Changing the table on which a query resolves returns results different than those returned by a simple query. To return expected results, the tables you replace must relate or join on a common field.

Chapter

3

Working with your data

This chapter contains:

- Using BIRT Analytics basic tools
- Using BIRT Analytics engineering tools

Using BIRT Analytics basic tools

Basic tools appear throughout BIRT Analytics to support common data analysis operations such as calculating and saving results, clearing work spaces, selecting data, and importing and exporting files. Engineering tools support creating new data fields that support your analysis of existing data values. This chapter explains the fundamental tools provided in BIRT Analytics.

Understanding the basic tools

A toolbar appears in the window for each analysis type. Some or all of the following basic tools are available in each window.

Calculate

Calculate runs and displays the analysis using the parameter values.

Export modes

Exporting Table view data

Press the Export button in the analysis toolbar. In this case the results table is exported from the Crosstab, Venn, Bubble or Profile analysis. Values are exported to a comma-separated values (CSV) file, a standard format supported by Excel and text editors such as Notepad ++.

Exporting Charts

The image icon displayed in the Chart view lets you export a chart as an image file.

Full export

The export icon displayed in the Chart view exports the full analysis (data and graphic) in PDF format.

Exporting a Crosstab to FastDB

:It is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option “Analytic DB” from the dropdown list of the “Export” tool found in the Crosstab toolbar..

Clear

Clear removes all entries from the window, without saving.

Convert

This option changes one analysis or indicator into another one with the same features.

Save or Save As

You can save all BIRT Analytics analyses you create using analysis tools or selections. The definition of the analysis is saved, but not the results themselves. Any modification to the database which affects its configuration is automatically applied when the saved analysis is run.

These analyses must always be saved in a folder. If a folder has not been created previously, it can be created at the time of saving the first analysis. You can access saved analyses from Data Tree using My Folders.

Any folder or subfolder is personal, unless you indicate otherwise and give viewing permission to other users. These permissions can be given for both folders and analyses.It is now possible to share data with groups instead of only with individual users. Groups can contain from one to any number of individuals. When new users are added to a group they automatically inherit the permissions granted to their group.

If you run a saved analysis, make a change to its configuration, and want to keep both the initial and modified versions, use Save As to save a new version of the analysis.

Applying a filter

Filters are used throughout BIRT Analytics and are based on data segments.

You usually can drag a discrete value directly to a filter area. For example, in an analysis of recent orders, you could drag the Customer Gender discrete value “female” to the analysis filter to see only orders placed by female customers.

Some tools offer more advanced filters.

About advanced filters

Crosstab, bubble, and map analyses support the following three types of filters: universal, target, and baseline.

A universal filter is applied before any change in resolution occurs. A target filter is applied after a change in resolution occurs. For example, to view only records for female customers, add as a universal filter: Gender equals female. If you add Gender equals female as a target filter and change the resolution from Customer to Household, only records that include households with females appear. Some of those households may also include males.

Target and baseline filters are used together to create comparative analyses. Be sure to use segments that can be compared. For example, compare one year with another or one population group with another. When calculating a comparative analysis, you can choose to display a measure as:

- Result
  The default for measures is to produce the count of records in both filtered segments. This is not directly useful for most comparisons but can be used as a total when creating calculated fields.
- Index
  Shows the degree to which compared groups differ using an indicator. An index value greater than 0 means that the baseline is as many times greater than the value shown by the index with respect to the target. An index value less than 0 means the reverse is true. The formula for Index is:

  ```
  (Target/Total) / (Baseline/Total)
  ```

- Difference
  Displays size differences between the baseline and the target as measured in units. A negative result means that the baseline has as many more values than the displayed number. A positive result indicates the opposite. The formula for Difference is:

  ```
  Target - Baseline
  ```

More about filters and resolution changes

If you are using filters, note that specific situations require certain filter types. You must use a target filter for pivoted analyses when there is a change in resolution between the axes and the measures in the direction N-to-1. You must use a universal filter with a non-pivoted table when there is a change in resolution between the axes and the measures in the direction N-to-1. When no such size disparity exists between axes and measures, the type of filter used for each analysis does not matter.

Consider creating a Crosstab using axes from one table and resolving the results in another. Use, as a filter, a segment from the source table for the axes.

For example:

- Universal filter. Apply the filter before carrying out the change in resolution. For example, the field low salary [axis 1: customer table, salary field] is used with the filter. Then, change the resolution to a different table, such as Households. The filter conditions are met by the same household and person. The resolution unit of the filter is the unit indicated by the axes of the crosstab, in this case Customers.
- Target filter. Apply the filter after carrying out the change in resolution. In other words, select the segments from the table to which the selected axes belong, and carry out the change in resolution for a table chosen as the resolution level for the results. Then, apply the filter. For example, low salary [axis 1: customer table, salary field] and the resolution is changed to Household. In this example, you see all households with at least one customer whose salary is low. The filter is applied. For example, [customer tables, gender field=female] gives a result qualitatively higher than the result from the universal filter. All households meeting the condition of low salary and female appear. The condition is not necessarily met by the same person. The filter's resolution unit is Households, the resolution table for the crosstab.
- Baseline filter. Selecting a target filter activates a baseline filter. Use a baseline filter for the purpose of comparison. For example, compare two periods of time using the following two filters: 2008 target and 2007 baseline.

To configure the table, you must first select the axis or axes by dragging to the appropriate space, then dragging the measures. By default, when you drag the axes, the value count for the table to which they belong appears.

Creating a parametric filter

Parameters used in filters should be included if a table is calculated or in a situation where you introduce a new data table in order to calculate the final output. The filter is determined by prompting you for the value when the analysis is calculated. You can use either a pre-set filter or a prompted filter, but not both.

Defining a selection

A selection is a segment of data, a set of values chosen for a specific purpose. A database is made up of values, and these values internally form groups that have similar features, with these segments being homogeneous. For example, an organization's customers can be gender=F or gender=M, but not both. You can combine these segments (gender=F, aged between 25 and 35, city=Barcelona, and average purchasing power, for example) and thus specify the target audience for a marketing campaign or promotional offer.

The BIRT Analytics Selections tool supports drag-and-drop configuration of elements in a selection. After configuring a selection, save it in Scratchpad or in a shared or private location for reuse. Calculate the selection to return all records meeting the criteria defined in the selection. As with full analyses, BIRT Analytics saves any defined parameters for each selection and runs them in real time.

For example, a simple selection may include only one query returning all rows from a single database table. To create a more complex selection, define multiple blocks. Each block may return records from a specific table or column, use a specific operator to compare values, or define ordering and grouping conditions. By combining elements in logical blocks, you refine the set of records the selection returns.

How to define a selection

1. Choose Analytics—Selections.
2. Choose Selections to open a new Selections window.
3. In the top-level block, type a name for the selection. As you type in your name, the name of the Selection window changes, displaying the name that you are typing, as shown in Figure 3-1.

Figure 3-1 Name and title are Home 3

4 From the Selections toolbar, drag a query element and drop it in a block.

5 From My Data, drag a table, column, segment, or report, and drop it in an element.

6 Choose Calculate to execute a selection. Results indicating a number of records selected appear for each element.

7 Click on "Save" in the toolbar menu and define the name here if you haven't already done it in Step 3. If you have already named your selection, the name appears in the Name field. You can change the name here if you wish. Then, choose a location in which to store the selection and define any desired sharing actions. Click on the "Create" button to save your selection. (See Figure 3-2).

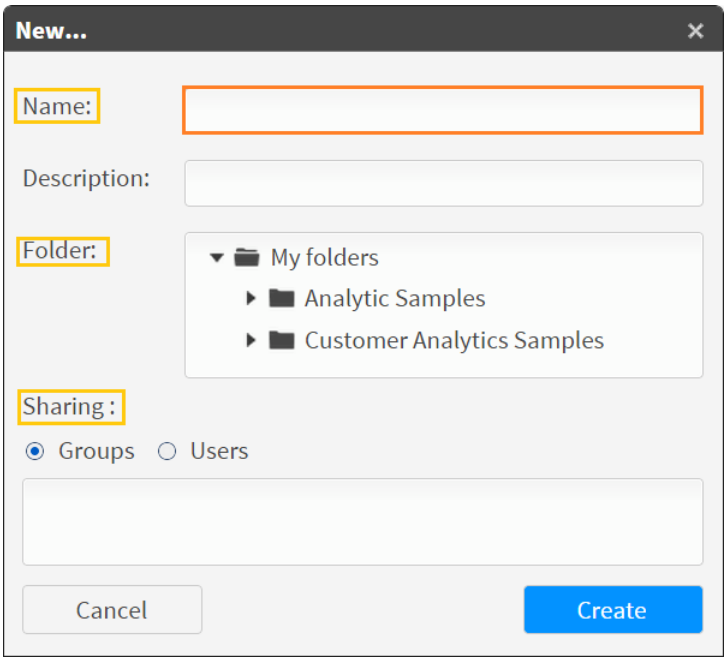


Figure 3-2 Saving your selection

8 Optionally, drag the selection name and drop it in Scratchpad.

For example, Figure 3-3 shows a simple selection returning all rows from the Customer table in the BIRT Analytics Demo database in which the customer is over 65 years old.

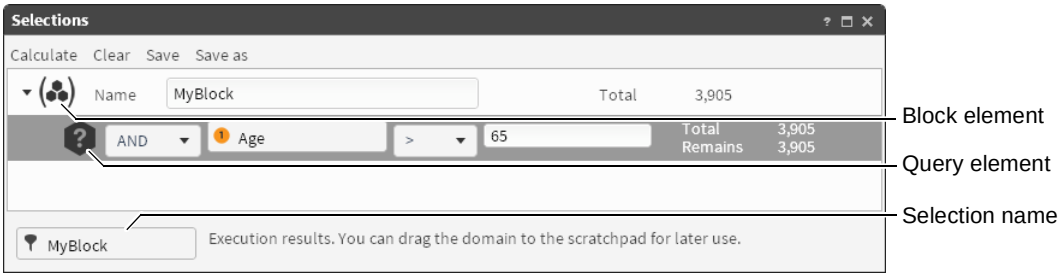


Figure 3-3 Examining basic elements in a simple selection

How to create a new query

1 Drag New query from Selections toolbar and drop it in a block.

2 Drag a field from the database, for example Gender, and drop it in the query element.

3 Select an operator from the following operators that appear in the drop-down list as shown in Figure 3-4.

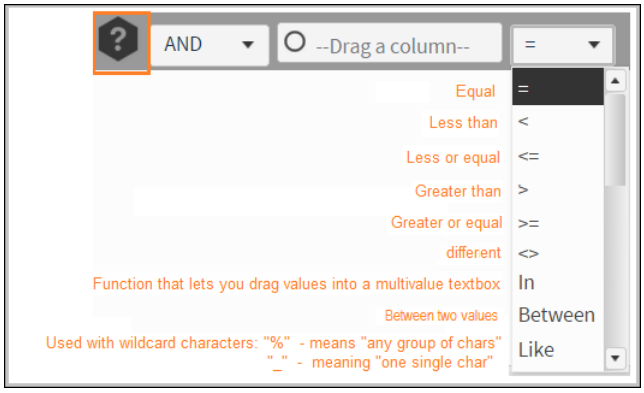


Figure 3-4 New Query list of available operators

4 Drag a value, for example Female, from Discrete values and drop it after the operator.

5 Choose Calculate. Then follow the same procedure for any other value you want to enter. Segments interact with each other by adding or excluding those records that do not match both values. For example, you have the value Female from the Gender field; if you drag the value 25 from the Age field, you can return:

- All women and everyone aged 25, regardless of gender
- All women aged 25

To accomplish this task, you select an operator from the drop-down list to the left of the dragged column. In the first example, the appropriate operator is Or. In the second example, it is And. By default, the And operator is used for values from different fields in the database. When the values are from the same field in the database, the default operator is Or. For example, if you try to combine the values Age=25 with Age=26, and the operator is And, the number of values that meet both of these conditions is zero.

You can also create a new query by dragging a segment from the data tree or the scratchpad to:

- A query in the selection. The new query is entered above.
- A block. The new query is entered at the end of the block.

To parameterize a query, right-click the query and choose Parameter.

How to create a new parametric query

A parametric query is a query based on a value determined at calculation time, which may be the default value of the parameter. You create a parametric query the same way you create a non-parametric query.

For example, create a selection to calculate the number of customers under 25 years of age with occupation janitor. You could have a query with age, occupation and another third query sentence. In this particular example, our third query sentence is parametric gender. You could include a default value for the parameter, such as Female. When you run the selection, you are prompted to type a parameter value or accept the default. The result is a selection parameterized by gender.

To modify the parameter prompt, right-click the query.

Adding a block

Use a block to group elements in a selection. For example, consider the following basic mathematical operation. If you want to determine the result of the sum of two values multiplied by the sum of two other values, group the sum operations as follows:

```
(34+89)x(23+65)
```

Now consider the following example as it applies to selections:

```
(Woman+salesperson) or (Woman+director)
```

If you do not include parentheses in this selection, you cannot be certain that the operators you use are invoked in the intended order. Hence one block is (Woman + salesperson) and another block is (Woman + director). A new selection has one block, in which you name the selection, by default. Every block must include a query element.

How to add a block

1. Drag a block element from Selections toolbar and drop it in an existing block.
2. Drag and drop additional elements in the block, as necessary.

Returning all rows from a table

To return all rows from a database table, add a simple query, using the "All" element.

How to return all rows from a table

1. Drag All from Selections toolbar and drop it in a block.
2. Drag a database table from My Data and drop it in the All element.
3. Optionally, select an operator from the list.
4. Choose Calculate.

Note: *the same action can now be performed by dragging a whole table from the Data Tree and dropping it on a selection item.*

Changing resolution

Adding a change resolution element to a selection or block changes the perspective of the selection. For example, to select customers who are both women and directors, you create a selection including query elements that return records for all customers who are female and directors. To see the households where the people who meet these conditions live, add a change resolution element to the selection.

A change resolution element cannot hold the top-level position in a selection. A change resolution element holding the last position in a selection causes the selection to return all records from the new table linked to all previous conditions. For example, the selection defined in the example shown in Figure 3-5 returns 868 records for households in which customers who are women directors live.

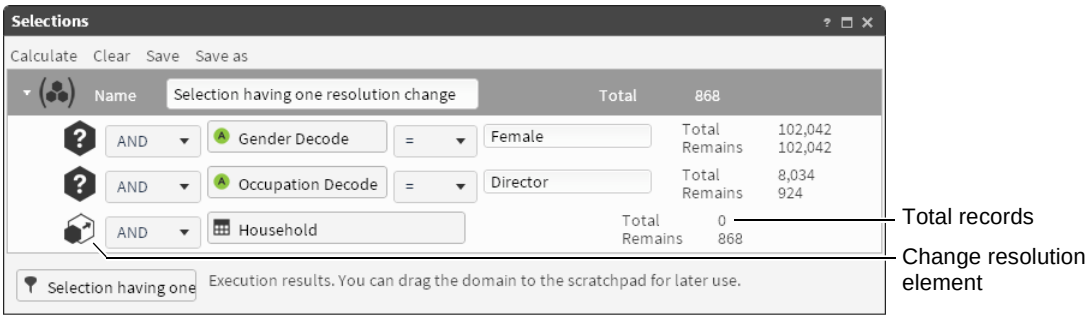


Figure 3-5 Examining a selection with one resolution change

For a contrasting example, add a change resolution element for each query element. This approach changes the perspective of each query in the selection. The selection in the example shown in Figure 3-6 returns all houses where customers who are female live, intersected with all houses where customers who are directors live; a total of 1009 records.

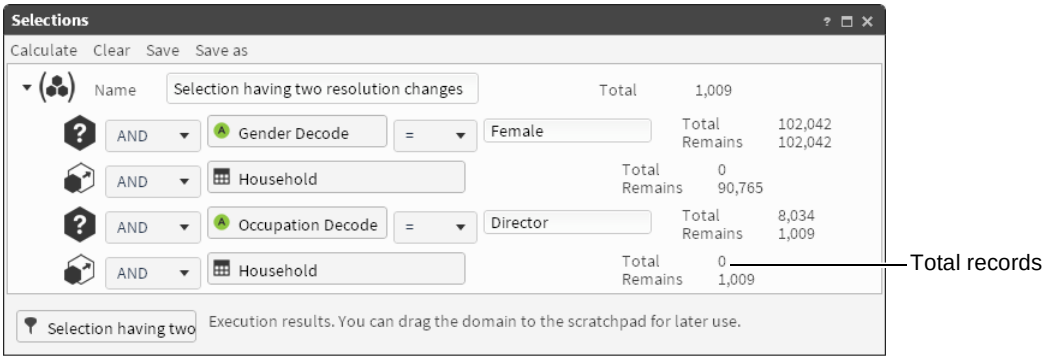


Figure 3-6 Examining a selection with two resolution changes

Returning all rows and changing resolution level

You may choose to create a selection using query elements that return records from different tables. Placing an All query element as the first, or the last element in a selection returns different numbers of records. Placing the All element first in a selection affects all single queries added after the All element by changing the resolution level to the table defined in the All element. An example showing a selection created in this way appears in Figure 3-7. The selection returns all houses, intersected with houses where female customers live, intersected with houses where director customers live.

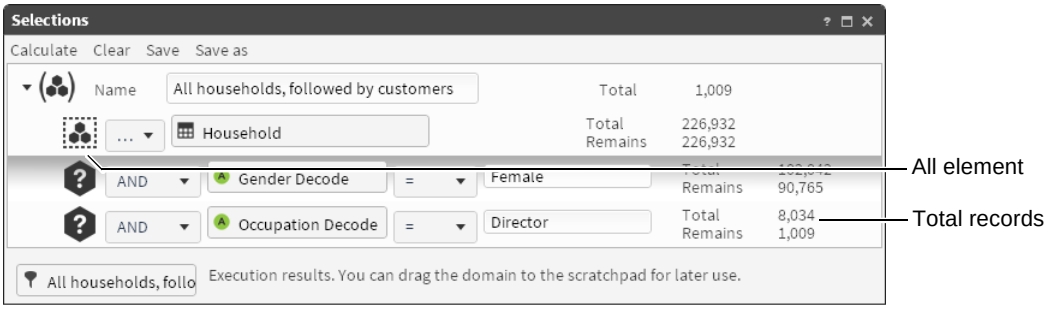


Figure 3-7 Selecting all records from a table first

Placing an All element last in a selection affects the selection by changing the resolution level to the table defined in the All element. In this case, the selection returns all female customers, intersected with director customers, intersected with all customers that have a house. In other words, this selection returns all female customers who are directors and have a house. In the BIRT Analytics Demo database, 868 records match this criteria. An example showing a selection created in this way appears in Figure 3-8.

Inverting a selection

To return the inverse set of values in a selection, drag Invert and drop it in a selection. For example, consider a selection that returns customer records containing Woman and Director; in other words, the set of women who are directors. Adding Invert to this selection causes this section to return all records in the database for customers who are not women directors.

Selecting discrete values

Add a Discrete Values element to a selection to return specific values. Drag Discrete Values from the Selections toolbar and drop it in a block. Then, drag a database field name from My

Data and drop it in the element. Choose Calculate to return the field values. For example, a discrete values element using the Age field returns seventy two values from the Age field.

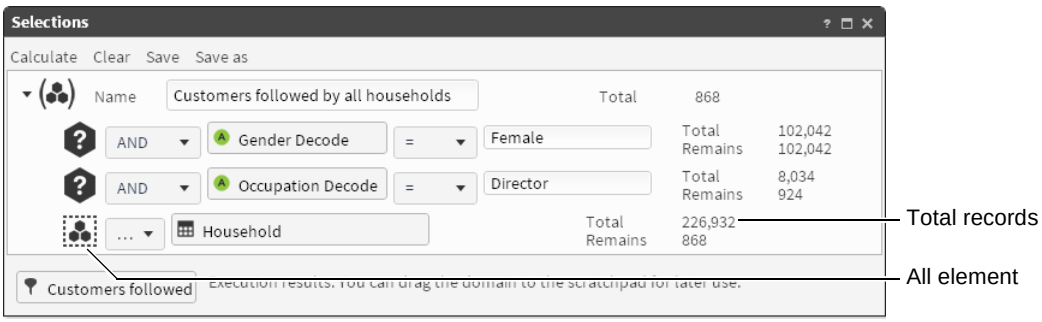


Figure 3-8 Selecting all records from a table last

Sorting values

Add a Sort element to a selection to sort the values in the selection. Drag a column and drop it in the Sort element, then select Ascending or Descending order.

Specifying a sample in a selection

Add a Sample element to a selection to return a defined range of N records for a block or a selection. Sample requires that you select a range parameter and number of values that define the sample. For example, select Top and type 10 to sample the top ten records of a selection.

Creating an inner selection

Use an Inner selection element to add an existing filter to a selection. For example, consider how to create a selection of female directors who are 50 years old, using an existing selection of female directors.

How to create an inner selection

1. Create a new selection and provide a name for the block.
2. Drag Inner selection from the toolbar and drop it in the block element.
3. From a shared location in My Folders, drag the Female directors selection and drop it in the Drag a report box in the inner selection element.
4. To the current selection, add a query element, Age=50.
5. Choose Calculate to return the number of female directors who are 50 years old.

Using import and export tools

Use these tools to import and export data and to create and delete links.

Using the import tool

Use this tool to import data to a new table created in the database currently loaded in the Engine. For example, this tool is useful when you want to create a Master table.

How to import a table

1. Select the source of the data to be imported. Options range from MS Access to text files.
2. For a text file, select file features and whether the first row contains the file header.
3. In the columns section, create the definition of the columns by selecting Get columns or Add columns. The definition of a column can be changed by double-clicking it. You should define as many columns as there are in the file.

4 Use the Column up and Column down buttons to change the position of a column.

Understanding links

Use links to delete existing relations between columns in different tables and to create new relations between columns. Creating links is important when working with the Engine. If the links have not been created or have been created incorrectly, it is not possible to work with various levels of resolution in the analysis. You cannot obtain reliable results when you mix columns from different tables in the database in a Crosstab or a Venn diagram.

Note: *In the case of links that have a 1 to 1 relationship, the first column will automatically become the Parent table and the second column will be the Child table.*

Using the export tools

BIRT Analytics provides two export tools. Export File exports a segment to a plain text file. Export Database exports a segment to a table in the database. Supported export file formats are: CSV (comma-separated values) file, a standard format supported by programs such as Excel and text editors including Ultraedit and Notepad++, and PDF file. Both tools provide Save and Save As options.

When you save an export definition, both the segment and the export configuration are saved in your personal folders. You can share your folders with others by granting viewing permission to a user or to a group or groups of users. Groups can contain from just one to any number of individuals. Permission granting is managed in the BA Admin tool.

How to export to a text file

1 In Enrichment—Import-Export, choose Export➛File.

2 Drag the segment you want to export from Scratchpad and drop it in Segment to export.

3 In Delimiter, select a column character separator: *tab, pipe, flat, comma, colon, semicolon, at, sharp, quote, plus, minus, apos, tilde.*

4 In End of line, select an end-of-line character: *CR, LF, CRLF,* or *LFCR.*

5 To save the file in the application server, choose *Deferred*. You can download this file from the application server by choosing *Downloads.*

6 To add a field containing the Uniform Resource Name (unique identifier) to the generated file, choose *URN.*

7 In Available columns, to select the fields to export, drag each field from the left pane to the right pane. To change the order of the fields, use the *up and down arrow buttons*.

How to export to a database

The export-to-database tool exports a segment to a new table in the analytical database engine. You must select the database in which you want to create the new table, the name of the new table, and the fields to create in the destination table.

Using downloads

This tool shows a list of deferred export files. This list shows the date, type, name, and file size. To download or delete an export file from the application server, you must open it.

Using BIRT Analytics engineering tools

BIRT Analytics engineering tools support creating new fields that you can include in your data analysis. To better quantify data, you can create fields that summarize, rename, and define expressions for existing fields within the database. You also can create ranges, groups, parameters, and ranks based on your existing data values. This chapter describes how to use

the following engineering tools: Aggregates, Decodes, Expressions, Numeric Ranges, Quantile, Parametric, and Ranking.

Aggregating values

Aggregates supports aggregating, or grouping, values in a data field. You can aggregate values from one field in any table. For example, create an aggregate using the Count function in the Order field of the Customer table to see how many orders each customer placed.

How to create an aggregate

1. In Enrichment—Engineering, choose Aggregates.
2. From My Data, drag the name of the table in which you will create an aggregate, and drop it in Create a new Aggregate based on, as shown in Figure 3-9.
3. In Name for new Aggregate, type the name of the new, aggregate field.
4. From My Data, drag the column that contains values to be aggregated, and drop it in Origin column.

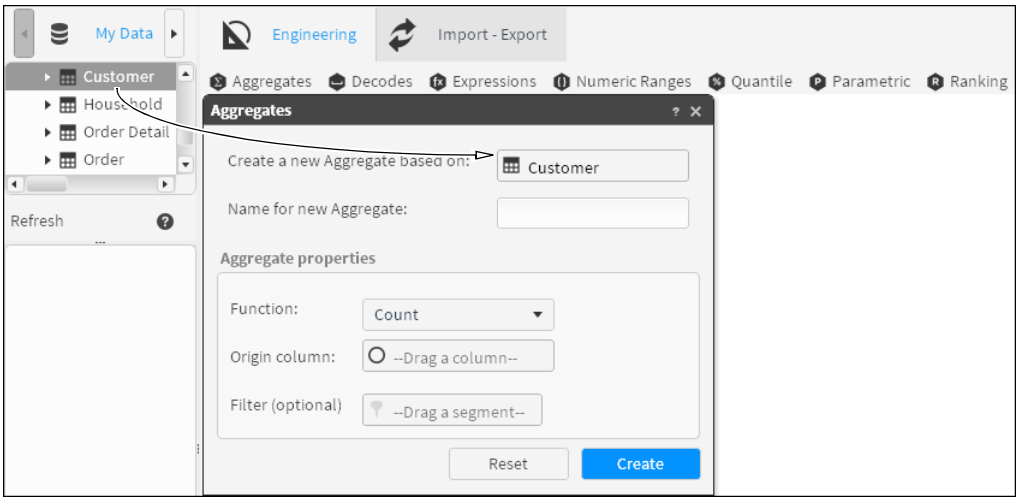


Figure 3-9 Assigning a table as a basis for an aggregate

5. In Function, select one of the following aggregate functions:
   - Average
   - Count
   - First
   - Last
   - Maximum value
   - Mean (integer)
   - Mean (real)
   - Minimum value
   - Quadratic sum
   - Standard deviation
   - Sum
6. Choose Create.
7. In My Data, review the new aggregate to verify your selections.

For example, to aggregate orders placed by Customers in London, select Count in Function, drag Order No. from Customers and drop Order No. in the Origin column.

How to aggregate totals for a specific group

1 In Enrichment—Engineering, choose Aggregates.

2 From My Data, drag a table on which to base the new aggregate, and drop it in Aggregates, as shown in Figure 3-10.

3 In Name for new Aggregate, type the name of the new, aggregate field.

4 In Aggregate properties—Function, select a function appropriate to aggregate values.

5 From My Data, drag a column that contains values to be aggregated, and drop it in Origin column, as shown in Figure 3-10.

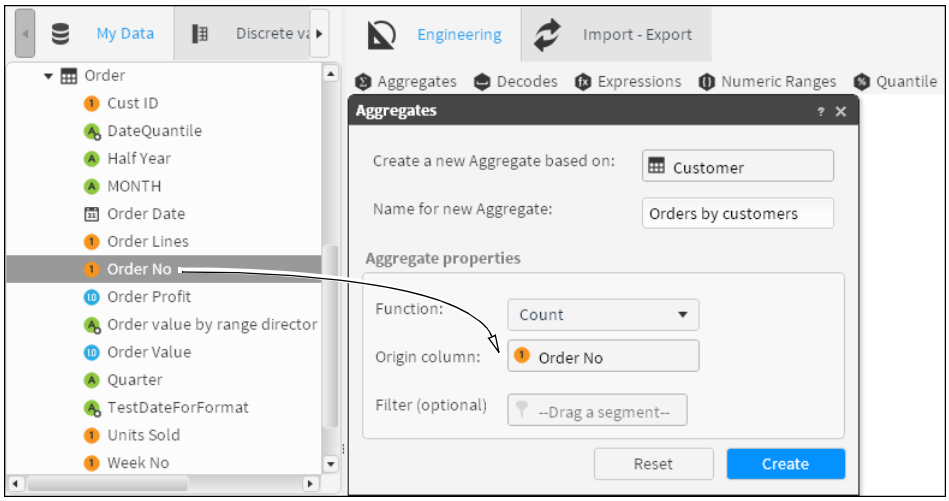


Figure 3-10 Defining properties for an aggregate

6 To create a filter that defines a specific group to be aggregated, double-click a column in My Data. From Discrete Values, drag a segment, and drop it in Filter (optional), as shown in Figure 3-11.

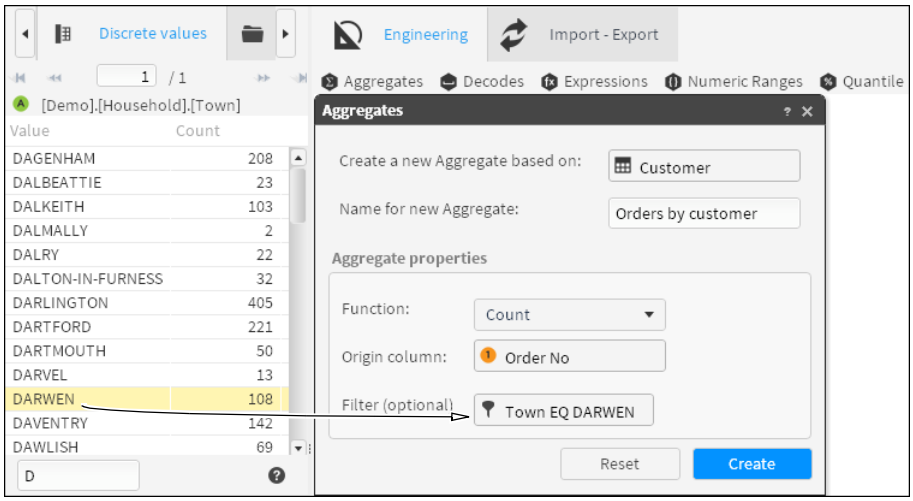


Figure 3-11 Adding a filter to an aggregate

7 Choose Create.

8 In My Data, double-click the newly created aggregate to view the results. Calculated values appear in Discrete Values, as shown in Figure 3-12.

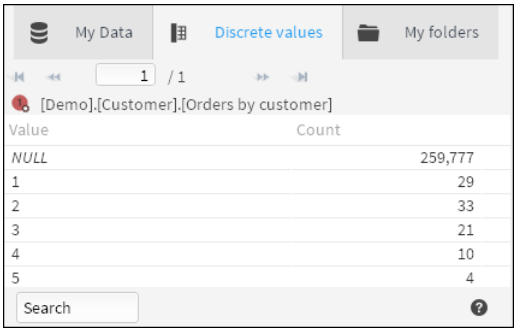


Figure 3-12 Viewing values in an aggregate

Decoding a field name

You can decode any field that contains up to 100 discrete values. For example, gender is usually encoded as M and F, or even as 0 and 1. Decodes supports changing the name to something more recognizable, such as Male and Female.

How to create a decode

1 In Enrichment—Engineering, choose Decodes.

2 From My Data, drag a column to decode, and drop it in Decodes, as shown in Figure 3-13.

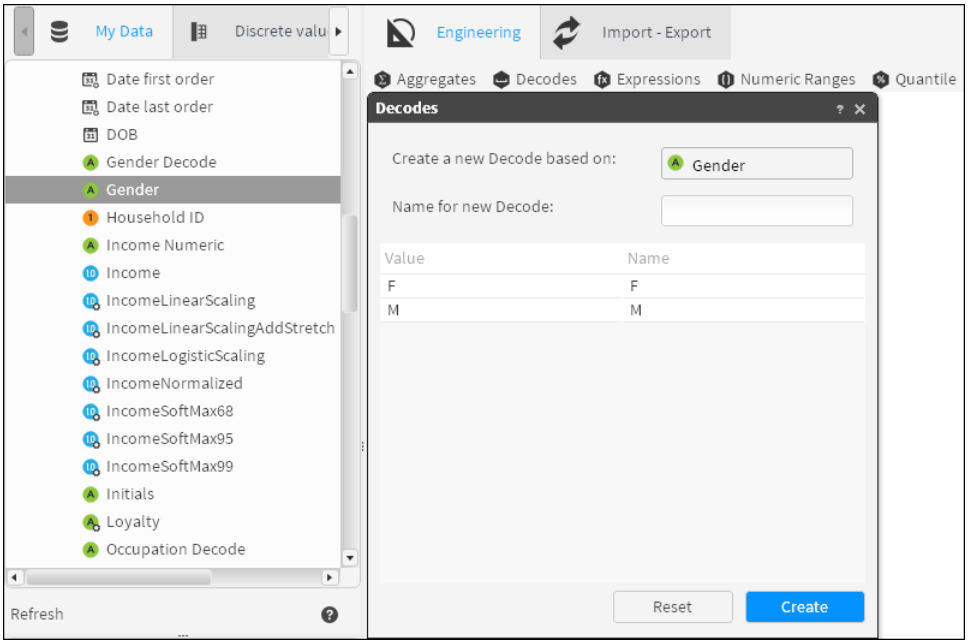


Figure 3-13 Defining a column to be decoded

3 In Name for new Decode, type a name for the decoded column.

4 To edit each field name, double-click a value in Name. Then, type any character, except the following characters, in the highlighted name field, as shown in Figure 3-14:

```
/ \ ° ª - accents, dieresis,?, !, *, @, #, &, $, o, ñ
```

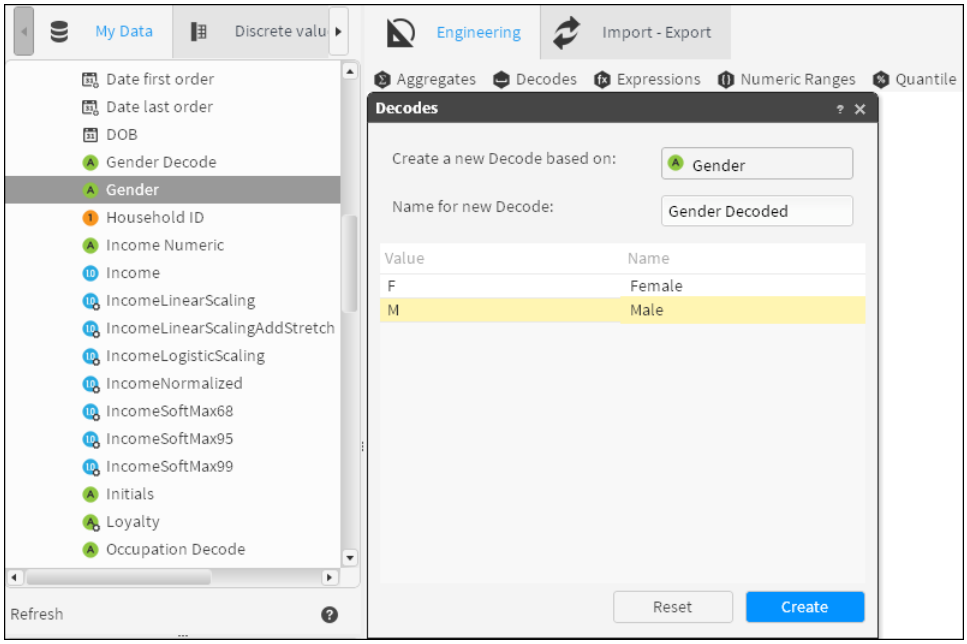


Figure 3-14 Providing a new name for a decoded column

5 Choose Create.

6 Double-click the newly created decoded field in My Data. Renamed values appear in Discrete Values, as shown in Figure 3-15.

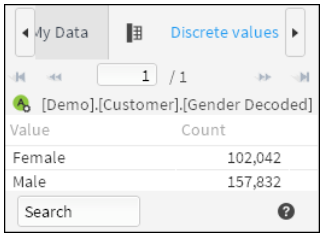


Figure 3-15 Viewing values in a decoded field

Working with expressions

Use an expression to create a calculated field from one, or multiple fields in a database. You can create a simple expression that concatenates strings by joining several fields or a complex expression that uses mathematical operations, such as logarithms. For expressions generating a decimal field, you can specify precision up to six decimal places. For example, a typical calculated field can display customer orders based on the year in which each order was placed.

To further demonstrate, you can use the following equations to accomplish the listed tasks:

- Concatenate Text (+): join two text fields.

  ```
  [Demo].[Customer].[Title]+[Demo].[Customer].[Surname]
  ```

- Concatenate (+): join two numeric fields or strings.

  ```
  STRING([CustomerID])+5+[DaysSinceLastOrder]-1
  ```

- Mathematical functions (+, -, *, /): numeric fields can be added, subtracted, multiplied, or divided.

  ```
  STRING([Demo].[Customer].[Cust_ID]+[DEMO].[Customer].[Household_ID])
  ```

- Boolean values: (logic 1, logic 0) assignment of Boolean values to fields, based on the conditions that are established.

```
[Demo].[Customer].[Edats]EQ"Adults">0 or
   [Demo].[Customer].[Edats]EQ"Joves">0
```

- Constants: a constant is a fixed value; in this case a constant function is one that gives a single value for a set of variables.

```
[Demo].[Customer].[Household_ID]=1
```

  This expression returns the value to position 1 in the Household_id field.

- Conditional Boolean value (Boolean if statements): complex conditions can be tested, such as whether the customer's age has a specific characteristic. Returns either true or false. Instead of returning true or false, the expression can return other values, such as yes or no.

```
If([DEMO].[Order].[OrderDate]>[Demo].[Customer].[DOB],1,0)
```

To display help for a function, double-click the function. Figure 3-16 shows the help for the AGE function.

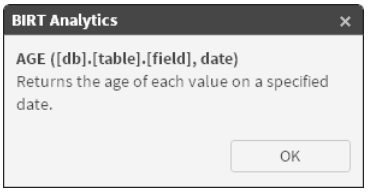


Figure 3-16 Help for AGE function

Expression grammar

The BIRT Analytics Loader supports the following operators described in Table 3-6.

Table 3-6 Supported operators and statements

Name	Description	Example
+	Concatenate two text columns.	`[Demo].[Customer].[Title]+[Demo].[Customer].[Surname]`
+	Concatenate Numeric columns or strings.	`STRING([CustomerID])+5+[DaysSinceLastOrder]-1`
+, -, *, /	Mathematical operators for adding, subtracting, multiplying, or dividing numeric columns.	`STRING([Demo].[Customer].[Cust_ID]+[DEMO].[Customer].[Household_ID])`
GT[>], LT [<],GE[>=], LE[<=],EQ [=], NE[<>], AND, OR, NOT	Assign Boolean values to columns, based on the conditions.	`[Demo].[Customer].[Edats]EQ "Adults">0 or [Demo].[Customer].[Edats]EQ" Joves">0`
Constants	Assign a constant to a table column.	`[Demo].[Customer].[Household_ID]=1`
Boolean If statement	Return a value depending on a condition. The condition can be simple or more complex. The return value can be any type, for example yes or no.	`If([DEMO].[Order].[OrderDate]>[Demo].[Customer].[DOB],1,0)`

Table 3-7 presents the complete list of the functions.

Table 3-7 Supported functions

Name and syntax	Description
ABS ([db].[table].[column])	Returns the absolute value of a numeric column, regardless of sign.
AGE ([db].[table].[column], date)	Returns the age of each value on a specified date.
ALLMONTHS ([db].[table].[column])	Returns the year and the month of each one of the values of a date column.
ALLQUARTERS ([db].[table].[column])	Returns the year and the quarter of each one of the values of a date column.
CEIL ([db].[table].[column])	Given a continuous numeric column, it returns the next integer value. For example, given the value 3.6, the CEIL function returns 4.
CHR ([db].[table].[column])	Given an integer numeric column, it returns the associated ASCII (American Standard Code for Information Interchange) value.
CODE ([db].[table].[column])	Returns the ASCII code for the first value of the text string. It only works for ASCII string columns, not for Unicode columns.
COUNT ([db].[table].[column])	Returns the number of records in a table.
DATE ([db].[table].[column], format)	Creates a date column converting input data to a date. You can use the following syntax: DATE(<string> | <unicode>,<date format:string>)
DATETIME ([db].[table].[column], format)	Creates a date column converting input data to a datetime. You can use the following syntax: DATETIME(<string>,<datetime format:string>) or DATETIME(<unicode>,<datetime format:string>)
DAY ([db].[table].[column])	Returns the part for the day in a date column.
DAYADD ([db].[table].[column], [db].[table].[column2])	Adds or removes the number of days required in a date column. The column returned is a date column.
DAYSTO ([db].[table].[column], date)	Returns the number of days between each value and a specified date.
DISTINCT ([db].[table].[column])	Returns the unique value or values in a column. For example, in a [Demo].[Customer].[Surname] column, it returns each surname stored in the column only once.
EXP ([db].[table].[column])	Returns the exponential value of each of the values of a numeric column.
FLOOR ([db].[table].[column])	Returns the largest integer number less than or equal to the array.
HOUR ([db].[table].[column])	Returns the time of each date.
IF (A, B, C)	Returns a conditional function. A is the condition to be applied, B is the value to return if the condition is true, C is the value to return if the condition is false.
INSTR ([db].[table].[column], “string”)	Returns an integer that specifies the initial position of the first appearance of a string in another one. First value is 0.
INT ([db].[table].[column], format)	Returns the integer number that forms part of the value in a column, or converts the column that contains the numbers into a numeric column. You can use the following syntax: INT(<string> | <unicode>, <thousands format:string>, <thousands grouping:integer | longint>)
ISNULL ([db].[table].[column])	Returns true for null values and false for not-null values.
ISNOTNULL ([db].[table].[column])	Returns true for not-null values and false for null values.
KURT ([db].[table].[column])	Returns the kurtosis of a column, which represents how values are distributed around the mean.

Table 3-7 Supported functions (continued)

Name and syntax	Description
LCASE ([db].[table].[column])	Returns a lowercase text string, turning uppercase characters into lowercase.
LEFT ([db].[table].[column], n)	Returns the left part of a string of characters with the specified number of characters - where n is the quantity of characters. The length of the strings in the column, resulting from an expression using the LEFT function, will be defined by the parameter provided in the function. The only exception that can occur is when the parameter is provided through a variable whose value is defined by a certain column. In this case the column result for LEFT will have a length equal to the original column.
LEN ([db].[table].[column])	Returns the number of characters in a text column, returning a number for each value in the column.
LOG ([db].[table].[column])	Returns the natural logarithm (logarithm in base e) of the values of the column.
LOG10 ([db].[table].[column])	Returns the decimal logarithm (logarithm in base 10) of the values of the column.
LONGINT ([db].[table].[column], format)	Returns the long integer part of a column, or converts the column that contains a number into a numeric column. You can use the following syntax: LONGINT(<string> | <unicode>, <thousands format:string>, <thousands grouping:integer | longint>)
LTRIM ([db].[table].[column])	Returns a string that is a copy of a string with no initial spaces (deletes any initial spaces).
MAX ([db].[table].[column])	Returns the maximum value of a column. The column used must always be numeric (continuous or discrete).
MD5 ([db].[table].[column])	Returns an MD5 hash of a free text or a text field
MEAN ([db].[table].[column])	Returns the average value of a numeric column (continuous or discrete).
MEDIAN ([db].[table].[column])	Returns the value that divides the values of a column into two equal parts.
MID ([db].[table].[column], number1, number2)	Returns a string that contains a specified number of characters in a string. Two numbers are required: the first indicates the position where the string to be returned starts and the second indicates the number of characters to return (optional). If Number1 (the position indicator) is greater than the number of characters in the initial string, it returns a 0-length string. The length of the strings in the column, resulting from an expression using the MID function, will be defined by the parameter provided in the function. The only exception that can occur is when the parameter is provided through a variable whose value is defined by a certain column. In this case the column result for MID will have a length equal to the original column.
MIN ([db].[table].[column])	Returns the smallest value of the parameters of a numeric column (continuous or discrete).
MINUTE ([db].[table].[column])	Returns the minutes part of a date and / or time column.
MODE ([db].[table].[column])	Returns the value that is most repeated in the values of a column (the most frequent value).
MONTH ([db].[table].[column])	Returns the month part of a date column.

(continues)

Table 3-7 Supported functions (continued)

Name and syntax	Description
NOW ()	Returns the current date and time in the engine's default format.
POWER ([db].[table].[column], number)	Returns all the values of the column to the specified second parameter (number).
PROPER ([db].[table].[column])	Returns a new column in which all records begin with an uppercase letter and the rest are in lowercase.
PROPERSENTENCE ([db].[table].[column])	Returns a new column in which the first letter of each word in the description of the variables is in uppercase.
QUARTER ([db].[table].[column])	Returns the quarter of the year for each of the records in the date column.
RAND (n)	Returns a column based on random (*n*). For example, RAND (120) in the Customers table returns a column in which customers are randomly divided into 120 groups.
REAL ([db].[table].[column], format)	Returns the real value of a column. The source column must be a numeric column, a text string containing text, or a string representing a numeric column. You can use the following syntax: REAL(<string> | <unicode>, <decimal format:string>, <thousands format:string>, <thousands grouping:integer | longint>)
REPLACE ([db].[table].[column], pattern, replacement)	Replaces the pattern text with the replacement text. Both pattern and replacement can be columns or text
RIGHT ([db].[table].[column],n)	Returns the right part of a string of characters with the specified number of characters - where n is the quantity of characters. The length of the strings in the column, resulting from an expression using the RIGHT function, will be defined by the parameter provided in the function. The only exception that can occur is when the parameter is provided through a variable whose value is defined by a certain column. In this case the column result for RIGHT will have a length equal to the original column.
ROUND ([db].[table].[column])	Rounds the input number to the nearest vakye containing no decimal positions. It creates a column of the same type as of the source column.
ROW ([db].[table].[column])	Returns a unique value, starting from 0, for each of the existing records in the column used to create the expression.
RTRIM ([db].[table].[column])	Returns a text string, an exact copy of the specified string without trailing spaces.
SECOND ([db].[table].[column])	Returns the seconds part of a date or time column.
SECSTO ([db].[table].[column], [db].[table].[column])	Returns the seconds elapsed between two columns with date or time format.
SGN ([db].[table].[column])	Returns the values grouped by -1 for negative values, 0 for invalid values, because not entered, for example, and +1 for positive values.
SHA1([db].[table].[column])	Returns an SHA1 hash of a free text or a text field.
SKEW ([db].[table].[column])	Returns the skewness value of the values of a column with respect to the mean value.
SQRT ([db].[table].[column])	Returns the square root of *n*, where *n* is the mean of the values of the column.
STDEV ([db].[table].[column])	Returns the standard deviation of the values of the column, which is the square root of the variance.

Table 3-7 Supported functions (continued)

Name and syntax	Description
STRING ([db].[table].[column], format)	Returns a text string that represents any other data type, except Unicode. Use any of the following syntaxes: STRING(<integer> | <longint> | <real>, <decimal format:string>, <thousands format:string>, <thousands grouping:integer | longint>) STRING(<date> <date format:string>[, <date separator:string>]) STRING(<datetime>, <datetime format:string>) STRING(<datetime>, <datetime format:string>, <date separator:string>, <datetime separator:string>, <time separator:string>, <time decimal separator:string>) STRING(<time>, <time format:string>) STRING(<time>, <time format:string>, <time separator:string>, <time decimal separator:string>)
STRIP ([db].[table].[column])	Returns the entered column deleting any spaces.
SUM ([db].[table].[column])	Returns a new column with the cumulative sum of the values of the original column.
SUMSQ ([db].[table].[column])	Returns a new column with the cumulative sum of each of the values of the original column squared.
TIME ([db].[table].[column])	Returns the time of a string, date-and-time, or time column. You can use the following syntax: TIME(<string> | <unicode>, <time format:string>)
TODAY ()	Returns current date in engine's format.
TRIM ([db].[table].[column])	Returns a text string that contains the copy of the specified string with no spaces either at the beginning or the end of the string.
UCASE ([db].[table].[column])	Returns a new column with text in uppercase.
UNICODE([db].[table].[column], format)	Returns a unicode string that represents any other data type. You can use the following syntax: UNICODE(<integer> | <longint> | <real>, <decimal format:string>, <thousands format:string>, <thousands grouping:integer | longint>) UNICODE(<date>, <date format:string>[, <date separator:string>]) UNICODE(<datetime>, <datetime format:string>) UNICODE(<datetime>, <datetime format:string>, <date separator:string>, <datetime separator:string>, <time separator:string>, <time decimal separator:string>) UNICODE(<time>, <time format:string>) UNICODE(<time>, <time format:string>, <time separator:string>, <time decimal separator:string>)
VAL ([db].[table].[column])	Returns a Unicode string that represents any other type of data.
WEEKDAY ([db].[table].[column])	Depending on the locale, returns the day of the week for each of the values. Default locale is en_US. For example, in the en_US locale: 1=Sunday. In es_ES locale: 1=Monday.
WEEKNUMBER ([db].[table].[column])	Returns the week number of a given date (as specified in ISO-8601).
YEAR ([db].[table].[column])	Returns the year of a date or date-and-time column.

Supported formats when working with DATE, TIME or DATETIME

The format is case sensitive, and these are the identifiers:

yyyy – represents the year with 4 figures

mm – represents the month with 2 figures

dd – represents the day with 2 figures

hh – represents the hour with 2 figures

MM – represents the minute with 2 figures

ss – represents the second with 2 figures

xxx – represents the millisecond with 3 figures

The character "_" represents any delimiter (any character)

The following formats are supported when working with the DATE, TIME or DATETIME instructions:

DATE

ddmmyyyy

mmddyyyy

yyyymmdd

dd_mm_yyyy

mm_dd_yyyy

yyyy_mm_dd

TIME

hhMMss

hhMMssxxx

hh_MM_ss

hh_MM_ss_xxx

DATETIME

yyyymmdd_hhMMss

yyyymmdd_hhMMssxxx

yyyymmdd_hh_MM_ss

yyyymmdd_hh_MM_ss_xxx

yyyy_mm_dd_hhMMss

yyyy_mm_dd_hhMMssxxx

yyyy_mm_dd_hh_MM_ss

yyyy_mm_dd_hh_MM_ss_xxx

ddmmyyyy_hhMMss

ddmmyyyy_hhMMssxxx

ddmmyyyy_hh_MM_ss

ddmmyyyy_hh_MM_ss_xxx

dd_mm_yyyy_hhMMss

dd_mm_yyyy_hhMMssxxx

dd_mm_yyyy_hh_MM_ss

dd_mm_yyyy_hh_MM_ss_xxx

mmddyyyy_hhMMss

mmddyyyy_hhMMssxxx

mmddyyyy_hh_MM_ss

mmddyyyy_hh_MM_ss_xxx

mm_dd_yyyy_hhMMss

mm_dd_yyyy_hhMMssxxx

mm_dd_yyyy_hh_MM_ss

mm_dd_yyyy_hh_MM_ss_xxx

How to create an expression

1. In Enrichment—Engineering, choose Expressions.
2. From My Data, drag a destination table for the field you want to create.
3. Type a name for the field, using any characters except:

   `/ \ ° ª - accents, dieresis,?, !, *, @, #, &, $, o, ñ`
4. Drag a function from the left pane and drop it in the function editor. The function appears in the editor, followed by parentheses.
5. Drag a field name from the destination table and drop it in the parentheses.
6. If necessary, edit the expression syntax.
7. Choose Create. A new calculated field appears in the destination table.

Creating numeric ranges

To create a unique field in which the values from a numeric field appear as a set of numeric ranges that you define, use Numeric Ranges. Two typical examples follow:

- From a numeric field that contains profit per order, you can create a group of numeric profit ranges. Label those ranges, for example: Very High, High, Medium, Low, and Very Low.
- From a numeric field that contains age values, you can create a group of age ranges, such as 10–20, 20–30, 30–40 and so on, that fit one or more types of data analysis.

How to create a field containing numeric ranges

1. In Enrichment—Engineering, choose Numeric Ranges.
2. Drag the appropriate field to Numeric Ranges, as shown in Figure 3-17.
3. Type a name for the new field.
4. In Band properties, set the following properties:
   - To create default ranges, type the highest field value in Upper bound, the lowest field value in Lower bond, and a number of ranges to create, in Divisions.
   - To create customized ranges, select Customized, as shown in Figure 3-18. Then, type a name for each range, and the upper and lower limit for each range. The lower limit must be equal to the lowest value in the range. The upper limit must be greater than the values in the range. For example, the range 15 to 20 includes values 15 through 19. The range 20 to 25 includes values 20 through 24.

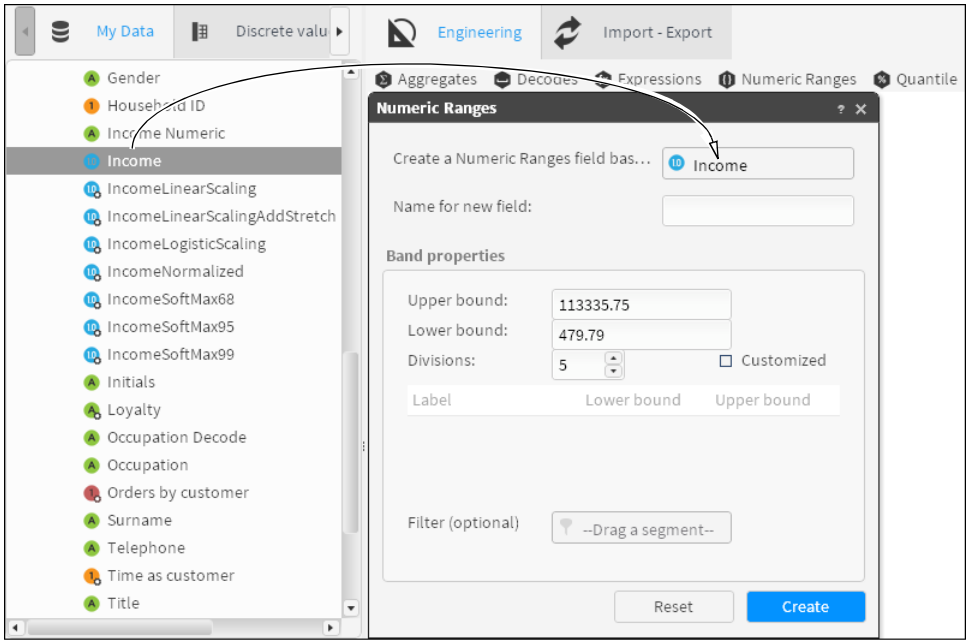


Figure 3-17 Dragging a field to use as a basis for a new numeric range

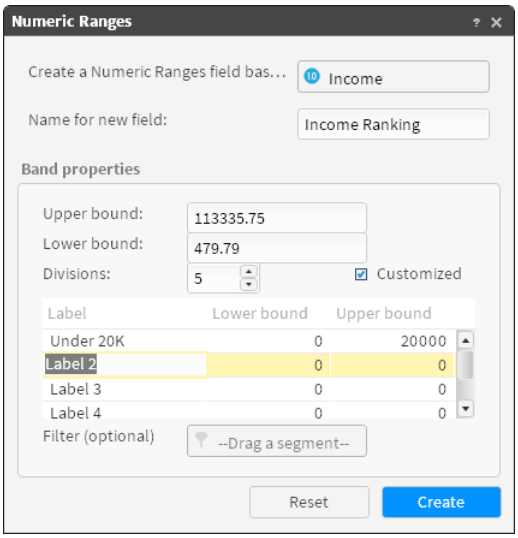


Figure 3-18 Creating customized numeric ranges

5 Optionally, to limit the number of values in the new field, drag and drop a segment in Filter (optional).

6 Choose Create.

Working with quantile ranges

A quantile range field is created from a numeric field. It creates groups that contain the same number of values. These groups are created by sorting the values in ascending order so that the first range contains the smallest values and the last range contains the largest values. For example, you can identify customers that generate the greatest profit.

How to create a quantile range field

1 In Enrichment—Engineering, choose Quantile.

2 Drag a numeric field from My Data, and drop it in New quantile based on.

3 In Name for new field, type a name for the quantile range field, as shown in Figure 3-19.

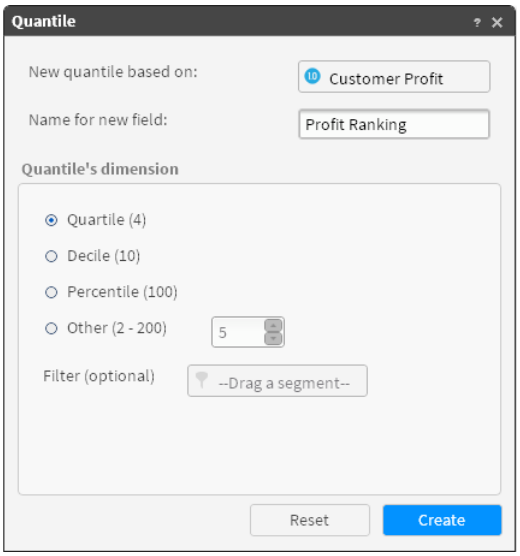


Figure 3-19 Creating a new quantile field

4 Select one of the following quantile types:

- Quartile. Divides the sample into four equal parts.
- Decile. Divides the sample into ten equal parts.
- Percentile. Divides the sample into one hundred equal parts.
- Other. Divides the sample into a selected number of two to two hundred equal parts.

5 Choose Create.

Understanding parametric columns

Parametric columns consist of query-based values. These queries are formulated with values or fields belonging to any table in the database as long as the tables are properly joined. The resulting field is created in the table indicated during the process, so the results are very different depending on the resolution marked by the selected table.

For example, you can determine sales of a particular product. For the customers table, the query returns the number of customers who bought one or more quantities of the product (customers account), while for the products table the query returns how many X products have been sold (products account). The number of products sold is probably greater than the number of customers who have bought them, because one customer may buy more than one unit of a product, but one unit of a product cannot be sold to more than one customer.

How to create a parametric query

1 In Enrichment—Engineering, choose Parametric.

2 Drag a table from My Data, and drop it in Parametric, as shown in Figure 3-20.

3 In Name for new Parametric, type a name for the new parametric.

4 Choose New Bin. Type a name for the bin.

5 Drag a segment from Discrete Values or Scratchpad and drop it in Filter, as shown in Figure 3-21 below.

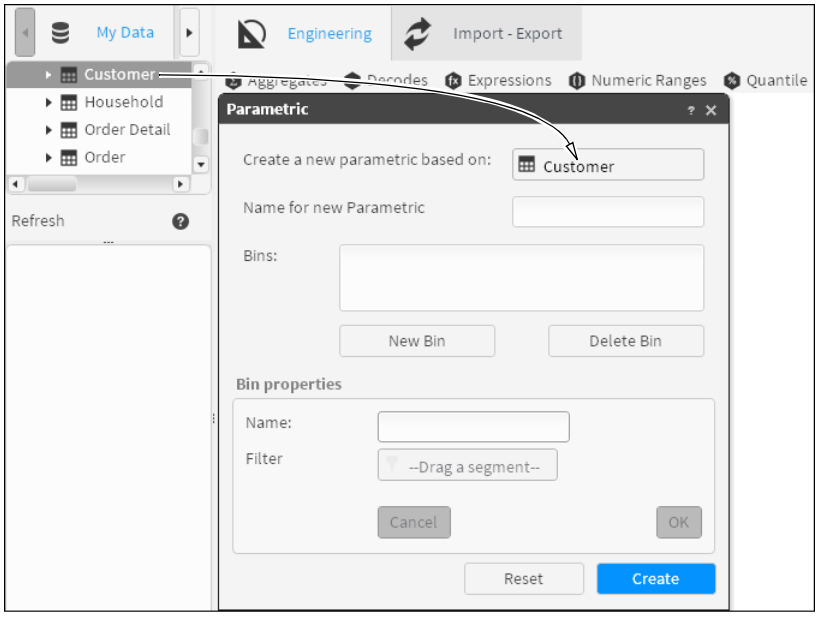


Figure 3-20 Setting the resolution table

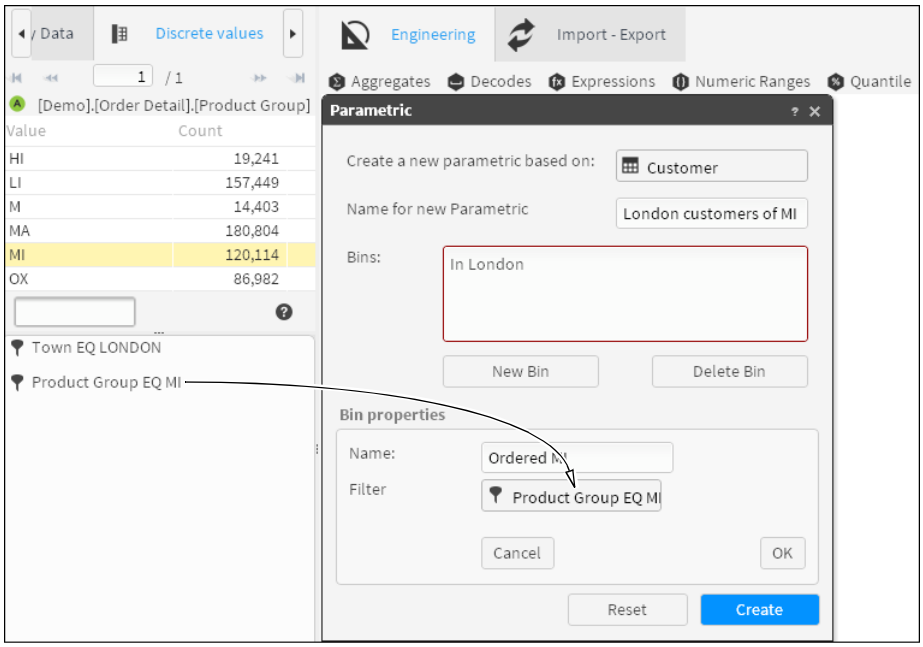


Figure 3-21 Setting a bin filter

6 Choose OK.

7 Repeat steps 4 through 6 to add more criteria to the query, if desired.

8 Choose Create.

To use the newly created parametric field, drag and drop it into Data Explorer. The parametric field itself does not appear in Data Explorer, because the same record may meet the condition of several values in the parametric field.

Understanding ranking

Use ranking to create a new column with the rank for each subset of records relating to the same key.These ranking columns can also be edited by right-clicking on the column in the Data Tree.

Ranking allows you to rank a table based on a specific column. For each subset of records associated with the same key in a table, a sequence number is assigned according to the relative value of a given field. Ranking is used, for example, when trying to identify the first products purchased by each customer, or the most recent product purchased.

How to create a ranking column

1. In Enrichment—Engineering, choose Ranking.
2. In New column name, type a name for the new column.
3. Select a ranking order, Ascending or Descending.
4. To select the key to rank, drag a column from My Data.
5. To select the field to use for ranking, drag another column from the same table, as shown in Figure 3-22. One table must include both key column and rank field.

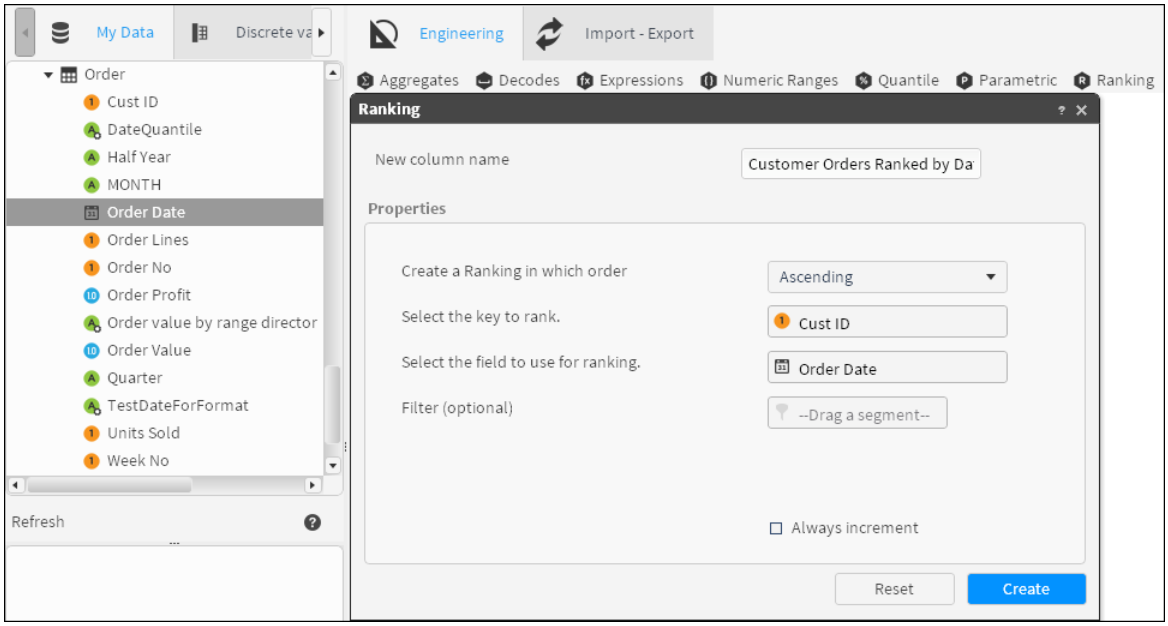


Figure 3-22 Selecting properties for a new ranking field

6. Optionally, to filter the ranked column, drag a segment from either Discrete Values or Scratchpad and drop it in Filter (optional).
7. In cases where two or more records have exactly the same value, select Always increment to ensure that each record is allocated a different rank. If you do not select Always increment, records with the same value have the same rank.
8. Choose Create. The new field appears in the table of rank elements. For example, you obtain 230,466 orders ranked with value 0, corresponding to the first purchases for each customer.

Chapter

4

Analyzing your data

This chapter contains:

- About analyzing your data
- Using crosstabs
- Using Venn diagrams
- Using Bubble analyses
- Using evolution
- Using profile analyses
- Using map analyses
- Using Pareto analyses

About analyzing your data

BIRT Analytics provides many ways to analyze data. The analyses make it possible to get new perspectives on your organization's information by crossing various fields, searching for characteristics that are or are not shared by groups, identifying suitable targets for each purpose, showing the evolution over time of different variables, and grouping values in different clusters.

BIRT Analytics supports the following types of data analysis: Crosstabs, Venn diagrams, Bubbles, Evolutions, Profiles, Maps, and Pareto. You can save the definition of the analyses (the means of generating them) for reuse and/or export them as PDF, CSV or BIRT (.rptdesign) files. Analyses exported as BIRT files can be edited in BIRT Designer and published on the BIRT iHub platform as well as in BIRT iHub F-Type.

When you run a saved analysis, any changes in the data appear in the current result of your analysis. You can also share your analyses with other users and user groups.

Analysis tool bars

The main window of each Analysis tool presents a tool bar. Although most operations are available to each analytics tool, some tools have fewer operations than others. In particular, the Evolution, Map and Pareto tools have neither the Export nor the Convert operations in their toolbar. The Convert operation is also not available in the Venn and Profile tools. The following list presents all available toolbar options:

- **Calculate:** Tool for calculating the analysis results in accordance with the set parameters.
- **Export:** Tool for exporting analysis results after they have been calculated. Depending on the type of analysis, the output format will be a CSV file (Comma Separated Values file) or a BIRT file (.rptdesign).

  - **Exporting Charts:** The image icon displayed in the Chart view lets you export a chart as an image file.

  - **Export icon**: Clicking on the Export icon, located on the upper right-hand side of the tool window, lets you export your analysis as a PDF file.
  - **Exporting a Crosstab to FastDB:** It is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option "Analytic DB" from the dropdown list of the "Export" tool found in the Crosstab toolbar.
- **Clear:** Tool for clearing the configuration window of the analysis, leaving it empty.
- **Convert:** Tool for converting one type of analysis to another. For example, in certain cases, you can convert a Crosstab analysis into a Bubble, an Evolution, or a Map analysis.
- **Save:** Tool for saving the analysis in a personal folder. If the analysis in question has been saved previously and a change has been made, the existing one will be overwritten and the changes recorded.
- **Save as:** Tool for saving the analysis in a personal folder. Here, if you select "Save as", no existing analysis will be overwritten and the file will be saved as a different analysis.

Using crosstabs

A crosstab is an analysis tool that allows you to cross different data fields either from within the same database table or different tables. For example, a sales manager can easily cross a "processed orders" field with a "type of articles" field to obtain (after adding calculated measures), the average profits generated or the sum of the purchase amounts, etc. The results

are displayed as both dynamic tables and graphics. See practical, hands-on examples in "Sample procedures for creating crosstabs."

Understanding Crosstabs

A *dimension* is an axis of the Crosstab -- i.e. one of the fields which is to be crossed with another field to provide an analysis of their data. When building a basic Crosstab analysis, you choose the database column or columns to be used as a dimension by dragging them into the *row* field and applying measures and filters.

It is also possible to pivot a Crosstab analysis by dragging the database column (to be used as a pivot) into the *column* field. This defines a special second dimension for the analysis, populating both X and Y axes - the first dimension "X" determined by a row(s) and the other dimension "Y" determined by the chosen pivot column. Pivoted results are presented according to the crossing of the values of the *pivot column* with the discrete values of the *dimension rows* and any applied measures and filters.

The *discrete values* in a database column that you choose to define as a dimension variable in your Crosstab are used as labels for dimension rows in the analysis. Options are made available for editing or clearing a discrete value by clicking on its corresponding dimension variable.

When dimension rows are created, cell values are automatically calculated for them through the application of a *default measure* (with a "count" operator) based on the database table that contains the chosen column. A Crosstab analysis must have at least one measure. (Otherwise no calculation operations can be done). When building your analysis you create/change measures to suit your needs.

Crosstab window environment

Before starting, you need to understand the Crosstab window environment. Clicking on "Crosstab" in the main "Analysis" tool window opens the Crosstab "Table" view where you build your Crosstab analysis. Its panels display basic information on how to start using the window immediately. You can build a simple Crosstab analysis directly in this window, in a few seconds, by simply dragging and dropping your chosen fields (columns) from the Data Tree on the left into the corresponding fields and panels. (See Figure 4-1.)

The main panel on the right will then display a column for each of your chosen fields and a final one for the calculated results. The column rows are labeled with the discrete values from your chosen dimension rows. (See Figure 4-2).

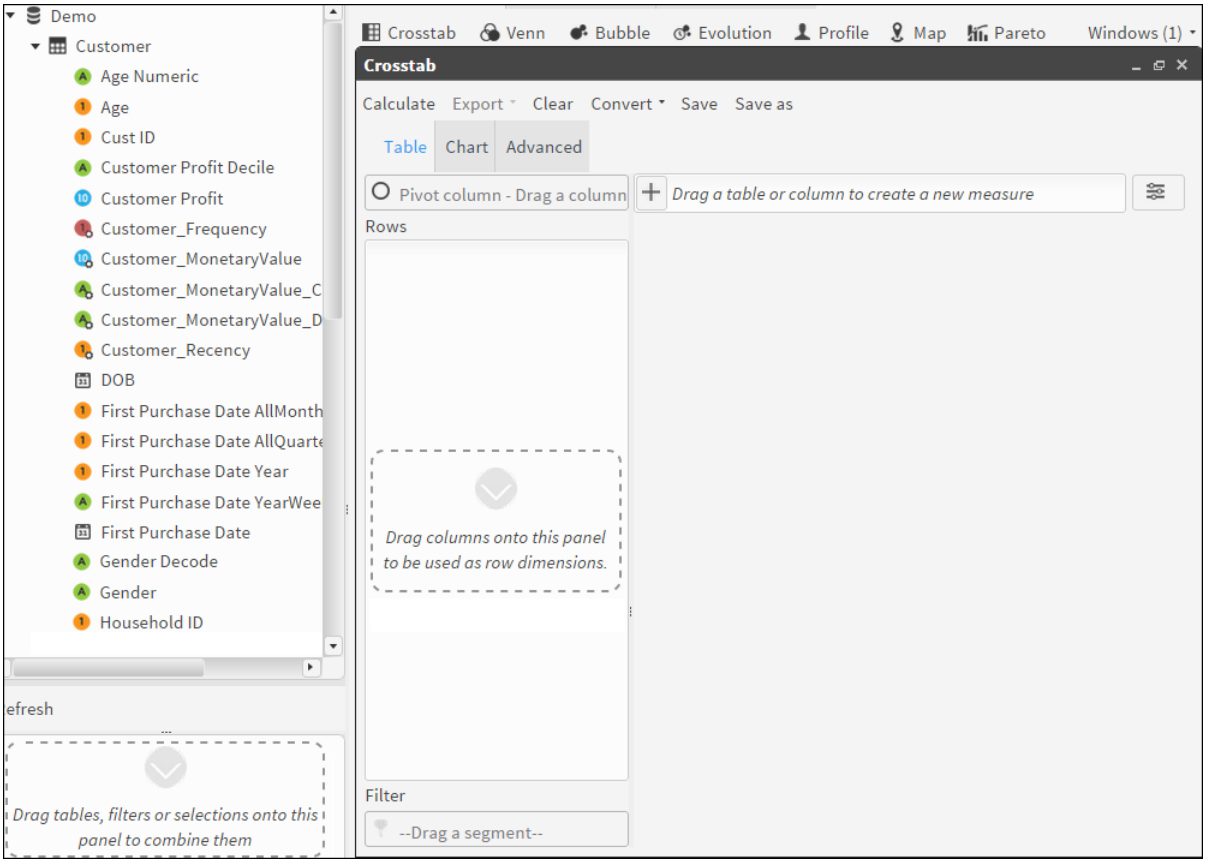


Figure 4-1 Crosstab Table view with explanatory panels

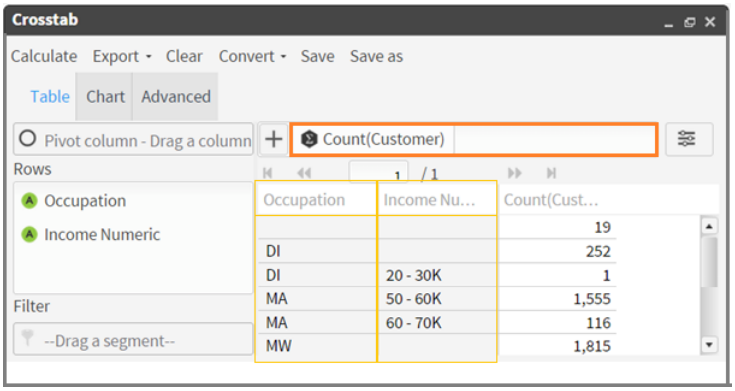


Figure 4-2 Table view of simple Crosstab analysis

In Figure 4-2, we also see that a default measure, “Count(Customer)”, has been automatically created when the first dimension was set. It appears in the “Measure” field near the top of the window. A Crosstab analysis must contain at least one measure before any calculation can occur. Clicking on a measure gives you access to an edit option. See Figure 4-3.

You can also edit the discrete values to be displayed by double-clicking on your chosen rows in the Rows panel. (See Figure 4-4).

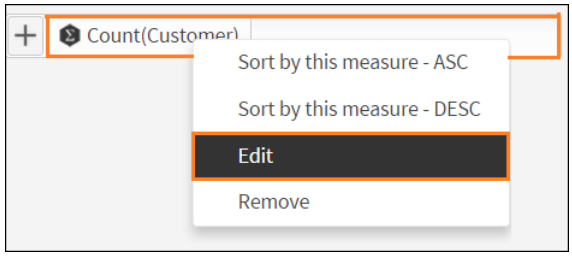


Figure 4-3 Measure options

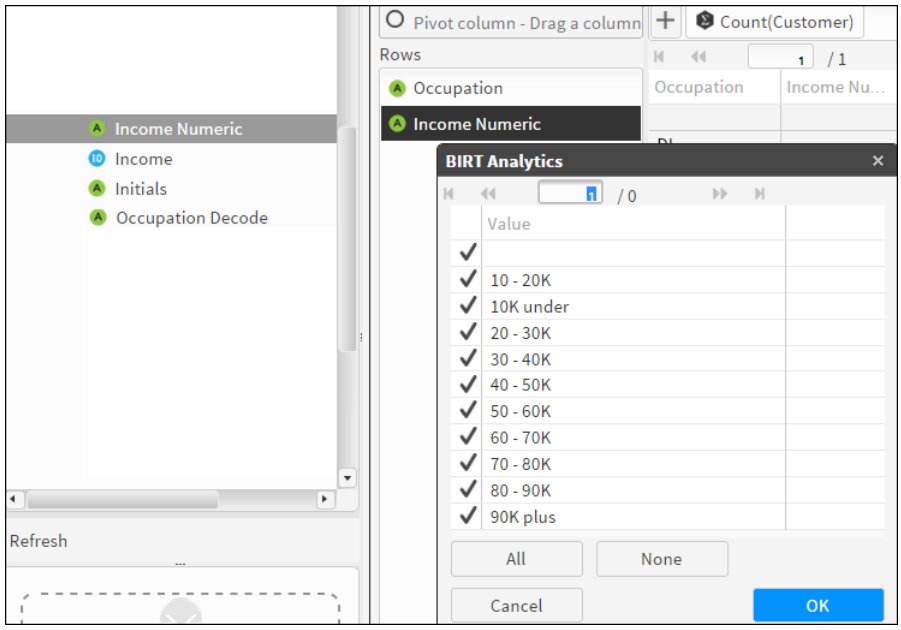


Figure 4-4 Editing discrete values in chosen dimension rows

Your results appeared automatically in the right-hand column as soon as you dragged your dimensions into their designated fields. This happens because "Autocalculate" is the default setting.

Note: *Clicking on the "Quick Options" icon in the main Crosstab window gives access to several options when building a Crosstab analysis. They include, among others, the "Autocaclculate" and the "Sort by measure" options. You can save time by turning off the "Autocalculate" option when building a high volume analysis with multiple rows. (See Figure 4-5).*

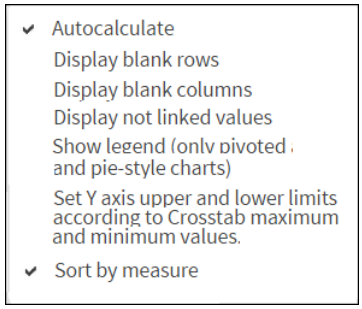


Figure 4-5 Quick Options list

Using the main viewing tabs in the Crosstab window

In order to be able to take full advantage of this tool, you also need to understand the usage of the Crosstab window view tabs "Table", "Chart" and "Advanced".

Table View

This is the default tab that opens when you click on “Crosstab” in the Analytics window (see figure 4.a at the beginning of this chapter). This tab provides you with an area for defining your dimensions (rows and pivot column) and for adding measures and filters. You simply drag and drop them into the corresponding fields.

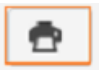
At least one measure must be defined in order to produce an output. The first dimension that you drag into a field sets a “Count” measure for the table to which the dimension belongs.

Chart View

This tab displays a graphic representation of your numeric results. Clicking on a chart section lets you send the specific selection information to the Scratchpad. Both Doughnut and Pie charts can be rotated by maintaining a click on them and dragging in the direction you want the chart to rotate.

To choose the type of chart you want displayed, click on the “Chart” icon.

You can export your chart as an image file by clicking on the “Image” icon.



To save it as a PDF file, Click on the “Export” icon.

Table 4-8 Basic Crosstab chart types: (by default – Columns)

Areas	Doughnut	Doughnut 3D
Bars	Lines	Lines 3D
Pie	Pie 3D	
Columns	Columns 3D	

Table 4-9 Pivoted Crosstab chart types: (by default – Columns 3D)

Areas	Areas with scroll		Stacked areas
Bars		Bars 3D	Stacked bars
Columns	Columns with scroll	Columns 3D	
	Stacked columns with scroll	Stacked Columns 3D	Stacked columns
Lines	Lines with scroll		

Advanced View

Clicking on the “Advanced” view tab opens the “Measures” tab by default. Here you have access to three other tabs: *Filters, Parametric filters* and *Options*.

Reminder: A measure is an operation that you choose to execute on the values that match the crossing of the fields in the rows and columns. As soon as you choose a field as an axis (setting a dimension or choosing a pivot column), our tool creates a measure using “Count” as the default operator.

Measures Tab

The “Measures” tab is used for advanced editing/creation and perfecting of the measures used in your Crosstab analysis. Upon opening this tab, you see a table of all the measures that have been set for the current Crosstab. You are also provided with brief instructions for defining measures in this tab. (See Figure 4-6).

Double -clicking on a measure in the Advanced tab opens it in the Measure Creation window, populated with the measure information, ready to be edited. The buttons at the bottom of the Advanced “Measures” tab offer four useful operations that can be performed when creating

measures, as shown in Figure 4-7.

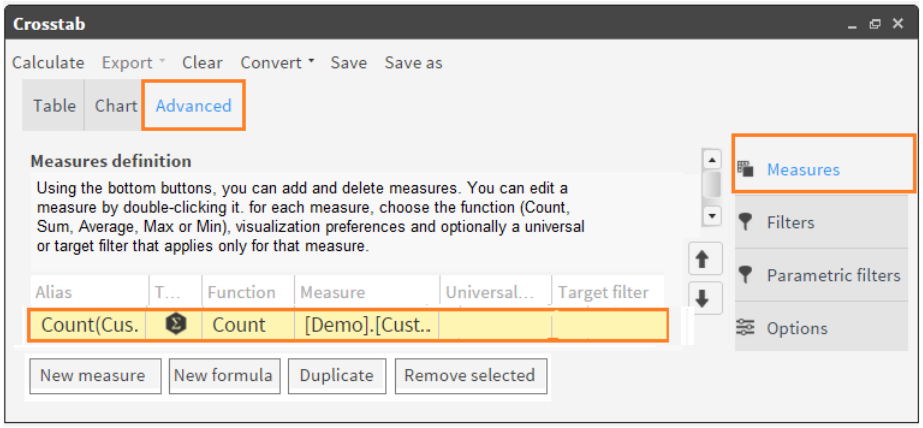


Figure 4-6 Crosstab Advanced Measures tab

Figure 4-7 Advanced Measures operations

New Measure Button

This button opens a window designed to help you build a new measure (See figure 4.7 below). In this window you can:

- Name your measure in the "Alias" field (mandatory).
- Choose your basic measure operation from the dropdown list in the functions field: *Count, Sum, Average, Max or Min*.
- Specify the field or table that will be operated on by dragging it from the Data Tree into the "Table" field beside the "Function" field.
- Decide what you want displayed using the dropdown list in the "Display as" field: *Result, Diff., Index*.
- Decide how you want the results displayed using the dropdown list in the "View" field: *Total, %Total, %Row and %Column.*
- Check or uncheck the "Visible" checkbox, depending on whether or not you want to display the results of your measure calculation in the "Table" view.
- Decide the number of decimal points to display using the dropdown list in the "Format" field.
- Add universal and target filters if required. Normally, you should only filter the measure by a specific category or segment (ie. specific group of customers, territory, type of product, etc.).
- Save your *new measure* and go back to the Advanced "measure" tab by clicking "OK". (See Figure 4-8.)

New Formula Button

This button gives access to a window where you can create more complex calculated measures, based on the simple ones that you have already created.

Example: *You could use the sum of the cost of company operations, divided by the number of operations carried out in order to obtain an approximation of company costs per type of operation.*

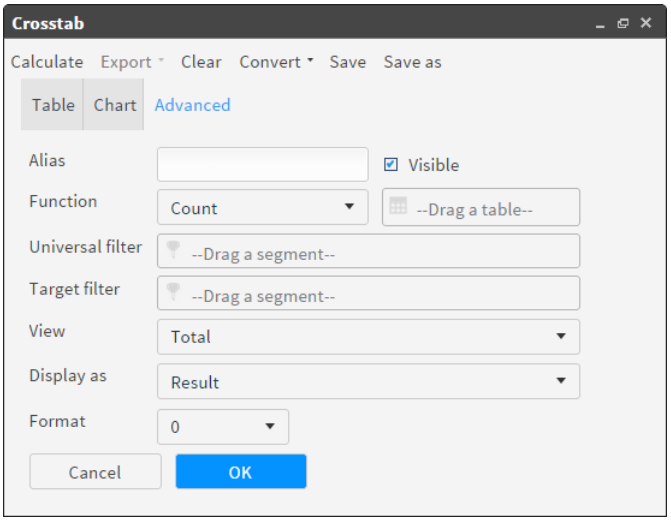


Figure 4-8 New Measure window in Crosstab

In the "New Formula" window (shown below in figure 4.8) you can:

- Name your measure in the "Alias" field (mandatory).
- Check or uncheck the "Visible" checkbox, depending on whether or not you want to display the results of your measure calculation in the "Table" view.
- Decide the number of decimal points to display using the dropdown list in the "Format" field.
- Choose the measure you want to operate on by clicking on it in the "Measures" panel on the left. (All measures that have been defined for this analysis will appear in this panel when you open the "New formula" window).
- Choose the advanced measures *operation* to be added to an existing operation from the sliding list in the functions (operators) field. All standard expressions and operations are available such as:
  - Mathematical functions (+-*/, floor, log10, exp)
  - Constants (mean, kurt)
  - Boolean conditionals (If (condition, true, false))
  - Date and time functions (daysto, age, year, month, secsto, time, datetime...)
  - Time functions (...)
  - Text functions (left, right, mid, replace...)
  - Data type conversion (date, string, integer)
- Build complex operations in the main workspace on the right.
- Save your new measure formula by clicking "OK". This also closes the "New formula" window. Click on it again if you have more new operations to add to your measure(s). (See Figure 4-9).

Note: *When two integers are divided, the result value is also and integer. If a real result is required, a real number should be used as a dividend or divisor. For example:*

- 100.0/3
- 100/3.0
- real (100)/3

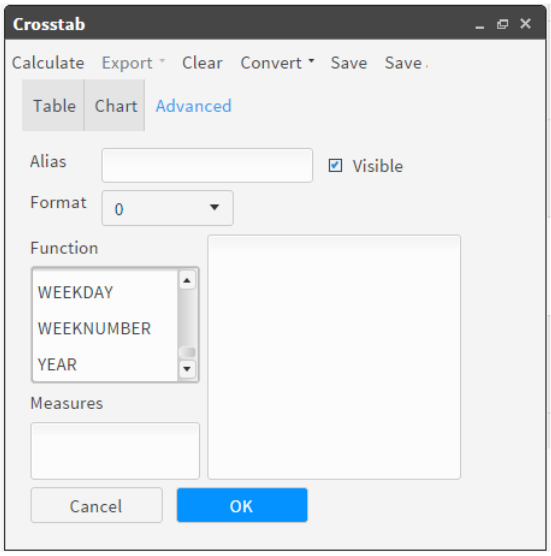


Figure 4-9 New Formula window

Duplicate Button

This button makes it possible to create a new measure based on an existing one (selected from the list of measures that appear in the Advanced "Measures" tab. It opens the "New measures" screen with its fields already populated with information concerning the selected measure. Enter the name for the new measure in the "Alias" field.

"Remove selected" Button

Just as its name indicates, this button lets you remove a selected measure from the Crosstab analysis.

Filters Tab

This tab opens a window used for applying filters to your analysis. Three types of filters are available: *Universal, Target* and *Baseline*.

You can also easily apply a filter to your analysis directly in the main Crosstab workspace by simply dragging and dropping a filter segment directly into the "Filter" field below the "rows" panel. For multiple or more complex filter additions, you need to use the *Filters* tab. (See Figure 4-10).

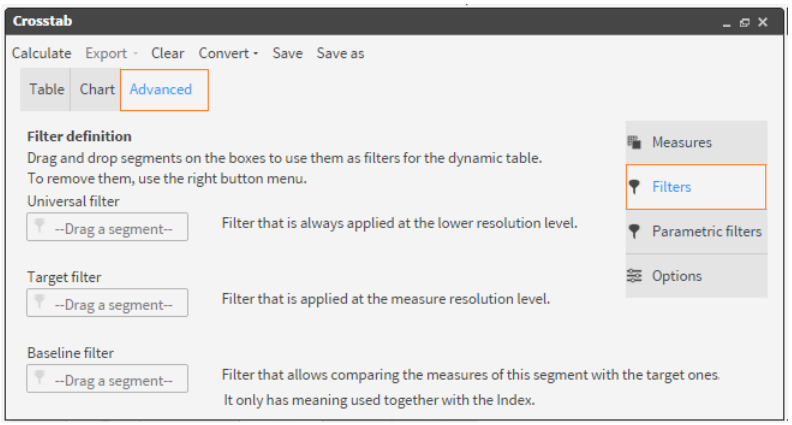


Figure 4-10 Filters tab in the Crosstab Advanced window

Using Filters

Filters are used throughout BIRT Analytics and are based on data segments. Choosing to use a Universal, Target or Baseline filter will depend on the situation. A Universal filter is applied to a non-pivoted analysis before any change of resolution occurs, whereas a Target filter is applied to a pivoted analysis after a change in resolution occurs.

Target and Baseline filters are used together to create comparative analyses. You must, of course, use segments that can be compared, such as comparing one year with another year or comparing one population group with another.

Understanding resolution change in Crosstabs

When building a Crosstab analysis requiring a change in resolution between dimensions (axes) and the measures in the direction N-to-1, intersection (or crossing) can take place in two different ways:

Type 1 (pivoted): with one variable in Rows and another one in Columns and with resolutions calculated separately before intersection takes place, as shown in Figure 4-11.

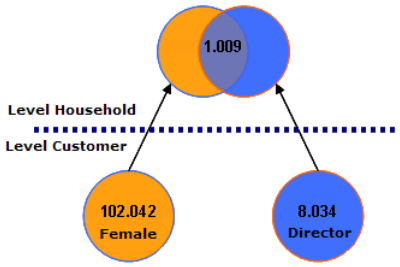


Figure 4-11 Pivoted crosstab intersection

Type 2 (all fields in rows): with 1 or more variables in Lines and none in columns and where intersection takes place before uploading the resolution, as shown in Figure 4-12.

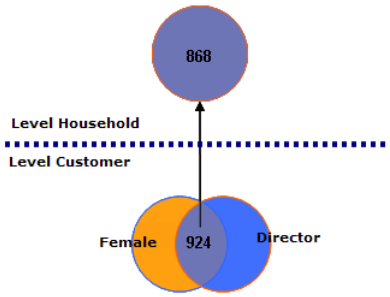


Figure 4-12 All fields in rows intersection

Parametric Filters Tab

This tab makes it possible to apply filters based on interactive selection of parameters when creating a Crosstab. The User interacts directly with the analysis during calculation by supplying values when prompted. Values in normal filters are pre-set before calculation, with no interaction possible.

Note: *You can choose to use either a pre-set filter or a prompted parametric filter when cross tabulating, but you cannot use both.*

Options Tab

The Options tab gives access to several different possibilities for building or displaying your analyses. Most are self-explanatory. The Autocalculate feature is selected by default. You will often need to turn it off. (See Figure 4-13).

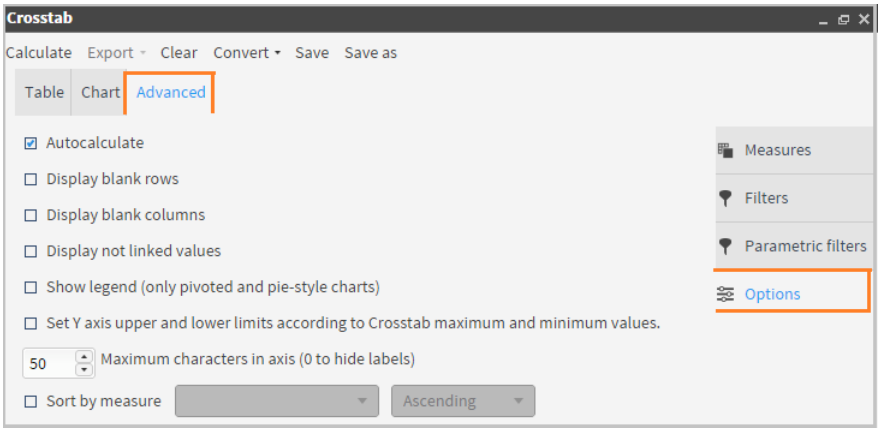


Figure 4-13 Options tab in the Crosstab Advanced view

Sample procedures for creating crosstabs

Our first sample shows you how to create a basic, filtered Crosstab analysis in a few seconds, doing everything in the default "Table" view. The second sample procedure shows you how to quickly create a simple pivoted Crosstab. The third and fourth procedures show you how to build more complex Crosstabs.

01 - How to create a very simple Crosstab Analysis

1 Go to the default "Table" view of the Crosstab analysis tool. Click on the Database icon directly above the Data Tree panel. The database symbol, along with the name of the database, appears in the panel. In this case we are using our "Demo" database. Click on the symbol to display the available tables in the database.
Note: *If you do not have access to our Demo database in your BA installation, you can use similar tables from your own company database to create this Crosstab.*

2 Then drag the "Occupation Decode" column from the "Customer" table in the Demo database tree and drop it into the "Rows" panel to start defining dimension rows. Now do the same thing with the "Gender Decode" column, dragging it into the "Rows" panel.

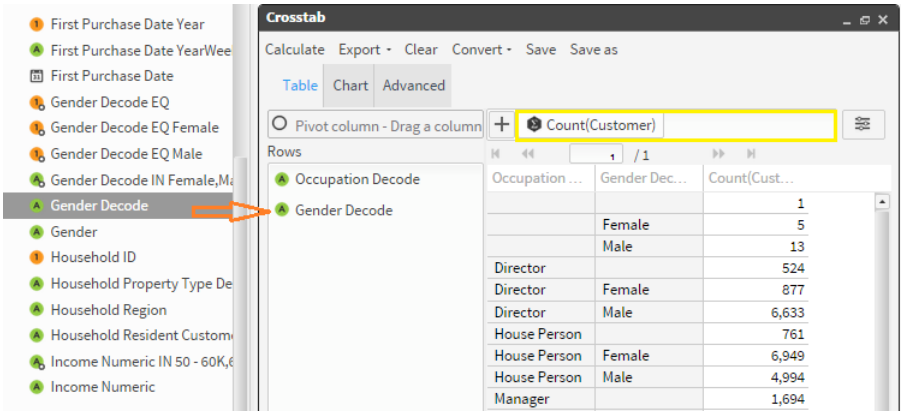
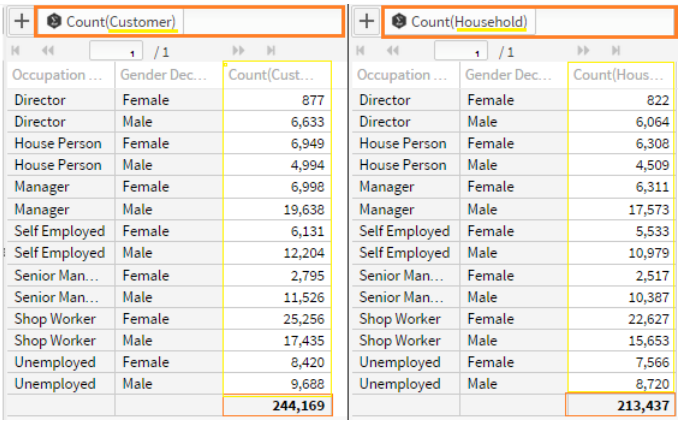


Figure 4-14 Table view of Crosstab results including Null values

3 See Figure 4-14 above. A default measure "Count(Customer)" now appears in the measure field in the Table view. It was set automatically when you defined your first dimension row. The measure is applied automatically on the dimension rows and the results are displayed in the main "Table" panel. They are grouped by occupation and gender based on the Customer table.

4 Remove the distracting Null value fields by clicking on “Occupation Decode” in the “Rows” panel to access the window for editing its discrete values. Uncheck the empty field at the top to remove the null values from your calculations and click “Accept”. Do the same thing to remove the null values from the “Gender Decode” dimension.

5 Click on the default measure and then choose “Remove”. While still in the main view, change the measure resolution by dragging the “Household” table from the Data Tree into the *measure* field where the Count(Customer) measure used to be displayed before we removed it. As the resolution is now “Household”, the results are now based on the “Households” table. (See Figure 4-15below). The results are not the same when counting *households* as they were when counting *customers*. This shows the importance of resolution changes.

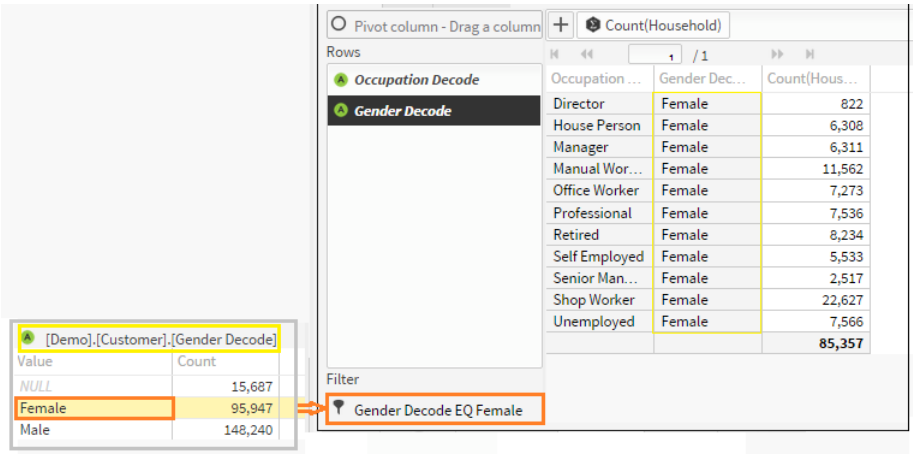


Occupation …	Gender Dec…	Count(Cust…
Director	Female	877
Director	Male	6,633
House Person	Female	6,949
House Person	Male	4,994
Manager	Female	6,998
Manager	Male	19,638
Self Employed	Female	6,131
Self Employed	Male	12,204
Senior Man…	Female	2,795
Senior Man…	Male	11,526
Shop Worker	Female	25,256
Shop Worker	Male	17,435
Unemployed	Female	8,420
Unemployed	Male	9,688
		244,169

Occupation …	Gender Dec…	Count(Hous…
Director	Female	822
Director	Male	6,064
House Person	Female	6,308
House Person	Male	4,509
Manager	Female	6,311
Manager	Male	17,573
Self Employed	Female	5,533
Self Employed	Male	10,979
Senior Man…	Female	2,517
Senior Man…	Male	10,387
Shop Worker	Female	22,627
Shop Worker	Male	15,653
Unemployed	Female	7,566
Unemployed	Male	8,720
		213,437

Figure 4-15 Comparing results

6 Now add a Universal filter to the analysis, (filtering by households with only females living in them) by dragging the “Female” segment from the “Gender Decode” discrete values accessed in the Data Tree and dropping it in the “Filter” field in the “Table” view. ***Note:*** *This filter has been applied before the resolution change (at the Customer level).* (See Figure 4-16 below for your results).



Occupation …	Gender Dec…	Count(Hous…
Director	Female	822
House Person	Female	6,308
Manager	Female	6,311
Manual Wor…	Female	11,562
Office Worker	Female	7,273
Professional	Female	7,536
Retired	Female	8,234
Self Employed	Female	5,533
Senior Man…	Female	2,517
Shop Worker	Female	22,627
Unemployed	Female	7,566
		85,357

Figure 4-16 Adding a Universal filter to the Crosstab analysis

7 This time apply the same filter after the resolution change (at the Household level) by adding a Target filter instead of the Universal filter. Do this by opening the “Filters” tab of the Advanced view and dragging our filter from the “Universal filters” field and dropping it into the “Target filters” field. Finally you need to remove the previously used Universal filter from its field by clicking on it and choosing “Clear”. (See Figure 4-17 below).

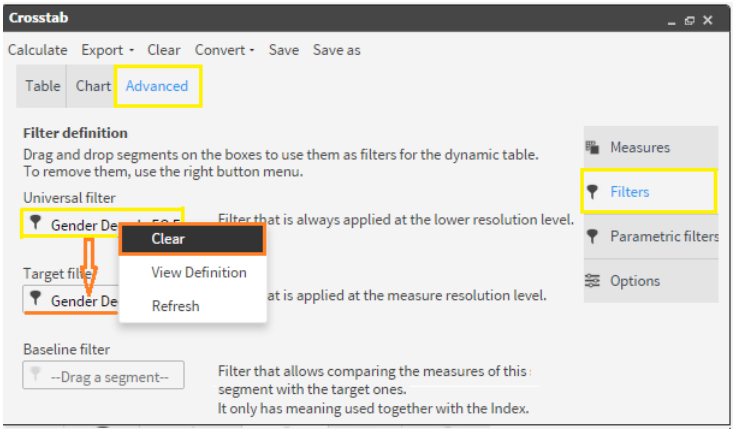


Figure 4-17 Replacing the Universal filter with a Target filter

8 Click "Calculate" in the Toolbar. Your new results are displayed in the Table view (See Figure 4-18 below).

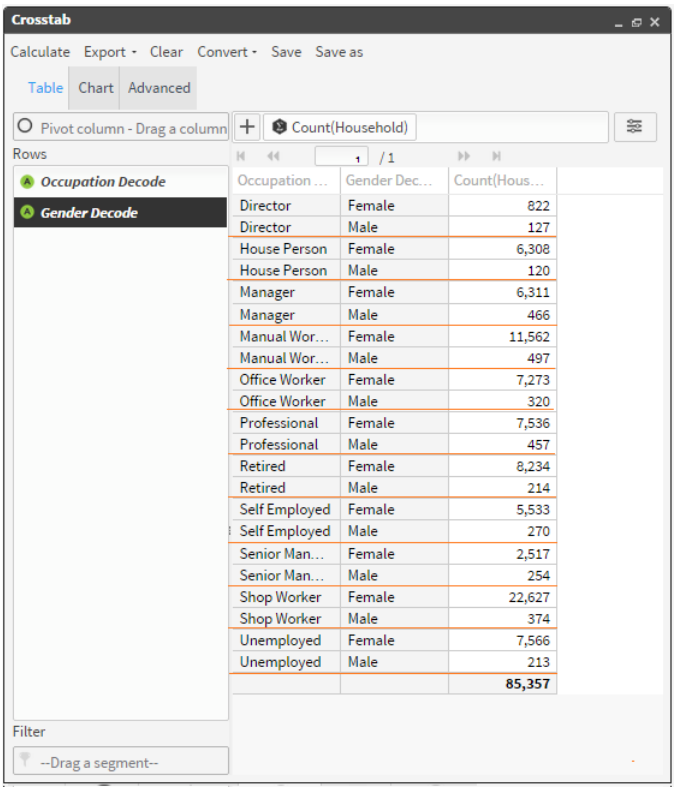


Figure 4-18 Table results after applying the Target filter

9 These results contain values that did not appear in the previous results where the filtering was done before the resolution change (Universal filter). Here, all the extra values grouped as "**Male**" represent households with males (having the designated occupation) AND where ***at least one female*** lives.

10 Open the "Chart" view to see the graphic display of the results (See Figure 4-19 and Figure 4-20 below) that show you the graphic results for both filtering situations

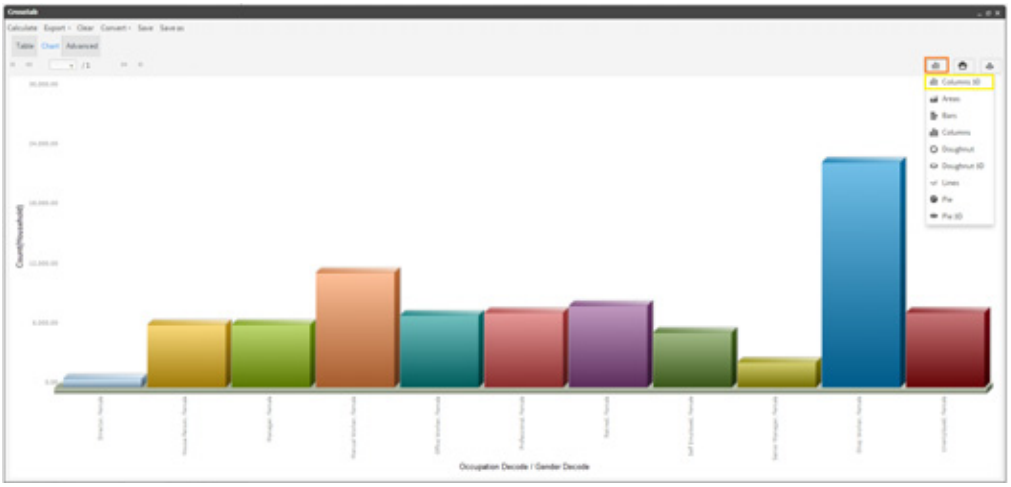


Figure 4-19 Chart view with Universal filter (applied before resolution change

Note: *We changed the type of chart display to "Column 3D" by choosing it in the "Chart" icon dropdown list.*

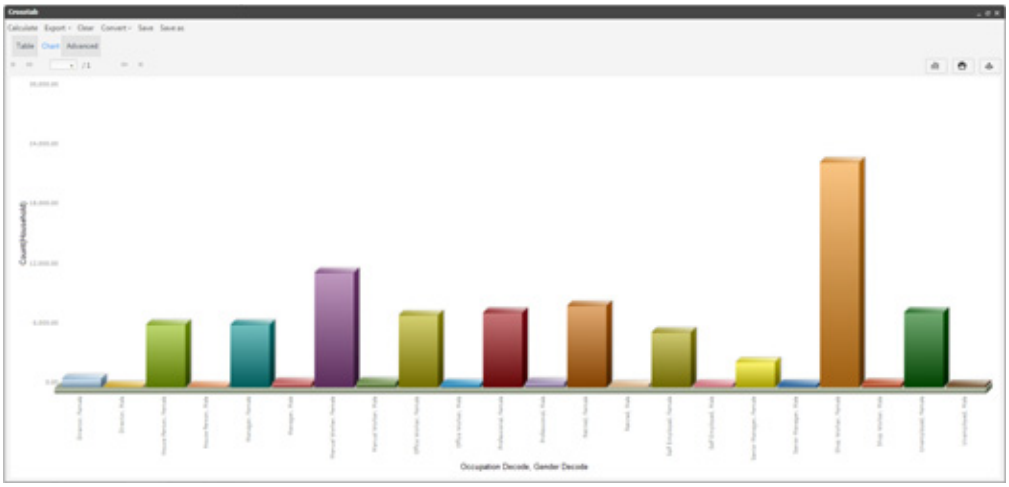


Figure 4-20 Chart view with Target filter (applied after resolution change).

You have now finished your analysis. Several possibilities are available to you, using the various tabs and icons on the Crosstab main screen, such as saving, exporting or even converting it to another type of analysis.

Note: *It is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option "Analytic DB" from the dropdown list of the "Export" tool found in the Crosstab toolbar.*

02 - How to create a simple Pivoted Crosstab Analysis

Here we will show how to make a pivoted Crosstab analysis showing a company's average profit from customers in each county in California, grouped by gender.

1 Open the main analysis window and click on "Crosstab" which opens the Crosstab Table view (by default).

2 Click on the Database icon directly above the Data Tree panel. The database symbol, along with the name of the database, appears in the panel. In this case we are using our "Demo" database. Click on the symbol to display the available tables in the database.
Note: *If you do not have access to our Demo database in your BA installation, you can use similar tables from your own company database to create this Crosstab.*

3 Define the first dimension by dragging the “County” column into the Row panel next to the Data Tree. Setting the first dimension row, using a column from the “Household” table, *automatically sets a measure* on that table with the default Count operator (See Figure 4-21 below).

Note: *Because the “Autocalculate” option is turned on by default, the counted Product Group results automatically appear in the main panel of the Crosstab Table view.*

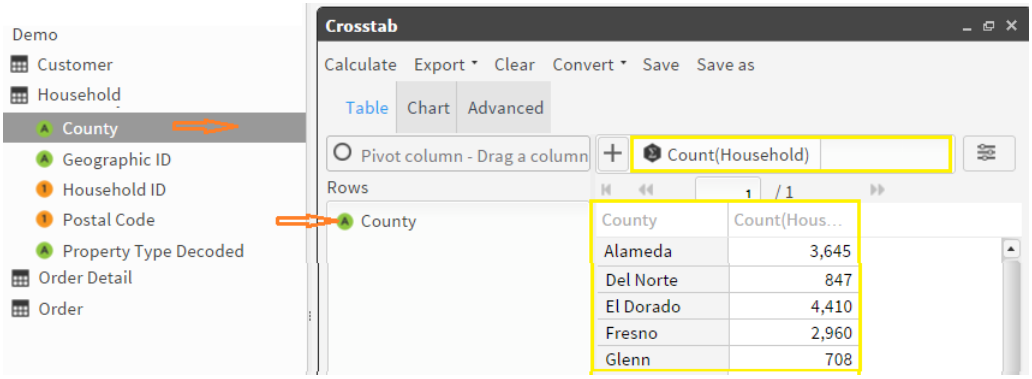


Figure 4-21 Defining a Dimension row

4 Next we define the second dimension by dragging the “Gender Decode” column into the “Pivot column” field at the top of the Table view. (See Figure 4-22 below). **Note:** *Setting the second dimension does not change the default “Count” measure that was set on the “Household” table when we created the “County” dimension row.*

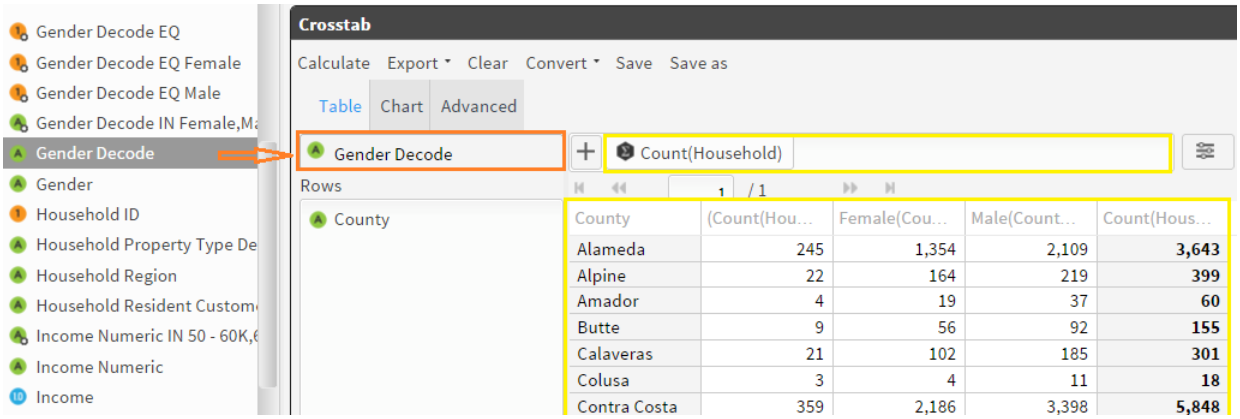


Figure 4-22 Defining a Dimension column (for pivoting)

5 While still in the “Tab” view, eliminate the default measure by clicking on it and choosing “Remove” in the dropdown menu that appears. Then click the [+] icon beside the “Measure” field. (See Figure 4-23 below).

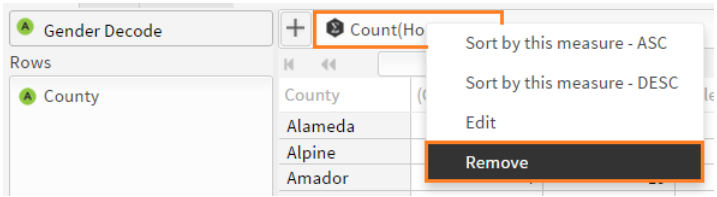


Figure 4-23 Removing the default measure

6 In the “New Measures” window that opens, it is mandatory to name your new measure. Because you wish to calculate the average profit per customer, enter “Mean profit” in the “Alias” field.

7 Next select the operator you want to use from the dropdown list in the “Function” field. Here you choose “Mean” because you want to calculate average profit. (See Figure 4-24 below).

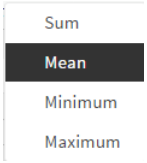


Figure 4-24 Choice of measure operator

8 The final results appear immediately in the “Table” view. Click on “Chart” to view them graphically. They appear as columns, the default view.
(See Figure 4-25 and Figure 4-26 below).

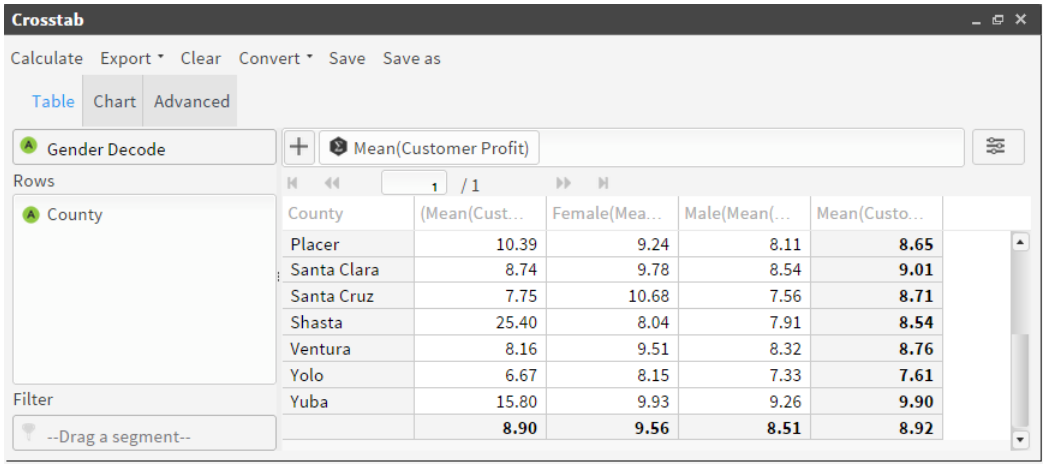


Figure 4-25 Table view of results

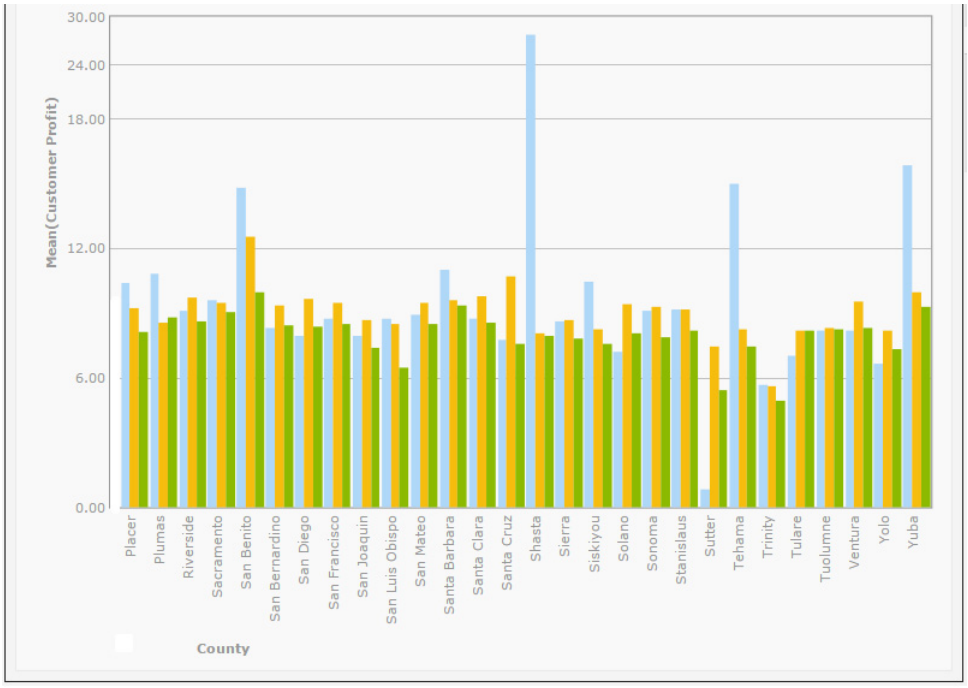


Figure 4-26 Chart view of results

9 You have now finished your analysis. Several possibilities are available to you, using the various tabs and icons on the Crosstab main screen, such as saving, exporting or even converting it to another type of analysis.

Note: It is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option "Analytic DB" from the dropdown list of the "Export" tool found in the Crosstab toolbar.

03 - How to create a Non-Pivoted Crosstab Analysis

The following steps will show you how to create a more complex Crosstab for an international company that will allow management to analyze their sales results in southern California. Our Crosstab analysis will find the **sales figures per product** as well as their **average profit from the sale of each product** in order to use these measurements to calculate the **total profit by product group in southern California.**

1 Open the main analysis window and click on "Crosstab" which opens the Crosstab Table view (by default).

2 Expand the database tables in the Data Tree. Define a dimension row by dragging the "Product Group" column into the Row panel next to the Data Tree. Setting the first dimension row, using a column from the "Order Detail table", automatically sets a measure on that table with the default Count operator.

Because the "Autocalculate" option is turned on by default, the counted Product Group results appear in the main panel of the Crosstab Table view. (See Figure 4-27 below).

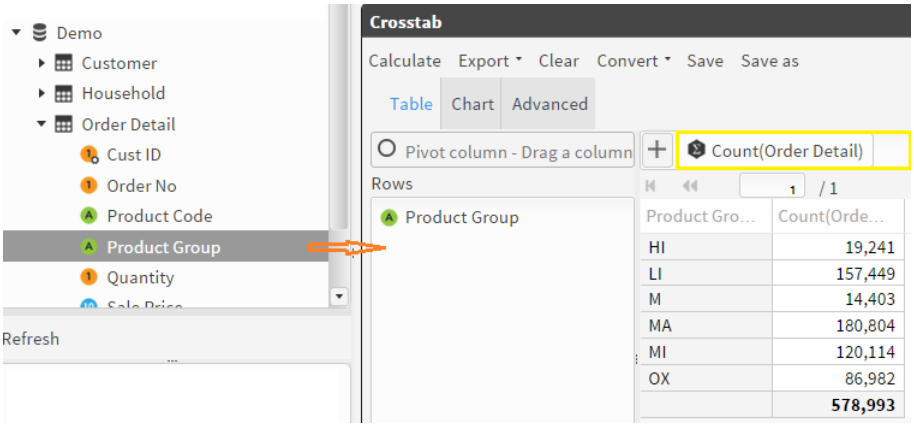


Figure 4-27 Creating the first dimension

3 **Step 3:** Edit the default measure by clicking on it in the *Measure* field and choosing "Edit" in the dropdown list that appears. (See Figure 4-28 below).

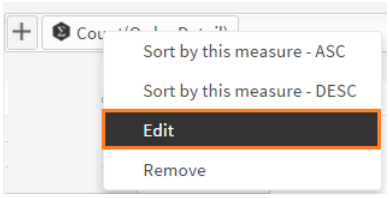


Figure 4-28 Measure field drop-down Options list

4 The "New Measures" window opens displaying the default *Count(OrderDetail)* measure ready for editing. In this case, change the name (Alias) to "Sales number". Then click "OK". (See Figure 4-29 below).

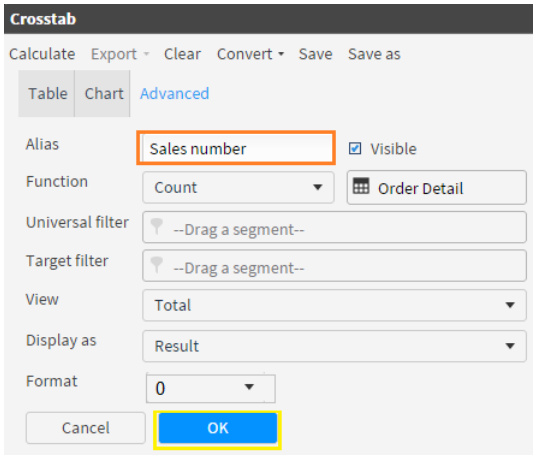


Figure 4-29 New Measure screen

5 Go to the default Measures screen in the "Advanced" view tab. It now displays your first measure with its new name "Sales number". Click the "New Measure" button at the bottom of the Measures tab to open the New Measures screen again, ready for you to add your second measure to the analysis. (See Figure 4-30 below).

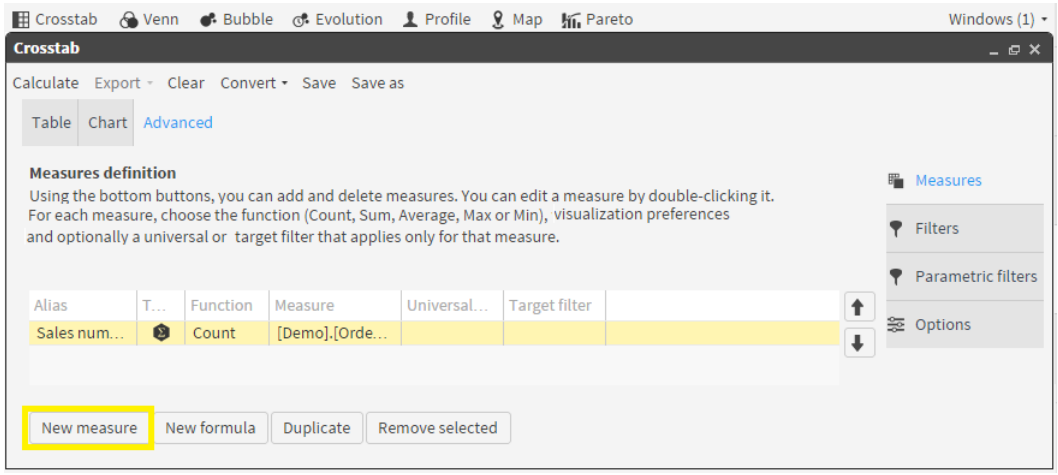


Figure 4-30 New Measure button

6 Our second measure will be used to calculate the *average profit from the sale of each product.* Start by naming the measure "Average profit". Then select the "MEAN" operator in the "Function" field. (See Figure 4-31 below).

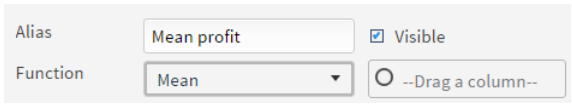


Figure 4-31 Creating a measure

7 "Drag a column" appears (see figure above) inviting you to drag the column to be operated on by the measure into the field next to the Function field.
Drag the "Line profit" column, from the "Order Detail" table in the Data Tree, into the "Drag a column" field. (See Figure 4-32 below).

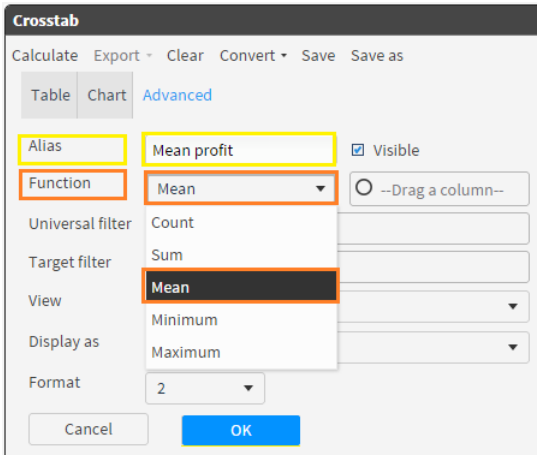


Figure 4-32 Setting the second measure in “Drag a column” field

8 Complete your measure by setting the number of decimal places to “2” in the “Format” field (near the bottom of the window) and click “OK” to save your measure and go back to the *Measures* tab in the *Advanced* view. See Figure 4.29 above.

9 Back in the Measures tab, where both your measures are now displayed, select the Mean profit measure and click on the “New Formula” button.

10 In the formula window that opens you will need to create a further calculation formula for your analysis (based on the Mean profit measure that you have just created). This is necessary to be able to obtain the total profit results. Name your formula “Total profit”.

11 Drag the “sales number” measure from the “Measures” field into the blank work area located on the right of the measures box, followed immediately by the “Mean profit” measure. Now place the multiplication symbol ”*” between the two measures being careful to stay outside of the brackets: [sales number]*[Mean profit]. (See Figure 4-33 below).

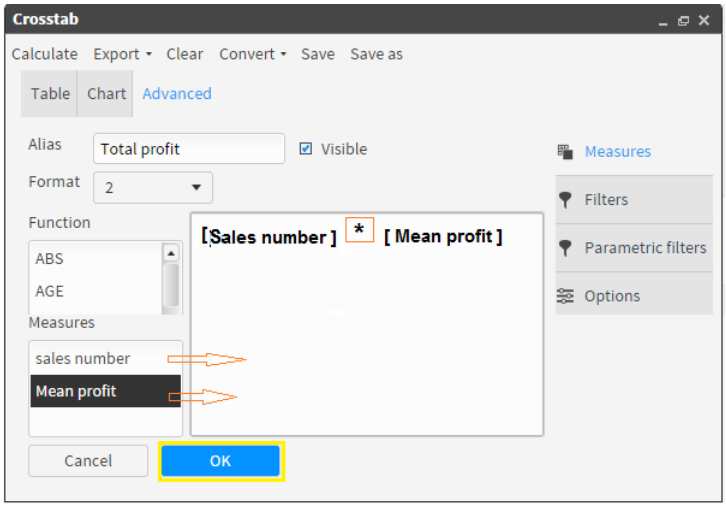


Figure 4-33 New Formula window

12 Set the “Format” field to “2” decimal places and click “OK” to save the calculated measure and to return to the Advanced view of your Crosstab.

13 Back in the "Advanced" view, open the "Filters" tab. Here you must drag the segment "Southern California" in to the "Universal" filter field. ***Note:*** *You obtain the desired segment in the Data Tree by opening the discrete values of the "Region" column located in the "Household" table of the "Demo" database.* (See Figure 4-34 below).

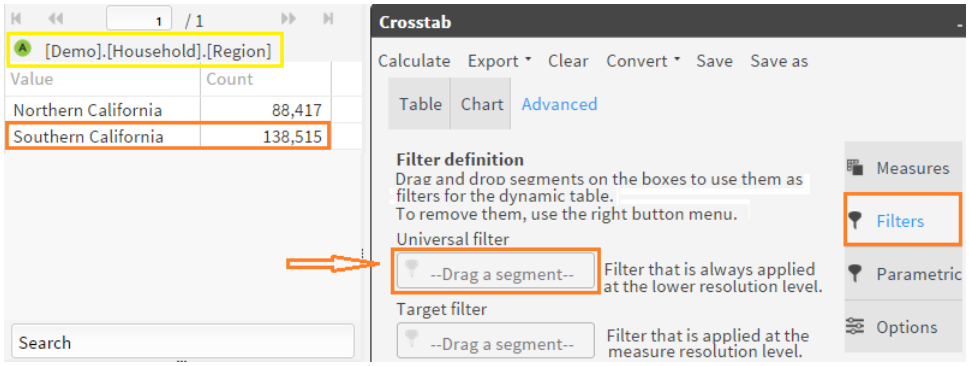


Figure 4-34 Creating a Universal filter

14 Click "OK" to save your filter. Go back to the Crosstab "Table" view to see your completed Crosstab analysis as a table or open the Chart view to see it graphically displayed. (See Figure 4-35 and Figure 4-36 below).

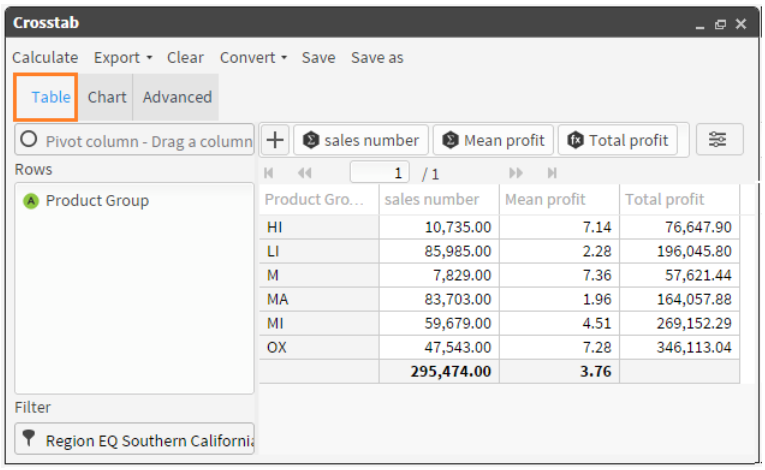


Figure 4-35 Table view of results

Figure 4-36 Chart view of results

You have now finished your analysis. Several possibilities are available to you, using the various tabs and icons on the Crosstab main screen, such as saving, exporting or even converting it to another type of analysis.

***Note:** It is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option "Analytic DB" from the dropdown list of the "Export" tool found in the Crosstab toolbar.*

04 - How to make Comparisons using a Crosstab Analysis

Here you build a crosstab analysis to determine a purchasing trend based on the difference between the number of orders placed in 2003 and 2004 crossed by occupation and gender.

1 Go to the main window of the Analysis "Crosstabs" tool. Click on the "Options" icon on the upper right and uncheck the "Autocalculate" box shown below in Figure 4-37.

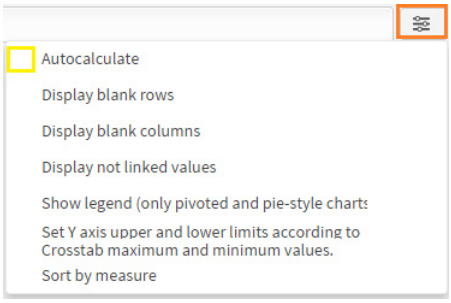


Figure 4-37 Autocalulate option (de-activated)

2 Create the first dimension by dragging the "Gender Decode" column (in the customer table) into the *pivot column* field as shown below in Figure 4-38.

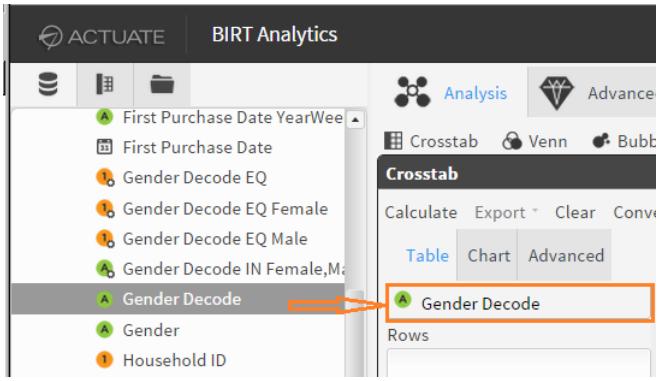


Figure 4-38 Setting the first dimension - Pivot column (Y-axis)

3 Create a second dimension by dragging the "Occupation Decode" column into the *rows* panel as shown below in Figure 4-39.

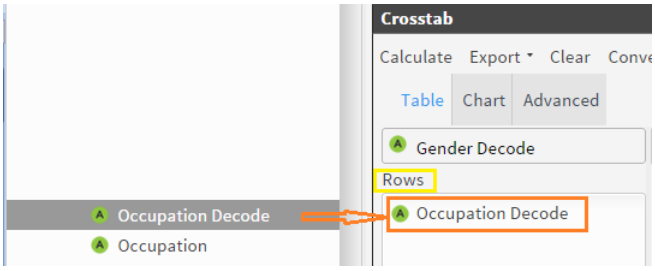


Figure 4-39 Setting the second dimension - Dimension Row (X-axis)

4 Remove the default measure "Count(Customer)" that was created when setting the first dimension. Click on it and choose "remove" from the dropdown list that appears (as shown below in Figure 4-40).

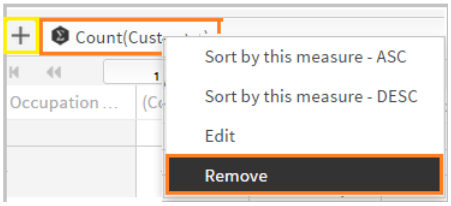


Figure 4-40 Removing the default measure

5 While still in the main "Table" view window, click on the **[+]** icon next to the *measure* field to create a new measure in the "Advanced" tab "New measure" window that opens. (See Figure 4-41). In this case the fields are already populated for the new measure that we are about to create in the next 4 steps.

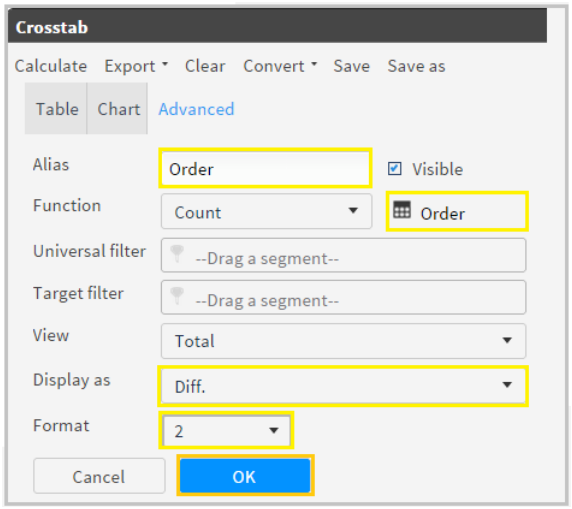


Figure 4-41 New Measure window with defined "Order" measure

6 Enter the measure name "Order" in the "Alias" field (mandatory field).Then Drag the "Order" table into the field next to the "Alias".(This changes the resolution of the measure compared to the previous default measure).

7 Click on the arrow in the "Display as" field and choose "Diff" in the dropdown list that appears. **Note:** This will display the difference in size (in units) between the Base and Target filters that we are soon going to create.

8 Click "OK" to save your measure. This opens the "Advanced' view tab that now displays your new measure.(See Figure 4-42 below).

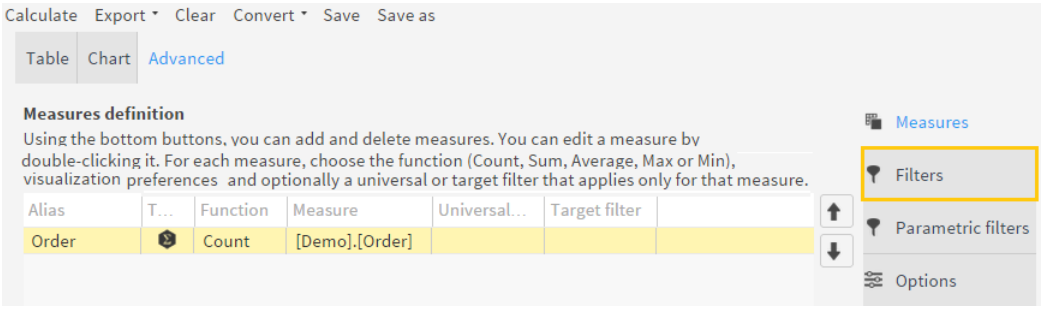


Figure 4-42 Advanced view Measures tab

9 Now click on the "Filters" tab in the "Advanced" tab. This opens the "Filter definition" window. (See Figure 4-43 below). ***Note:*** *Defining filters requires finding or creating the necessary segments to be dragged into the chosen "Filter" fields in the "Filter definition" window.*

In this case we will need to use the "Range selection" tool in order to create 2 segments – one for 2003 and the other for 2004.

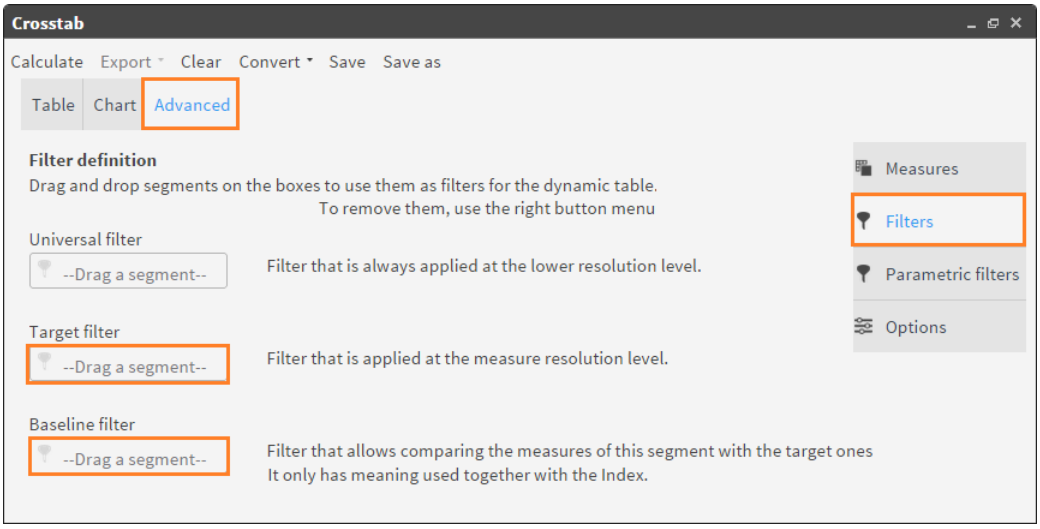


Figure 4-43 Filter Definition window

10 Make a right-click on the "Order Date Year" column in the "Order" table and choose "Select" in the dropdown list that opens. This opens the "Select" window where you can specify ranges based on years.

Set the range in this window to go from "Between" 2004 "and" 2004, as shown above. Click "OK" to close the "Selection" window and send the new range segment (year 2004) into the Scratchpad. (See Figure 4-44 below).

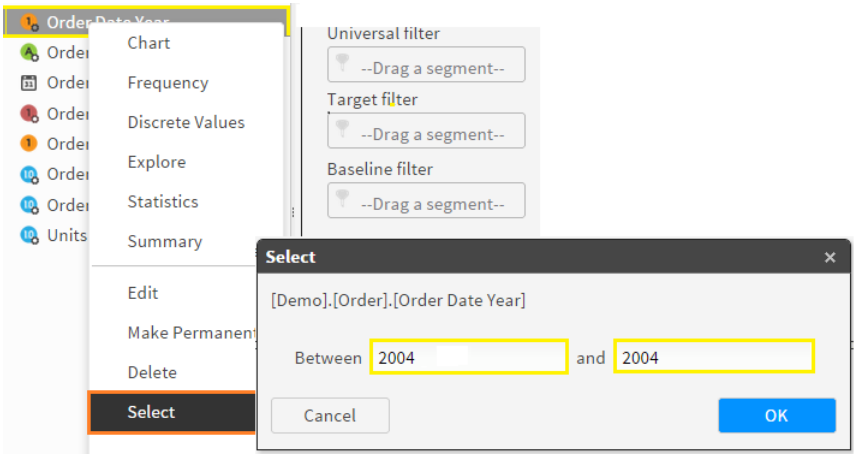


Figure 4-44 Preparing filter segments using the "Select" ranges tool

11 Drag this segment (year 2004) from the Scratchpad into the "Target" filter field in the "Filter creation" window. (See Figure 4-45 below).

12 Repeat the operations done in steps 10 and 11 to create your Base filter. This time setting your range from "Between" 2003 "and" 2003. Then click "OK" to close the window and send the new range segment to the Scratchpad.

13 Drag this second segment (year 2003) into the "Base" filter field in the "Filter definition window. (See Figure 4-45).

14 Now that both filters have been specified, click "Calculate" in the toolbar at the top of the window. The results are now displayed in the Table view. (See Figure 4-46below).

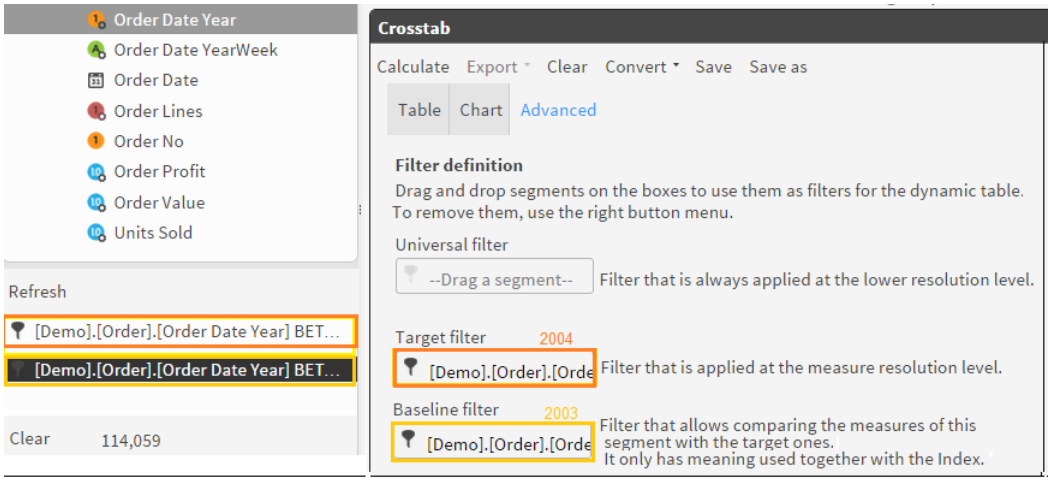
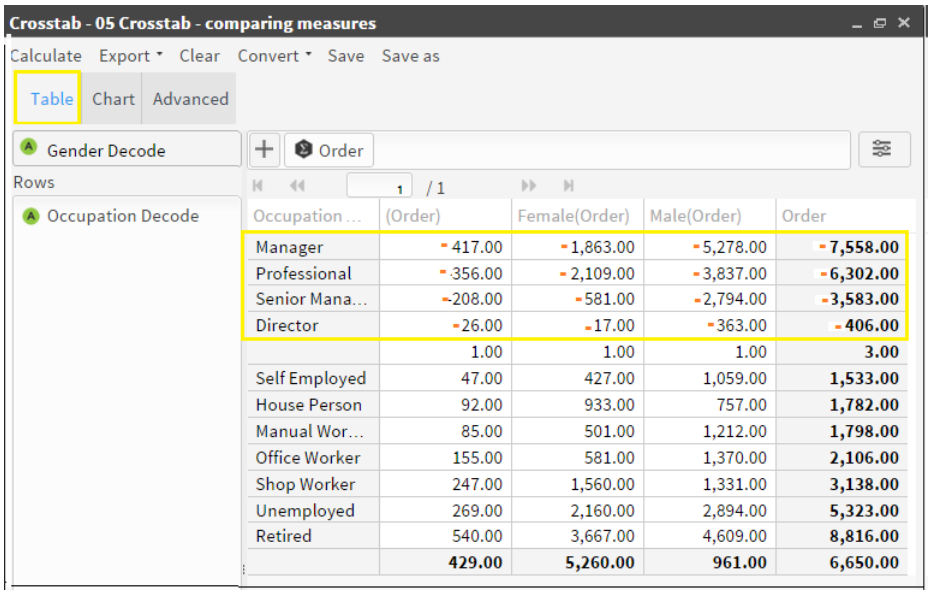


Figure 4-45 Dragging the year segments into the corresponding filter fields

Note: *There are several negative results. A negative result indicates a reduction in orders placed in 2004 (Target filter) compared with orders placed in 2003 (Base filter).*



Occupation ...	(Order)	Female(Order)	Male(Order)	Order
Manager	-417.00	-1,863.00	-5,278.00	**-7,558.00**
Professional	-356.00	-2,109.00	-3,837.00	**-6,302.00**
Senior Mana...	-208.00	-581.00	-2,794.00	**-3,583.00**
Director	-26.00	-17.00	-363.00	**-406.00**
	1.00	1.00	1.00	**3.00**
Self Employed	47.00	427.00	1,059.00	**1,533.00**
House Person	92.00	933.00	757.00	**1,782.00**
Manual Wor...	85.00	501.00	1,212.00	**1,798.00**
Office Worker	155.00	581.00	1,370.00	**2,106.00**
Shop Worker	247.00	1,560.00	1,331.00	**3,138.00**
Unemployed	269.00	2,160.00	2,894.00	**5,323.00**
Retired	540.00	3,667.00	4,609.00	**8,816.00**
	429.00	**5,260.00**	**961.00**	**6,650.00**

Figure 4-46 Table view of your calculated analysis

15 Click on the "Chart" view to see a graphic display of the results, shown as columns (default chart). If you want the results displayed in a different type of chart, click on the "Charts" icon and choose your chart in the dropdown list. (See Figure 4-47 below).

16 You have now finished your analysis. Several possibilities are available to you, using the various tabs and icons on the Crosstab main screen, such as saving, exporting or even converting it to another type of analysis.

Note: ***I****t is also possible to export a Crosstab analysis directly to the FastDB engine, creating a new table in the database. This is done by selecting the new option "Analytic DB" from the dropdown list of the "Export" tool found in the Crosstab toolbar.*

05 - How to change the type of analysis results to be displayed

You can easily change the way your information is displayed in an analysis by simply changing the "Result" parameter in a defined measure.

Figure 4-47 Chart view of your analysis

1. Open the Crosstab analysis that you have just created above and click on the defined measure in the "Measure" field in the Crosstab "Table" view and choose "Edit" from the dropdown list. This opens the measure in the "Measures definition" window.
2. Click on the "Result" box at the bottom and choose "Index" from the dropdown list that appears. This will display the difference between the compared groups. (Target/Total) / (Baseline/Total).
3. Click "OK" and then click on "Calculate" in the toolbar at the top of the window. This opens the "Table" view where your results are now displayed very differently with no negative values. (See Figure 4-48).

Table | Chart | Advanced

Gender Decode | Order

Rows: Occupation Decode

Filter: --Drag a segment--

1 / 1

Occupation ...	(Order)	Female(Order)	Male(Order)	Order
Manager	0.56	0.54	0.53	**0.54**
Senior Mana...	0.56	0.62	0.55	**0.57**
Professional	0.59	0.52	0.59	**0.57**
Director	0.82	0.89	0.81	**0.82**
Manual Wor...	1.06	1.05	1.10	**1.08**
Shop Worker	1.23	1.14	1.17	**1.16**
Self Employed	1.05	1.14	1.21	**1.18**
Office Worker	1.27	1.16	1.22	**1.20**
House Person	1.42	1.51	1.59	**1.54**
Unemployed	1.48	1.53	1.69	**1.60**
	0.00	1.88	1.41	**1.88**
Retired	2.20	1.92	2.28	**2.10**
	1.00	**1.00**	**1.00**	**1.00**

Figure 4-48 New Table results with no negative values

4. Go back to the Chart view and click on the chart icon to choose 3D columns for your display. (See Figure 4-49 and Figure 4-50 below). Your chart appears as a 3D column display.

   Note: *The way the results are displayed now, values > 1.0 indicate that more orders were placed in 2004 than in 2003.*

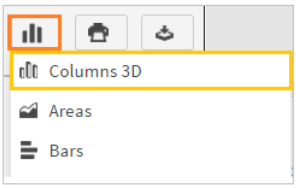


Figure 4-49 Choosing your Chart display

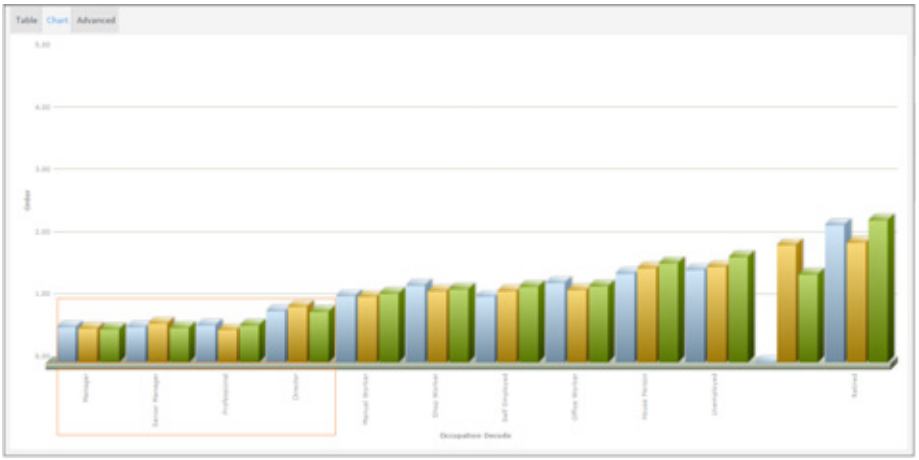


Figure 4-50 New 3D column Chart display of your analysis results

Using Venn diagrams

A Venn diagram provides an analysis of data by crossing different fields. A Venn analysis identifies coincidences and differences between N segments. Up to five different segments appear on the graphic display in Chart.

On Advanced, in addition to the resolution table, you can set Filter and Format.

- Filter
  You can include a filter at the analysis level, which means that a Venn diagram can be constrained to show results for a specific and smaller universe. For example, include the customers who have purchased one, two, or three product types, but only in Madrid.
- Format
  You can format your results to show a count of values that match the criteria you selected, a percentage of the total number of records included the analysis, or a global percentage, which means that each percentage is based on the total possible records from the resolution table being analyzed.

01 How to create a Venn diagram analysis

1. Drag a segment from Discrete Values or from Scratchpad and drop it in Chart in Venn, as shown in Figure 4-51 below.
2. Drag additional segments to Chart, up to a maximum of five. The Venn diagram is drawn as you drop each item into Chart.

   For example, *to identify cross-selling opportunities, examine customers who have purchased different categories of products.*

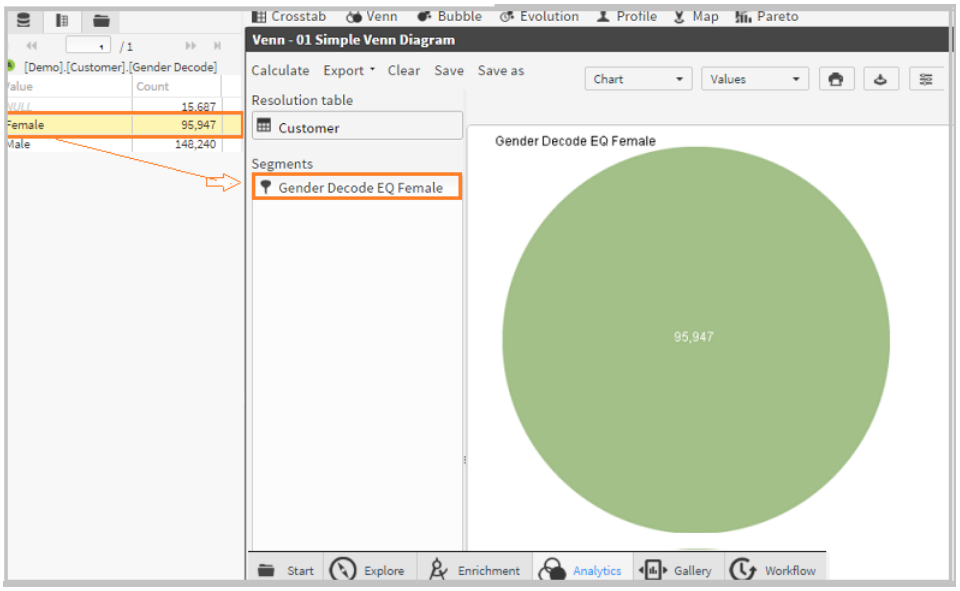


Figure 4-51 Dragging a discrete value to a Venn analysis

3 Using the Venn analysis, cross the data segments representing three different product categories. Zero coincidences occur, because a product is of one type or another, but not both, as shown in Figure 4-52.

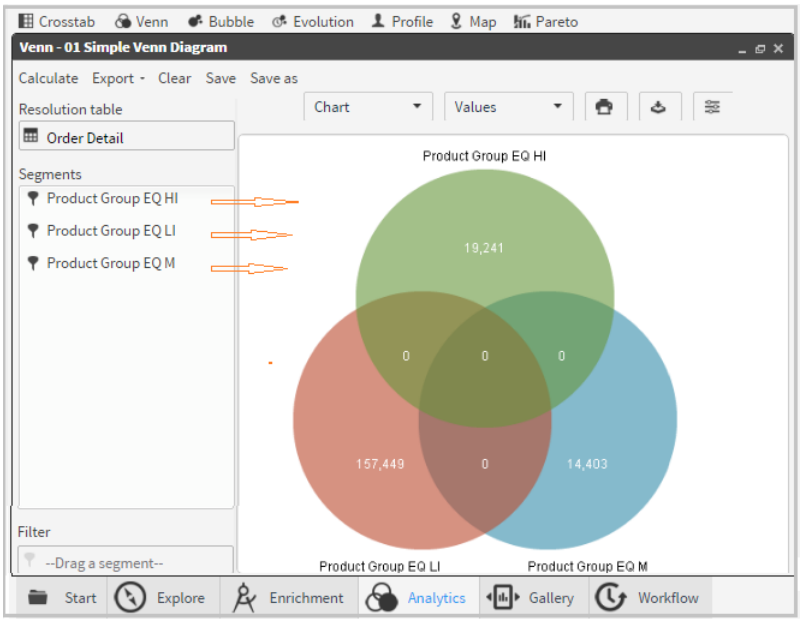


Figure 4-52 Showing three discrete values with no coincidences

02 How to change the resolution of a Venn diagram analysis

A Venn diagram can show coincidences or differences at other levels of database resolution. For example, resolving the Venn diagram at the customer level shows coincidences between the two or three segments, including customers purchasing one, two, or three products.

1 Change the resolution by dragging a different table to Resolution Table, as shown in Figure 4-53.

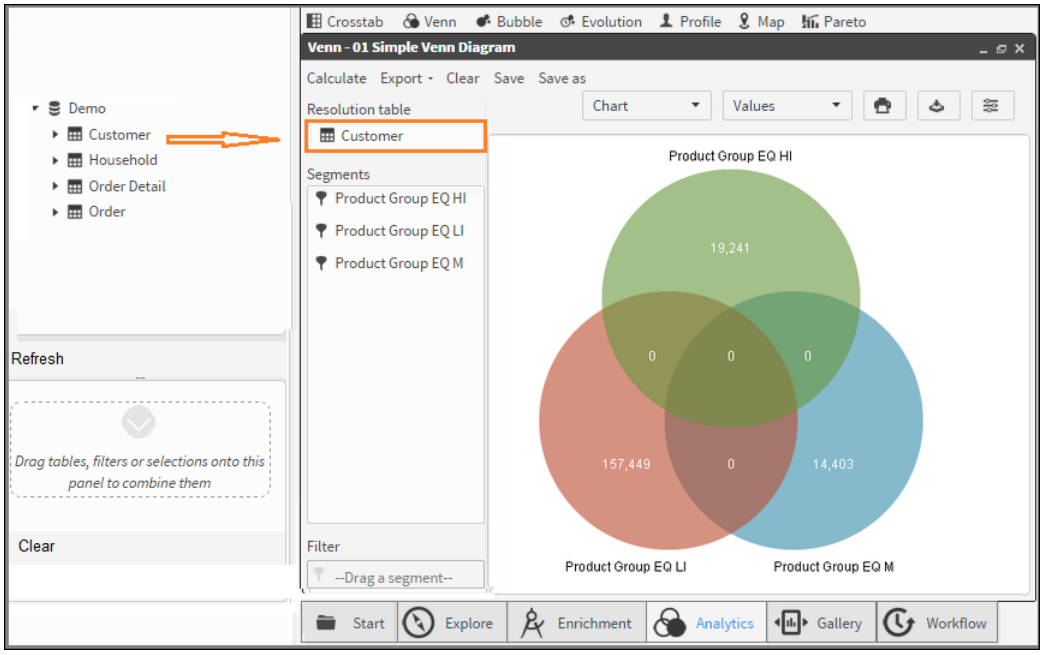


Figure 4-53 Changing the Venn diagram's resolution

2 Choose Calculate to redraw the chart after changes. The resulting chart is shown in Figure 4-54.

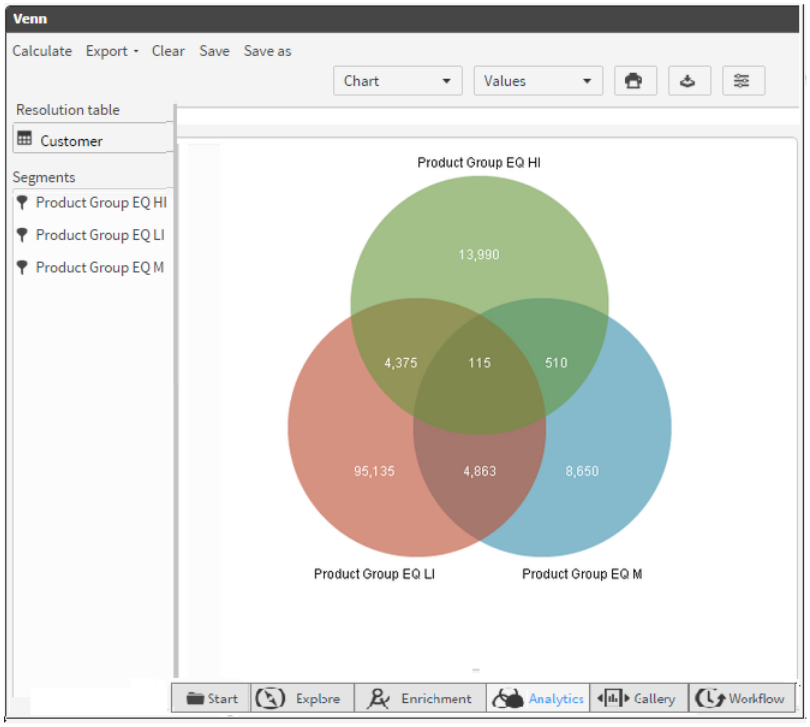


Figure 4-54 Showing a Venn diagram with a changed resolution table

Using Bubble analyses

A bubble diagram shows the distribution of categorical data across two axes of numeric variables. You can also view results on a data table.

Use a third variable to set the size of each bubble. To configure a bubble analysis, set two or three variables: two on the axes and an optional third to size the bubbles.

Instead of showing measures by attributes or categorical variables, as in the case of crosstabs, a bubble diagram shows groups or categories according to numeric variables, providing a different way of showing numeric data versus categorical data. For example, you can examine customers by age, order amounts, and average salary.

Drag a field to use as a categorical variable, such as profession, gender, age ranges, or number of children. Next, select continuous or numeric variables, such as age, salary, or purchase amount. The continuous variables are the axes of the analysis. For each variable, choose a function, such as count, sum, average, maximum, minimum, and a display type, such as percentages or totals. You may optionally choose a universal or target filter that applies only to this variable or measure.

You can set filters on a bubble analysis at either the measure level or at the complete analysis level. For example, you can set a partial filter when you are analyzing customers by age, and you need to view the order average for only the last month. You also, however, want to check the average salary for these age ranges, so you do not want the filter to be applied at the global-analysis level, only at the order-measure level.

You can use parametric filters as analysis filters.

Results can be sorted in ascending or descending order by measures.

The check box Show legend is only available for pivoted and pie-style charts, but there is also an option to set the maximum number of characters to display in chart labels. Hide chart labels by setting the maximum number of characters to 0.

The legend configuration is bound to the chart configuration, so if the chart configuration (type, rows, columns, or filters) changes, the legend changes. When you select a pie-style chart, clicking a value label separates the corresponding piece of the pie, but when the bubble is multidimensional, clicking a value label hides its graphical representation.

You can convert this type of analysis into a crosstab analysis.

01 How to create a Bubble Analysis

1. Open the main analysis window and click on “Bubble” which opens the Bubble Table view (by default).

2. Click on the Database icon directly above the Data Tree panel. The database symbol, along with the name of the database, appears in the panel. In this case we are using our “Demo” database. Click on the symbol to display the available tables in the database.
   Note: *If you do not have access to our Demo database in your BA Analysis installation, you can use similar tables from your own company database to create this Crosstab.*

3. Define the *categorical* continuous variable by dragging the “Age Numeric” column from the Data Tree into the “Row” panel in the main Bubble window. A message that **you must specify two continuous variables to calculate the analysis** appears, (as shown in Figure 4-55). Click “OK” to continue defining your variables. ***Note:*** *Defining your first variable sets your first measure based on the table that contains the chosen column and on the default “count” operator. In this case “Count(Customer)”. It appears in the field beside the* ***[+]*** *icon, as you can see in Figure 4-56 above.*

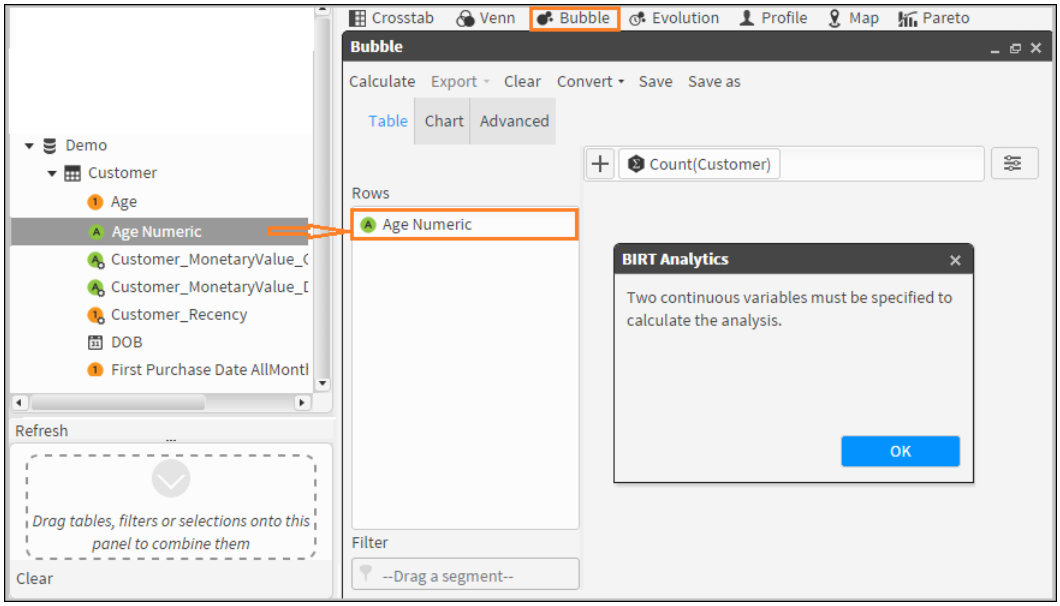


Figure 4-55 Defining a continuous variable for a Bubble analysis

4 Next we define the **second continuous variable** by dragging the "Income" column into the main panel of the Bubble Table view. A pop-up list of operators appears. Choose "Sum". (See Figure 4-56 below).

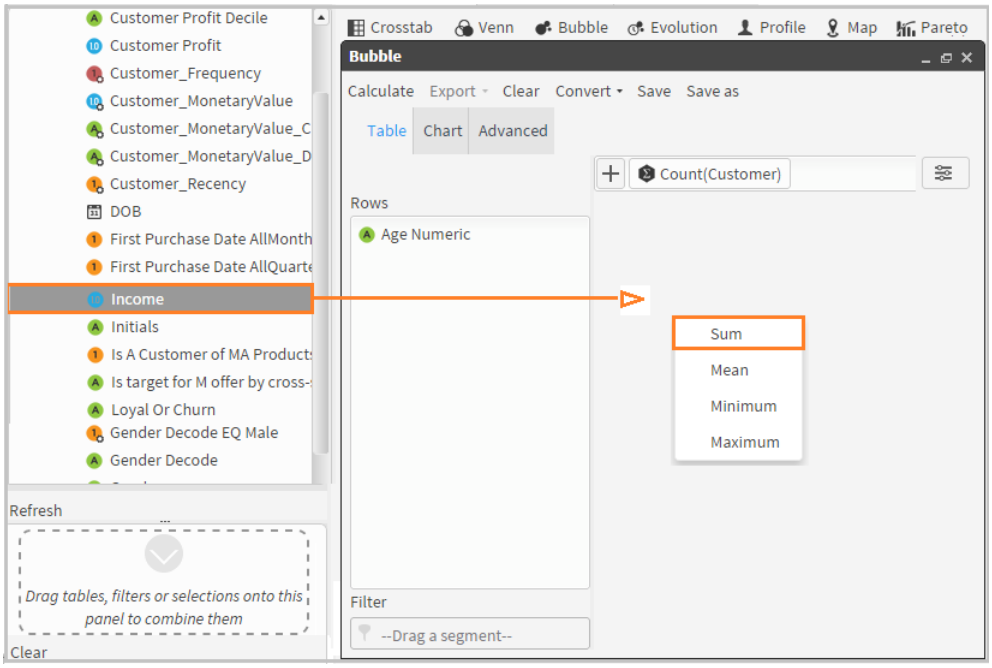


Figure 4-56 Defining the second continuous variable for a bubble analysis

5 Defining your second variable sets a second measure, also based on the Customer table and the choice of your operator "Sum". The new measure appears next to the first measure in the *Measure* field of the Bubble Table view. ***Note:*** *Because the "Autocalculate" option is turned on by default, the calculated results of the customer count (grouped by age) and the corresponding total incomes now appear in the Bubble Table view.* (See Figure 4-57 below).

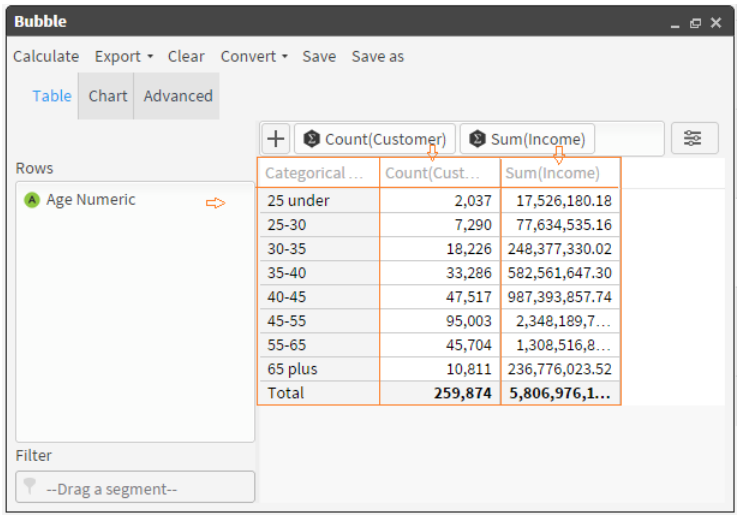


Figure 4-57 Calculated results in the Table view

6 It is mandatory to name your new measure. Because you wish to calculate the average order values per customer age group, enter "Average Order Value" in the "Alias" field and choose "Mean" in the *functions* dropdown list below the "Alias" field.

7 To set the third continuous variable, open the "New Measure" window by clicking on the [+] icon to create a new measure based on order values. The results of this third measure will determine the size of the bubbles in the analysis chart.

8 Drag the "Order value" column from the "Order" table in the Data Tree and drop it in the *Drag a column* field beside the *functions* field. Set the number of decimal places to 2 using the "Format" field dropdown list. Click "OK". (See the settings in Figure 4-58 below).
Note: *the "Visible" box has been checked by default. If it is unchecked, the results will be calculated but they of will not appear.*

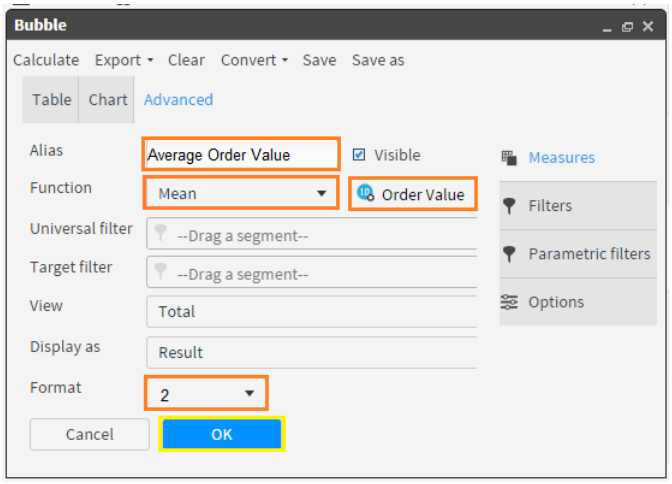


Figure 4-58 Defining a new measure for a bubble analysis

9 The "Measures" screen in the Advanced view tab opens again. It now displays all three calculated measures that have been created. (See Figure 4-59 below).

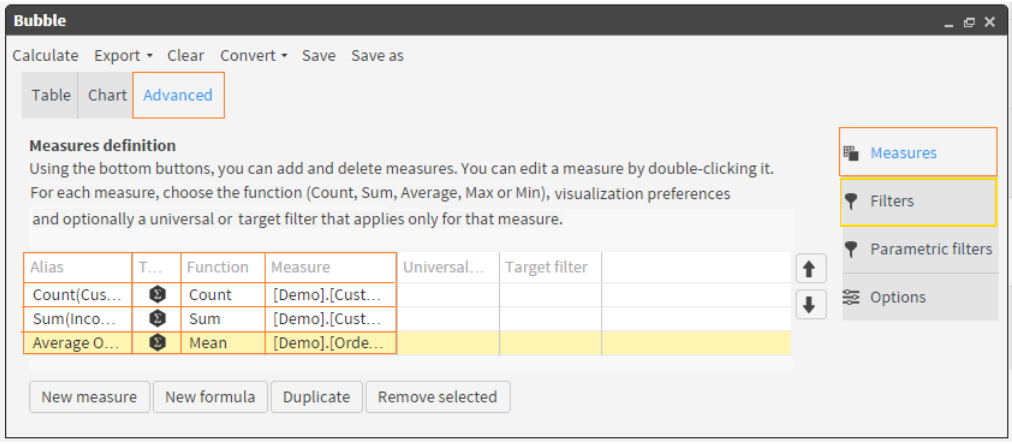


Figure 4-59 Measures list in the Advanced view Measures tab

10 Click "Calculate" in the Bubble menu toolbar. You can now see all your results in the "Table" view that opens. Click on the "Chart" tab to see your results displayed as a Bubble chart. (See Figure 4-60 and Figure 4-61 below).

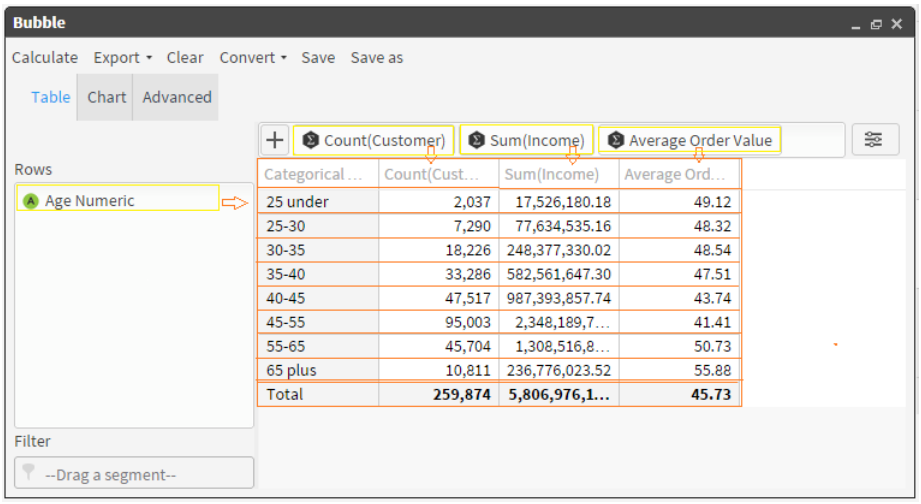


Figure 4-60 Results of the three measures grouped by age

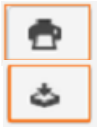
Figure 4-61 Results displayed with Average Order Values setting bubble size

Saving and exporting a bubble analysis

Your Bubble analysis can now be saved or exported or even converted to a Crosstab:

Saving: Click on “Save” or “Save as” in the menu toolbar to save your analysis in your Analysis folder or elsewhere.

Exporting: Click on “Export” in the menu toolbar to export your analysis as a CSV file.



- Click on the “Export” icon to export your analysis as a PDF file.
- Click on the “Image” icon to export your chart as an image file.

Converting: Click on “Convert” in the menu toolbar where you can convert your Bubble a Crosstab analysis. See Figure 4-62 below with your Bubble analysis now displayed as a Crosstab analysis.

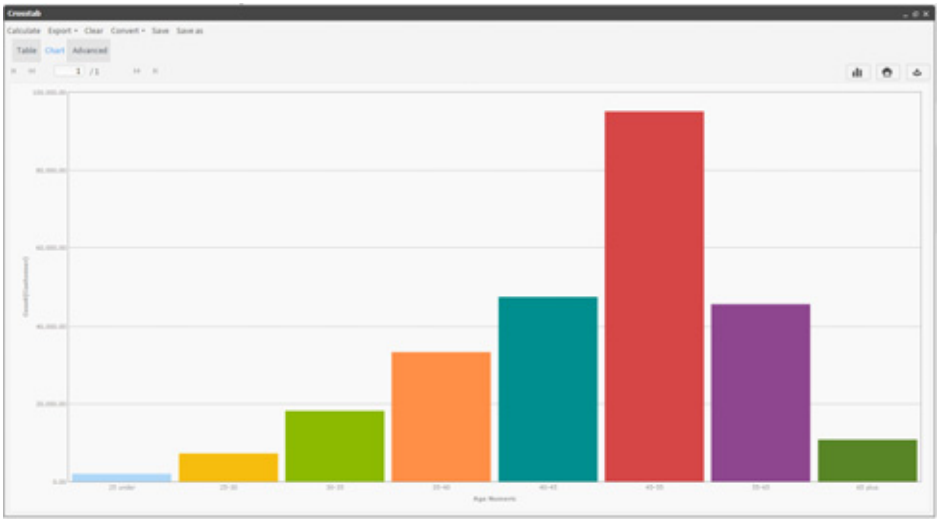


Figure 4-62 Bubble analysis after conversion to a Crosstab analysis

How to duplicate an existing measure

You can easily duplicate an existing measure in order to create a new one (based on the existing one) that you can use with a different column or table, depending on the type of operator used. This is done by using the “Duplicate” button located at the bottom of the “Measures” tab in the “Advanced” view.

1. Select the desired measure in the Advanced view table and click on the “Duplicate” button. This opens the measure in the *New measures* screen with all the necessary fields populated.
2. Enter a new name in the “Alias” field.
3. Drag the new column (or table) to be used from the Data Tree into the “Drag a column or table” field and click “OK”. Your new measure appears in the Measures list in the “Measures” tab of the “Advanced” screen.
4. Click “Calculate” in the Toolbar when you are ready to apply it to the analysis.

02 How to change the Bubble Chart display

The way that your results are displayed in the Chart view can be significantly changed by changing the order of the measures in the Advanced “Measures” tab.

1 Open the Bubble analysis that you have just created in the previous procedure and then open the Advanced tab where all your measures are displayed in a table.

2 Select the Count(Customer) measure and use the down arrow button to move it to the bottom of the table so that it now appears in the table as the 3rd and last measure. (See figure Figure 4-63 below).

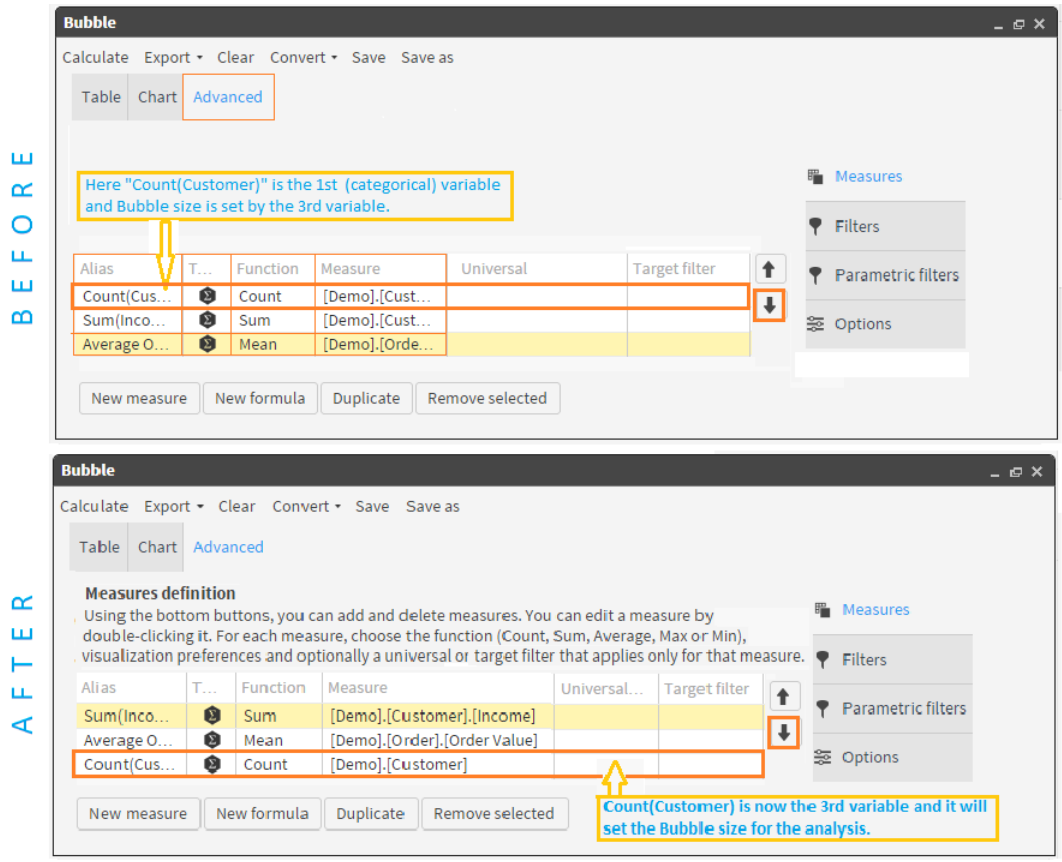


Figure 4-63 Changing the order of Bubble measures

3 Click "Calculate" in the menu toolbar at the top. The Sum(Income) measure is now the categorical measure and the third measure "Count(Customer)" will now determine the size of the bubbles. Now the largest bubble will represent the age group containing the most customers. (See figure Figure 4-64 below).

Figure 4-64 New display with age group size determining bubble size

03 How to add a Filter to a Bubble analysis

1 Open your previous Bubble analysis that we built together – using the "Age Numeric" column as the categorical variable.

2 In the Table view, drag the filter segment "Director" from the discrete values of the "Occupation Decode" column in the Data Tree and drop the segment into the main Table view panel. This populates the "Filter" box and the filtered results appear, replacing the previous results. (See Figure 4-65).

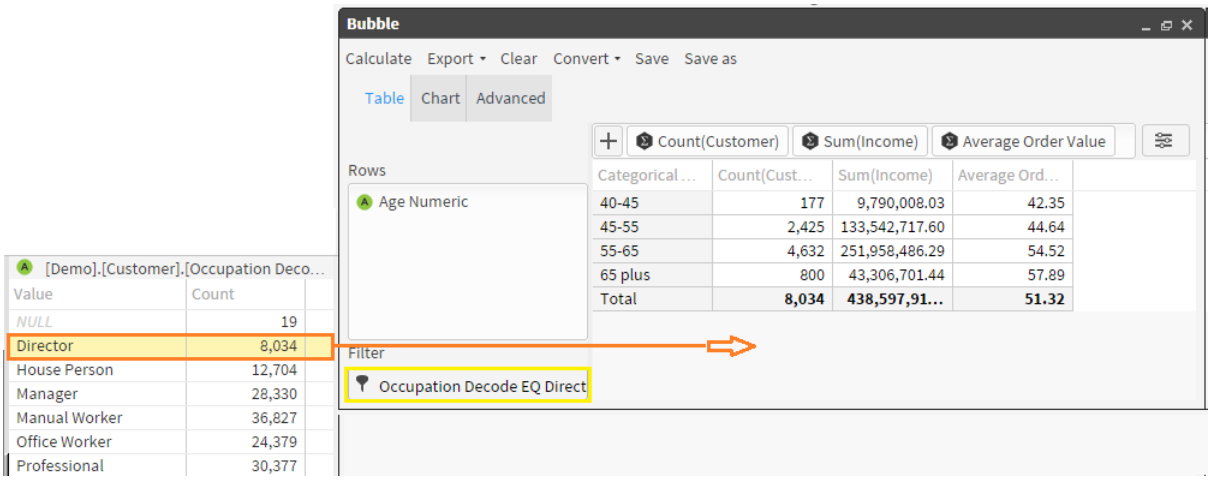


Figure 4-65 Filtering a Bubble analysis

3 Open the Chart view to see the new Bubble chart filtered by "Director" – now showing only 4 bubbles. (See Figure 4-66 below).

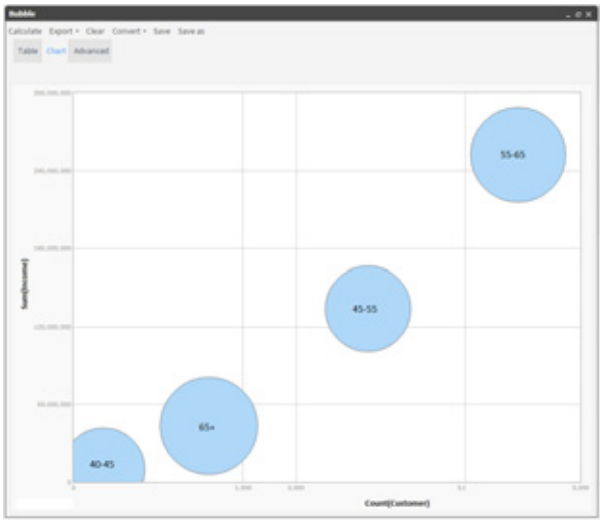


Figure 4-66 Chart view of the analysis filtered by Directors

04 How to display multiple segments in a Bubble Chart

1 Go to the "Table" view of the Bubble analysis window and re-open your saved "First Bubble" analysis. The first categorical variable "Age Numeric" is displayed in the "Rows" panel.

2 Define a second categorical variable by dragging the "Gender Decode" column from the Customer table in the *Data Tree* and dropping it into the "Rows" panel.

The Table view immediately displays the newly calculated measure values, this time grouped by gender. They also include Null values (See Figure 4-67).

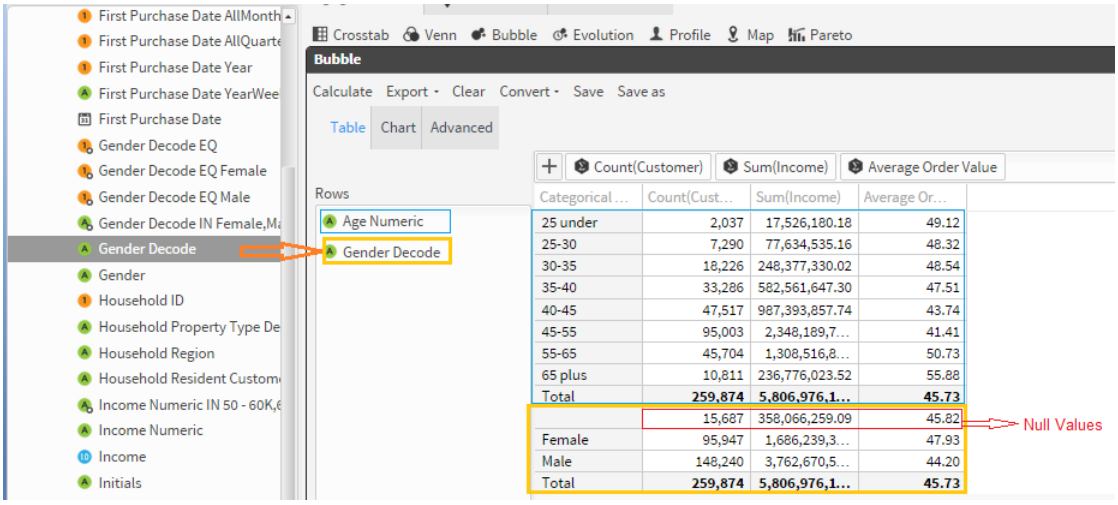


Figure 4-67 Defining a second categorical variable for your analysis

3 To remove the Null values, first double-click on the "Gender Decode" variable in the "Rows" panel. This opens a window displaying its discrete values. Then de-activate the checkmark next the empty row. (See Figure 4-68 below).

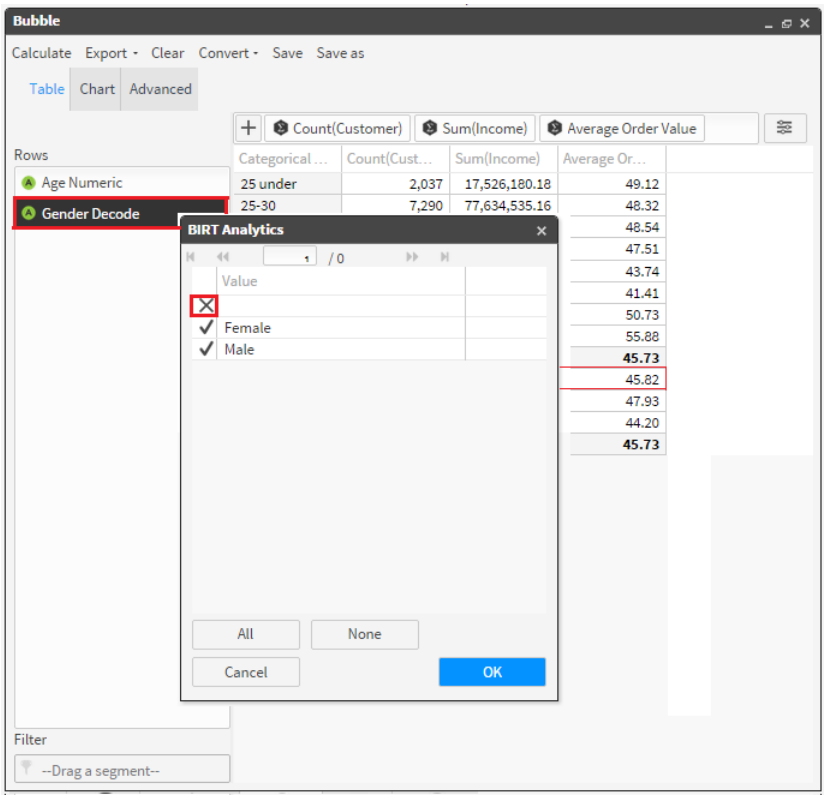
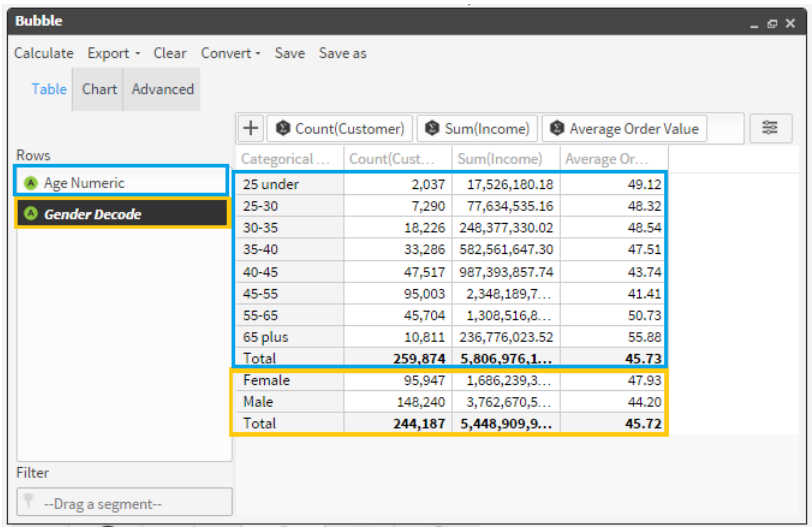


Figure 4-68 Removing Null discrete values from the Gender Decode variable

4 Click "OK" to see your Bubble results displayed without any distracting Null values. (See Figure 4-69 and Figure 4-70 below).



Categorical ...	Count(Cust...	Sum(Income)	Average Or...
25 under	2,037	17,526,180.18	49.12
25-30	7,290	77,634,535.16	48.32
30-35	18,226	248,377,330.02	48.54
35-40	33,286	582,561,647.30	47.51
40-45	47,517	987,393,857.74	43.74
45-55	95,003	2,348,189,7...	41.41
55-65	45,704	1,308,516,8...	50.73
65 plus	10,811	236,776,023.52	55.88
Total	**259,874**	**5,806,976,1...**	**45.73**
Female	95,947	1,686,239,3...	47.93
Male	148,240	3,762,670,5...	44.20
Total	**244,187**	**5,448,909,9...**	**45.72**

Figure 4-69 Table results showing both variables

Figure 4-70 Bubble analysis Chart results for both variables

Using evolution

An evolution analysis shows a progression of data over time. You can examine the behavior of certain measures in different periodic scenarios. For example, examine how the sales of some product families evolve over a period of months. The product family field is the categorical variable under study, while the different scenarios are determined by the different values in the month of purchase variable.

Define the following components to create an evolution analysis:

- Categorical variable: Field that contains the categories whose behavior is to be analyzed, or in other words, the variable under study.

  To indicate the variable you want to work with, drag the field from Data Tree to the control that is in the top part of the configuration form, as shown in Figure 4-71.

Figure 4-71 Control without variables

After dragging the variable to the control, it checks the number of discrete values and shows all of the divisions as well as the single values that it contains, as shown in Figure 4-72. Each green strip corresponds to a product type.

Figure 4-72 Control with variables

You can drag to this control only fields with a maximum of 100 discrete values. In the analysis, a sphere with a specific color selected at random represents each of these categories.

If you want to delete one or more of the categories from the analysis variable, click the box that represents the category. If you position the cursor over each of the green boxes, a tag appears, displaying the value of the option.

You can also select or deselect all of the boxes by using the buttons located just below the discrete values control (all, none).

- Measures: Evolution can display up to a maximum of three measures, two of which position the spheres on the axes of the *x*- and *y*- coordinates, while the third (optional) defines the relative measure of the sphere inside the group.

  To define the measures of the axes, drag the numeric fields from Data Tree to the vertical bar (*y*-axis) or to the horizontal bar (*x*-axis).

  When you drag a column to the *x*-axis, the latter changes color, indicating that it is ready to accept a column. After you drop the field, the control shows the applicable functions available in accordance with the type of column.

  To use the third measure (sphere measure), drag the column you want to use to the icon that is located in the top-right part of the form.

  If the cursor is positioned over the definition controls of the measures (axes, measure control), the current operation as well as the field involved in it are displayed.

- Transition variable: It is necessary to define what data tree variable to use as a base for creating transitions. To indicate the field, you drag it to the discrete values control at the bottom of the form.

  This control operates in the same way as the categorical variable control and enables you to indicate which elements to use in the analysis.

How to create an evolution

1. In Analytics—Analysis, choose Evolution.
2. From Data Tree, drag a column to use as the categorical variable, as shown in Figure 4-73.

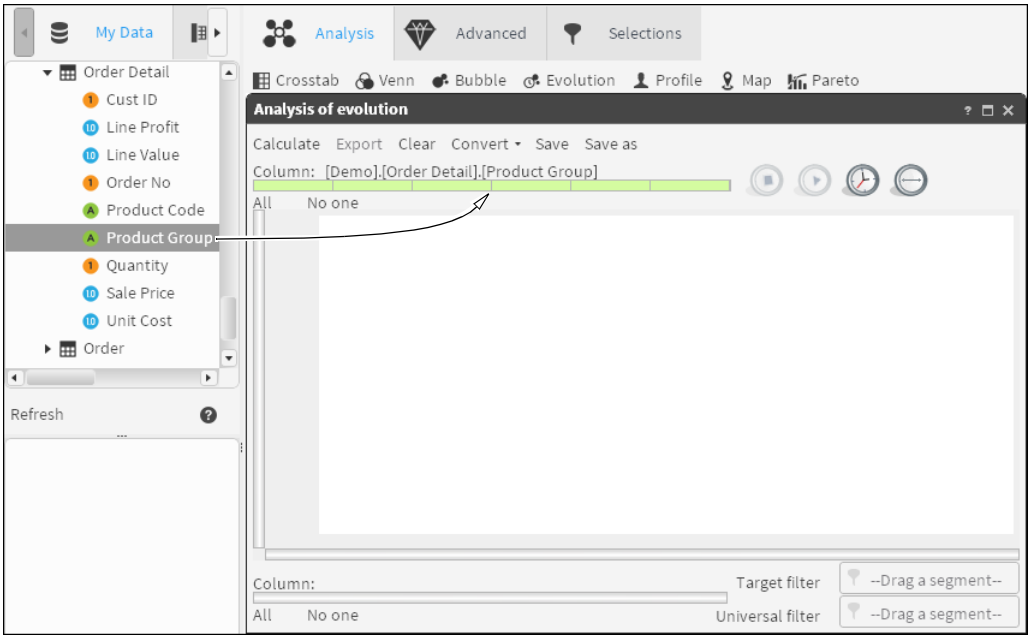


Figure 4-73 Defining a categorical variable for an evolution analysis

3 From Data Tree, drag a column to use as the transition variable, as shown in Figure 4-74.

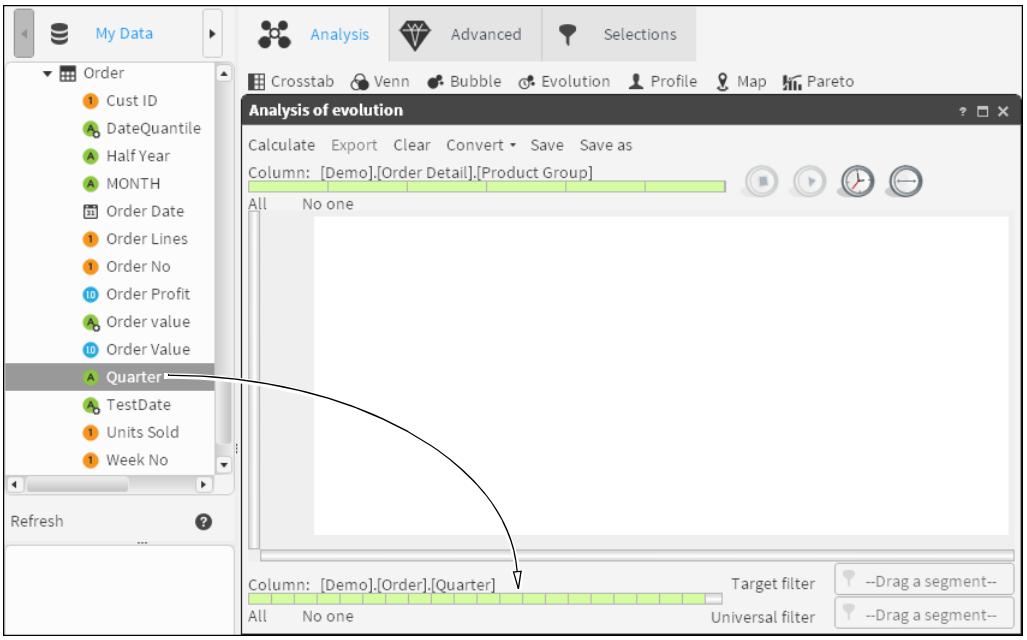


Figure 4-74 Defining a transition variable for an evolution analysis

4 From Data Tree, drag columns onto the *x*-axis and *y*-axis. Choose appropriate functions for each axis. Figure 4-75 shows the *y*-axis being added as mean order profit with the x-axis in blue, to indicate that a column has already been applied.

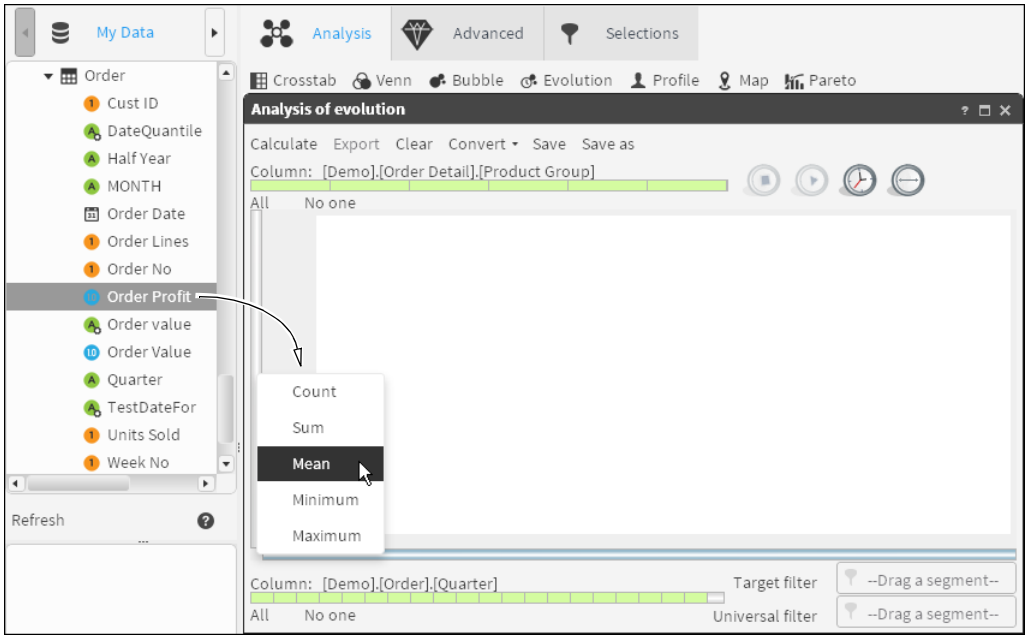


Figure 4-75 Defining *y*-axis properties for an evolution analysis

5 Choose Calculate. Choose Play to show the evolution over time. Figure 4-76 shows the evolution analysis.

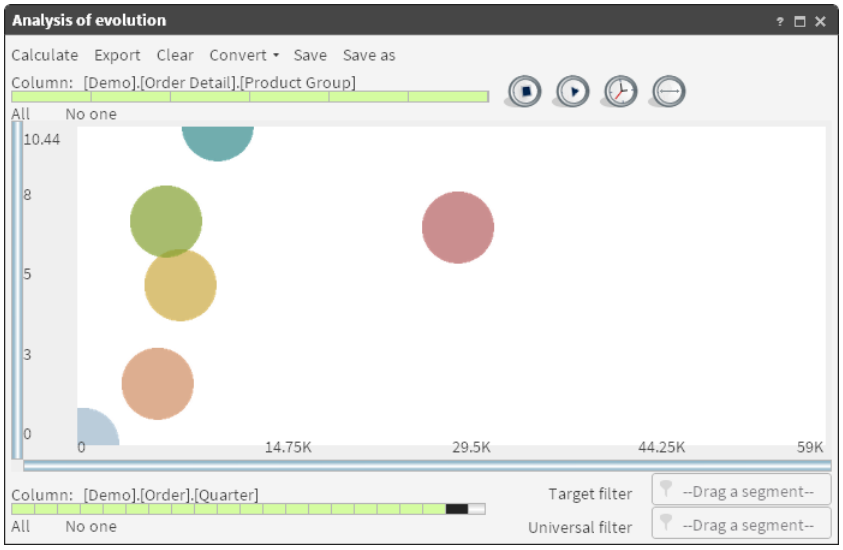


Figure 4-76 Viewing an evolution analysis

6 To indicate a third measure, from Data Tree, drag a column to the far right icon.

7 To modify the transition of time, select the clock icon and choose an interval. For example, an interval of 0.25 is faster than an interval of 3.

More about viewing an evolution

When evolution is defined, BIRT Analytics internally creates a crosstab in which the categorical variable that you want to study is the row dimension (in the case of the example that accompanies this explanation, the product category) and the transition variable (the temporary variable) is the column dimension. The crosstab also includes the two or three defined measures (x, y, and measure).

The internal result of the crosstab is similar to Table 4-1.

Table 4-1 Internal result of crosstab

	Scenario 1			Scenario 2			Scenario 3		
Category 1	x1	y1	measure1	x2	y2	measure2	x3	y3	measure3
Category 2	x1	y1	measure1	x2	y2	measure2	x3	y3	measure3

The graphic representation of the evolution analysis consists of data extracted from each of the scenarios, displayed as spheres. The animation extracts data from each scenario and displays each category as a sphere in a particular color, positioning it in a way that takes into account the maximum and minimum values of the measures for other categories.

During execution, you can hover the cursor over one of the spheres to display the name of the category and the values for each of its measures for the current transition. A transition is the category of the transition variable that is displayed by an evolution.

As the transitions occur, the active scenario of the transition variable appears.

It is possible to modify the time between transitions by choosing the time icon and selecting the speed.

The analysis definition form also allows you to set the filters for the calculation.

Recommendations

- To make this analysis useful, the number of discrete values of the categorical variable must be as low as possible.
- The analysis is deemed to exceed the maximum if it occupies more than one page.
- To learn more about internal calculations, consult the crosstab help utility.

For example, consider the question, “How do the sales of a product family change over the months?” To answer this question, create an evolution indicating the group of products as a categorical variable and the month when the order is placed as a transitional variable. Possible measures are the orders count (x-axis) and the average profit (y-axis), which enable you to see quickly the group of products that sells most over time, and the group of products that produces the most profit. You can convert this type of analysis into a crosstab analysis.

Using profile analyses

A profile analysis groups values and determines their relatedness to another group, called the profile segment. This analysis helps you draw a profile of a group of values from attributes selected in Data Tree.

In advertising, a target group refers to the people to whom the advertising is directed. The target group is equivalent to a profile segment. If a profile segment is chosen, the profile carries out a comparison with each of the attributes that are selected based on all the database’s values, including the analysis segment itself. In other words, a request is made for an analysis of the significance of the attributes to define the segment, versus all other values in the database.

Alternatively, if a base segment is selected in addition to the profile segment, the analysis is based on the values that are part of that chosen segment. In the first case the segment is defined. In the second case, whether the attributes are suitable for showing differences between both groups (profile and base segment) is defined. Three different scenarios are possible:

- Customers who provide higher profits vs. customers who provide lower profits (no overlapping of values).
- Customers who have responded to a campaign via e-mail vs. all who have responded (the segment to be profiled is part of the base).
- Customers who have bought product X this year vs. customers who bought product X last year. There are some values that are part of both segments, profile and base, but not all.

There are several indicators that measure the significance of the attributes to define the analysis segment, including Z-score. The Z-score determines whether the difference between two proportions is statistically significant. In the case of profile, this determination is carried out between the group to be analyzed and the group considered Rest, values that belong to the attribute whose significance you want to measure but not to the analysis segment.

The Z-score uses the binomial for unequal variances, whose formula is:

```
Z = (p1-p2)/sqrt((p1*(1-p1)/n1 + (p2*(1-p2)/n2))
```

- p1 is the sampled proportion and n1 is the number of sampled values.
- p2 is the proportion in Rest and n2 is the number of Rest values. Table 4-2 shows calculated Z-scores.

Table 4-2 Calculating Z-scores

	Men	Rest	Z-score
M	34.4%	19.0%	171.31
X	27.9%	16.6%	132.66
W	17.0%	9.8%	102.52
S	20.7%	54.6%	-336.01
	682,493	317,507	

In other words:

- You would first choose the characteristics or attributes whose significance you want to measure to determine the analysis segment.
- Significance means those attributes or characteristics that meet the highest possible number of cases within the segment analyzed.
- You set the characteristics or attributes of those whose significance you want to measure to define or characterize the analyzed segment:
  - How is this significance measured?
  - Using a Z-score indicator.
  - This indicator is obtained by applying a statistical formula.
  - The higher the Z-score of the attribute, the higher the defining capacity, or relatedness, of the segment analyzed.

A Z-score higher than 0 can be considered significant to describe a segment, but values below 1.96 cannot be considered statistically significant. Only values above this threshold can be considered significant. You set attributes of which you want to measure the significance, which means that other values that are not chosen for the study may be the ones that are truly statistically significant to define the analysis segment.

How to create a Profile analysis

1. In Analytics—Analysis, choose Profile.
2. In Profile, choose Advanced.

3 To create a segment, drag one or more values from Discrete Values to Scratchpad. Alternatively, use an existing segment. The segment of data represents a profile group against which you test the significance of various attributes.

4 If necessary, rename the segment appropriately.

5 Drag the segment from Discrete Values and drop it in Segment filter, as shown in Figure 4-77.

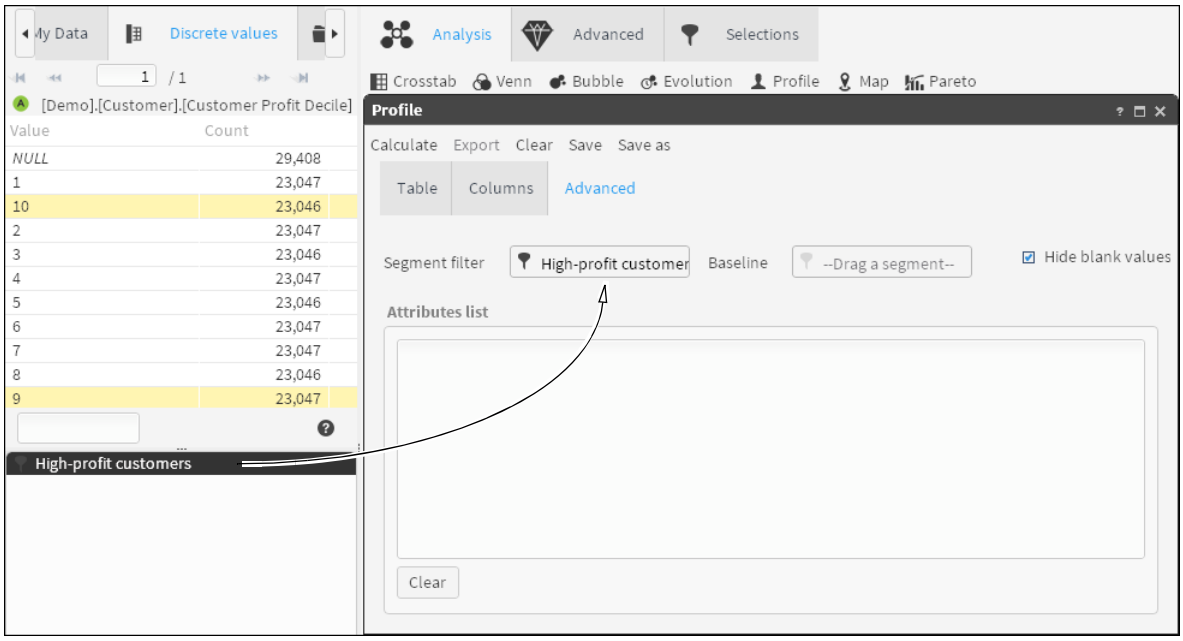


Figure 4-77 Adding a segment filter

6 Drag fields from Data Tree into Attributes list, as shown below. Using the Profile analysis, you test these attributes for how significantly they contribute to the profile. (See Figure 4-78

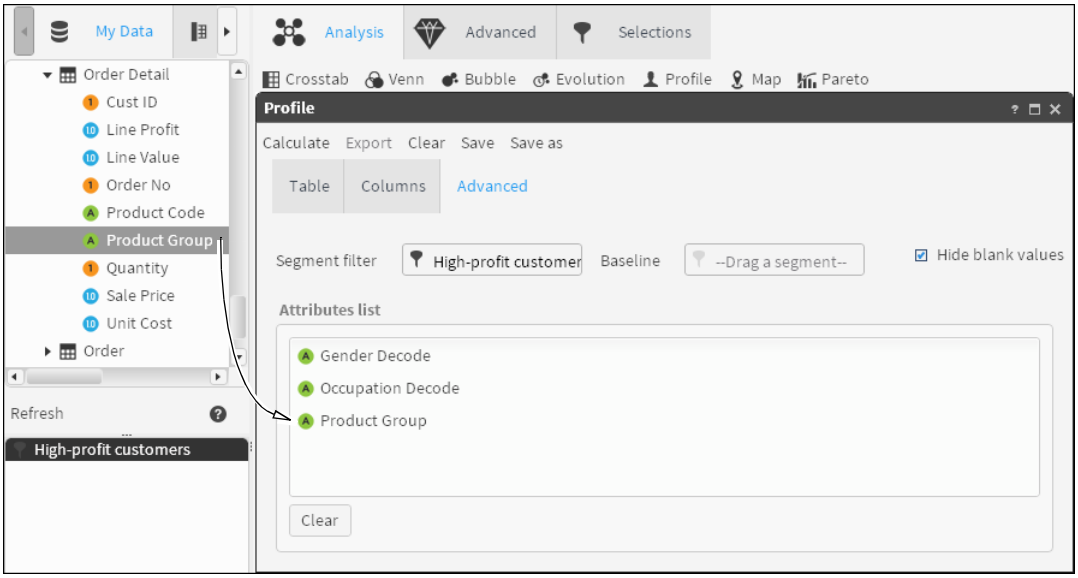
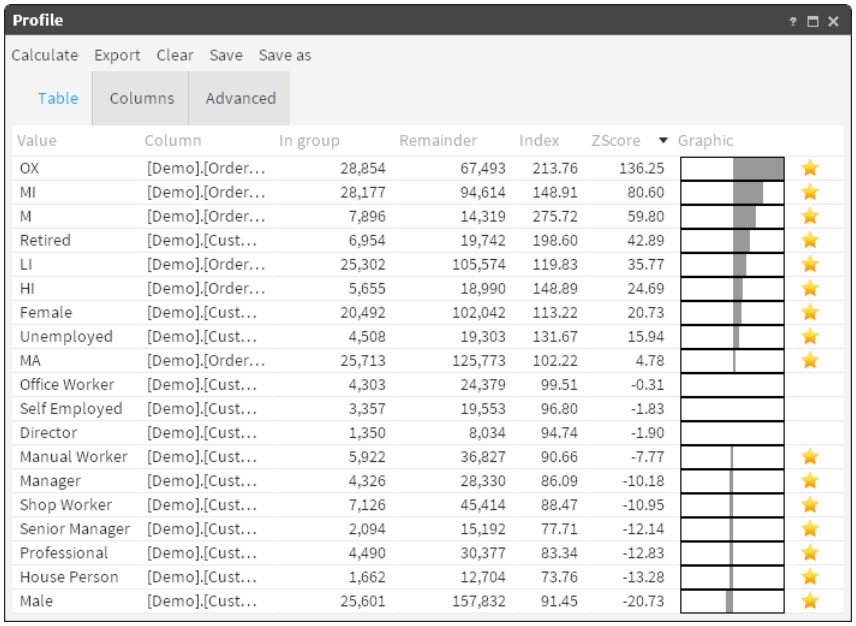


Figure 4-78 Adding profile attributes

7 Choose Calculate.

8 Examine the results on Table. The gray bars in the Graphic column show which discrete values from the list of attributes contribute most significantly to the profile, as shown in Figure 4-79. ***Note:*** *the gray bars represent the Z-score, and the column is sorted by default from most significant to least significant*.



Value	Column	In group	Remainder	Index	ZScore	Graphic
OX	[Demo].[Order...	28,854	67,493	213.76	136.25	
MI	[Demo].[Order...	28,177	94,614	148.91	80.60	
M	[Demo].[Order...	7,896	14,319	275.72	59.80	
Retired	[Demo].[Cust...	6,954	19,742	198.60	42.89	
LI	[Demo].[Order...	25,302	105,574	119.83	35.77	
HI	[Demo].[Order...	5,655	18,990	148.89	24.69	
Female	[Demo].[Cust...	20,492	102,042	113.22	20.73	
Unemployed	[Demo].[Cust...	4,508	19,303	131.67	15.94	
MA	[Demo].[Order...	25,713	125,773	102.22	4.78	
Office Worker	[Demo].[Cust...	4,303	24,379	99.51	-0.31	
Self Employed	[Demo].[Cust...	3,357	19,553	96.80	-1.83	
Director	[Demo].[Cust...	1,350	8,034	94.74	-1.90	
Manual Worker	[Demo].[Cust...	5,922	36,827	90.66	-7.77	
Manager	[Demo].[Cust...	4,326	28,330	86.09	-10.18	
Shop Worker	[Demo].[Cust...	7,126	45,414	88.47	-10.95	
Senior Manager	[Demo].[Cust...	2,094	15,192	77.71	-12.14	
Professional	[Demo].[Cust...	4,490	30,377	83.34	-12.83	
House Person	[Demo].[Cust...	1,662	12,704	73.76	-13.28	
Male	[Demo].[Cust...	25,601	157,832	91.45	-20.73	

Figure 4-79 Examining profile results

Using map analyses

A map analysis supports visualizing data on a geographic map. Analysis results appear as a chloropleth map on which each predefined region is assigned a color or shade. Each shade corresponds to the magnitude of data values for that region. For example, a map analysis assigns each region in a country map a different shade of red that corresponds to the number of unemployed customers in each region. In this example, a region that appears dark red has a high number of unemployed customers.

A map analysis assigns, by default, increasingly darker shades of red to regions having values of correspondingly higher magnitude. Optionally, create a custom set of color shades. When selecting custom colors, consider the following points:

- Darker colors typically represent data items with a higher magnitude.
- To make clear, distinct data ranges, select a distinct color shade for each range. Test map analysis results to verify that the map conveys meaning when viewed in color or in grayscale shades.
- To provide effective contrast between data ranges, limit the number of colors to seven, if possible.

BIRT Analytics provides multiple maps with region names encoded. You can also add new maps. If necessary, you can use a decode field to match the names in the map to the region names in your data.

How to add a measure to a map analysis

1 In Analytics—Analysis, choose Map.

2 In Map, choose Definition, as shown in Figure 4-80.

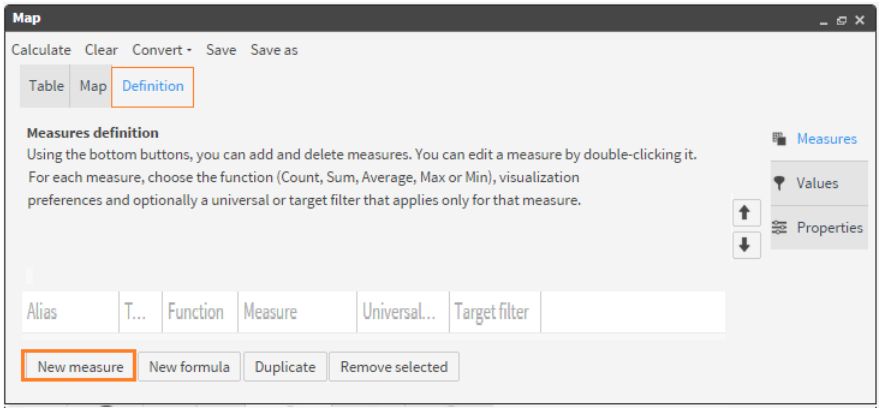


Figure 4-80 Map definition view

3 In Measures, choose New measure.

4 Type a descriptive alias.

5 To define the new measure, drag a table from My Data and drop it as shown in Figure 4-81.

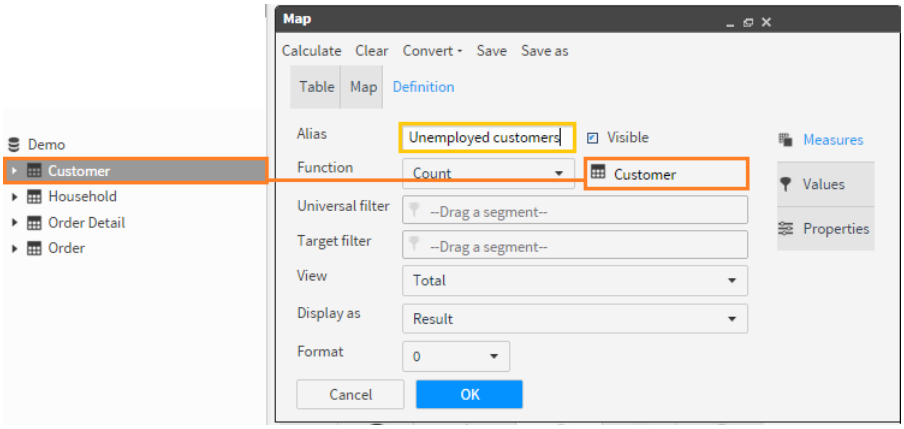


Figure 4-81 Adding a table to the measure definition

6 To filter the measure, drag a segment from Discrete Values and drop it in Universal filter, as shown in Figure 4-82.

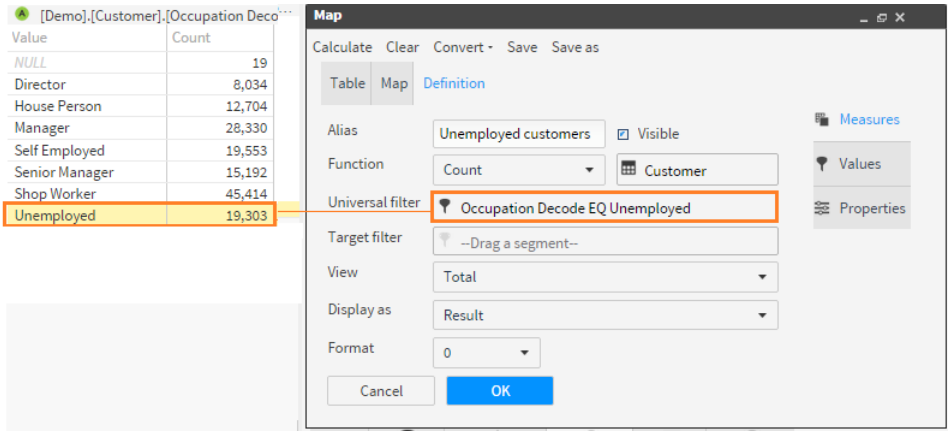


Figure 4-82 Filtering the measure

7 Choose the check mark. A defined measure appears in Measures definition.

How to generate map analysis results

1 Choose Values.

2 To link geographic data to a measure, drag a column that contains geographic data from My Data and drop it in Geographical Column, as shown in Figure 4-83.

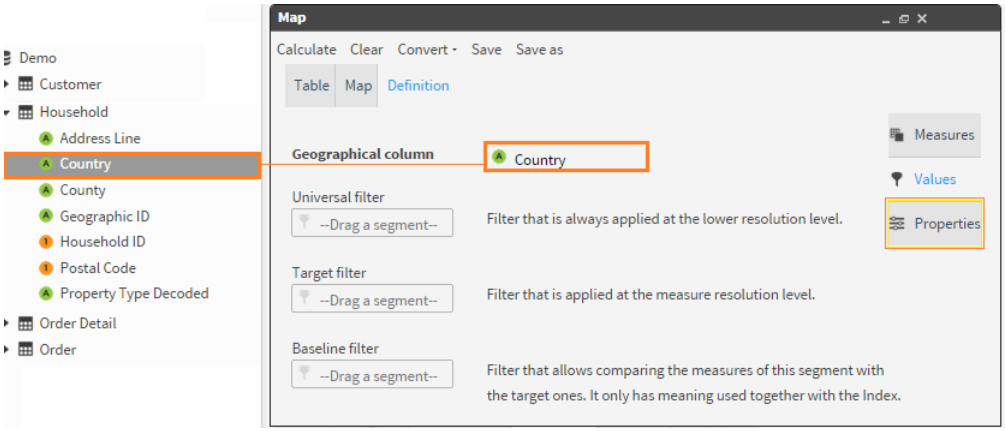


Figure 4-83 Linking geographic data to the measure

3 Choose Properties.

4 In Select map, select a map name that matches the geographic region defined for the measure. If no available map appears to match the geographic region for your analysis, ask a BIRT Analytics administrator to load the appropriate map.

5 Choose Calculate.

How to view map analysis results

- To view graphical results, choose Map, if necessary.
- To view data values that correspond to each color appearing in Map, choose Legend.
- To see numerical results of the map analysis, choose Table.
- To see analysis results appearing in Map on a crosstab, select Convert. Then, choose Crosstab.

Using Pareto analyses

A Pareto analysis represents Pareto's 80-20 theory with available data. Pareto's theory states that:

- A minority of the population (approximately 20%) bears 80% of something.
- The remaining majority group (approximately 80%) bears 20% of something.

For example:

- 20% of clients are responsible for 80% of turnover.
- 80% of turnover comes from 20% of the product catalog.

You can use this theory to explore the relationship between a numerical and a categorical variable. For example, you can analyze the accrued benefit of customers (continuous numerical variable) through the grouping of customer's accrued benefit into deciles (categorical variable).

How to create a Pareto analysis

1 In Analytics—Analysis, choose Pareto.

2 In Pareto, choose Advanced, as shown in Figure 4-84.

Figure 4-84 Pareto—Advanced

3 Drag a categorical field from Data Tree to Attribute.

4 Drag a related numeric field from Data Tree to Measure. The resulting entries are shown in Figure 4-85.

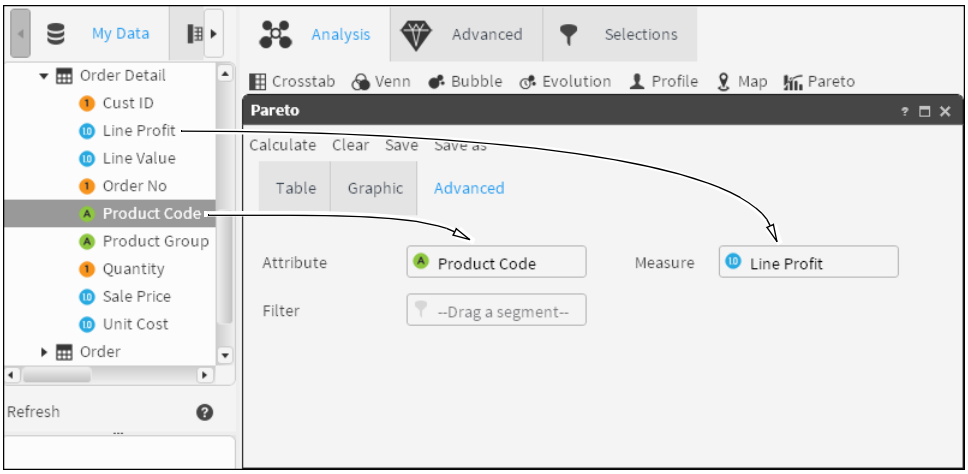


Figure 4-85 Dragging an attribute and a measure to Pareto

5 Add a filter to limit the results, if necessary. Drag a segment from Discrete Values or Scratchpad to Filter, as shown in Figure 4-86.

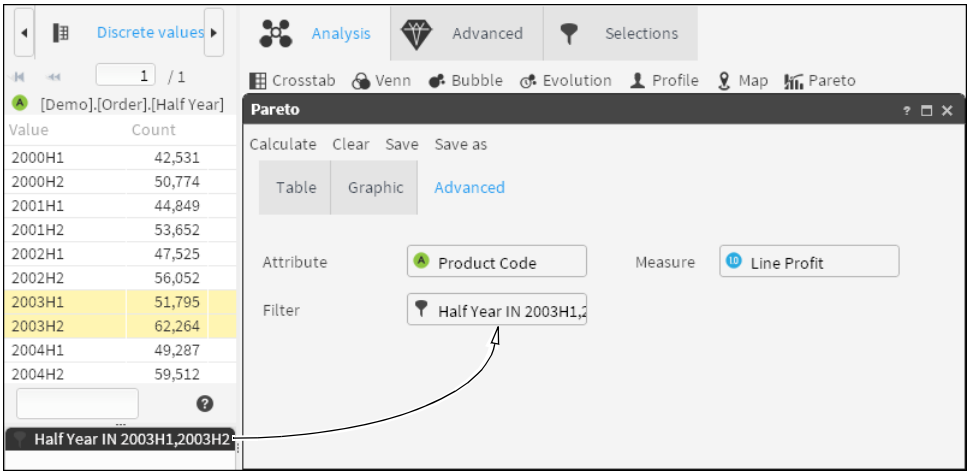


Figure 4-86 Dragging a filter to Pareto

6 Choose Calculate.

7 Examine the results on Graphic.

For example, select a numeric field as a measure, and a quantile rank of the same field as an attribute. This quantile groups its values into n equal groups. The analysis displays a Pareto curve. You can see if the Pareto analysis satisfies Pareto's theory by looking at the growth curve. The graph shows cumulative percentages.

Table shows, sorted by the amount, the number of records, both amount and percentages, cumulative and cumulative percentages.

Double-clicking Count (record number), the records appear in Scratchpad. By choosing the Cumulative Count value, you can display the record and all previous records.

Chapter

5

Visualizing your data

This chapter contains:

- Using the Gallery
- Working with a Canvas

Visualizing your data

BIRT Analytics provides many ways to visualize data. By defining data measures, refining their properties, and assigning a visual indicator to display the data measures, you use BIRT Analytics to visually compare and contrast data values.

Using the Gallery

Gallery supports creating visual indicators that display data values as graphic visualizations. For each indicator, you define one or multiple data measures. All BIRT Analytics indicators provide the following two features:

- Graphic displays a visual preview of the indicator.
- Definition supports defining data measures, values, and properties that determine how data, shapes, and colors appear on the indicator.

To view an indicator on Graphic, you must create valid measures using Definition. This section describes how to visualize data in the following ways:

- Using a dial, meter, label, sphere, cylinder, and funnel indicator
- Converting data measures for one indicator to another

Using a dial

A dial indicator displays data values using a needle that points to data values and range boundaries appearing on a semi-circular dial. To compare values with defined bounds, define data measures, values, and properties for a dial using the following features on Definition:

- Measure: Add, edit, duplicate, or delete measures and formulas.
- Values: Assign maximum, mean, and minimum values to a dial by selecting an appropriate measure. To indicate multiple values, use Add dial to create a new needle, then set alias, color, and size properties for each new needle.
- Properties: Choose values to appear as boundaries for defined ranges, and assign a unique color or shade to each range.

You can covert a dial into a meter, sphere, or cylinder indicator.

How to create measures to display on a dial indicator

1. In Gallery—Dial, choose Definition.
2. To create a measure for the dial, choose New measure.
3. In Alias, type a name for the new measure.
4. Select Maximum as the function for the new measure. From My Data, drag a column on which to base the measure and drop it next to Maximum, as shown in Figure 5-1.

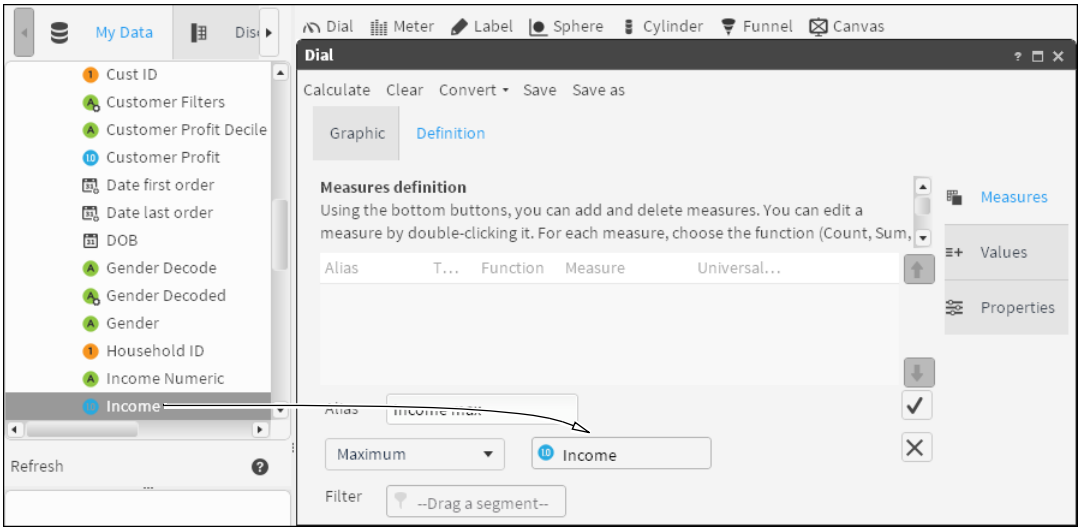


Figure 5-1 Defining a new measure for a dial

5 To create a filter, double-click a column in My Data. Drag a field from Discrete Values and drop it in Filter, as shown in Figure 5-2.

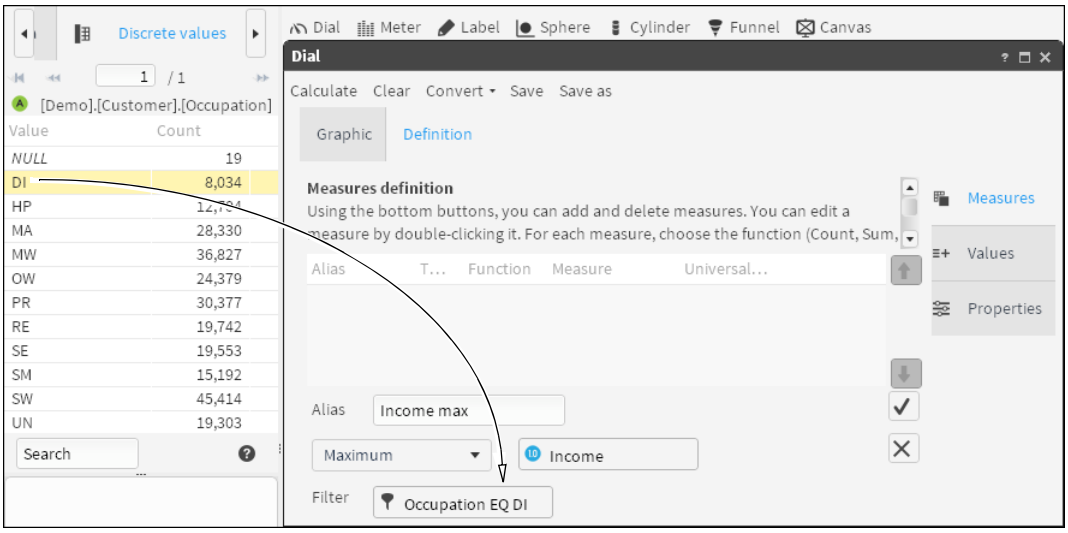


Figure 5-2 Adding a filter to a measure

6 Choose the check mark to create the new measure.

7 To create another measure with properties similar to the first, select a measure and choose Duplicate.

8 Type a name and change the function to Minimum, as shown in Figure 5-3.

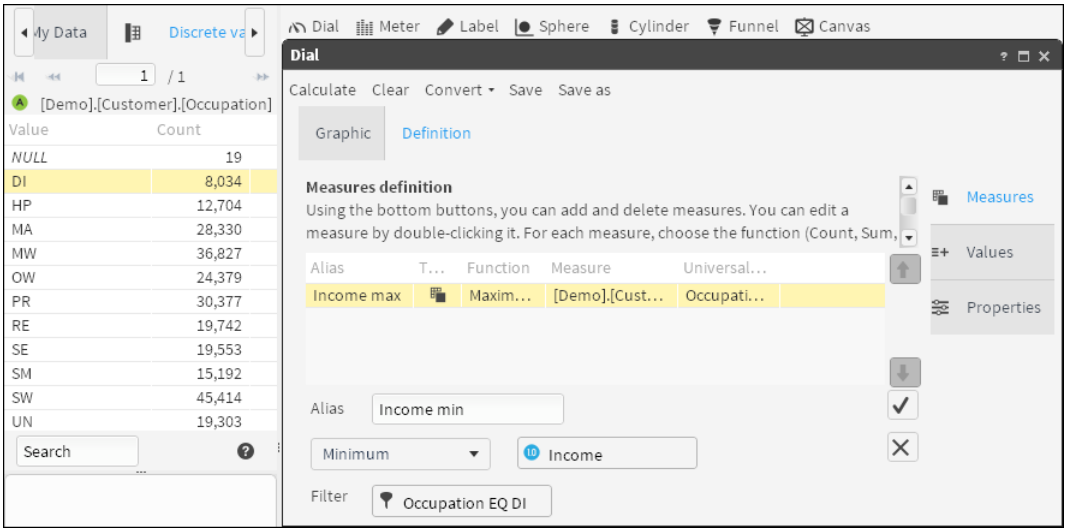


Figure 5-3 Creating a new measure based on a measure having similar properties

9 Choose the check mark to create the new measure.

10 Repeat steps 7–9, changing the function to Mean.

How to add formulas and thresholds to a dial indicator

1 Choose New formula to create a formula.

2 In Alias, type a name for the formula, and type a formula for a lower threshold in the dialog box below Alias. In this example, the formula is a value.

3 Choose Validate. At the message that the formula definition is valid, shown in Figure 5-4, choose OK.

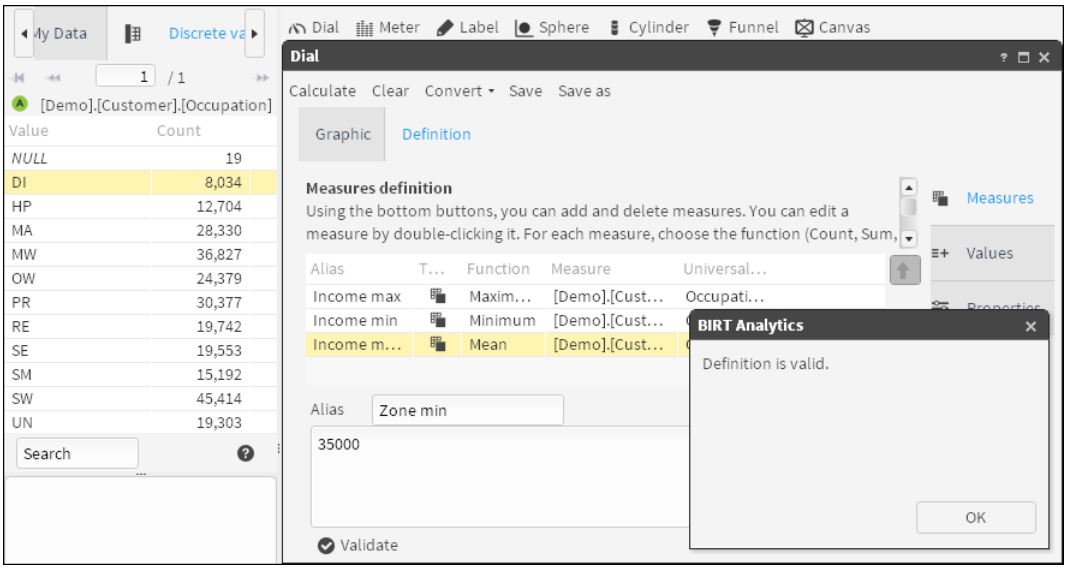


Figure 5-4 Validating the definition for a formula

4 To create the new formula, choose the check mark.

5 Repeat steps 12–4, entering a formula for a higher threshold.

6 Choose Values.

7 For each value appearing on the dial, choose an appropriate measure from the list. For example, choose measures that correspond to minimum, mean, and maximum values, as shown in Figure 5-5.

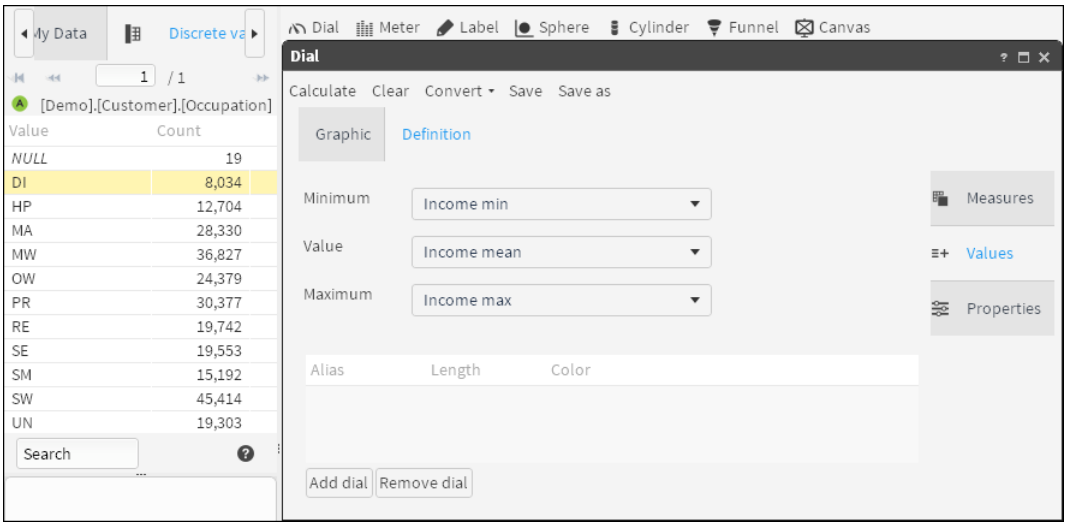


Figure 5-5 Assigning measures to values that appear on a dial

How to add properties to the dial indicator

1 Choose Properties.

2 Choose Add range.

3 Choose a measure or formula to indicate the minimum and maximum thresholds for the range, as shown in Figure 5-6.

4 In Color, open the palette and select a color for the range. Choose OK.

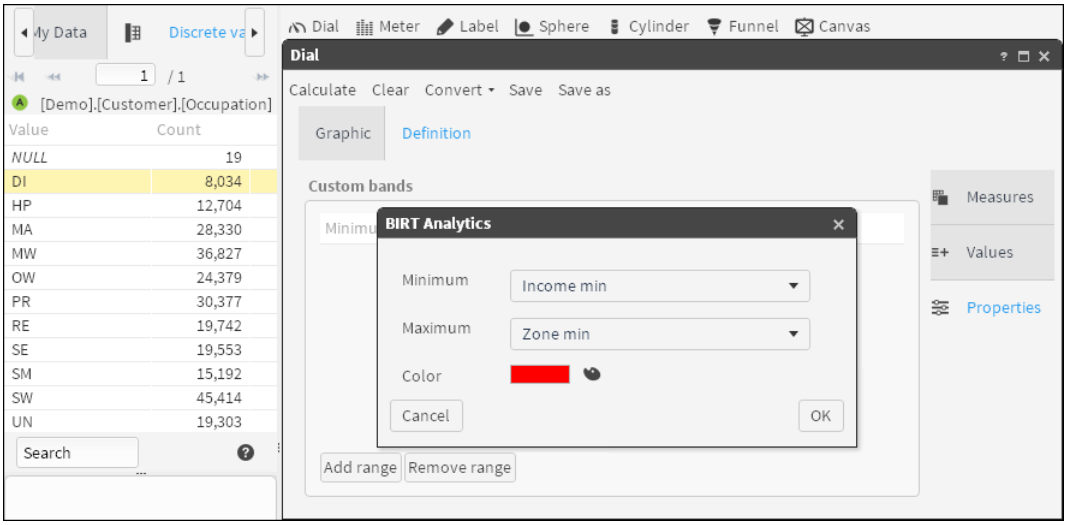


Figure 5-6 Configuring properties for a dial

5 Repeat steps 2–4 until you have created three ranges.

6 Choose OK.

7 Choose Calculate to view the dial. To see the average value, hover the cursor over the needle, as shown in Figure 5-7.

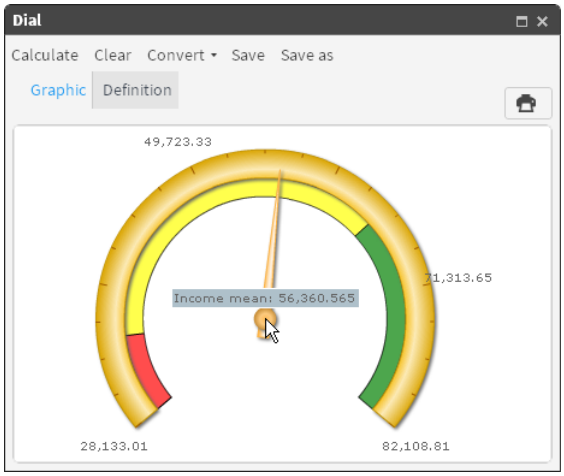


Figure 5-7 Viewing the value of an average on a dial

How to configure a dial so that min, max, and average values for two similar groups appear

1. In Data Tree, choose My folders.
2. Locate and open a previously saved dial. Choose Calculate. The dial appears on Graphic, as shown in Figure 5-8.

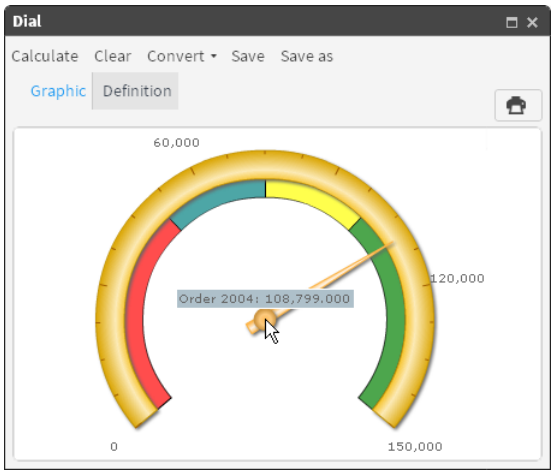


Figure 5-8 Calculating results for a dial

3. Choose Definition.
4. Select a measure and choose Duplicate.
5. In Alias, type a name, then replace the filter. To replace a filter, drag a segment, drop it in Filter, as shown in Figure 5-9, then choose Replace.
6. Choose the check mark to create the new measure.
7. Choose Values.
8. Choose Add dial.
9. In Alias, select the measure shown to open a pull-down menu and choose the newly created measure, as shown in Figure 5-10.
10. In Color, open the palette and select a color for the range. Choose OK.
11. Choose Calculate to view the dial. The dial now represents two similar groups. Hover the cursor over each needle to see and compare values represented by each needle, as shown in Figure 5-11.

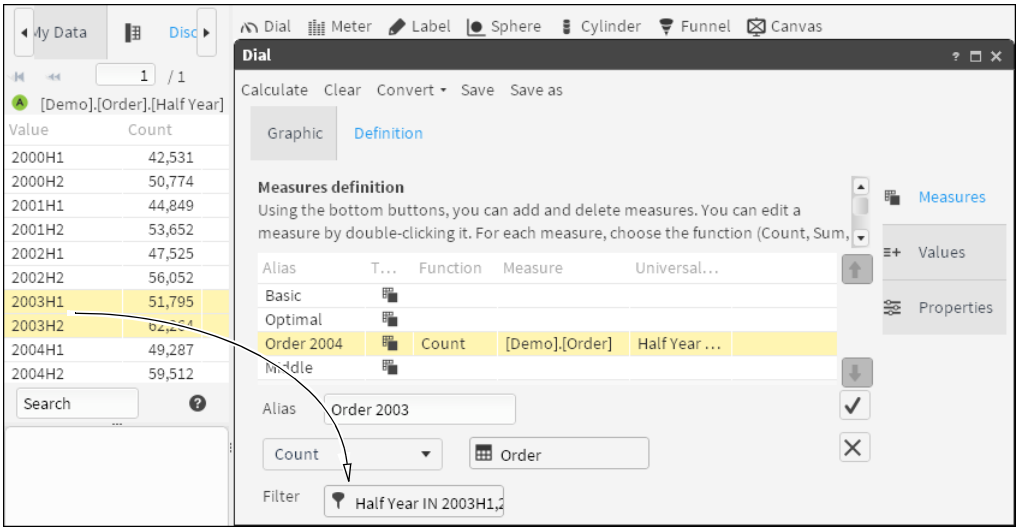


Figure 5-9 Replacing the filter in a duplicated measure

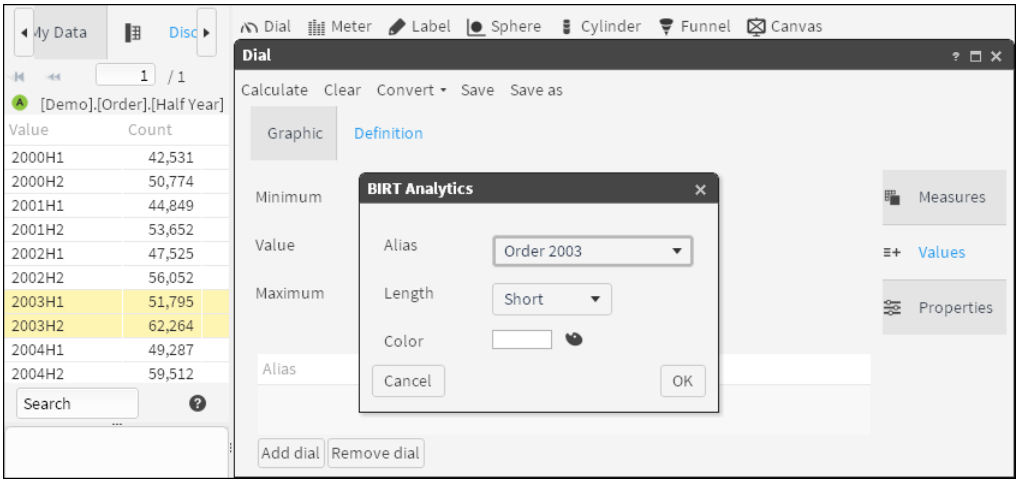


Figure 5-10 Adding a new measure to a dial

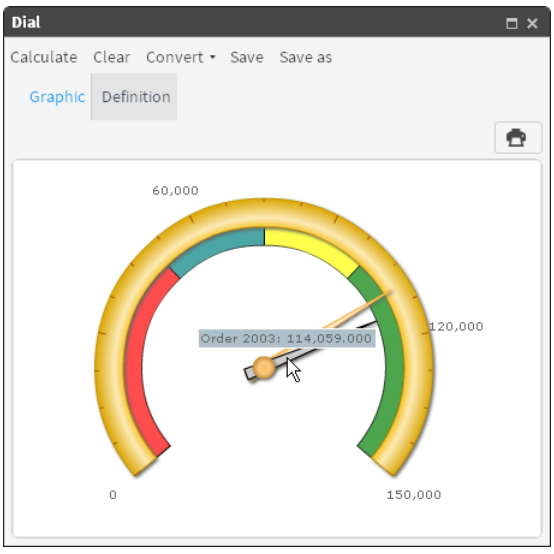


Figure 5-11 Viewing a dial representing values from two similar groups

Converting data measures to another indicator type

Data measures that you define for one BIRT Analytics indicator may appear on another indicator. Convert lists indicators for which a calculated indicator supports conversion.

How to convert a dial to a meter

1 In Data Tree, select My Folders.

2 Locate and open a previously saved dial. Choose Calculate. The dial appears on Graphic, as shown in Figure 5-12.

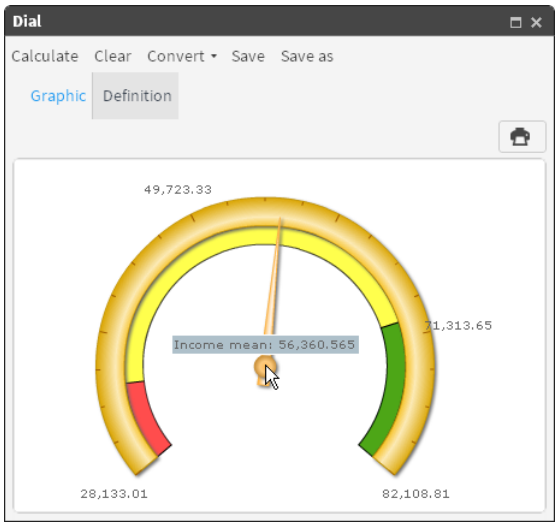


Figure 5-12 Visualizing a data measure using a dial

3 Select Convert, then choose Meter.

4 Choose Definition, then view the Measures, Values, and Properties tabs.

5 Choose Calculate. The new meter appears on Graphic, as shown in Figure 5-13.

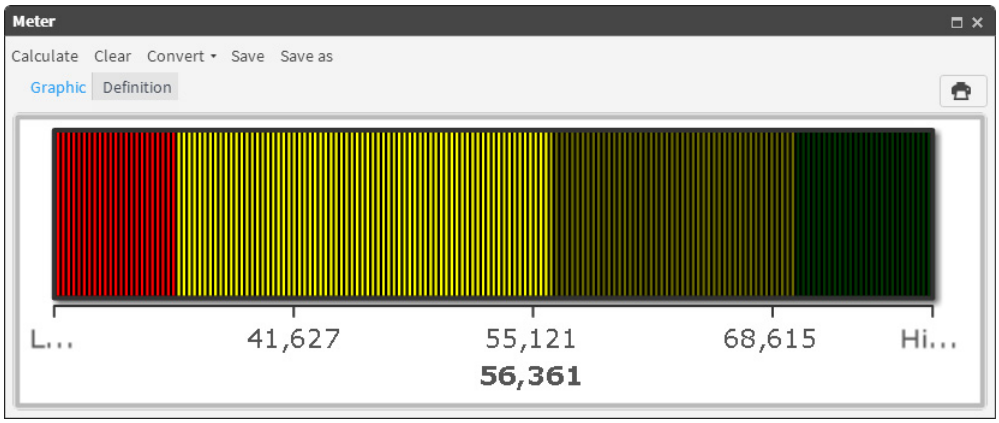


Figure 5-13 Visualizing a data measure using a meter

How to convert a dial graph to a sphere

1 For a calculated dial, select Convert, then choose Sphere.

2 To view Measures, Values, and Properties, choose Definition.

3 Choose Calculate. The new sphere appears on Graphic, as shown in Figure 5-14.

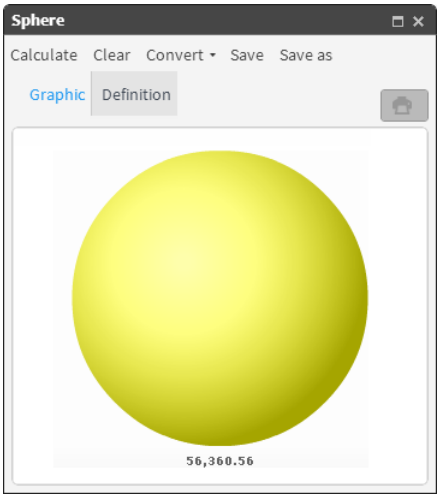


Figure 5-14 Visualizing a data measure using a sphere

How to convert a dial graph to a cylinder

1. For a calculated dial, select Convert, and choose Cylinder.
2. To view Measures, Values, and Properties, choose Definition.
3. Choose Properties.
4. Using the palette, choose a color, as shown in Figure 5-15.

Figure 5-15 Formatting properties of a data visualization gadget

5. Choose Calculate. The cylinder appears on Graphic, as shown in Figure 5-16.

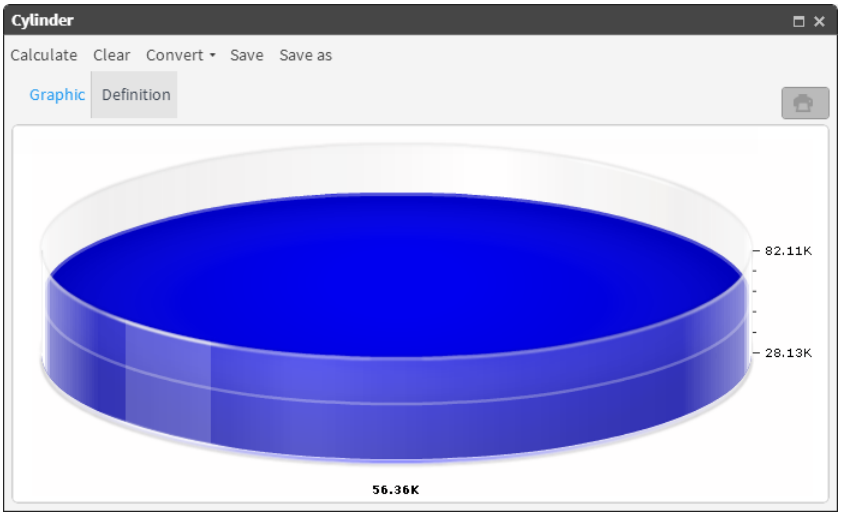


Figure 5-16 Visualizing a data measure using a cylinder

Using a meter

A meter is an indicator that uses a color bar display to show if the data are within established parameters. The following tabs are in the Definition tab:

- Measure tab: Use the buttons to add or delete measures and formulas. To edit an existing measure, double-click it.
- Values tab: Type maximum and minimum values and the value to be represented by measures.
- Properties tab: Type the values to be used as the limits of the areas to be represented. Choose the color.

You can convert a meter into a dial, sphere, or cylinder indicator.

How to create measures to display on a meter

1. In Gallery—Meter, choose Definition.
2. To create a measure for the meter, choose New measure.
3. In Alias, type a name for the new measure.
4. Select Maximum as the function for the new measure. From My Data, drag a column on which to base the measure and drop it next to Maximum, as shown in Figure 5-17.

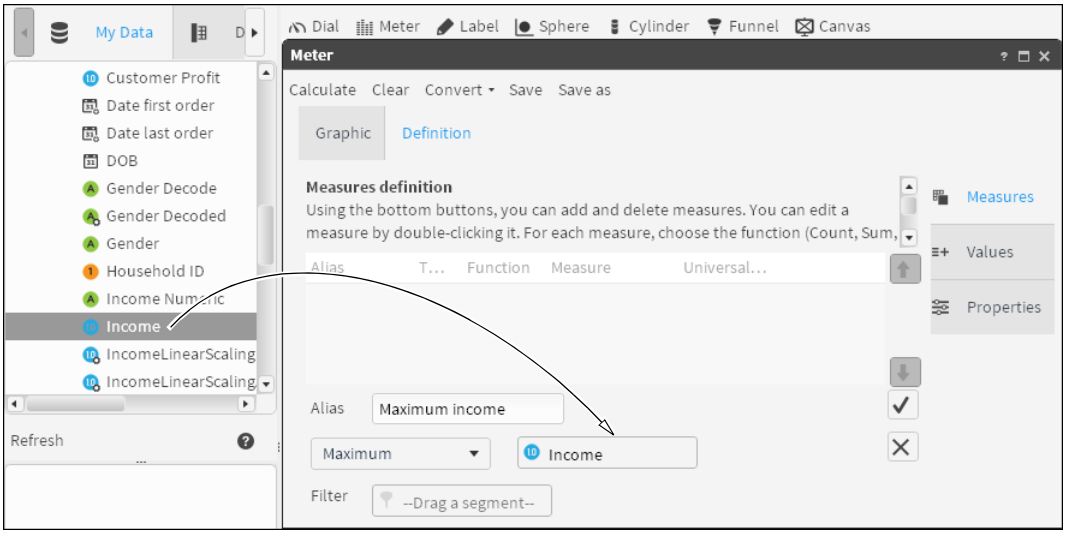


Figure 5-17 Defining a new measure for a meter

5. To create a filter, double-click a column in My Data. Drag a field from Discrete Values, and drop it in Filter, as shown in Figure 5-18.
6. To create the new measure, choose the check mark.
7. To create another measure with properties similar to the first, select a measure, and choose Duplicate.
8. In Alias, type a name and change the function to Minimum, as shown in Figure 5-19.

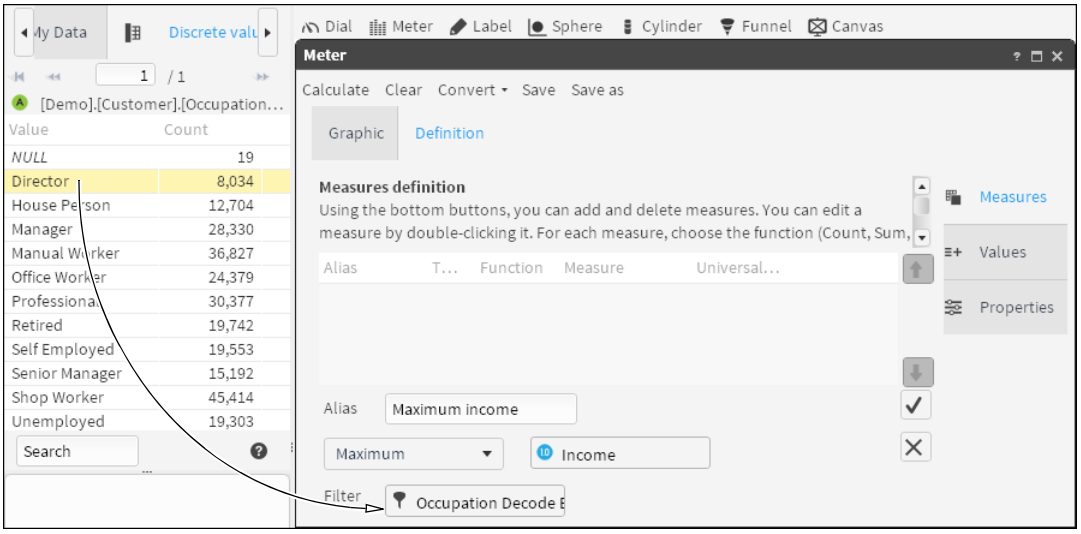


Figure 5-18 Adding a filter to a measure

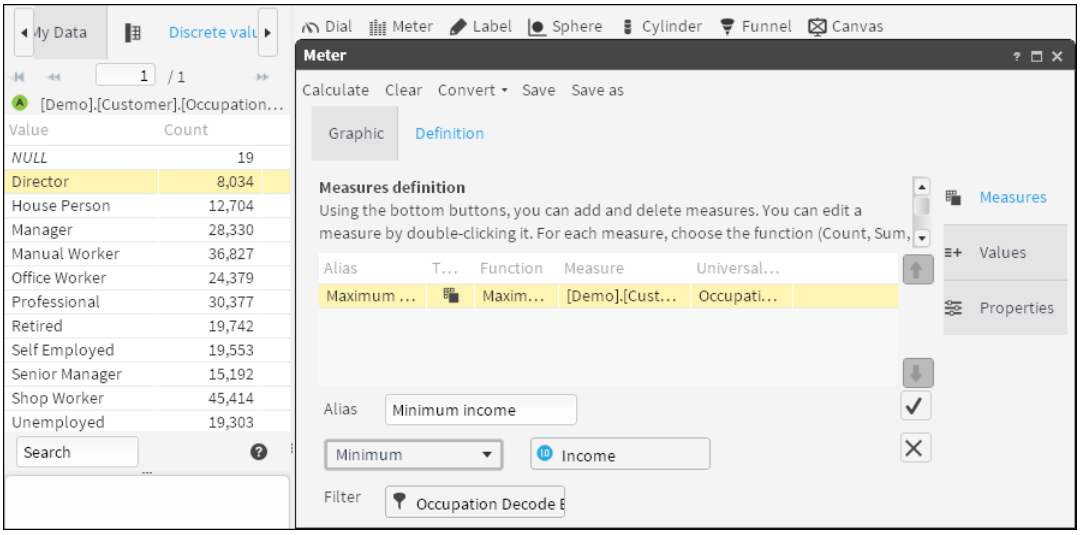


Figure 5-19 Creating a new measure based on a measure having similar properties

9 To create the new measure, choose the check mark.

10 Repeat steps 7–9, changing the function to Mean.

How to add formulas and thresholds to a meter

1 To create a formula, choose New formula.

2 In Alias, type a name for the formula, then type a formula for a lower threshold in the dialog box below Alias. In this example, the formula is a value.

3 Choose Validate. At the message that the formula definition is valid, shown in Figure 5-20, choose OK.

4 To create the new formula, choose the check mark.

5 Repeat steps 11–4, entering a formula for a higher threshold.

6 Choose Values.

7 For each value to appear on the meter, select an appropriate measure from the list. For example, select measures that correspond to minimum, mean, and maximum income values, as shown in Figure 5-21.

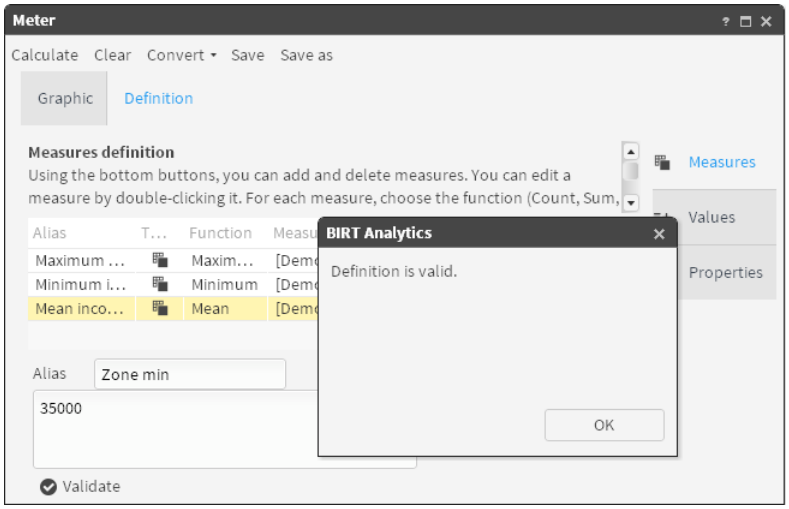


Figure 5-20 Validating the definition for a formula

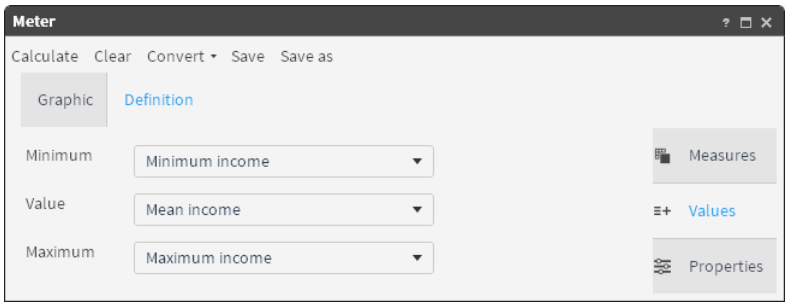


Figure 5-21 Assigning measures to values that will appear on a meter

How to add range properties to a meter

1. Choose Properties.
2. Choose Add range.
3. To indicate the minimum and maximum thresholds for the range, choose a measure or formula, as shown in Figure 5-22.
4. In Color, open the palette and select a color for the range. Choose OK.

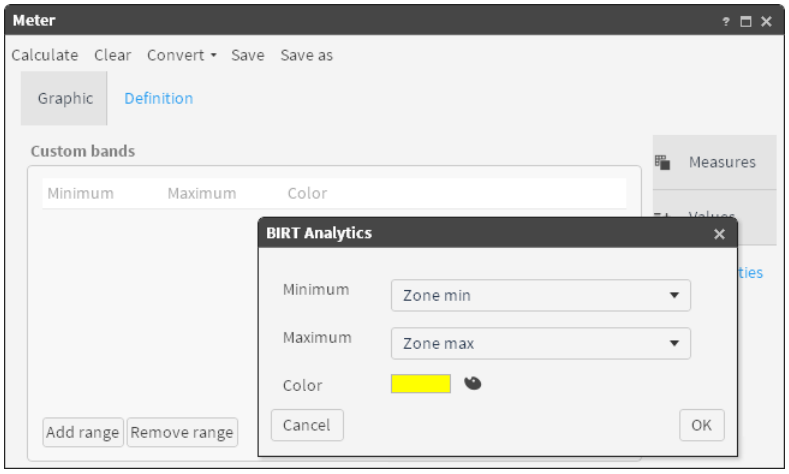


Figure 5-22 Configuring properties for a meter

5. Repeat steps 2–4 until you have created three ranges.

6 Choose Calculate to view the meter. Different color shades indicate different ranges, as shown in Figure 5-23.

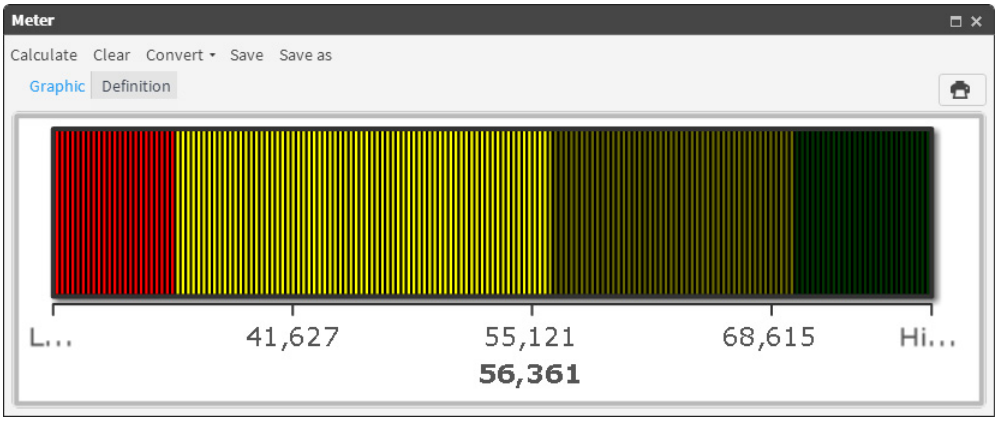


Figure 5-23 Viewing data measures on a meter

Using a label

A label indicator displays the values of defined measures as text. Use labels to create comments that can be included in the canvas.

How to create a measure for a label

1 In Gallery—Label, choose Definition—Measures.

2 Choose New measure.

3 In Alias, type a measure name.

4 From the list of function names, select a function name.

5 From My Data, drag an item on which the selected function will perform and drop it next to the selected function name. For Count, you must drag a table. For all other functions, drag a column.

6 From Discrete Values or Scratchpad, drag a segment and drop it in Filter, as shown in Figure 5-24.

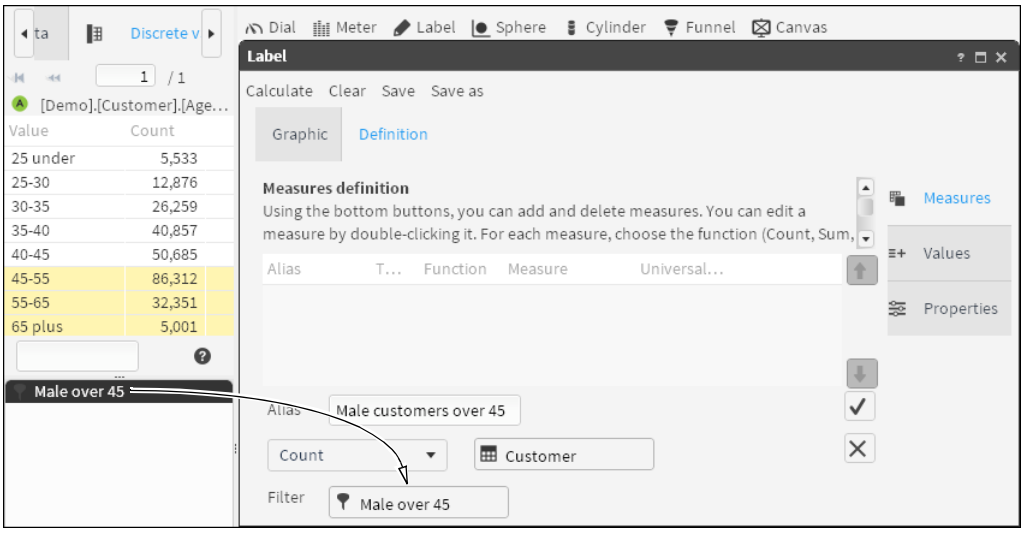


Figure 5-24 Adding a segment from Scratchpad as a filter for a label

7 Choose the check mark.

How to create a new, duplicate measure

1 Select an existing measure.

2 Choose Duplicate.

3 In Alias, type a unique name for the new measure.

4 Replace the function, column, or filter.

5 Choose the check mark.

How to define a label

1 In Label—Definition, choose Values.

2 In Text to show, type text that provides context for the label.

3 Right-click, then select a measure. The selected measure appears at the cursor location. The example in Figure 5-25 shows selecting a third measure for a text label.

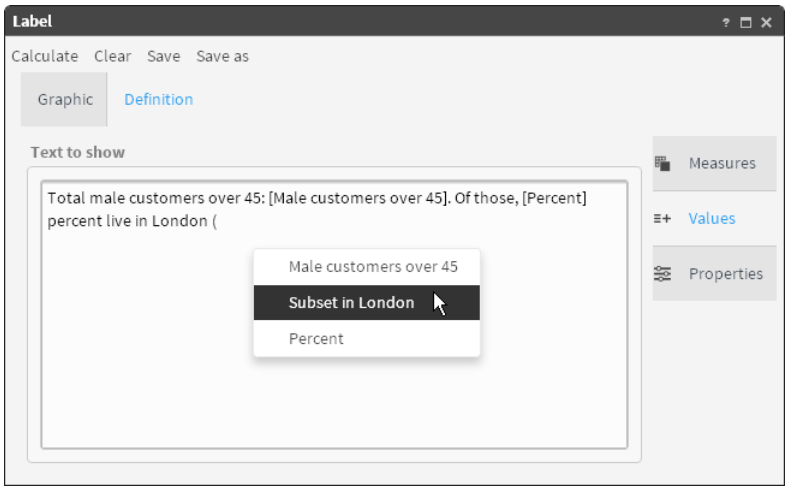


Figure 5-25 Selecting a measure for a label

4 Choose Properties. Then, select color, size, and font properties for the label text, as shown in Figure 5-26.

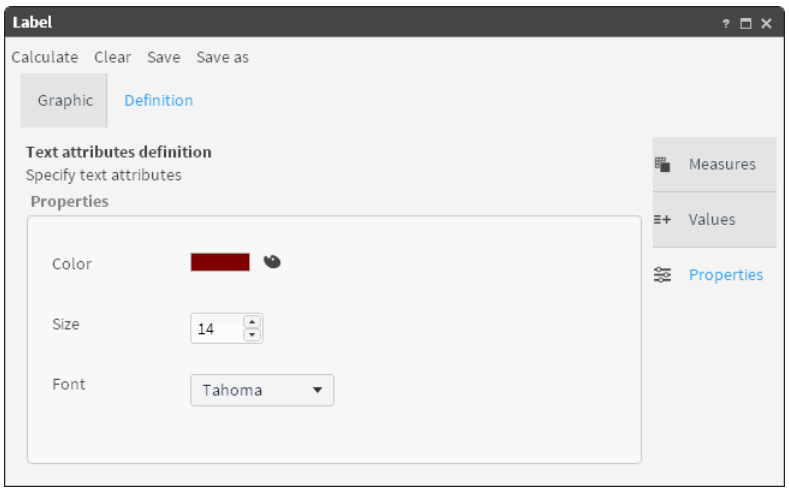


Figure 5-26 Selecting text properties for a label

5 To create the label, choose Calculate. The label appears in Graphic, as shown in Figure 5-27.

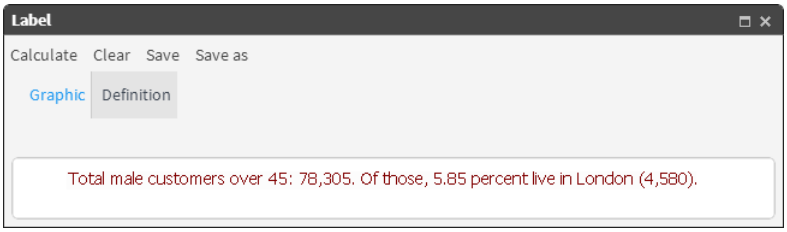


Figure 5-27 Previewing a complete label

6 To keep the label for future use, choose Save.

Using a sphere

A sphere displays data graphically using a round, three-dimensional shape that changes color based on the limit area of the value it represents. The following tabs are in the Definition tab:

- Measure tab: Use the buttons to add or delete measures and formulas. To edit an existing measure, double-click it.
- Values tab: Type the maximum and minimum values and the value to be represented by the measures.
- Properties tab: Select the values to be used as the limits of the areas to be represented. Choose the color.

You can convert a sphere into a dial, meter, or cylinder indicator.

How to add measures to a new sphere

1 In Gallery—Sphere, choose Definition.

2 Choose New measure.

3 Type a name for the new measure.

4 Select a function for the sphere. From Data Tree, drag a column on which to base the measure, as shown in Figure 5-28.

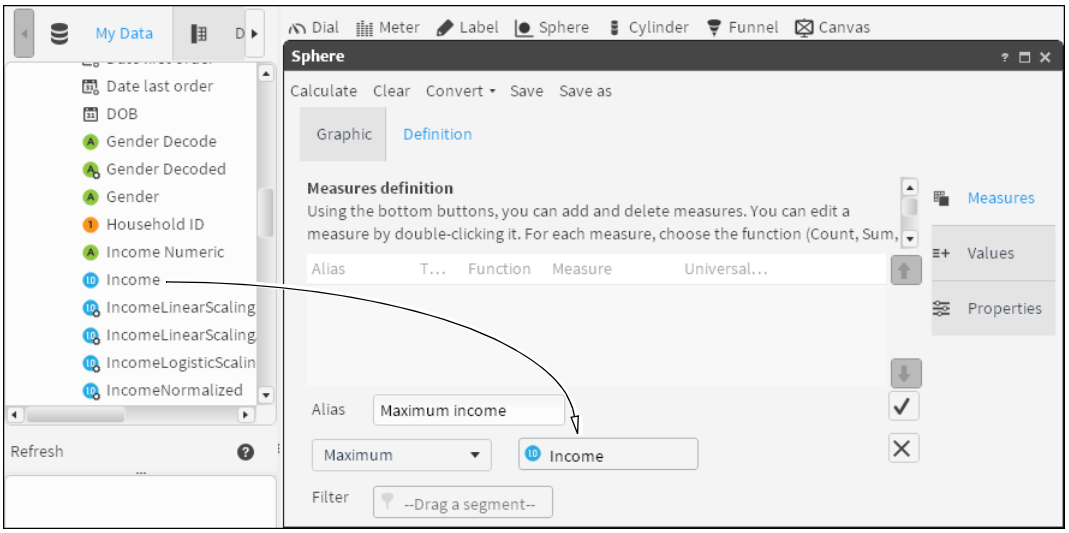


Figure 5-28 Defining a measure for a sphere

5 To create a filter, double-click one of the columns in Data Tree. In Discrete Values, drag a field and drop it in Filter, as shown in Figure 5-29.

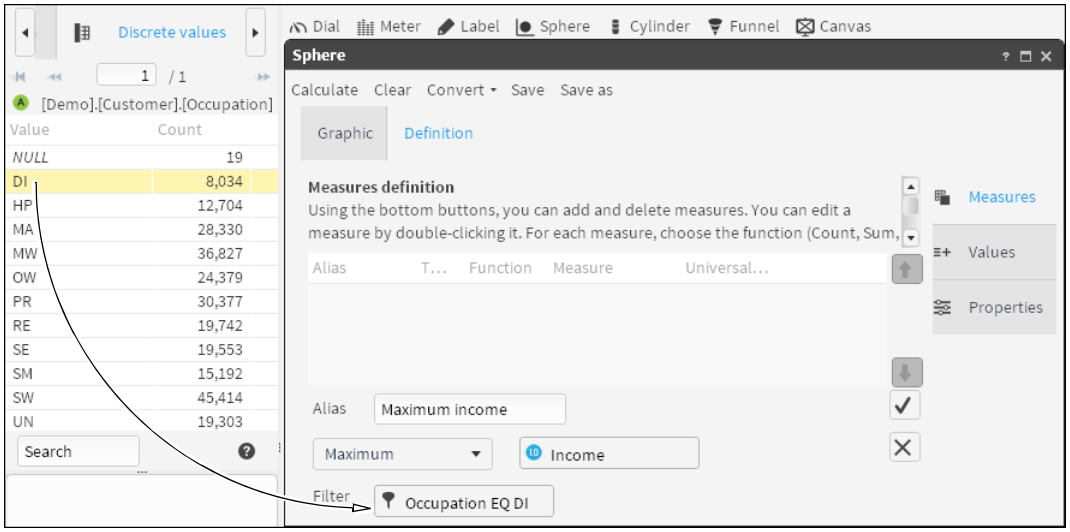


Figure 5-29 Adding a filter to a measure

6 To create the measure, choose the check mark.

7 To create another measure having properties similar to an existing one, select a measure and choose Duplicate.

8 Type a measure name and change the function to Minimum, as shown in Figure 5-30.

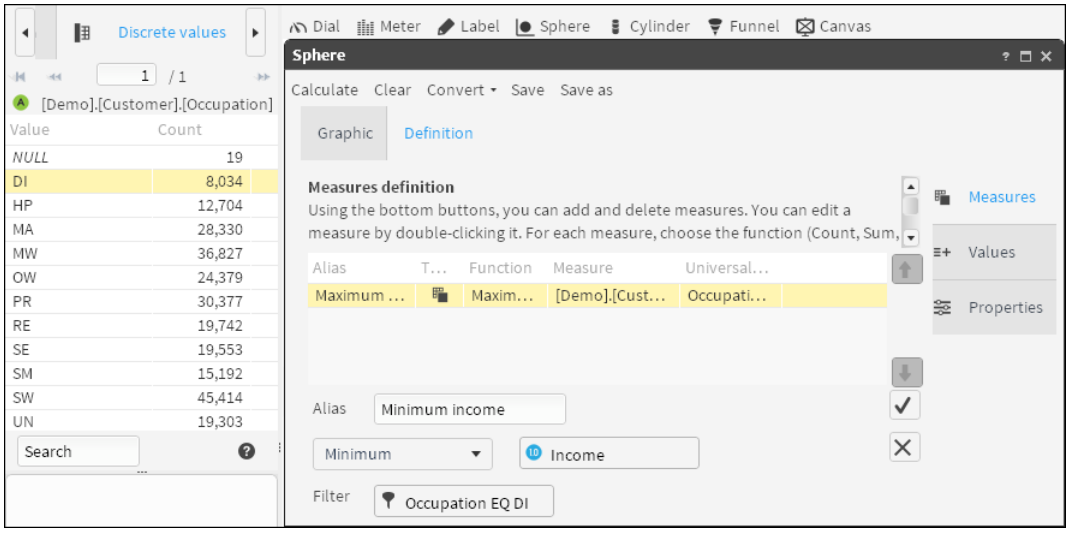


Figure 5-30 Creating a new measure having properties similar to an existing measure

9 To create the new measure, choose the check mark.

10 Repeat steps 7–9, changing the function to Mean.

How to add a formula to a sphere

1 Choose New formula.

2 Type a name for the formula, and enter a formula for a lower threshold in the dialog box below.

3 Choose Validate to ensure the function is valid. Choose OK.

4 To create the new function, choose the check mark.

5 Repeat steps 1–4, entering a formula for a higher threshold.

How to create ranges to display in a sphere

1 Choose Values.

2 Using the pull-down menus, select measures to indicate minimum, average, and maximum values to be used in the sphere, as shown in Figure 5-31.

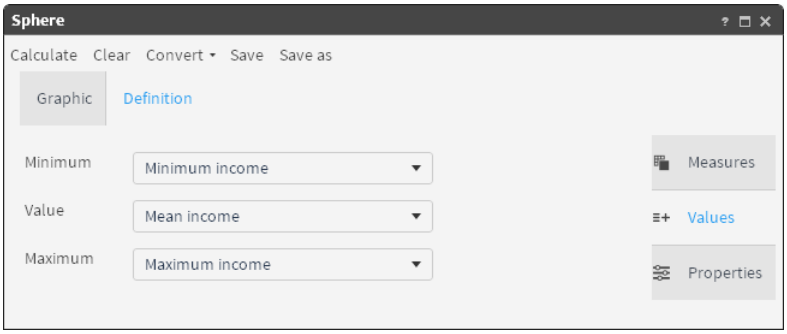


Figure 5-31 Assigning measures to values that appear in a sphere

3 Choose Properties.

4 Choose Add range.

5 Using the pull-down menus, select measures to indicate a minimum and maximum value to be used, as shown in Figure 5-32.

6 In Color, use the palette to apply a color to each range you create.

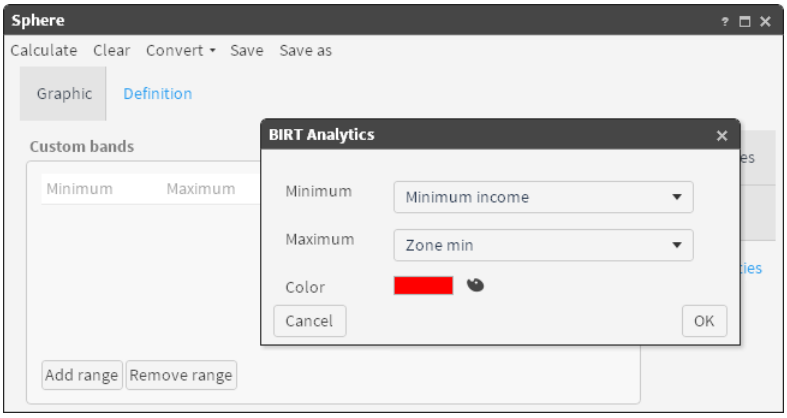


Figure 5-32 Configuring properties for a sphere

7 Repeat steps 4–6, until you create three ranges.

8 Choose OK.

9 Choose Calculate. The sphere appears on Graphic, as shown in Figure 5-33.

Using a cylinder

A cylinder is an indicator that shows the volume of a cylinder depending on the value it represents. The following tabs are in the Definition tab:

- Measure tab: Use the buttons to add or delete measures and formulas. To edit an existing measure, double-click it.
- Values tab: Type the upper and lower limits of the indicator. Choose the measure to be represented.
- Properties tab: Choose the color of the cylinder.

You can convert a cylinder into a dial, meter, or sphere indicator.

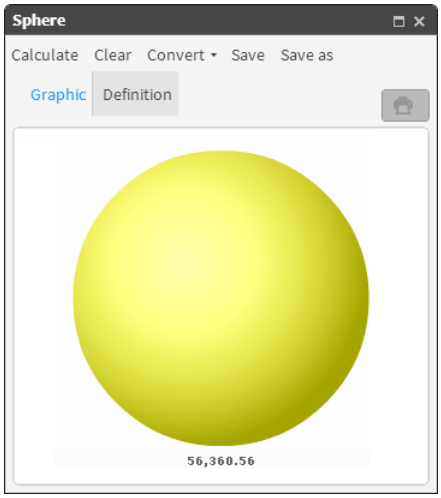


Figure 5-33 Visualizing a data measure using a sphere

How to add measures to a new cylinder

1. In Gallery—Cylinder, choose Definition.
2. Choose New measure.
3. Type a name for the new measure.
4. Select a function for the cylinder. From Data Tree, drag a column on which to base the measure, as shown in Figure 5-34.

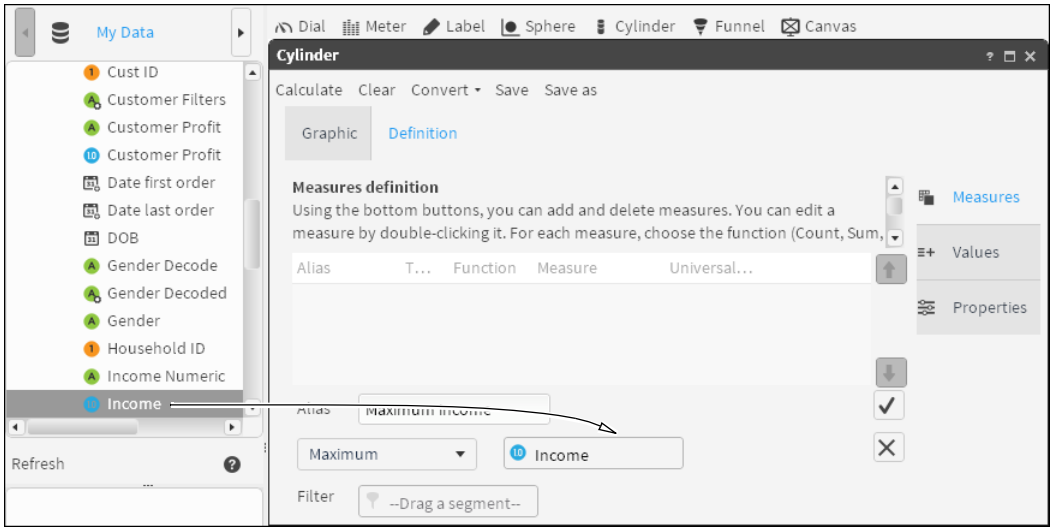


Figure 5-34 Defining a measure for a cylinder

5. To create a filter, double-click one of the columns in Data Tree. In Discrete Values, drag a field and drop it in Filter, as shown in Figure 5-35.
6. To create the measure, choose the check mark.
7. To create another measure with properties similar to an existing one, select a measure and choose Duplicate.
8. Type a measure name and change the function to Minimum, as shown in Figure 5-36.

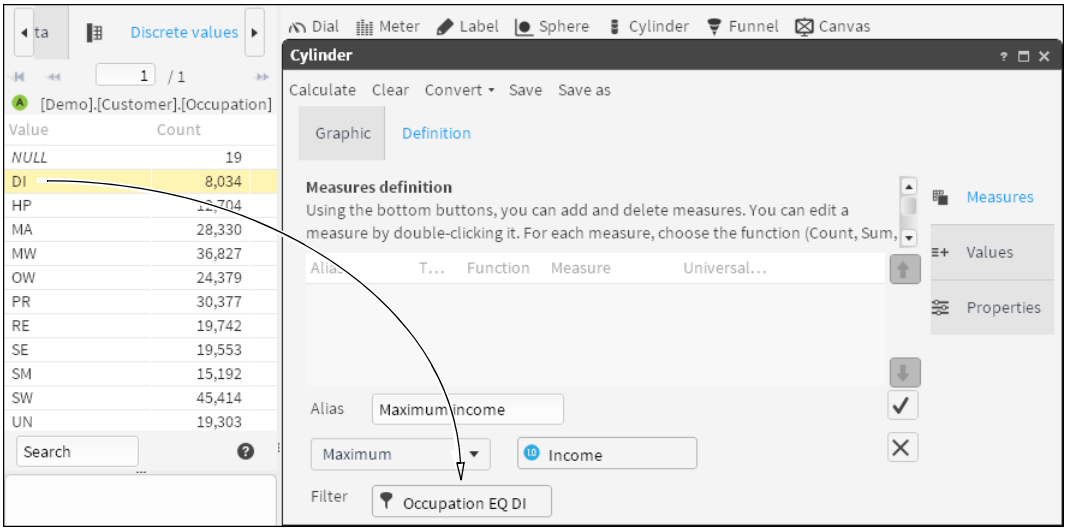


Figure 5-35 Adding a filter to a measure

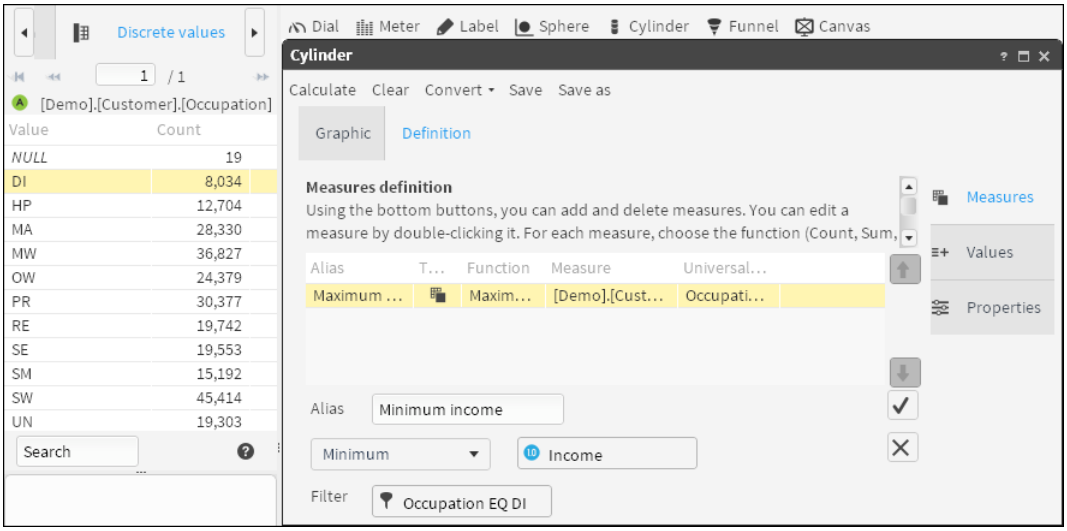


Figure 5-36 Editing the definition of a duplicate measure

9 To create the new measure, choose the check mark.

10 Repeat steps 7–9, changing the function to Mean.

How to add values and properties to the cylinder

1 Choose Values.

2 Using the pull-down menus, choose measures that indicate minimum, average, and maximum values to appear on the cylinder, as shown in Figure 5-37.

3 Choose Properties.

4 In Filling color, use the palette to apply a color to the range you created, as shown in Figure 5-38.

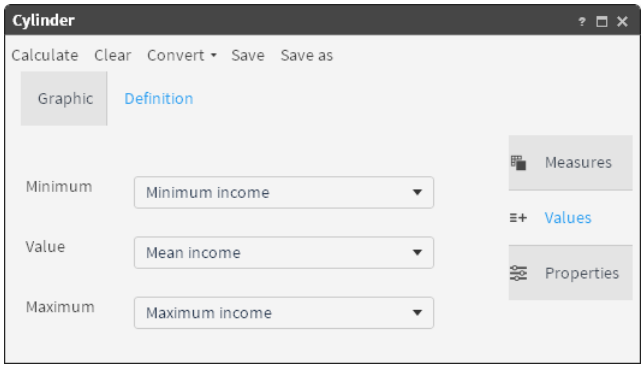


Figure 5-37 Assigning measures to values that appear on a cylinder

Figure 5-38 Configuring properties for a cylinder

5 Choose Calculate. The cylinder appears on Graphic, as shown in Figure 5-39.

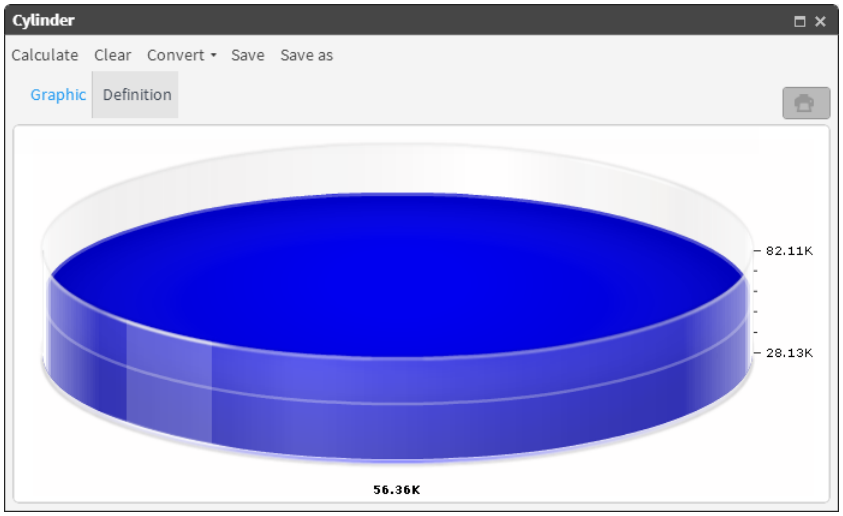


Figure 5-39 Visualizing data measures using a cylinder

Using a funnel

A funnel displays simplified data in graphic form using several, stacked slices. Each slice represents a group of data. A slice that you select on the funnel separates from the rest of the funnel to highlight the visualization. The following tabs appear in the Definition tab:

- Measure tab: Use the buttons to add or delete measures and formulas. To edit an existing measure, double-click it.
- Values tab: Provide the data to be used in the segmentation and measure.
- Properties tab: Select whether to display values or percentages.

How to add measures to a new funnel

1 In Gallery—Funnel, choose Definition.

2 Choose New measure.

3 Type a name for the new measure.

4 Select a function for the funnel. From Data Tree, drag a column on which to base the measure, as shown in Figure 5-40.

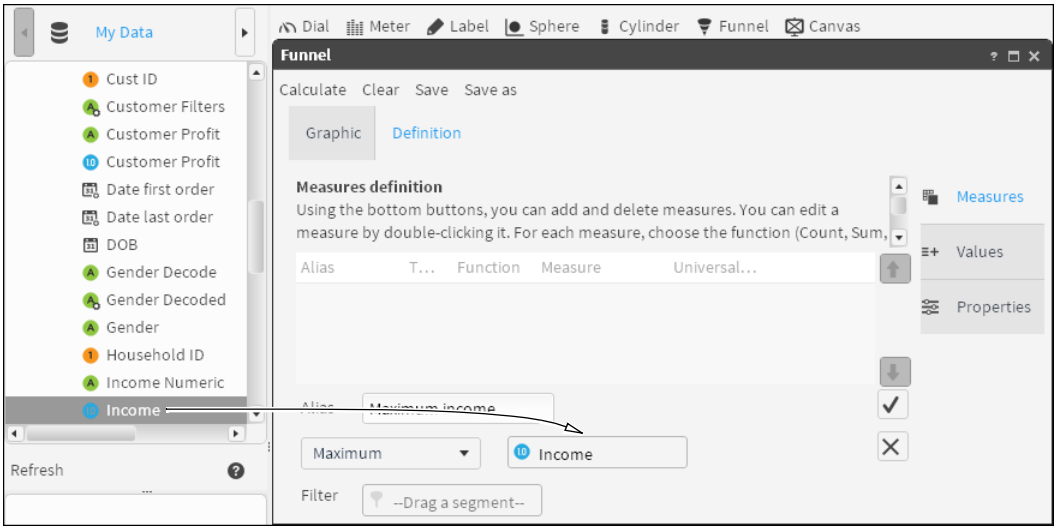


Figure 5-40 Defining a measure for a funnel

5 To create a filter, double-click one of the columns in Data Tree. From Discrete Values, drag a field and drop it in Filter, as shown in Figure 5-41.

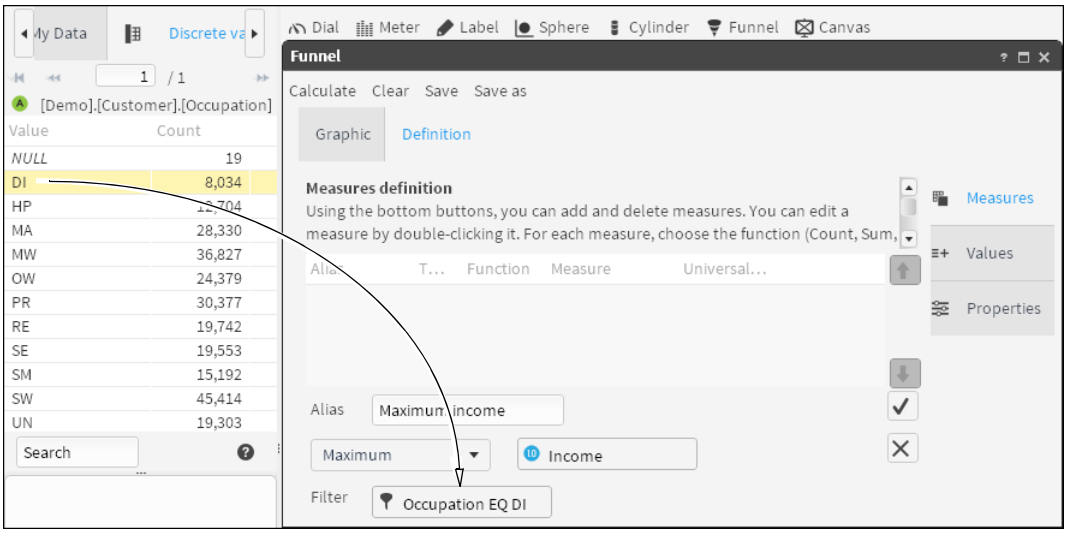


Figure 5-41 Adding a filter to a measure

6 To create the measure, choose the check mark.

How to add values and properties to a funnel

1 Choose Values.

2 Drag a column from My Data and drop the column in Section, as shown in Figure 5-42.

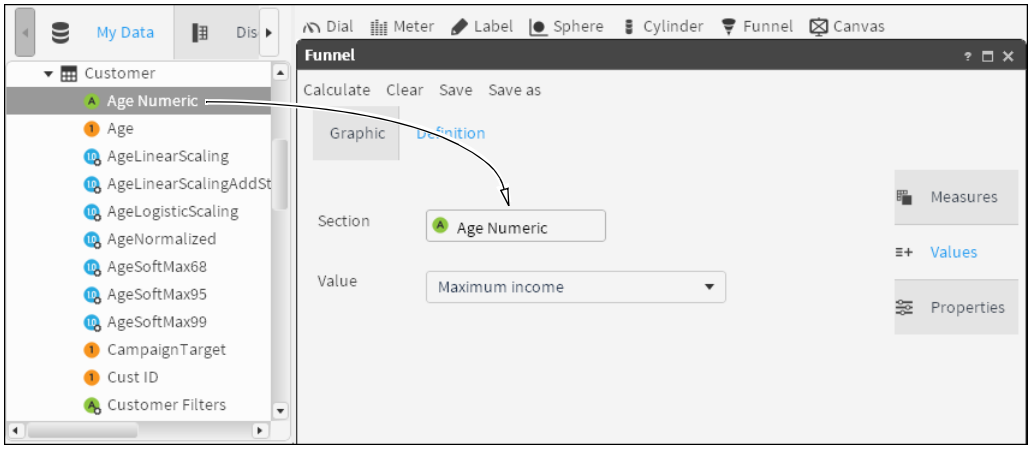


Figure 5-42 Assigning a measure to a value that appears on a funnel

3 Select a measure on which to base a funnel.

4 Choose Properties.

5 In View As, choose an option to represent the data. For example, Figure 5-43 shows choosing to view data as values.

Figure 5-43 Configuring properties for a funnel

6 Choose Calculate. The funnel appears on Graphic. To see the values of a section, hover the cursor over the section, as shown in Figure 5-44.

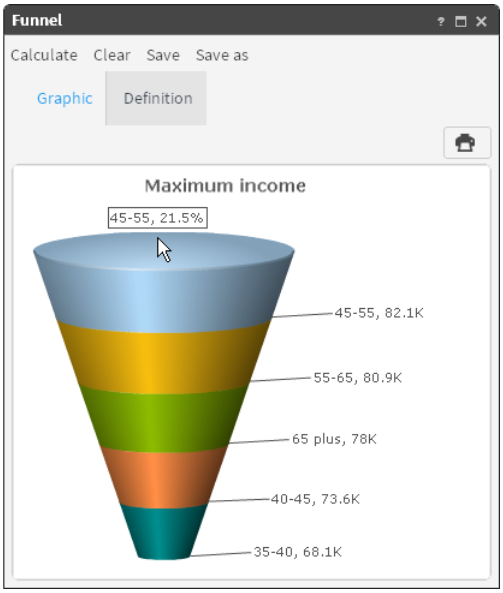


Figure 5-44 Visualizing a data measure using a funnel

Working with a Canvas

Canvas supports getting summaries of a series of previously defined indicators, for example a collection of indicators that simultaneously displays information about the number and amount of orders. Canvas has two modes, view and edit. The view mode, shown in Figure 5-45, does not allow you to make any modifications to content. The edit mode enables you to add or delete one or more components in the command box.

To configure an analysis, select the unlock icon to access the edit analysis mode. Then drag and drop the saved analyses or indicators from the analysis tree to the window to add them to the command box. These components are placed sequentially in the column.

Figure 5-45 Canvas

When you drag a component, arrows appear indicating the area where you can drop it, as shown in Figure 5-46.

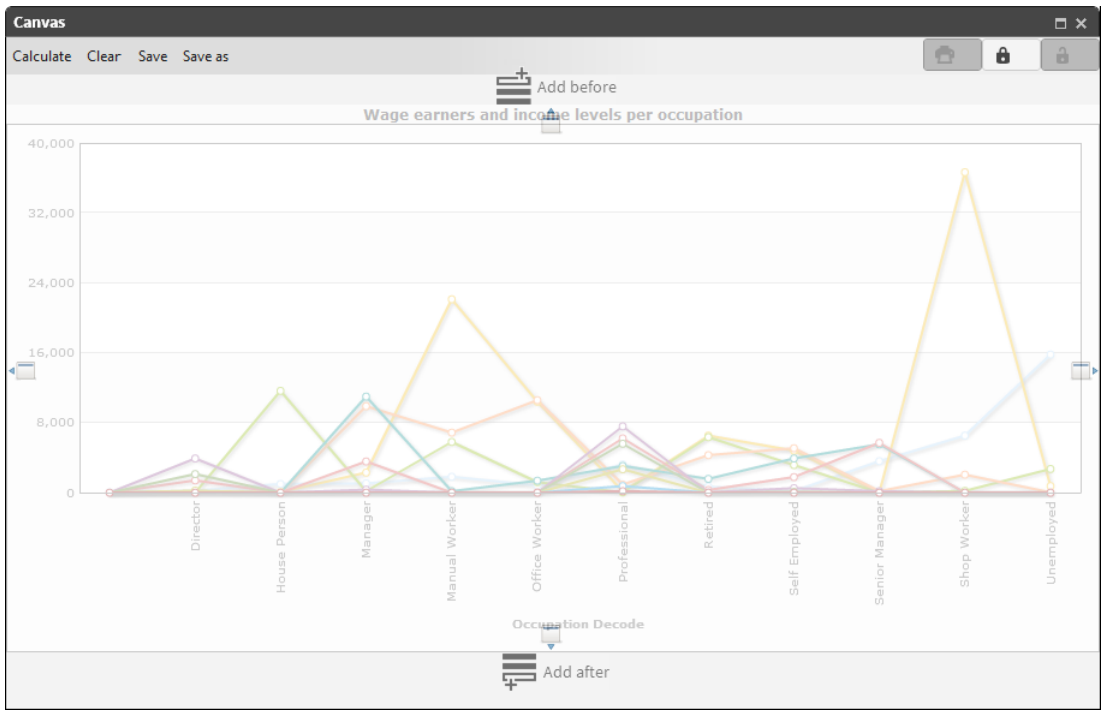


Figure 5-46 Arrows indicating where to drop a component

When you drag a new component to Add before or Add after, the analysis is dropped in a new row, as shown in Figure 5-47.

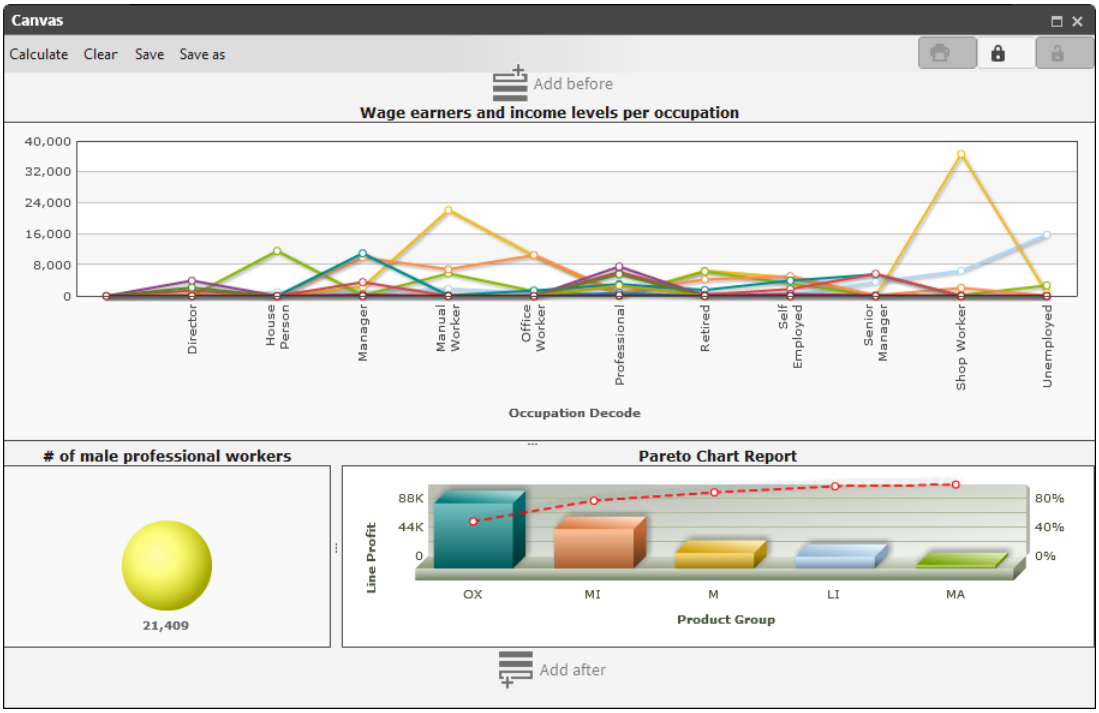


Figure 5-47 Analysis dropped in a new row

To change the size of the design areas, you use the sliders that appear in those areas. You can move canvas components by dragging them to another area or delete them by right-clicking the component and choosing Remove. In edit mode, you can modify components by right-clicking the component and choosing Open. Reopen the canvas to view saved changes.

When you finish placing components in the canvas and you have set its size, click the lock icon to change to query mode and finalize the canvas.

Any previously defined indicator analyses can be added to the canvas. It is also possible to add the graphic results from a crosstab, a bubble diagram, a Venn diagram, map, or Pareto. In the case of a crosstab or bubble, you can choose the display form as a table or graph.

Chapter

6

Identifying and predicting data trends

This chapter contains:

- Understanding data mining and predictive analytics
- Preprocessing - Preparing data for mining
- Understanding Clustering
- Understanding Forecasting
- Understanding decision trees
- Understanding association rules
- Understanding correlation
- Understanding linear regression
- Understanding logistic regression
- Understanding Naive Bayes classification

Understanding data mining and predictive analytics

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. This information can be converted into *knowledge* about historical patterns and future trends. For example, sales information can be analyzed in light of promotional efforts to provide knowledge of future consumer buying behavior.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. It involves the following processes:

- *Anomaly detection*: The identification of unusual data records.
- *Association rule learning:* The identification of relationships between variables. For example, association rule learning can help a supermarket to determine which products are frequently bought together through the analysis of its point-of-sale data.
- *Clustering:* Discovering groups and structures in the data that are similar.
- *Classification:* Determining the class of an object based on its attributes. For example, an email program classifies e-mails as legitimate or as spam.
- *Correlation and Regression:* Finding data relationships and applying functions that model the data with the least error.
- *Summarization:* Providing a more compact representation of the data set, including visualization and report generation.
- *Sequential pattern mining and prediction:* Finding sets of data items that occur together frequently in some sequences. Sequential pattern mining is the basis for web user analysis, stock trend prediction, DNA sequence analysis or using a history of symptoms to predict disease.

BIRT Analytics harnesses all these processes to provide a complete analysis of your data. Its predictive analytics functions enable, among others, time-series predictions and the production of short-term demand forecasts.

The use of predictive analytics has been very useful in the following domains:

- Customer relationship management (CRM)
- Clinical decision support systems.
- Debt Collection
- Cross-selling
- Customer retention
- Direct marketing
- Fraud detection
- Risk management
- Insurance Underwriting

For more information about data mining and predictive analytics, see the following pages:

`http://en.wikipedia.org/wiki/Data_mining`
`http://en.wikipedia.org/wiki/Predictive_analytics`

Preprocessing - Preparing data for mining

I BIRT Analytics preprocessing operations provide normalization, scaling and Boolean column creation as well as remapping operations that prepare a data column to meet

conditions required by each data mining algorithm. Preprocessing operations simply apply a relevant mathematical operation to values in a chosen data set.

Understanding Boolean column creation

Some data mining operations require Boolean columns. BIRT Analytics preprocessing tools make it possible to create Boolean columns from an existing column. Each individual new column presents one of the values from the original column as a Boolean value (0/1).

For example, a column containing only a binary response with (0/1) female values can be created from a gender column (with both male and female values as well as a neutral value).

How to create Boolean columns

1. In the Advanced tab of the Analytics tool, open the Preprocessing tab.
2. Click on *Create Boolean columns (0/1)* to open the Create Boolean columns window.
3. Drag the Gender column from the Customer table in the Data Tree and drop it into the Column field in the Create Boolean columns window.
4. Click Create. Three newly created columns appear in the Data Tree: one for females (Gender EQ F) one for males (Gender EQ M) and another for the neutral value (Gender EQ).

Standardizing data in a column

Preprocessing a column of data values having a distribution that differs from a standard, or normal, distribution before applying a data mining algorithm can produce a more useful result. For example, you can compare data sets that have different scales and units by standardizing the data so that it falls in the 0 to 1 range. Test scores are often calibrated by percentile, with most scores falling in the 25th to 75th percentile.

Figure 6-1 shows the distribution of raw, or non-standardized, data for age and income. Ages fall in the 19 to 93 range, while incomes fall in the 479.79 to 111571.4 range. To compare these distributions, you must standardize the data.

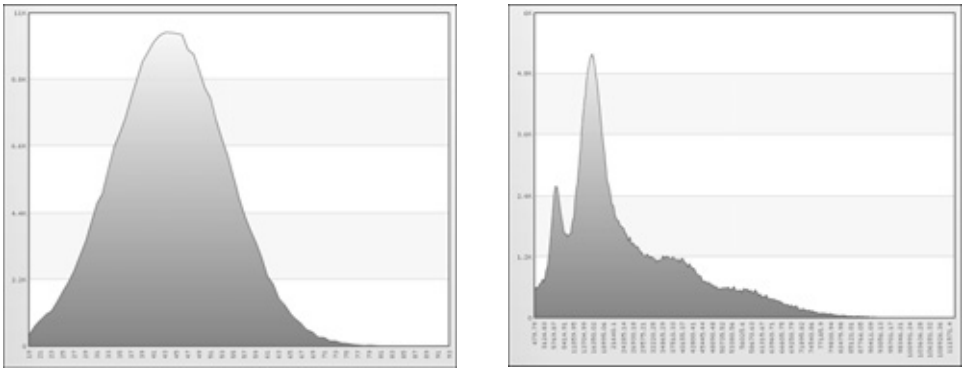


Figure 6-1 Distribution of raw data for age (left) and income (right)

There are four ways to standardize the data in a column:

- Normalization
- Linear scaling
- Logistic scaling
- Softmax scaling

In each case, a new column is created to contain the standardized data.

Understanding normalization

Normalization calculates the mean of all values in a column. Each value in the new column compares in the same way to values in a standard, normal distribution. The operation subtracts the mean value from each value in the column, then divides the difference by the standard deviation. The formula is:

$y = (x - \text{mean}\{x_1, x_N\}) / (\text{stdv}\{x_1, x_N\})$

Standard deviation shows how much variation there is from the average (mean), or expected value. A low standard deviation indicates that the data points tend to be very close to the mean. A high standard deviation indicates that the data points are spread out over a large range of values.

Figure 6-2 shows normalized data for age and income. The values on the horizontal axis represent the number of standard deviations from the mean. The standard deviation of the mean is 0.

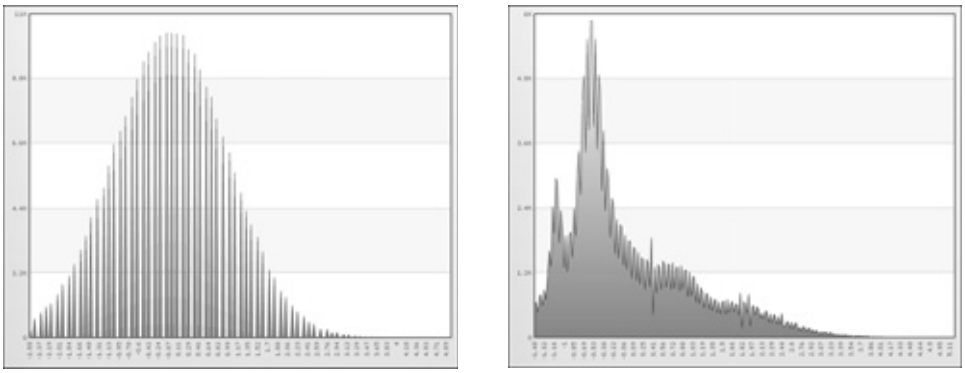


Figure 6-2 Normalized data for age (left) and income (right)

Understanding linear scaling

Standardization by linear scaling is useful when values in a column have the following characteristics:

- Many similar values
- Very extreme minimum and/or maximum values that are highly concentrated

The more the data are clustered, the better the result obtained by linear regression study. Maximum and/or minimum values are calculated to be suitable to perform linear regression analysis. Linear scaling supports two options:

- Use the original minimum and maximum values.
- Add a stretch to the minimum and maximum values. This option adds, according to the percentage indicated, a minimum and maximum value before the data is preprocessed.

The formula is:

$y = (x - \min\{x_1, x_N\}) / (\max\{x_1, x_N\} - \min\{x_1, x_N\})$

Figure 6-3 shows linear scaling with the original minimum and maximum values for age and income.

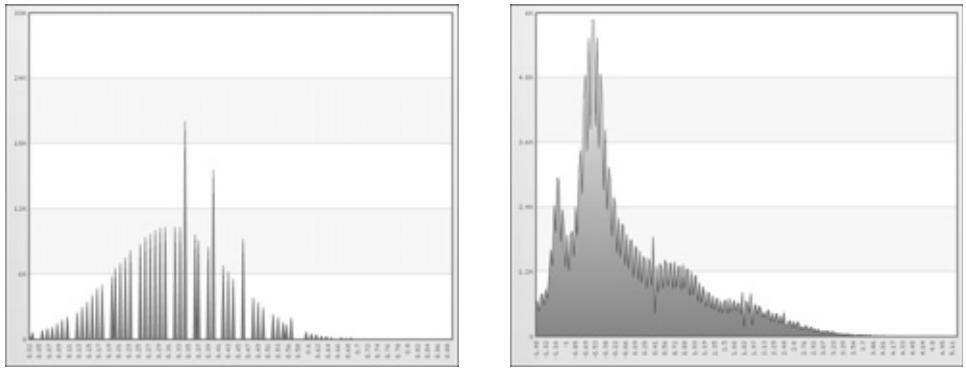


Figure 6-3 Linear scaling with original minimum and maximum values for age (left) and income (right)

Figure 6-4 shows linear scaling with a stretch to the minimum and maximum values for age and income.

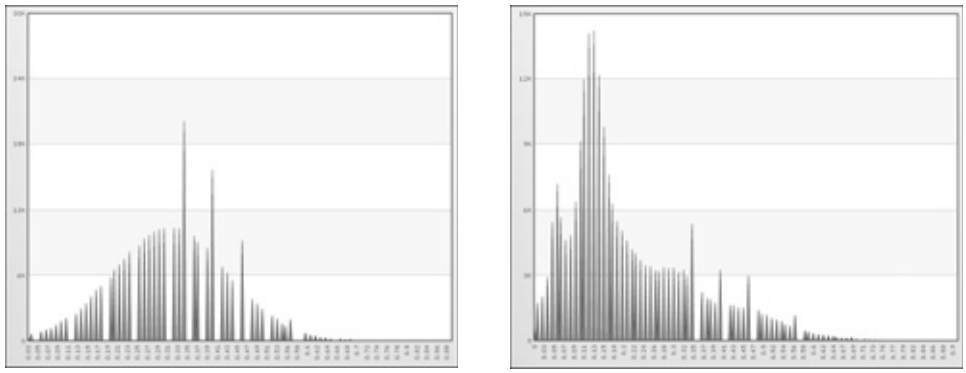


Figure 6-4 Linear scaling with a stretch to the minimum and maximum values for age (left) and income (right)

Understanding logistic scaling

Standardization by logistic scaling is used to recode the variable of study for use in a logistic regression. Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of values) based on one or more predictor variables. The equation used is:

$$P(n) = 1/(1 + e^{-n})$$

where n represents the values in the column. This equation analyzes the values to form a logistic model.

Understanding Softmax scaling

Softmax scaling standardization is a nonlinear transformation that reduces data ranges for the values in a column as much as possible. The objective is to achieve the minimum and maximum values asymptotically. In other words, the low-end and high-end values gradually approach the minimum and maximum values without ever reaching them.

If you choose Softmax scaling, you can set the confidence level to 68%, 95%, or 99%. The lower the confidence level is, the shorter the intervals and the greater the probability of error will be. The formula is:

$$x' = x - E(x) / \lambda(\sigma_x/2\pi)$$

- x represents the values in the column.
- E(x) is the average value.

- λ is the confidence level.
- σ_x is the standard deviation of the study variable.
- π is the number 3.14159…

Figure 6-5 shows Softmax scaling at 68% for age and income.

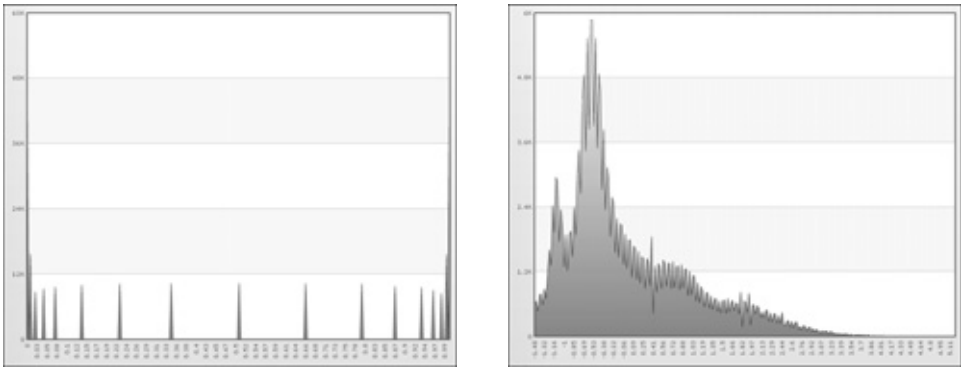


Figure 6-5 Softmax scaling at 68% for age (left) and income (right)

Figure 6-6 shows Softmax scaling at 95% for age and income.

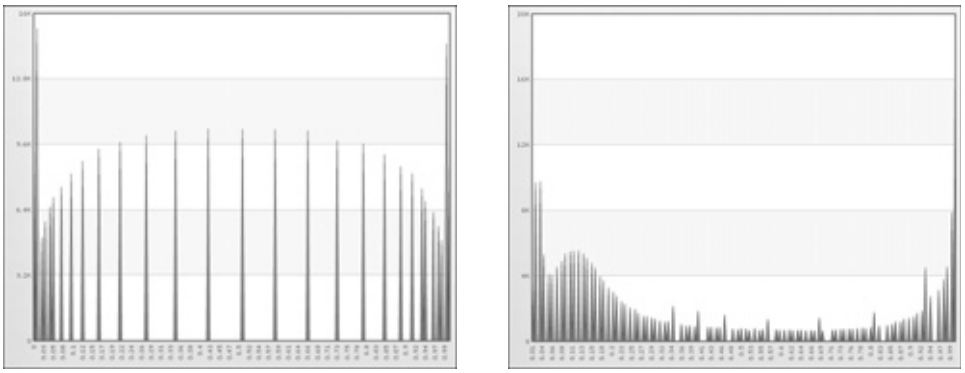


Figure 6-6 Softmax scaling at 95% for age (left) and income (right)

Figure 6-7 shows Softmax scaling at 99% for age and income.

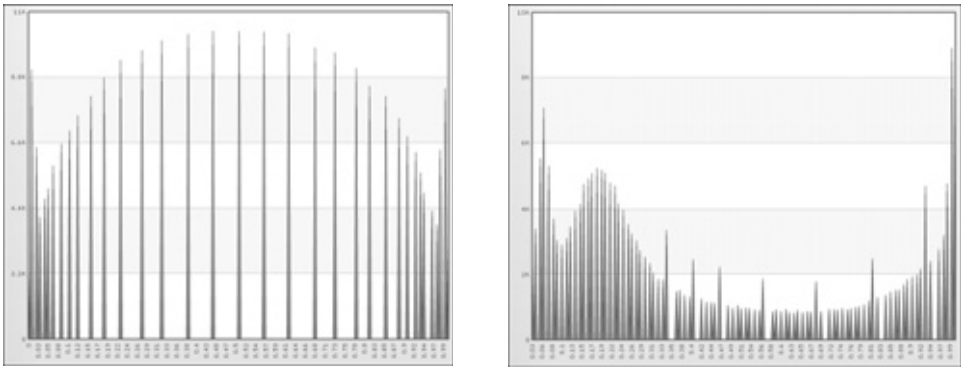


Figure 6-7 Softmax scaling at 99% for age (left) and income (right)

How to standardize the data in a column

1. In Analytics—Advanced, choose Preprocessing➛Standardize Column.
2. Drag the column you want to standardize from My Data and drop it in Column in Standardize column, shown in Figure 6-8.
3. In New column, type the name of the new column.

4 Choose Normalize, or choose one of the scaling algorithms: linear scaling, logistic scaling, or SoftMax scaling. If you choose linear scaling or SoftMax scaling, you can customize the scaling.

5 Choose Create. The column appears in My Data.

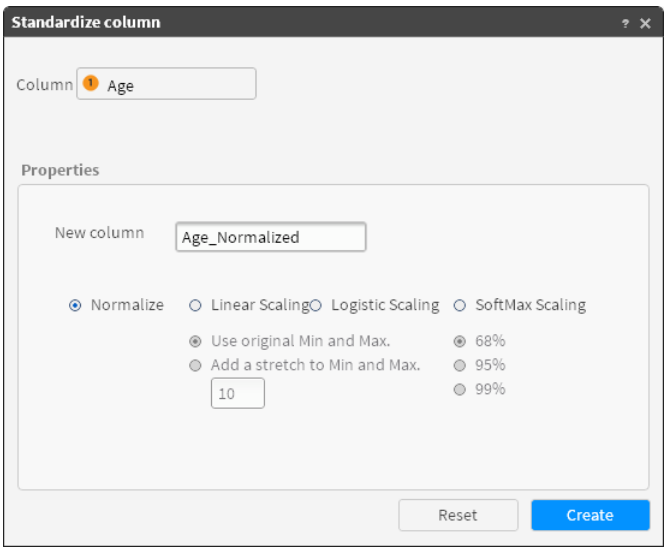


Figure 6-8 Standardizing the data in a column

Remapping a column

If you want to use a column that contains strings in an analysis, you must remap the strings to numbers. For example, Figure 6-9 shows the strings in the Occupation column on the left and the numbers in the remapped Occupation column on the right.

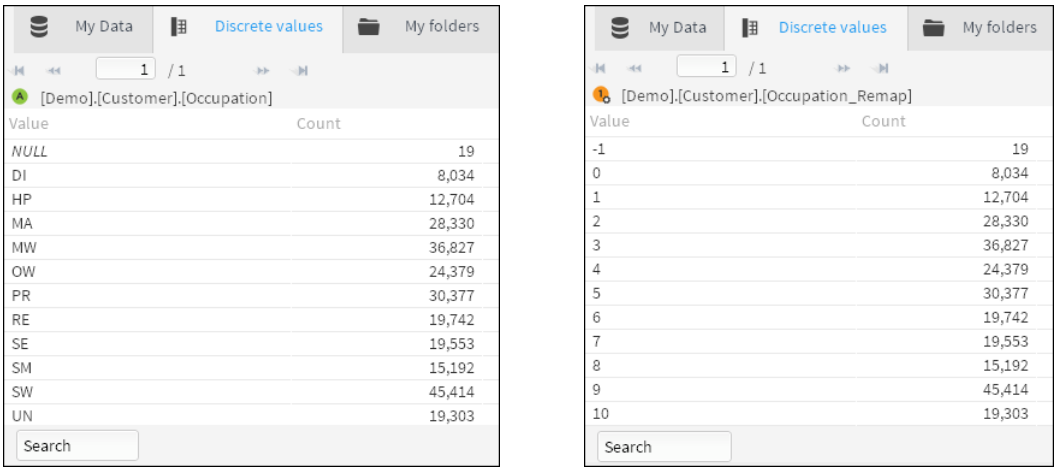


Figure 6-9 Occupation column (left) and remapped Occupation column (right)

How to remap a column

1 In Analytics—Advanced, choose Preprocessing➛Remap Column.

2 Drag the column you want to remap from My Data and drop it in Column in Remap column, shown in Figure 6-10.

3 In New column, type the name of the new column.

4 Choose Create. The column appears in My Data.

Figure 6-10 Remapping a column

Understanding Clustering

Clustering enables you to organize data based on variables you specify. Usually, the clustering algorithm produces segments of data that help you identify groups with the largest number of attributes in common. In other words, clustering provides an idea about the similarity and differences between records in the same group.

After BIRT Analytics applies this algorithm, a new field is created in the selected table to group records into a specified number of clusters (N). Because each record is given a value for the clustering, you can see a count of records in each cluster in the Data Tree's Discrete Values view. For example, customers grouped in the same category or cluster may have common demographic features.

You must use continuous variables because clustering calculates the distance between values to set up a group, and only fields with continuous values work for clustering. Continuous means that there are many discrete values. Categorical variables, or fields with few discrete values like gender or occupation, do not work.

To set up a clustering model, create a training process.

How to set up a training process

1. Choose Parameters and specify the following:
   - Domain: The segment of data from the database. All linked tables are automatically added.
   - Confidence level: The representative sample size to create the groups.
   - Clusters: The number of groups.
   - The attributes or categories to create the groups. Only fields containing continuous values are available to choose. Add them by dragging from the list on the left to the area on the right.
2. Choose Train.

How to use the results

When training finishes, Results contains a list with all the groups that have been created, the number of records in each group, and the mean of each attribute used to set up the groups. Note that every mean value acts as a centroid of the group.

1. Review the results. After you save the cluster, you cannot train it again.
2. Save the trained cluster in My Folders, to an existing folder, or a newly created one.
3. Select the saved cluster in My Folders. Right-click and choose Open.
4. On K-Means, drag a new segment and drop it in Domain.

5 Type the new target column name.

6 Choose OK. The new column appears in the selected table.

Understanding Forecasting

Forecasting is a method of extrapolating or predicting data based on time. BIRT Analytics forecasting uses the Holt-Winters method, iteratively applying a formula to produce a time series and a forecast. This formula uses a weighted average of data prior to time t to provide a result for time t.

This method consists of three components: the level, trend, and seasonal component.

For example, to forecast the number of orders to be received during the next 12 months, you would perform the following tasks:

- Select the data you wish to observe, the *y*-axis of the time series.
- Select the time interval for the series, the *x*-axis of the time series.
- Execute pre-analysis, if necessary.
- Set model parameters. These may include:
  - Number of projections to make
  - Whether the time series has seasonality
  - Seasonal periodicity
  - Whether or not to replace outliers
  - Level, trend, and season smoothing parameters
- Execute forecast.

More about outliers

Outliers are observations that appear to deviate markedly from other members of the sample in which they occur. When running pre-analysis, values that are more than two standard deviations away from the sample mean are considered outliers. Outliers are replaced by the sample mean. Generally, outliers should be replaced so that they do not bias any projections.

The value of the previous observation replaces any null values encountered in the sample. If the null value encountered is the first observation of the sample, the value of the nearest non-null observation replaces the null value.

How to create and execute a forecast

To get a forecast of the number of orders to be received in the next 12 months, select Parameters and complete the following procedure:

1 From My Data, drag the Order table and drop it in Measure. Leave the default operation as Count.

2 Drag the Month column and drop it in Column, in the Dimension area.

3 To find out if your data contains a seasonal element, select Has seasonality and set Seasonal periodicity to 12. With monthly data, there are 12 observations in each cycle until the seasonal pattern is repeated.

4 Set other values, as required:

- Filter: Constrain the data by the specified selection.
- Number of projections: The number of observations to make into the future.

- Replace outliers: To eliminate outliers in the original data.
- Format: To specify the number of decimal places to use when displaying the forecasted data.
- Autocalculate: To calculate the smoothing parameters for the model automatically.
- Level, Trend and Season smoothing: If Autocalculate is not selected, you can set the smoothing parameters for the model manually. Smoothing parameters are values from 0.1 to 0.9. The smaller the smoothing parameter is, the less weight given to the most recent observations in the time series will be, making the series smoother.

5 Execute pre-analysis:

1. To analyze the time series before making a prediction, choose Pre-Analysis.
2. On Pre-Analysis, analyze the results and decide:
   - Whether to replace any outliers in the time series. The number of outliers appears next to Number of outliers. The coordinates of any outliers appears on the graph with a red line.
   - How much confidence you can place in the projections. Evidence for non-randomness is a 5-star assessment. An assessment of three or more stars is considered statistically significant. The less random the results are, the more confidence you can have in the reliability of any projections made. Keep in mind, however, that any randomness may be attributed to seasonal patterns in the time series.
   - Whether to account for seasonality in the forecast and, if so, the seasonal periodicity. Evidence for seasonality is a five-star assessment. An assessment with three or more stars is considered statistically significant. If seasonality is indicated, you select Has seasonality when making a forecast, using the same seasonal periodicity that you used to run the pre-analysis.

6 To run the forecast, choose Calculate.

How to use forecasting results

After calculating the results, you can analyze and save the forecast.

1 On Forecast, analyze the graphical results:

- A yellow line represents the forecast with a blue line representing the original data. Hover the cursor over a chart co-ordinate to display its value. All projections are labeled numerically on the *x*-axis, starting from 1.
- If outliers were replaced, then the original data shown displays the value used to replace the outlier.
- You can also zoom in to any part of the chart to get a more detailed perspective. Select the area of the graph on which to zoom. Zoom out by right-clicking in the chart and selecting Zoom Out Chart. Choose Reset Chart to show all data points.
- To make comparisons between different time periods, overlay different parts of the chart. Choose Switch to pin mode and select an area to float, then drag the floating area over another area of the chart to make the comparison. To return to the original state, choose Switch to Zoom Mode.
- The smoothing parameters used for the model, including those selected automatically, appear at the top of the tab.
- To export the forecast to a PDF document, choose the export icon. The generated document contains the chart exactly as shown in the Forecast tab, along with the data in tabular format.

2 On Table, review the forecasted data and the original data in tabular format. The forecasted values are labeled numerically, starting from 1.

3 To save the forecasting analysis, choose Save. When saving, you have the possibility to share your analysis with others by granting viewing permission to a user or to a group or groups of users. Groups can contain from just one to any number of individuals. When new users are added to a group they automatically inherit the permissions granted to their group.

Understanding decision trees

A decision tree predicts the class of an object based on the object's attributes. For example, you can use a decision tree to predict whether a passenger on the *Titanic* survived, based on the passenger's gender, age, and number of siblings, as shown in Figure 6-11. The numbers under each leaf show the probability of survival for a passenger with the specified attributes and the percentage of passengers represented by the leaf. Figure 6-11 tells us:

- A female passenger had a 0.73 chance of surviving.
- A male passenger over the age of 9 had only a 0.17 chance of surviving.
- For male passengers age 9 or younger, the probability of survival depends on the number of siblings the passenger had. If the passenger had more than two siblings, he had only a 0.05 chance of surviving. If the passenger had two siblings or fewer, he had a 0.89 chance of surviving.

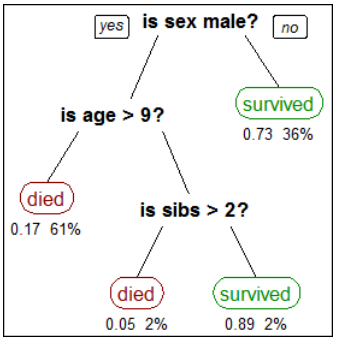


Figure 6-11 Decision tree showing survival probabilities for passengers on the *Titanic*

Figure 6-14 shows a decision tree that predicts whether a worker is low income, medium income, or high income, based on the worker's occupation and gender.

Training and testing a predictive model

A decision tree must be trained on sample data. The decision tree learns with each successive application of the predictive model. Some patterns found by data mining algorithms, however, are invalid. Data mining algorithms often find patterns in the training set that are not present in the general data set. This is called overfitting. To solve this problem, test the predictive model on a set of data different from the training set. The learned patterns are applied to the test set and the resulting output is compared to the desired output. For example, a data mining algorithm that distinguishes spam from legitimate e-mails is trained on a set of sample e-mails. Once trained, the learned patterns are applied to a test set of emails. The accuracy of the predictive model is measured from how many e-mails it classifies correctly.

Understanding the confusion matrix

A confusion matrix tabulates the results of a predictive algorithm. Each row of the matrix represents an actual class. Each column represents a predicted class. For example, consider a

classification system that has been trained to distinguish between cats, dogs, and rabbits. Figure 6-12 shows how frequently the predictive algorithm correctly classifies each type of animal. The sample contains 27 animals: 8 cats, 6 dogs, and 13 rabbits. Of the 8 cats, the algorithm predicted that three are dogs, and of the six dogs, it predicted that two are cats and one is a rabbit. The confusion matrix shows that the algorithm is not very successful distinguishing between cats and dogs. It is, however, successful distinguishing between rabbits and other types of animals, misclassifying only two of 13 rabbits. Correct predictions are tallied in the table's diagonal. Non-zero values outside the diagonal indicate incorrect predictions.

		Predicted class		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

Figure 6-12 Confusion matrix showing the actual class and predicted class for animals

Understanding sensitivity and specificity

Sensitivity, also called the true positive rate, measures the proportion of actual positives that are correctly identified as such, for example the percentage of sick people who are correctly identified as having a condition. Specificity, also called the true negative rate, measures the proportion of negatives that are correctly identified as such, for example the percentage of healthy people who are correctly identified as not having a condition. A perfect predictor is 100% sensitive, in other words predicting that all people in the sick group are sick, and 100% specific, in other words not predicting that anyone in the healthy group is sick. For any test, however, there is a trade-off between the measures. For example, in an airport security setting where one is testing for potential threats to safety, scanners may be set to trigger on low-risk items such as belt buckles and keys, low specificity, in order to reduce the risk of missing objects that pose a threat to passengers and crew, high sensitivity.

How to create a decision tree

1. In Analytics—Advanced, choose Decision tree.
2. Drag the appropriate table from My Data and drop it in Domain in Decision tree, shown in Figure 6-13.
3. In the left pane of Domain columns, expand the database and the appropriate tables.
4. Drag the appropriate columns from the left pane and drop them in the right pane. The columns specify the attributes used to assign classifications in the decision tree.
5. Create the classifications:
    1. Choose New classifier.
    2. In Alias, type the name of the classifier.
    3. Choose the color picker button and select a color. The name and the color visually identify the classification in the decision tree. For example, you can define a High income classification identified by the color green.
    4. Drag a segment from Discrete values and drop it in Drag a segment. The segment specifies the condition for the classification. For example, you can define a High income classification as income over 80000.
    5. Choose the check mark. The classification appears in Classifications.
    6. Create the remaining classifications. Figure 6-14 shows three classifications: High income, Medium income, and Low income.

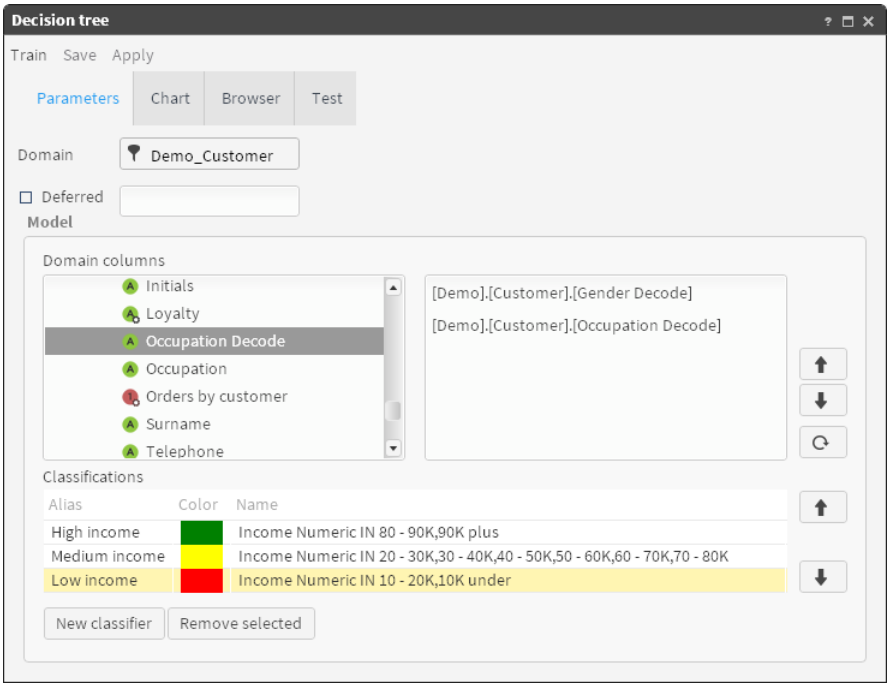


Figure 6-13 Specifying the parameters for a decision tree

6 Choose Train. A graphical representation of the decision tree appears in Chart. The classifications appear with the names and colors you specified.

Figure 6-14 shows part of a decision tree for the Customer table domain with domain columns Occupation Decode and Gender Decode. The decision tree predicts whether a worker is low income, medium income, or high income, based on the worker's occupation and gender. For professionals, the income classification does not depend on gender. For office workers, however, the classification for men is medium income, while the classification for women is low income.

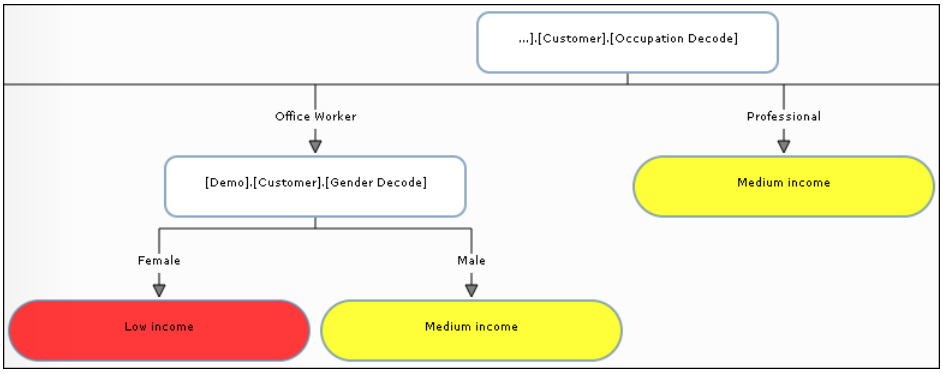


Figure 6-14 Graphical representation of a decision tree

7 Choose Browser. A tabular representation of the decision tree appears.

8 Expand the tree. Figure 6-15 shows a tabular representation of the decision tree, shown graphically in Figure 6-14.

Tree	Color	Class	Slice	Segment error	Total error
Total			100%	16.5155%	16.5155%
[Demo].[Customer].[Occupation Decode]					
Director		Medium income	3.4598%	0.0257%	0.0009%
House Person		Low income	5.2026%	0%	0%
Manager		Medium income	12.1265%	8.9896%	1.0901%
Manual Worker		Low income	15.5659%	20.1502%	3.1366%
Office Worker			10.478%	43.4491%	4.5526%
[Demo].[Customer].[Gender Decode]					
Female		Low income	3.7145%	41.7953%	1.5525%
Male		Medium income	6.7635%	44.3568%	3.0001%
Professional		Medium income	12.0768%	4.1121%	0.4966%
Retired		Low income	8.4796%	32.5067%	2.7564%
Self Employed		Medium income	8.6174%	41.1959%	3.55%
Senior Manager		Medium income	5.1448%	0%	0%
Shop Worker		Low income	17.2922%	5.2706%	0.9114%
Unemployed		Low income	1.5565%	1.3425%	0.0209%

Figure 6-15 Tabular representation of a decision tree

9 Choose Test.

10 Drag the appropriate table from My Data and drop it in Domain.

11 Choose Test. Figure 6-16 shows the test results.

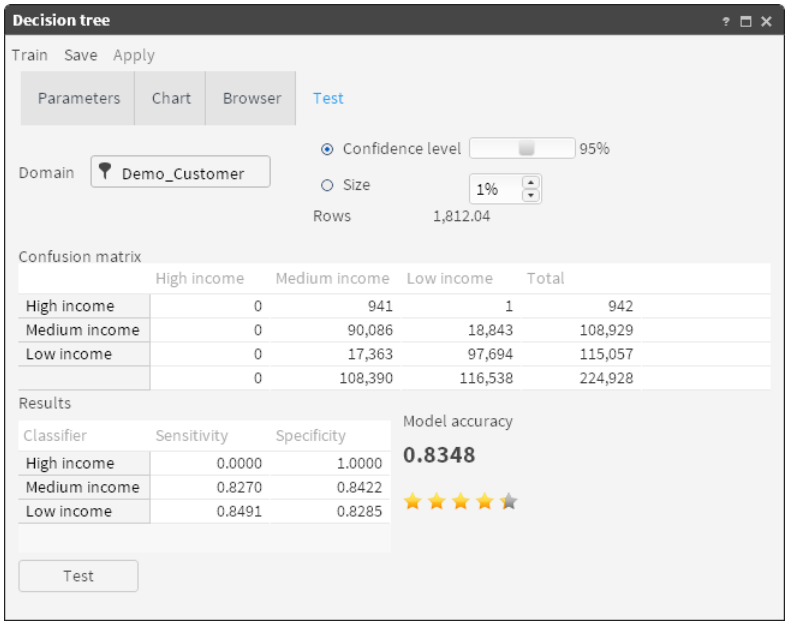


Figure 6-16 Test results, including the confusion matrix

Understanding association rules

Association rules find patterns in very large amounts of point-of-sale data. An association rule is of the form "If a shopper purchases Item A and Item B, the shopper also purchases Item C." For example, association rules might indicate that if a shopper buys onions and potatoes on a trip to the supermarket, she is likely also to buy hamburger meat. Such information can be used as the basis for decisions about marketing activities, such as promotional pricing and product placements.

To identify association rules, you must provide the transaction ID and the transaction items for a very large number of purchases. The Transaction Id and Transaction Item columns must belong to the same database table. You must also provide values for Minimum Support and Minimum Confidence:

- Minimum Support is the minimum percentage of transactions that must contain all the items in the rule. By default, it is set to 10%.

  ```
  Support (A,B) = Transactions (A,B)/Total transactions
  ```

- Minimum Confidence measures how well a rule predicts the right-hand side after the "then" clause of the rule (If A,B then C) by comparing how often the right-hand side (consequent) appears when the condition on the left-hand side (antecedent) is met. By default, this is set to 65%.

  ```
  Confidence (A,B-> C) = Support (A,B,C)/Support (A,B)
  ```

Use the default values for Minimum Support and Minimum Confidence for the first trial. Depending on the results, you can increase or decrease these values.

Optional parameters include:

- Filter: You can define a domain to restrict the number of transactions used in the trial.
- Antecedent Item Range: Minimum and maximum number of items that the left-hand side of the rule must contain.
- Filter for the antecedent items.
- Consequent Item Range: Minimum and maximum number of items that the right-hand side of the rule must contain.
- Filter for the consequent items.

The results of a trial are displayed in a table. The antecedent and consequent items appear in the first two columns. There are also several numeric columns that measure how well the rule predicts the consequent. The columns that appear in the Results tab are:

- Antecedent: Items on the left-hand side of the rule.
- Consequent: Items on the right-hand side of the rule.
- Support (%): Measures the percentage of transactions that contain all the items in the rule.
- Confidence (%): Measures how well the rule predicts the consequent by determining how often the consequent appears when the condition of the antecedent is met.
- Lift: Also called improvement. Measures how well the rule predicts the consequent. A lift greater than one indicates that the items in the rule appear together more than expected.
- Leverage: Measures how well the rule predicts the consequent. The calculation method is different from lift, but the purpose is the same.
- Opportunity: A star indicates that the lift is greater than one.

Items in the antecedent and consequent are enclosed in double quotation marks and separated by a comma. Numeric rule measures such as support, confidence, and lift are rounded to two decimal places.

How to find association rules

1. In Analytics—Advanced, choose Association Rules.
2. Drag the appropriate column from My Data and drop it in Column in Transaction ID.
3. Drag the appropriate column from My Data and drop it in Column in Transaction Item. Accept the default values for Minimum Support and Minimum Confidence, as shown in Figure 6-17.

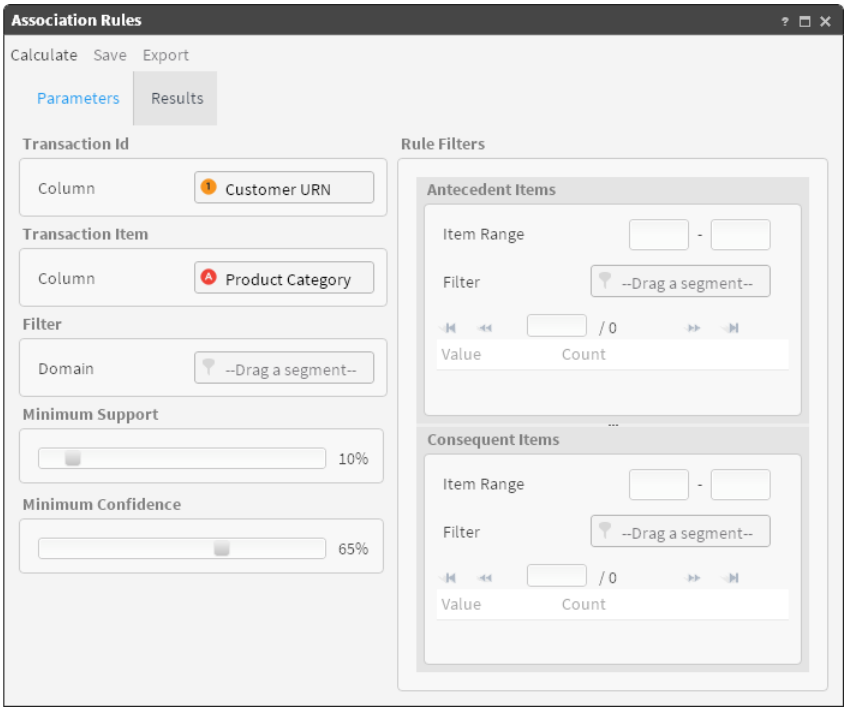


Figure 6-17 Default values for Minimum Support and Minimum Confidence

4 Choose Calculate. The association rules appear in the Results tab, as shown in Figure 6-18.

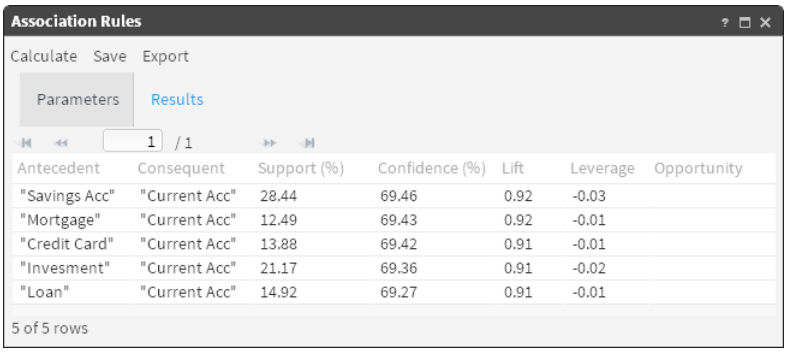


Figure 6-18 Results using default values for Minimum Support and Minimum Confidence

5 If the default values for Minimum Support and Minimum Confidence do not yield many opportunities, choose Parameters and decrease these values. In Figure 6-19, Minimum Support and Minimum Confidence are decreased to 5% and 30%, respectively.

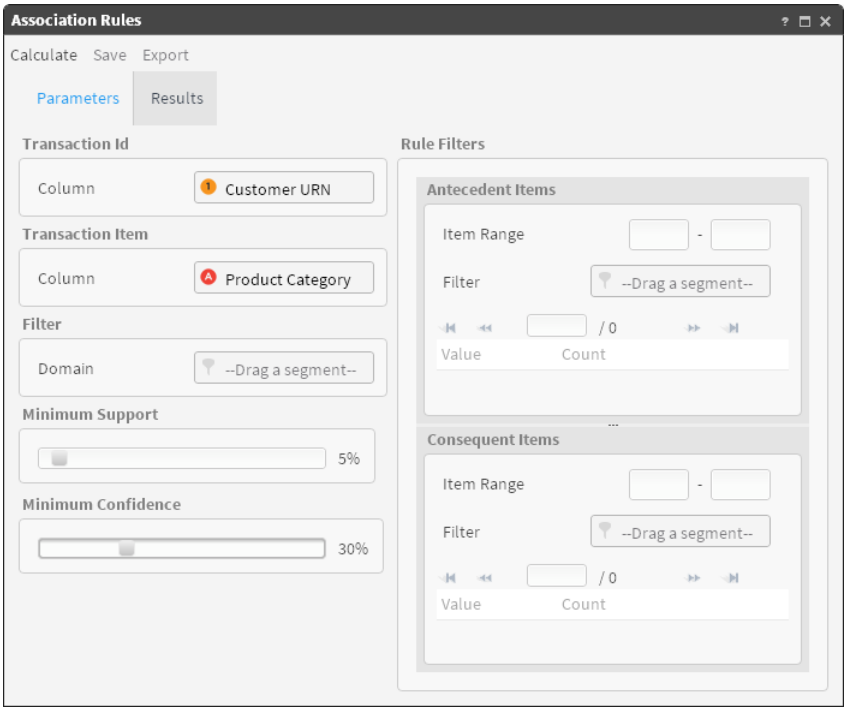


Figure 6-19 Decreasing the values for Minimum Support and Minimum Confidence

6 Choose Calculate. The association rules appear in the Results tab, as shown in Figure 6-20. Decreasing the values for Minimum Support and Minimum Confidence yields several opportunities, indicated by a gold star.

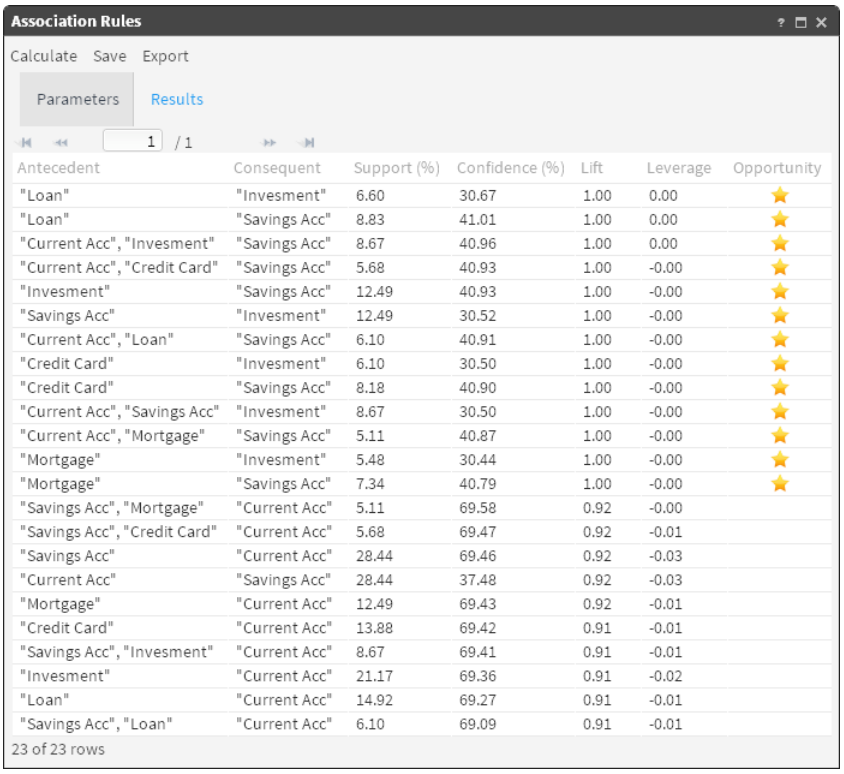


Figure 6-20 Results using decreased values for Minimum Support and Minimum Confidence

Understanding correlation

A correlation measures the dependence relationship between two or more continuous sets of data. Once a domain segment is provided, BIRT Analytics calculates the distinct correlation coefficients (using Pearson's correlation coefficient) between all the possible combinations of pairs of sets of data.You must use continuous variables because this correlation coefficient needs this kind of data to run all its calculations.

Understanding the correlation matrix

A correlation matrix tabulates the results of the correlation coefficient between pairs of variables provided. Each cell shows the correlation coefficient for a certain combination of column and row.Each row and column represents the distinct continuous sets of data compared.

The value of the Pearson's correlation coefficient goes from -1 to 1. Results near to 1 indicate a direct dependency. If one value grows so does the other. Results are close to -1 indicate a decreasing linear dependency, also known as anti-correlation. Values approaching zero indicate that the data sets are "uncorrelated".

How to create a correlation matrix

1. In Analytics—Advanced, choose Correlations.
2. In the Parameters tab, drag the desired database table and drop it in the Domain field.
3. In the Independent variables pane, expand the database and the appropriate tables.
4. Drag and drop the appropriate columns (specifying the continuous variables for your correlation calculations) from the left pane into the right pane and click on Train. (See Figure 6-21 below).

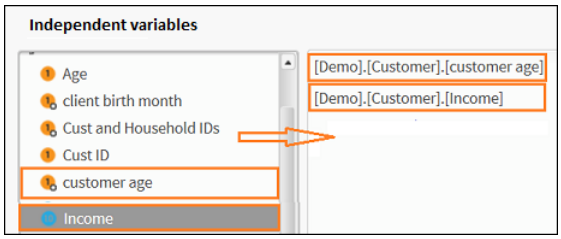


Figure 6-21 Correlations - Parameters tab

A correlation matrix automatically appears in the Results tab, as shown in Figure 6-22.

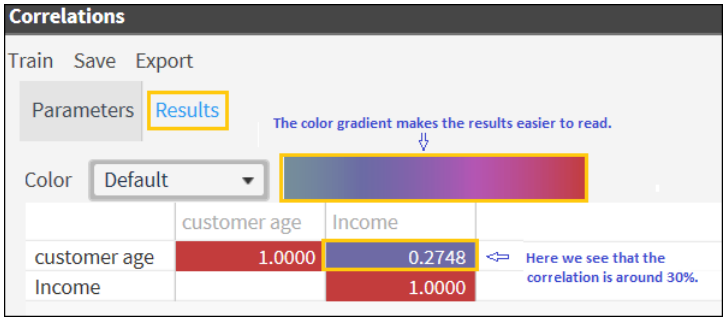


Figure 6-22 Correlations - Results tab

Understanding the difference between correlation and linear regression

Correlation and Linear Regression are often used together but they are not the same.

Correlation quantifies the degree to which two variables (X and Y) are related. It can be computed and interpreted for *any* two variables. It does not fit a line through the data points. It involves computing a correlation coefficient (**r**) that tells you *how much one variable tends to change when the other one changes.*

Linear regression *finds the best line* that predicts Y from X. It is usually used when X is a variable you manipulate (time, concentration, etc.)

With correlation you do not have to think about cause and effect. It does not matter which of the two variables you call "X" and which of the variables you call "Y". You get the same correlation coefficient if you swap the two.

The decision of which variable you call "X" and which you call "Y" matters in regression. You will get a different best-fit line if you swap the two. In other words, the line that best predicts Y from X is not the same as the line that predicts X from Y (even though both lines will have the same value for $\mathbf{r^2}$).

In correlation both X and Y are measured. With linear regression, the X values can be measured or can be a variable controlled by the user.

Relationship between results

Correlation:

- computes the value of the Pearson correlation coefficient **- r**
- has value ranges from **-1** to **+1**

Linear regression:

- quantifies the goodness of fit with $\mathbf{r^2}$

Understanding linear regression

Linear regression attempts to model the relationship between variables by fitting a linear equation to observed data. One or more variables are considered to be independent predictors or explanatory variables while the other variable is considered to be a dependent variable.

Linear regression is mainly used for:

- Forecasting a dependent variable
- Quantifying the level of relationship between a dependent variable and independent predictor variables

Before attempting to fit a linear model to observed data, you first need to determine whether or not there is a relationship or significant association between the chosen variables.

Displaying the data on a scatter plot helps to determine the strength of relationship between two variables. If the scatter plot does not show any increasing or decreasing trends then you should not bother trying to fit a linear regression model to the data.

Least-Squares Regression

The BIRT Analytics Linear Regression tool uses the most effective and well-known method for fitting a regression line. It is called the OLS (Ordinary Least-Squares) algorithm. It calculates the best-fitting line for the observed data by minimizing the sum of the squares of

the vertical deviations from each data point to the line (if a point lies exactly on the fitted line its vertical deviation is zero). Because the deviations are first squared and then summed, there are no cancellations between positive and negative values. This minimizes error in the dependent variable.

How to make a linear regression

The following procedure show you how to create a linear regression using the BIRT Analytics tool:

1. Click on the *Analytics* button on the home screen.
2. Open the *Advanced tab* and click on *Linear regression*.
3. Expand the database and tables in the Data Tree pane on the left.
4. Drag the table that you want to analyze from the *Data Tree* on the left and drop it into the *Filter field*. This table represents the Domain of the variables to be studied. In this case drag and drop the *Customer table* into the *Filter field*.
5. Again from the *Data Tree*, drag the column that will be used as the dependent variable (on which predictions will be made) and drop it into the *Dependent variable* field.
   In this case drag and drop the *Total Orders* variable. This action automatically populates the *Table* field on the right.
6. Expand the tables that appear in the left pane of the workspace. Then drag from this pane the columns that you want to use as independent predictor variables and drop them into the empty pane on the right. In this case drag and drop the *Total products* variable.
   You can compare your results to Figure 6-23 below.

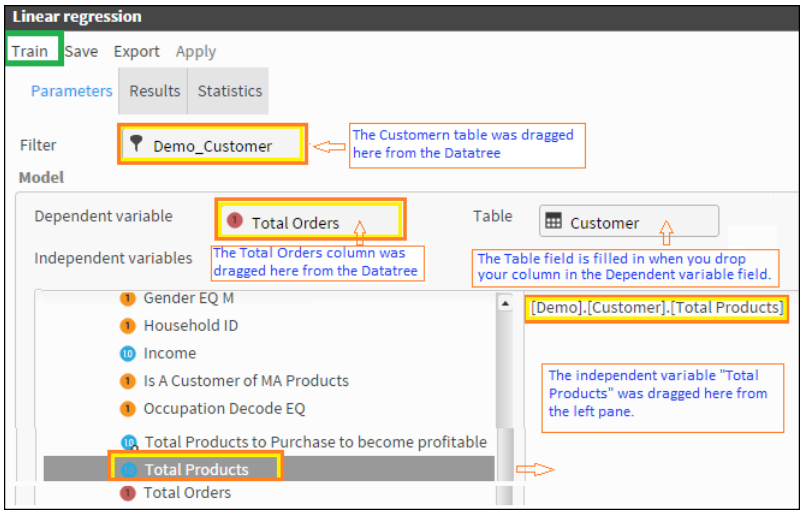


Figure 6-23 Creating a linear regression

7. Choose *Train* to make the calculation.
   This opens the Results tab that gives you the equation that you can apply to any of the Predictor values to predict the dependent variable. There is a graphic display of the function (including the plotted values) which makes it easy to see at a glance that, in this case, it is a very close fit.A set of one or more (or none) gold stars indicate the goodness of fit of your equation. (See Figure 6-24 below)

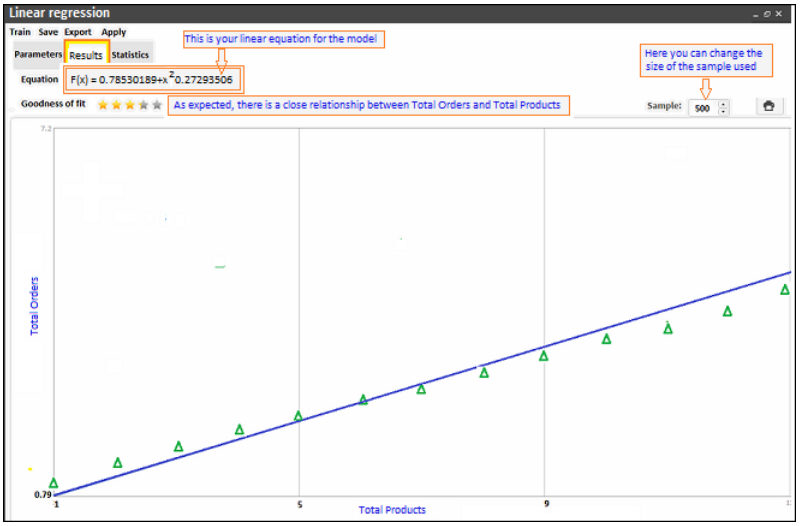


Figure 6-24 Linear regression - Results tab

Understanding advanced statistical values in the Statistics tab

To view more advanced statistical results you can open the Statistics tab for your model. (See Figure 6-25 below).

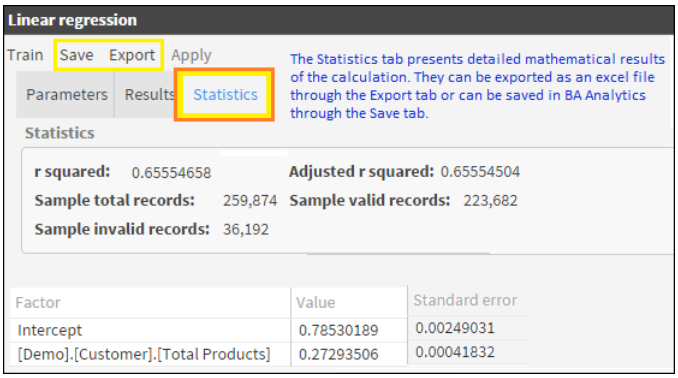


Figure 6-25 Linear regression - Statistics tab

These results are the coefficients that accompany each independent predictor and the intercept. The *Interceptor* is the value of the dependent variable when all the predictor values are zero.

Each coefficient has additional associated parameters such as:

- *Standard error*
- *t-stats* - associated with the *Student-T distribution test* - higher values imply that the coefficient is not zero.
- *P-value* - this value shows the results of the hypothesis test as a significance level. Values of less than 0.5 imply that the coefficient is not zero
- *Upper and lower confidence levels* - Assuming that the error in the prediction of the dependent variable is normally distributed, BIRT Analytics is able to calculate the confidence interval of the linear regression. The results are two linear functions with coefficients and intercepts for both the 95% upper confidence level and the 95% lower confidence level, also known as the 95% confidence bands.

For the global results of the linear regression, the statistics that measure the goodness of fit are:

- $\mathbf{R^2}$ - *(R-Squared)* - Also known as a coefficient or determination, it indicates how well data provided fits with the linear regression model that has been calculated. Values close to zero indicate no linear relationship.
- *Adjusted R-Squared*- Sometimes R-Squared suffers an increase in its value due to the addition of extra predictors without improving the fit. Adjusted R-Squared will always be less or equal to $\mathbf{R^2}$
- The Linear regression tool also shows the total records used from the Domain. Invalid records come from those having null values.

Understanding logistic regression

In the previous section we described the linear regression model used for predicting relationships between data variables. Although logistic regression also measures relationships based on observed data, it is a much more complex model than linear regression. Logistic regression tries to produce a realistic *binary* result concerning the likelihood of something occurring (ie. the odds of success) - predicting probability from observed data variables.

Our model predicts a binary response from a binary predictor (explanatory/independent variable) that is used to predict the outcome of a dependent variable based on one or more predictor variables. In other words, it estimates the parameters of a qualitative response model.

In our model the dependent variable is binary which means the number of available categories is two. It measures the relationship between a categorical dependent variable and one or more independent variables, which are usually continuous. Logistic regression uses probability scores as the predicted values of the dependent variable.

Logistic regression is used in many fields. In the medical field it is often used to predict whether a patient has a particular disease, such as diabetes or coronary heart disease, based on observed patient characteristics such as: age, gender, body mass index, relative weight, blood cholesterol levels, etc.

In marketing it is used to predict customer propensity to purchase a particular product or to cease a subscription, etc. In economics it is used to predict the rise or fall of unemployment over a coming period. In business and banking it was recently used to predict the likelihood of homeowners defaulting on their sub-prime mortgages (with results that encouraged disastrous behavior in this case). Logistic regression results only predict the odds of success but they cannot guarantee it.

Basic principles

In binary logistic regression the outcome is coded as “0” or “1” which leads to the most straightforward interpretation. If a particular observed outcome for the dependent variable is a relevant possible outcome (often called a “success” or a “case”), it is coded as “1”. The contrary outcome (called a “failure” or “noncase”) is coded as “0”.

Logistic regression predicts the odds of “success” (or the “case”) based on the values of the independent variables (predictors). Mathematically speaking, these odds are defined as *the probability that an outcome is a success (or case)* divided by *the probability that the outcome is a failure (or noncase)*

Logistic regression takes the natural logarithm (or “logit”) of the odds of the dependent variable being a success to create a continuous criterion as a transformed version of the dependent variable. This “logit” transformation is referred to as the “link” function in logistic regression. Although the dependent variable in logistic regression is binomial, the “logit” is the continuous criterion upon which linear regression is conducted.

Once transformed, the 'logit' of success is then fit to the predictors using linear regression analysis. The predicted value of the logit is converted back into predicted odds via the inverse of the natural logarithm, the exponential function.

The observed dependent variable in logistic regression is a zero-or-one variable. However, the logistic regression estimates the odds, *as a continuous variable,* that the dependent variable is a success.

A categorical prediction can be based on the computed odds of a success, with predicted odds above some chosen cut-off value being translated into a prediction of success.

How to make a logistic regression

The following procedure show you how to create a logistic regression using the BIRT Analytics tool. [In this case we will determine the probability of a customer being female, using as a predictor the total number of purchases made by the customer.]

1. Click on the *Analytics* button on the home screen.
2. Open the *Advanced tab* and click on *Logistic regression*.
3. Prepare the segment you want to analyze in the *Scratchpad* by dragging in the desired discrete values from your chosen column in the *Data Tree*. In this case use the discrete values from the gender column for your segment.
4. Drag the newly created segment "GENDER IM F,M" into the *Filter field* from the *Data Tree*. Now two panels appear in the workspace under "*Independent variables*". (See Figure 6-26 below.)

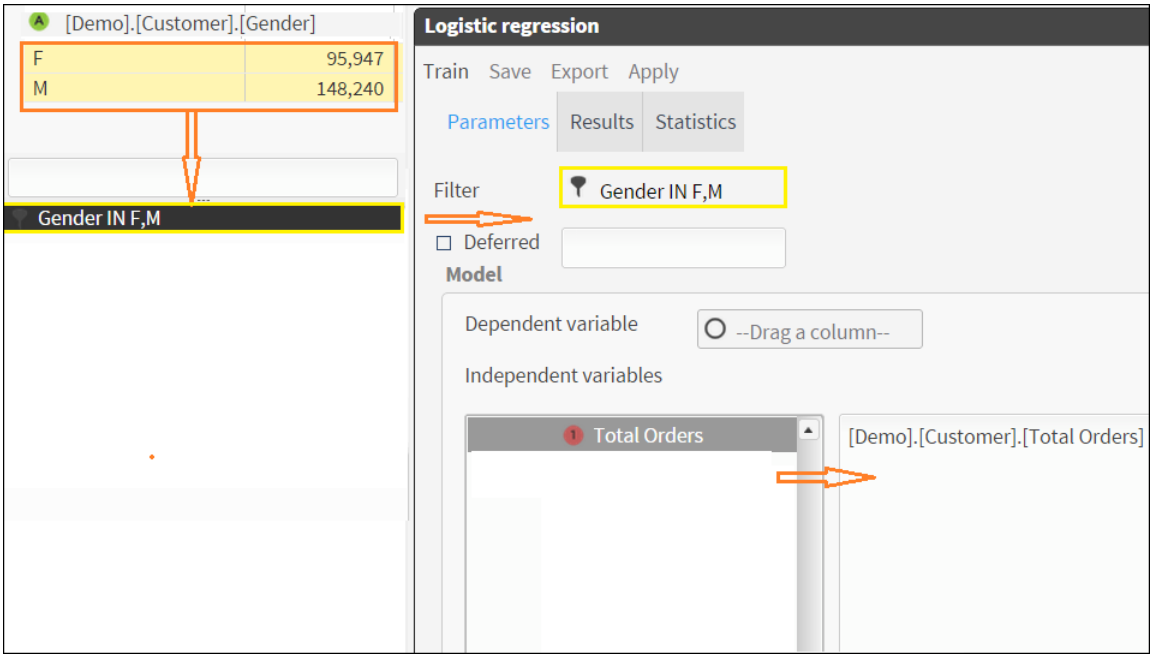


Figure 6-26 Logistic regression - preparing a segment for the Filter field

5. Again from the *Data Tree*, drag the column *Gender decode EQ Female* and drop it into the *Dependent variable* field. (Predictions will be made on this variable). This action automatically enters the "*Customer*" table in the *Table* field on the right. This table represents the *Domain* of the variables to be studied. (See Figure 6-27 below.)

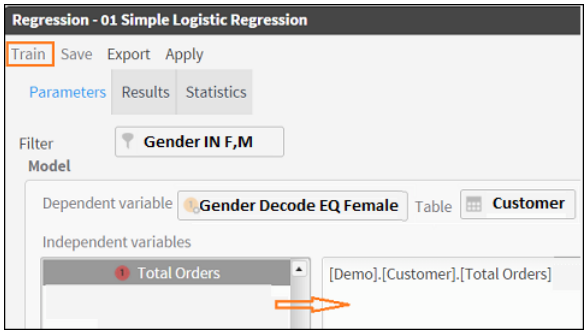


Figure 6-27 Logistic regression - Setting the dependent and predictor variables

6 Expand the tables that appear in the left pane of the workspace. Then drag from this pane the *Total Orders* column that will be used as an independent predictor variable and drop it into the empty pane on the right. (See Figure 6-27 above.)

7 Choose *Train* to make the calculation.
This opens the Results tab that gives you the equation that you can apply to any of the Predictor values to predict the dependent variable. The results of the calculation are the coefficients that accompany each predictor and the intercept.

A set of five gold stars indicates that the goodness of fit of your equation is very high. There is also a graphic display of the function, as shown in Figure 6-28 and Figure 6-29.

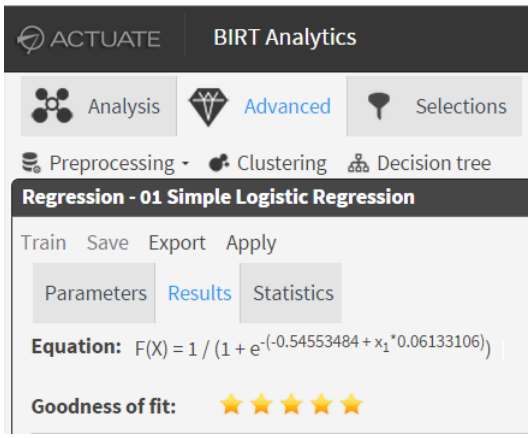


Figure 6-28 Logistic regression - results tab

Understanding advanced statistical values in the Statistics tab

For more advanced results, the *Statistics* tab lets you analyze each calculated coefficient of the equation, their goodness of fit and relevance of the model. The model can be saved and applied to predict values of the dependent variable using the calculated equation.

Each coefficient has additional associated parameters such as:

- *Standard error*
- Odds ratio. This ratio quantifies how strongly the presence or absence of certain property is associated with the presence or absence of another property in a given domain. As bigger is the ratio, better is the relationship between dependent variable and the independent related to the coefficient.
- Odds Upper and Lower Confidence Level (95%). It has the same calculation as confidence level for a domain mean, but it's calculated on the natural log scale. It gives two functions to define the confidence interval or band.

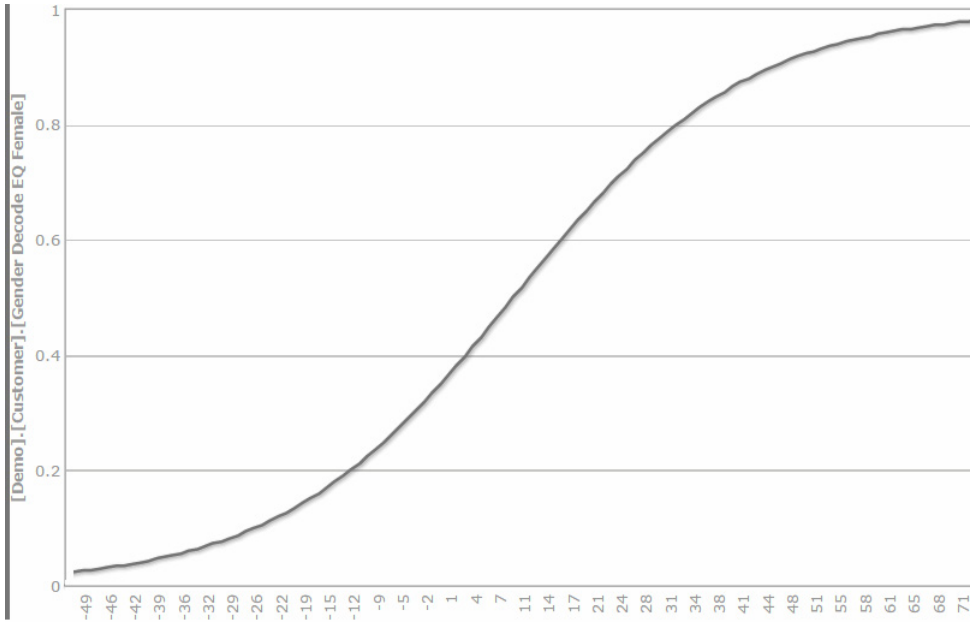


Figure 6-29 Logistic regression - graphic display of the function

- Log likelihood p value. The p-value shows the results of the hypothesis test as a significance level. In that case smaller values than 0.5 are taken as evidence that the coefficient is nonzero.
- Significance Level. Based on a distinct range of significance values for p-value, it is possible to classify the level of significance of this coefficient. It is a range from 0 to 5, where 0 means no significance and 5 means a highly relevant significance level.

For the global results of logistic regression, the statistics that measure goodness of fit are:

- *Chi Squared test*. Also known as the likelihood ratio test, it's an asymptotically distributed Chi Squared test with certain degrees of freedom. As bigger is its value, better is the goodness of fit of the model.
- *Chi Squared p*-value. Is the statistical significance testing from the Chi Squared test. The p-value is the probability of obtaining the observed sample results (or a more extreme result) when the null hypothesis is actually true. So, when p-value is very small (less than a certain threshold), it tells that the modeled data is inconsistent with the assumption that the null hypotheses is true. In other words, this hypothesis could be rejected, so the modeled data could be accepted as true.
- *Log likelihood.* It's the logarithm of the likelihood ratio. This will always be negative, with higher values (closer to zero) indicating a better fitting model.

The tool also shows the total records used from the total selected in the Domain. Invalid records come from those that have null values in some variables.

Understanding Naive Bayes classification

Naive Bayes classification is a type of simple probability classification based on the Bayes algorithm, with a strong hypothesis of independence (called naïve"). It implements a naïve Bayesian classifier, belonging to the family of *linear classifiers.*

In probability theory and statistics, the *Bayes algorithm* relates current probability to prior probability. It is important in the mathematical manipulation of conditional probabilities. Bayes' rule can be derived from more basic axioms of probability, specifically conditional probability.

This algorithm was first used as a method for *text categorization*, in order to identify documents as belonging to one category or another (such as spam or legitimate, sports or politics, etc.)with word frequencies as the features. It is still competitive in this domain. It also finds useful application in the field of *automatic medical diagnosis*.

Naïve Bayes classifiers are highly scalable, requiring a number of parameters that are *linear* in the *number of variables* (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes *near time*, rather than by time consuming *iterative approximation* as used for many other types of classifiers.

Basic principles

The underlying probability model of the Naïve Bayes classification algorithm is best described as a model with *statistically independent* characteristics. A naïve Bayesian classifier supposes that the existence of a characteristic for a given class is independent from the existence of the other characteristics for that class.

For example, a fruit can be considered as an apple if it is red and round. Even if these characteristics can often be connected in reality, a naïve Bayesian classifier will determine that the fruit is an apple by considering i*ndependently* these characteristics of (A) *color* and (B) *shape*.

Bayes' theorem is stated mathematically as the following equation:

$$P(A\ B) = \frac{P(B|A)P(A)}{P(B)}$$

where ***A*** and ***B*** are events.

- ***P*(*A*)** and ***P*(*B*)** are the probabilities of ***A*** and ***B*** independent of each other.
- ***P*(*A* | *B*)**, a *conditional probability*, is the probability of ***A*** given that ***B*** is true.
- ***P*(*B* | *A*)**, is the probability of ***B*** given that ***A*** is true.

In many practical applications, the estimation of the parameters for the naïve Bayesian models is based on *maximum likelihood*. In spite of their "naïve" design model and its extremely simplistic basic hypotheses, the naïve Bayesian classifiers demonstrated an acceptable level of efficiency in many complex real-world situations.

Advantages

The main advantage of the naïve Bayesian classifier is that it requires *relatively little* training data. The necessary parameters for the classification are the *averages* and the *variances* of the various variables. Indeed, the hypothesis of independence of variables does not require knowing more than the variance of each variable for every class, without having to calculate a *covariance* matrix.

The naive Bayes classifier has several properties that make it surprisingly useful. In particular, the decoupling of the class *conditional feature distributions* means that each distribution can be independently estimated as a *one-dimensional* distribution. This helps alleviate problems stemming from the difficulties of dimensionality, such as the need for data sets that scale exponentially with the number of features.

Relation to Linear and Logistic Regression

In statistics, *Bayesian linear regression* is an approach to *linear regression* in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

In the case of discrete inputs (indicator or frequency features for discrete events), naïve Bayes classifiers form a *generative-discriminative* pair with (multinomial) *logistic regression classifiers:* each naïve Bayes classifier can be considered to be a way of *fitting* a probability model that *optimizes* the *joint likelihood* $\mathbf{p(C,x)}$, while logistic regression *fits* the same probability model to *optimize* the *conditional* $\mathbf{p(C|x)}$.

The link between the two can be seen by observing that the decision function for naïve Bayes (in the binary case) can be rewritten as "predict class $\mathbf{C_1}$ if the odds of $\mathbf{p(C_1|x)}$ exceed those of $\mathbf{p(C_2|x)}$.

Expressing this in log-space gives:

$$\log \frac{p(C_1|\mathbf{x})}{p(C_2|\mathbf{x})} = \log p(C_1|\mathbf{x}) - \log p(C_2|\mathbf{x}) > 0$$

The left-hand side of this equation is the log-odds, or *logit*, the quantity predicted by the linear model that underlies *logistic regression*. Since naïve Bayes is also a linear model for the two "discrete" event models, its parameters can be re-set as a *linear function:*

$$b + \mathbf{w}^\top x > 0$$

Obtaining the probabilities is then a matter of applying the logistic function to b + wTx, or in the multiclass case, the softmax function.

Although discriminative classifiers have lower asymptotic error than generative ones, in many practical cases naive Bayes can *outperform logistic regression* because it reaches its asymptotic error faster.

How to make a Naive Bayes classification

1. Click on the *Analytics* button on the home screen.
2. Open the *Advanced tab* and click on *NaIve Bayes*. This opens the screen where you set your parameters. (See Figure 6-30).
3. In the "Parameters" tab, set the Training domain by dragging the appropriate table from the Data Tree and dropping it into the "Filters" field. Here we chose: *[Demo].[Customer].[Customer Profit Decile] EQ 1 OR 10.*
4. In the left-hand panel of the "Domain columns" field, expand the Database tree and its appropriate tables. Then drag the desired Domain columns from this expanded tree into the right-hand panel of the field. These columns specify the attributes used to assign classifications in Naive Bayes. Here we chose: Age Numeric, Gender Decode, Income Numeric and Occupation Decode.
5. We are now ready to create our first classifier by clicking on the "New classifier" button. This opens the configuration fields in the lower right-hand side of the screen. Now drag the segment "*Customer Profit Decile EQ10*" into the "*drag a segment*" field.
6. Type the name of your classifier in the "*Alias*" field and choose your color via the color palette icon. Here we named our classifier "*High Profit*" and choose a rusty brown for its color via the icon beside the color block. (See Figure 6-31 below). Now click "*OK*".

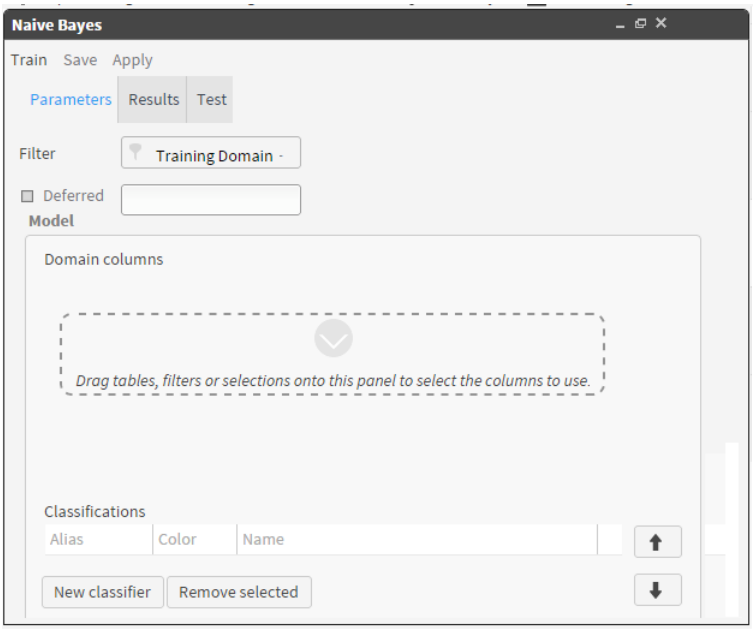


Figure 6-30 Configuration screen in the Parameters tab

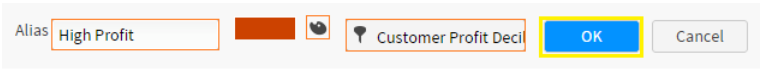


Figure 6-31 "New classifier" configuration fields

7 Repeat the operation to set the second classifier "*Customer Profit Decile EQ1*" naming it "*Low Profit*", choosing a teal green color and click "*OK*" again. Now you are back in the main configuration screen that now displays your parameters. (See Figure 6-32 below).

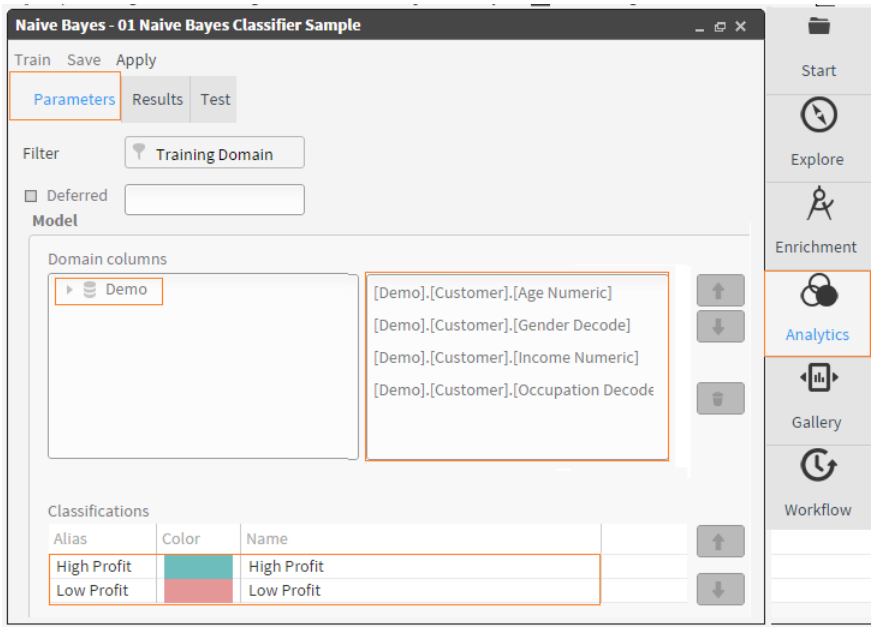


Figure 6-32 Configured parameters for Naive Bayes analysis

8 Click "Train" in the toolbar to see your results (See Figure 6-33 below).

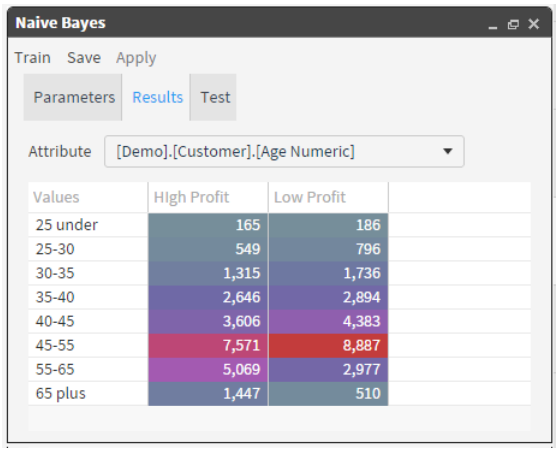


Figure 6-33 Naive Bayes analysis results

9 Now test the accuracy of your results in the "*Test*" tab by dragging your classifier segments into the "Domain" field and clicking "Test" at the bottom of the screen. (See Figure 6-34 below).

Note: *To obtain the most useful results, it would be wise to find the best classifier by comparing the same Domain, Classifiers and Domain Columns in both a Decision Tree and a Naive Bayes analysis.*

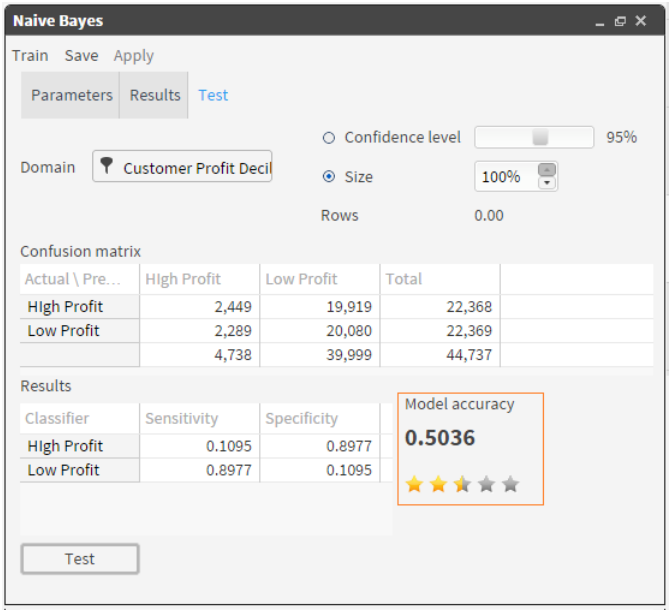


Figure 6-34 Test results including the confusion matrix

Useful Guidelines when building Naive Bayes classifications

- When training in Naive Bayes, the classifier domains must be disjoint, otherwise an error message is displayed. (See Figure 6-35). This means that each register in the Filter domain has to be linked with only one classifier.

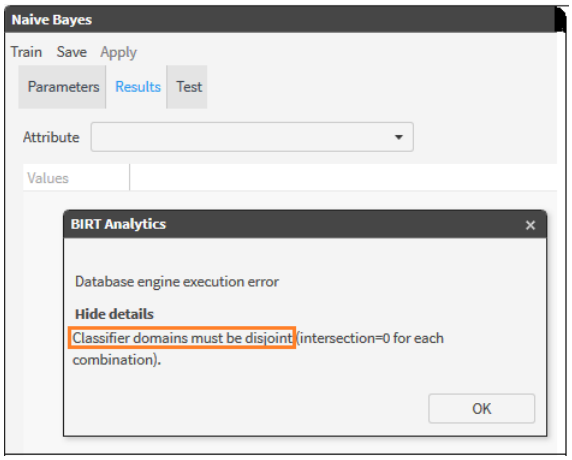


Figure 6-35 Alert received if chosen classifiers intersect

- Although no official size limitation exists concerning the number of discrete values that a Domain column can contain, the User may sometimes receive the following Alert when using a particular high volume Domain column in an analysis: "**Not enough memory**."

  The alert also indicates the number of bytes requested compared to the number of bytes available (for example: 10000000 bytes requested, but only 3797856 are available).

  If this message appears, it is necessary to use a Domain column with fewer discrete values, such as replacing an "Income" column with an "Income numeric" or an "Age" column with an "Age numeric" column, etc.

Chapter

7

Managing campaigns

This chapter contains:

- Understanding campaigns
- Configuring campaign elements
- Planning a campaign
- Running a campaign

Understanding campaigns

A campaign is a set of tasks, defined for specific population segment and completed during a defined time period to accomplish a specific goal. For example, a typical business campaign defines a set of communication tasks that channel information to a segment of customers or prospects. The most typical campaign generates advertising messages to customers in a selected market segment. Common goals of an advertising campaign are web site visits and online purchase transactions made by customers.

BIRT Analytics supports automating campaign tasks associated directly with events or conditions that occur in your database. For example, a forecast analysis predicts a seasonal percentage increase in purchases by customers. Also, association rules modeling shows what additional items a purchasing customer typically buys. By defining a campaign strategy that includes seasonal timing and targeted messaging, your company can effectively persuade a customer to buy an additional item or upgrade to one having a higher profit margin.

Further refining this idea, you can design specific messaging delivered to select market segments, based on data that you collect from that specific segment. For example, analyzing the profile of a customer who responds by purchasing one item enables your website to offer suggestions about similar products to other customers having a similar profile.

Through various supported media, you may also reach members of the following markets:

- Electoral voters
- Financial services clients
- Grocery store customers
- Hospital patients

This chapter explains how BIRT Analytics supports planning, creating, managing, and automating activities that comprise sophisticated business campaigns.

Configuring campaign elements

A campaign includes the following elements:

- Workflows that define campaign roles
- Permissions required to perform campaign tasks
- Stages that group tasks in a workflow
- Cells that define a campaign activity for a specific segment
- Properties that define specific activities
- Segments of data on which cells operate
- Media appropriate to communicate with the segment
- Resolution tables for history and response records
- Strategies that identify campaign goals

To configure a campaign using BIRT Analytics, use Campaign Workflow to define workflows, stages, permissions, and a resolution level for a campaign.

Creating a campaign workflow

A workflow defines a role required to complete campaign tasks and activities. A workflow includes one or multiple stages and users. In a workflow stage, each user is assigned permissions required to complete tasks. For example, the role Campaign Manager may

include two users, one with permission to initiate an activity such as generating messages to a group of customers. In the same workflow, another user may have permission to receive responses and initiate new messages.

How to create a workflow

1 Choose Workflow—Campaign workflow.

2 In Campaign Workflow, select Configuration, and choose Workflow.

3 In Workflow, choose Create, as shown in Figure 7-1.

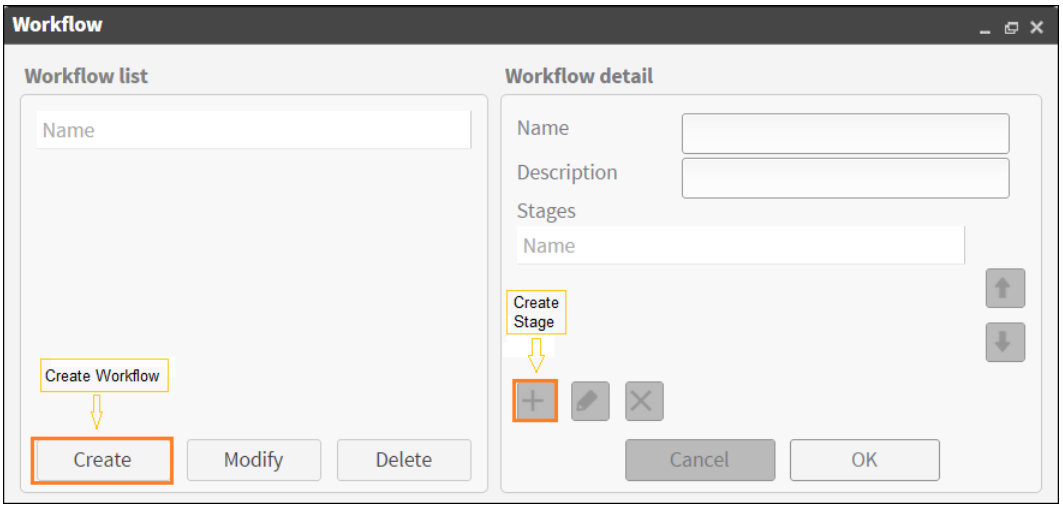


Figure 7-1 Creating a new workflow for a campaign

4 In Workflow detail—Name, type a name of a role that completes campaign tasks, such as Campaign Manager.

5 In Workflow detail—Description, describe the purpose or goal of this workflow. For example, name and describe the goal of Campaign Manager.

6 Create at least one stage in the workflow. You must create all stages in a workflow before executing a campaign.

7 In Workflow, choose OK.

Creating a stage

A workflow includes at least one stage. A stage supports defining users as task owners and assigning to each task owner the permissions required to perform tasks. Using multiple stages supports grouping sets of tasks and activities in a workflow. You must define all stages in a campaign workflow before executing the campaign. Executing a campaign locks the campaign workflow, preventing modifications.

Use stages to organize a workflow. For example, a typical workflow has two stages:

- Stage1, Define/Validate
- Stage2, Load/Response

Different users may have defined tasks in each stage. Organizing a workflow using a logical sequence of stages supports one role, such as Campaign Manager, which is responsible for completing tasks in multiple stages of a campaign.

For example, in a typical company, one group owns data collection tasks, another group owns data validation tasks, and another group owns data modeling and analysis tasks. A workflow uses stages to sequentially group tasks in a campaign. Stages support timely completion of tasks triggered by time and tasks triggered by events.

How to create a stage in a workflow

1 In Workflow detail, click on the “create” icon.

2 In New Stage—Name, type a name that identifies the stage in the workflow.

3 You have the possibility of sharing tasks with individual users or with user groups. In the “Users” tab, select either “Groups” or “Users” to display the list of available Groups or Users in the “New Stage” tab. Here we chose “Groups”. In New Stage, choose OK. (See Figure 7-2 below).

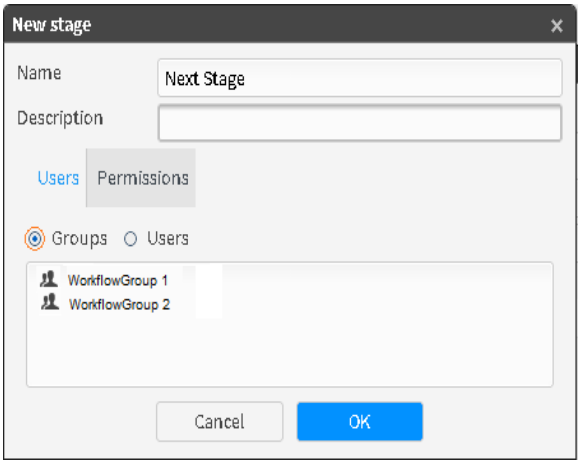


Figure 7-2 Creating a new stage

Note: *Before being able to share your work with a User or a User Group, you must first create the Group and assign the Users, via the BA Admin tool. Whenever a new user is added to a User Group, he or she automatically inherits all existing permissions that have been given to that group.*

About assigning permissions

A workflow requires at least one stage. For practical purposes, you must define users and permissions sufficient to complete tasks in each stage. To demonstrate assigning permissions and users in different workflow stages, consider the following procedure, for example.

How to assign permissions on stages in a workflow

To create three stages in a workflow called Campaign Manager:

1 Create stage1, called generate messages.

2 In stage1, assign users Admin and User1 all permissions except Execution, Response, and Validation actions. To assign all permissions for a stage to one user:

   1 In New Stage, in Users, select a user, then choose Permissions.

   2 In Permissions, select any permission name, then choose Select.

   3 To deselect a permission, in Selected for a permission name, double-click the green check mark.

   4 To assign a user all permissions except three, select all permissions, then double-click each of those three, as shown in Figure 7-3.

3 Create stage2, called Validate.

4 In stage2, assign a user only the Validation action permission.

5 Create stage3, called Execution and Response.

6 In stage3, assign Execution action and Response action permissions to Admin and User2.

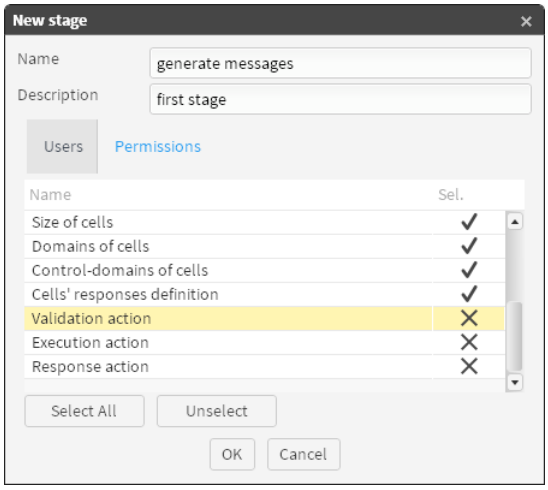


Figure 7-3 Adding user permissions to a new stage

7 In Workflow, choose OK.

Defining a campaign resolution level

The resolution level for a campaign identifies a database column in which records used to generate campaign activities are stored. For example, a campaign that contacts all customers typically uses customer ID records stored in the Customers table to generate activities. Defining a resolution level for a campaign links history and response records generated by the campaign to the records used to generate campaign activities.

How to define a resolution level for a campaign

1 In Campaign Workflow, select Configuration and choose Resolution level.

2 In List of levels, choose Create.

3 In Level detail, type a name that describes a column in which to store the history and response data for the campaign.

4 From My Data, drag a column on which to resolve the campaign and drop it in Record Identifier, as shown in Figure 7-4. The column selected as Record Identifier must be discrete. In other words, each value must appear in only one column record.

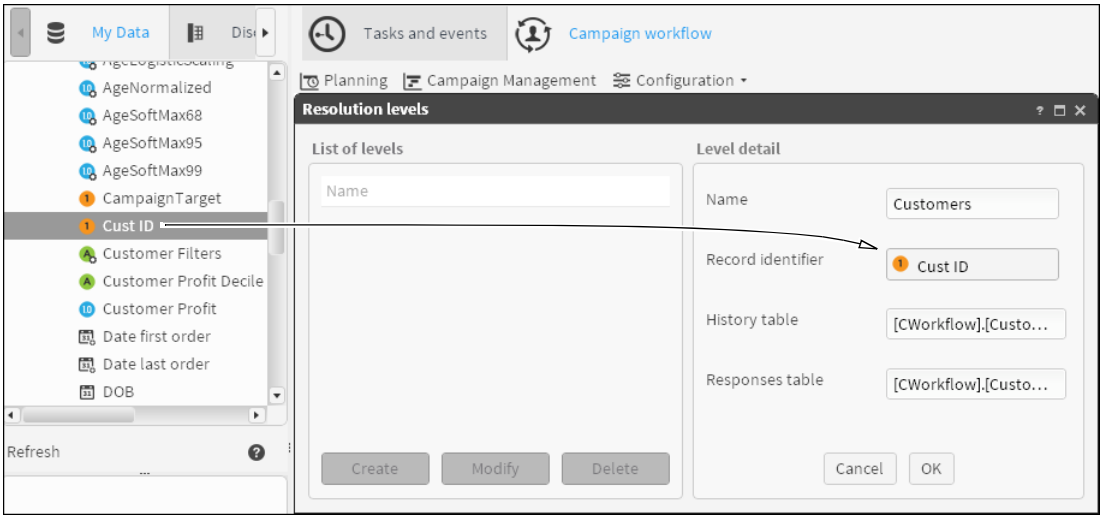


Figure 7-4 Creating a resolution level for a campaign

5 Choose OK. Column names for storing campaign history and response records appear in a new table named CWorkflow, as shown in Figure 7-5.

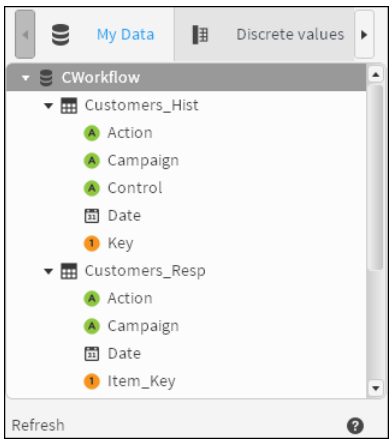


Figure 7-5 Creating a resolution table for a campaign

6 In Resolution levels, choose OK.

Defining a media condition

A media condition associates a media type and a selection or segment of records. Defining specific media conditions supports assigning a specific media condition to a campaign cell. For example, associating the email media type and a selection of valid email address records enables a cell to send email messages to valid recipients.

How to define a media condition

1 In Campaign Workflow, select Configuration➛Media.

2 In Media list, choose Create.

3 Define the following media details, as shown in Figure 7-6:

- In Name, type characters that establish a media name for selection in a campaign cell.
- In Description, describe this media type and purpose.
- Select a listed icon to associate with this media name.

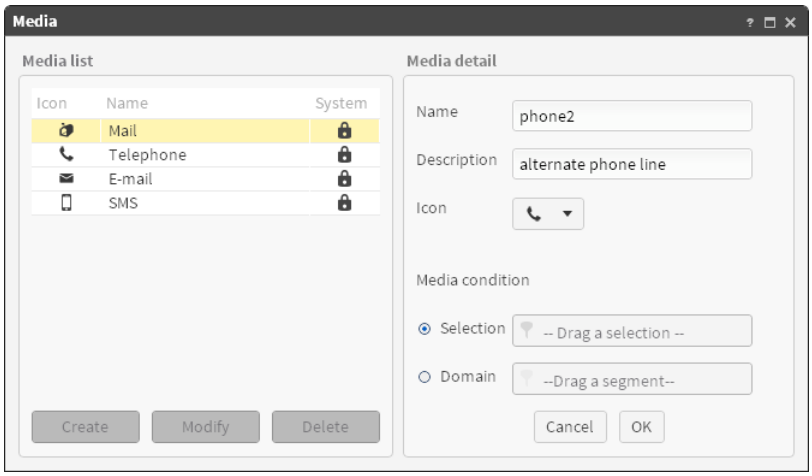


Figure 7-6 Defining media details

4 In Media condition, define a media condition using one of the following options:

- Choose Selection. Drag a selection from Data Tree and drop it in Selection.
- Choose Domain. Drag a segment from Data Tree and drop it in Domain.

For example, a media condition can limit a specific phone to call only the first one hundred phone records. Create a segment containing records 1 through 100 in [Customers].[Telephone]. Drag the First 100 Phone Numbers segment and drop it in Domain, as shown in Figure 7-7.

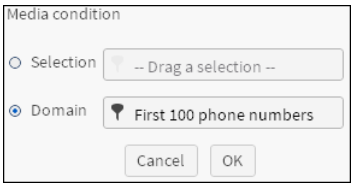


Figure 7-7 Defining a media condition

Choose OK.

Defining an action goal

You can associate an action with each campaign cell that describes a goal for the cell. In Campaign Workflow, action is a descriptive term only. Campaign Workflow includes the following default actions:

- Marketing
- Informative
- Collect

For example, assign the Informative action to a cell that has a goal of sending advertising messages. Alternatively, define a new action goal, such as Prospect.

How to define an action goal

1. In Campaign Workflow, select Configuration➛Actions.
2. In Actions list, choose Create.
3. Define the following action details, as shown in Figure 7-8.
   - In Name, specify an action name for selection in a campaign cell.
   - In Description, specify a goal for this action.

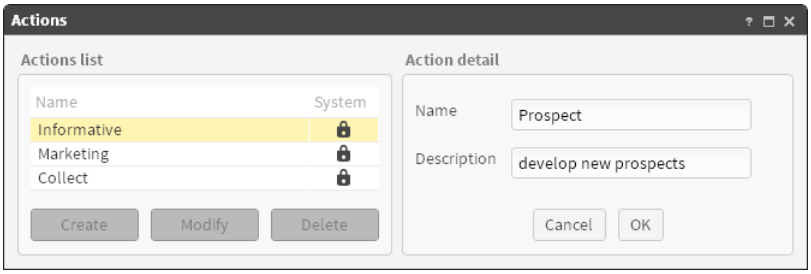


Figure 7-8 Defining action details

4. Choose OK.

Planning a campaign

Planning a campaign involves organizing a set of goal-oriented activities. Creating campaign structure by defining a strategy and cells establishes a manageable path toward completing

the campaign. Defining campaign properties completes the campaign structure. Campaign workflow planning supports completing multiple activities toward a campaign goal.

This section describes how to define campaign strategies, properties, and cells.

About campaign properties

You define a campaign by setting the following properties:

- General Data
  A code, name, and description that identify a campaign
- Dates
  Calendar dates on which activities in a campaign begin and end
- Resolution
  Database table that stores records on which a campaign runs
- Domains
  A selected segment of data on which a campaign runs
- Responses
  Records generated as a result of campaign activities

Creating a strategy

A campaign strategy names and describes one or multiple campaigns. A strategy supports grouping multiple campaigns. BIRT Analytics Campaign Workflow Planning supports creating a branched hierarchy containing strategies and campaigns.

How to create a strategy for a campaign

1. In Campaign Workflow, select Planning.
2. Right-click Strategies and choose New Strategy.
3. In New Strategy, type a name and description of a unique strategy, as shown in Figure 7-9.

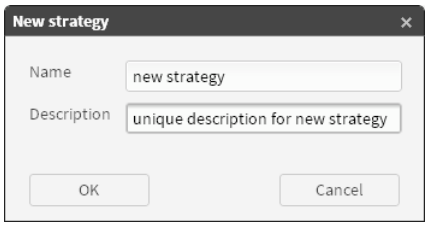


Figure 7-9 Creating a new strategy for a campaign

4. Choose OK.

Creating a campaign

Creating a campaign requires a strategy. As a best practice, establish a strategy, then create one, or multiple campaigns within the scope of that strategy. Modify the scope of a strategy as necessary, to accommodate expanding campaign activities.

Note: *For convenience, when creating or editing a campaign, you can resize the upper* List of Actions *panel and/or the lower* Campaign properties *panel.*

How to create a new campaign

1. Choose Campaign Workflow—Planning.
2. In Planning, right-click a strategy and choose New Campaign.

3 In Campaign Detail—Campaign's Properties—General Data, shown in Figure 7-10, type the following information:

- In Code, specify only one campaign.
- In Name, type a name for this campaign code.
- Optionally, describe the campaign name and code.

Figure 7-10 Adding general properties for a new campaign

4 In Campaign's Properties—Dates, shown in Figure 7-11, select the following options:

- In Start Date and End Date, select values representing dates on which campaign activities begin and finish.
- Optionally, select Number of times. Then, select a number to set how many times to repeat all campaign activities between the defined start and end dates.

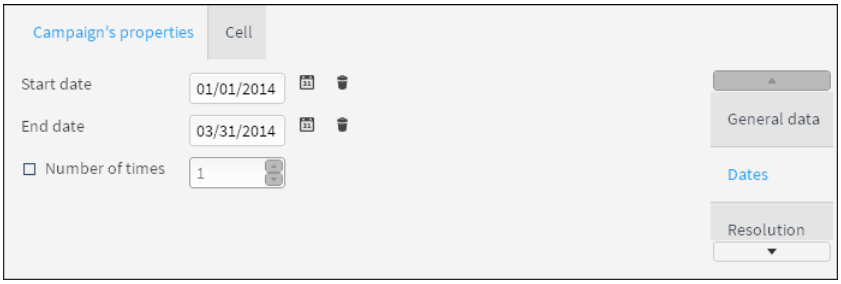


Figure 7-11 Adding a date range for a new campaign

5 In Campaign Properties—Resolution, select a resolution level, as shown in Figure 7-12.

Figure 7-12 Selecting a resolution level for a new campaign

6 Optionally, to limit the resolution level of a campaign, select De-duplicate and choose a column name from the list.

For example, consider a campaign designed to contact all customers by setting the resolution level to Customers table. To limit customers contacted to one-per-household, limit the campaign resolution by selecting De-duplicate and selecting Household ID.

7 In Campaign's Properties—Domains, define any selections to include or exclude in the campaign, as shown in Figure 7-13.

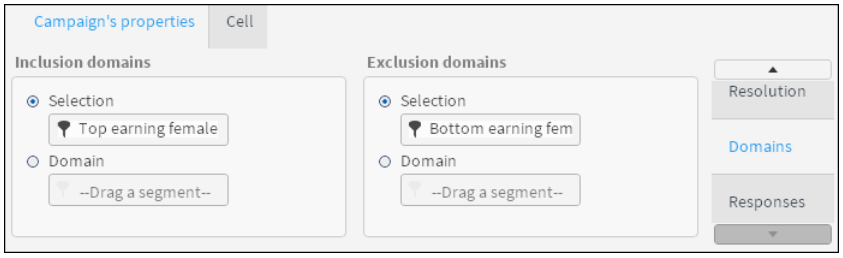


Figure 7-13 Defining a selection to exclude for a new campaign

8 In Campaign's Properties—Responses, select Deadline for Responses. Then, select a date value to set a deadline for responses to be included in this campaign, as shown in Figure 7-14.

Figure 7-14 Setting a response collection deadline for a new campaign

9 In Campaign detail, choose Save.

Saving a campaign locks property modification for all users except the user creating the campaign.

How to unlock a campaign

To unlock a campaign for editing by other users, choose Check in, as shown in Figure 7-15.

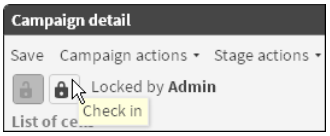


Figure 7-15 Unlocking a campaign for modification

No blocked appears for an unlocked campaign, as shown in Figure 7-16.

Figure 7-16 Examining an unlocked campaign

How to exclude a segment of records from a campaign

1 In Campaign Properties—Domains, select Domain in Exclusion domains, as shown in Figure 7-17.

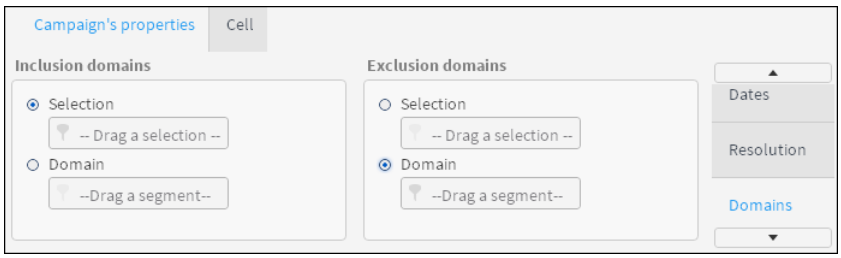


Figure 7-17 Selecting exclude domain for a campaign

2 From Scratchpad, drag a data segment and drop it in Exclusion domains—Domain.

For example, to exclude all male directors from a campaign, create a segment that includes only male directors. Then, drag the Male Directors segment from Scratchpad and drop it Exclusion domains, as shown in Figure 7-18.

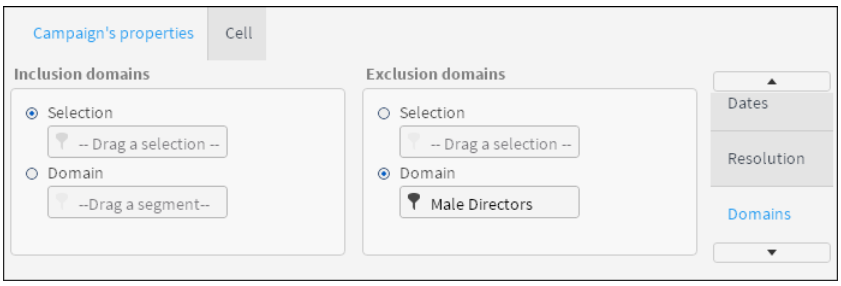


Figure 7-18 Excluding a segment from a campaign

In Campaign detail, choose Save.

How to define a segment as a campaign domain

1 In Campaign Properties—Domains, select Domain in Inclusion domains, as shown in Figure 7-19.

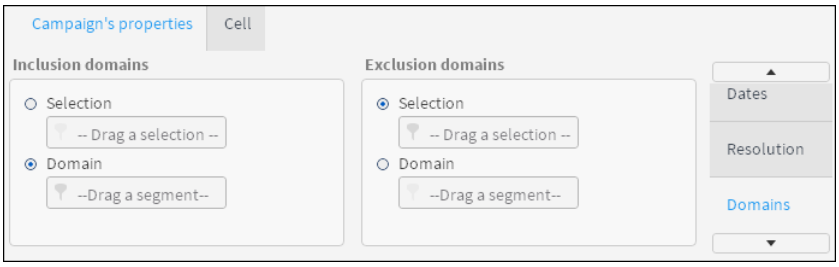


Figure 7-19 Selecting include domain for a campaign

2 From Scratchpad, drag a data segment and drop it in Inclusion domains—Domain.

For example, to define the domain of a campaign to include only those customers who buy product MA:

1 In My Data, find the Product Group MA value in the Order detail table.

2 Drag Product Group MA and drop it in Scratchpad, as shown in Figure 7-20.

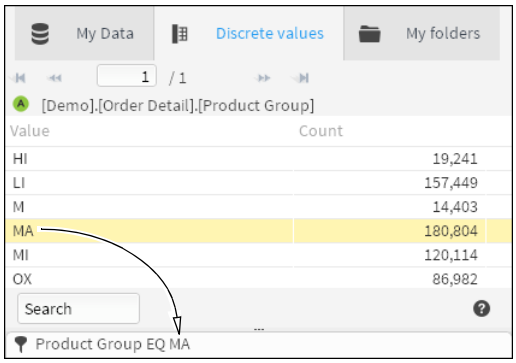


Figure 7-20 Selecting a segment for a campaign domain

3 Change the resolution of Product Group EQ MA segment to Customers, as shown in Figure 7-21.

- In Scratchpad, right-click Product Group EQ MA and choose Change resolution level.

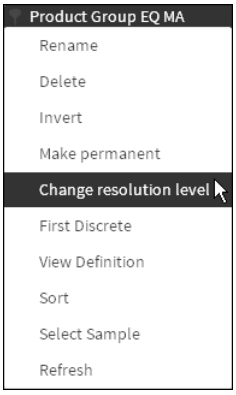


Figure 7-21 Changing the resolution level for a segment

- In Change resolution level, select Customer table and choose OK.
- Rename the segment Customers who buy MA, as shown in Figure 7-22.

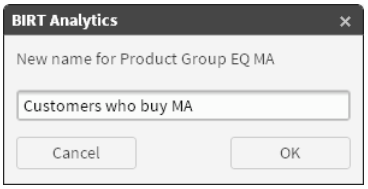


Figure 7-22 Renaming a segment

4 In Campaign Properties—Domains, in Inclusion Domain, select Domain.

5 Drag a segment from Scratchpad and drop it into Inclusion domains—Domain, as shown in Figure 7-23.

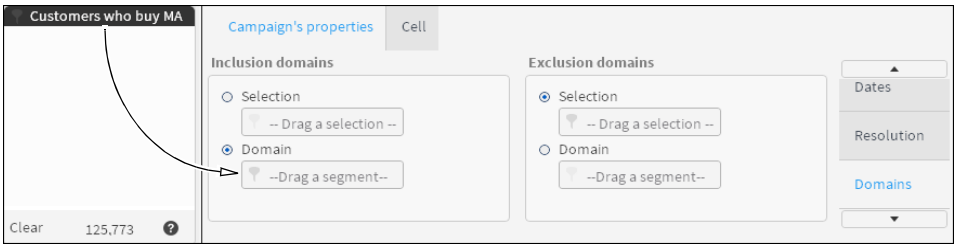


Figure 7-23 Including a segment as a campaign domain

3 In Campaign detail, choose Save.

About campaign cells

A campaign cell defines an action executed according to defined properties for all records in a segment. The action is defined to accomplish a specific goal within a campaign. For example, you can set cell properties that generate a specific text message to each male customer who purchased an Android phone. A cell may be set to execute at a specific time or coincident with a specific event during a campaign. A cell also may collect response data associated with records in a segment, according to business rules defined as cell properties.

To modify or delete an existing cell, in Campaign Detail—List of cells, select a cell and choose Modify or Delete, as shown in Figure 7-24.

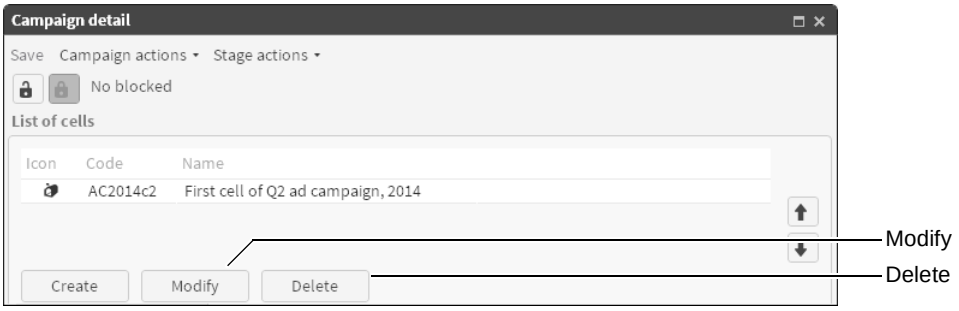


Figure 7-24 Modifying a campaign cell

How to create a cell in a campaign

1. In Campaign Detail, in List of cells, choose Create.
2. In Cell—General data, provide descriptive information for the cell, as shown in Figure 7-25.
    1. In Code, specify a cell for the campaign.
    2. In Name, type a name for this campaign cell.
    3. In Description, describe this campaign cell.
    4. In Media, select a media type.
    5. In Action, select an action type.

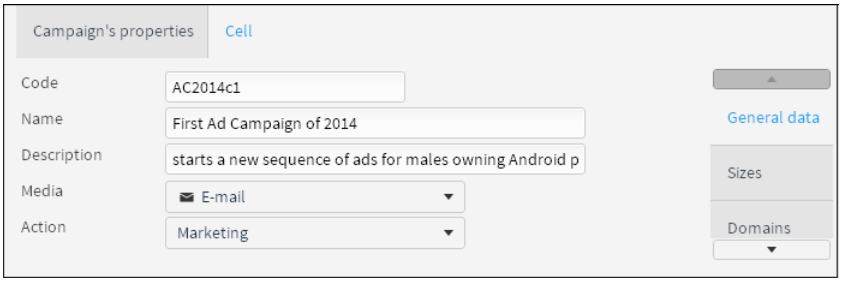


Figure 7-25 Adding descriptive information to a campaign cell

3. In Cell—Sizes, limit the size of the campaign using the following steps, as shown in Figure 7-26:
    1. In Minimum count, set a low value in the range of records for which the cell executes.
    2. In Maximum count, set a high value in the range of records for which the cell executes.
    3. In Sample method, select the type of method by which the cell selects records.
    4. In Remaining, select one of the following options:
        - Pass on. This option saves cell result records for use in subsequently executed cells.
        - Exclude from others. This options prevents saving cell result records.

Figure 7-26 Adding size and sampling properties to a campaign cell

4 In Cell—Domains, define the campaign cell domain. To specify a selection or domain for the cell, drag a selection or a segment from Data Tree and drop it in Selection or Domain, as shown in Figure 7-27.

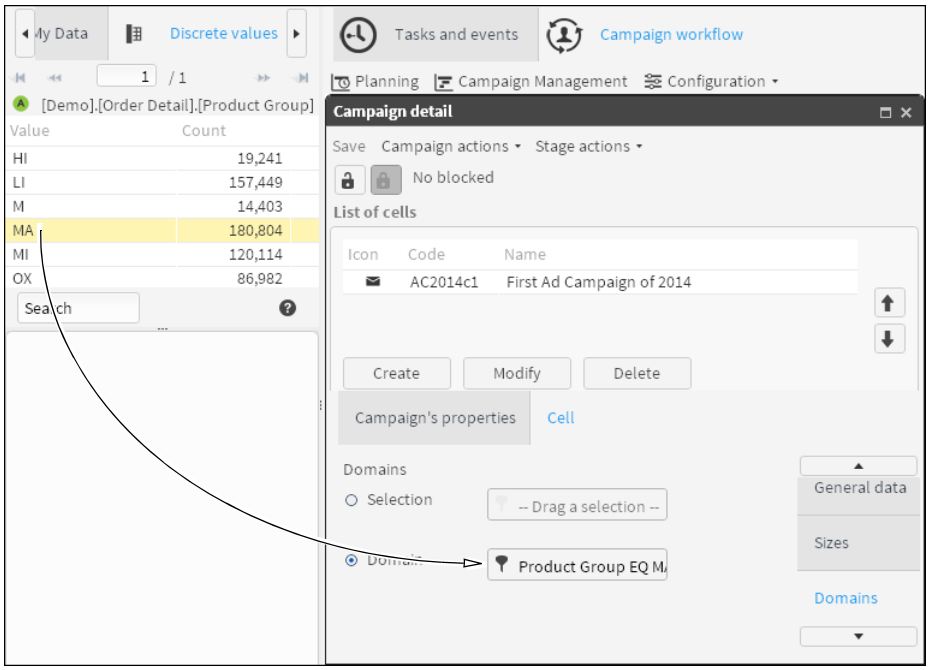


Figure 7-27 Defining the domain for a campaign cell

5 In Cell—Control, define a control cell and set parameters. Select a control type, sample size, and sample method, as shown in Figure 7-28.

Figure 7-28 Defining a control cell

To determine the effectiveness of a cell, compare cell results to the control cell.

6 Optionally, in Cell—Responses, select Implied response rules. Then, define a business rule using a segment from which responses generate, as shown in Figure 7-29.

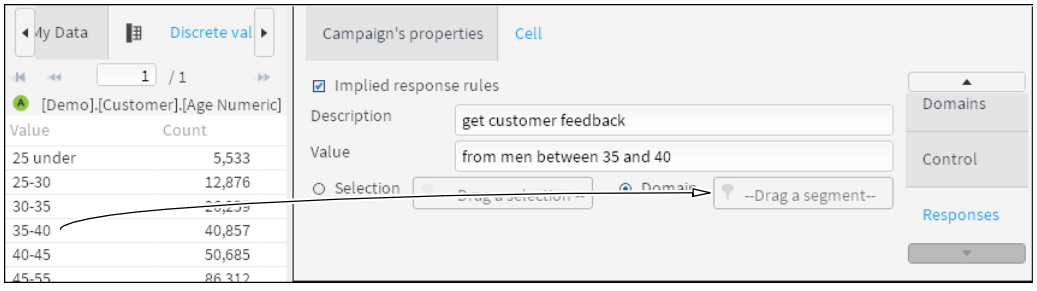


Figure 7-29 Adding a response rule to a campaign cell

7 Choose Save.

Running a campaign

Running a campaign includes the following activities:

- Planning
- Configuring
- Starting
- Managing
- Executing
- Reviewing

This section describes how to start, manage, and execute a campaign that is already planned and configured.

Starting a campaign

Before starting a campaign, lock the campaign to block any changes to campaign properties. Then, start the stage and workflow that contains the first campaign activity or action.

How to start a campaign

1. Choose Campaign Workflow—Planning.
2. In Planning, expand Strategies.
3. Select and expand a specific strategy, if necessary.
4. In a specific strategy, *double-click* on a campaign to open it for editing. You can also do this by making a *right-click* on it and choose Modify campaign.
5. To prevent changes to campaign properties, in Campaign detail, choose Lock.
6. Select Stage actions and choose Start campaign.
7. In Workflow, select the workflow that includes the first action in the campaign.
8. Choose OK. The current workflow and stage appear in Campaign detail, as shown in Figure 7-30.

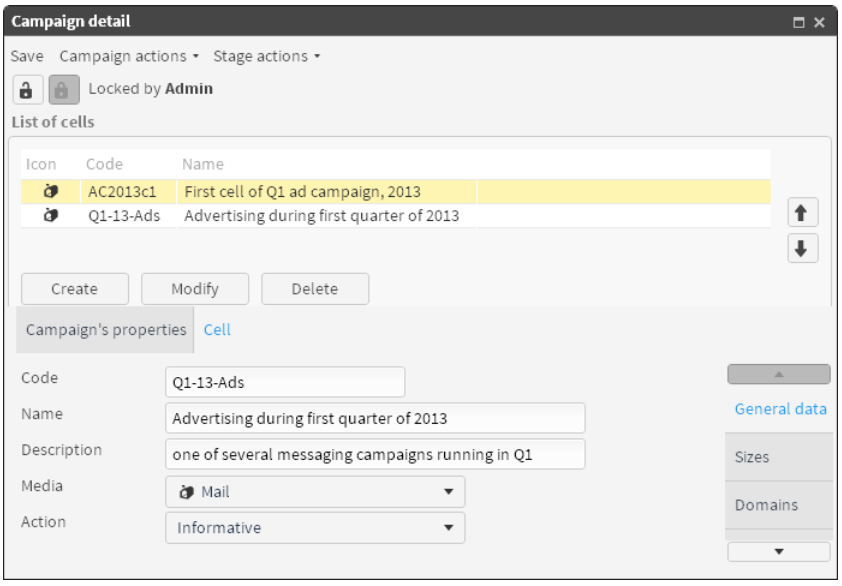


Figure 7-30 Examining the Campaign Manager workflow in stage 1 of a campaign

Managing campaign stages

You can manage and monitor campaign stages and workflows manually. You can also automate many actions in a campaign by definition. The following procedure summarizes the processes completed in a typical campaign.

1. Start campaign.
2. Start initial workflow.
3. Start initial stage.

   Cells defined in a stage execute, according to defined properties.
4. Start next stage.
5. Start next workflow.
6. Repeat steps 4–5 until all tasks in all stages and workflows have completed.
7. Execute campaign.
8. Evaluate campaign results.

How to manage stages in a campaign

1. In Campaign detail, select Stage actions and make one of the following choices:
   - To start a subsequent stage, choose Next stage.

     If you do not have permissions required to execute a stage, choose Accept in the message describing the permissions you lack.
   - To review a preceding stage, choose Previous stage.
   - To review a list of transitions between stages, choose Show History.
2. To close Campaign History, choose OK.

How to evaluate a campaign

1. Log in to BIRT Analytics as a user who has the Evaluate permission.
2. In Campaign detail, select Campaign actions and choose Evaluate.
3. In Results of the evaluation, examine the following levels of information:
   - To see details of records generated during completed stages, choose Table.
   - To see a summary of records generated by each completed cell, choose Chart.

Viewing campaign summaries

Campaign management supports viewing summary information using a Gantt chart display. This view presents a time-line background on which multiple campaigns appear. Options include viewing a summary of all campaigns, or only campaigns having current activity.

How to view campaigns on a Gantt chart

1. Choose Campaign Workflow—Campaigns Management.
2. In List of campaigns, select All campaigns, then choose Refresh.
3. To view information about a selected campaign, select one of the following options:
   - To limit campaigns appearing on the Gantt chart to only active ones, in List of campaigns, select Current.
   - To view information about a campaign, hover the cursor over a campaign icon on the chart.

- To view details for a campaign, double-click the image for a single campaign on the Gantt chart.

Executing a campaign

Executing a campaign completes the set of campaign activities. Executing a campaign loads all data generated to date by campaign activities into the history and response database tables defined for the campaign.

How to save campaign results and responses in a database table

1. After all stages in all workflows complete, select Campaign actions and choose Execute.
2. In Execution options, to manage loading data in the Workflow history table, select one of the following options:
   - To create a file containing all campaign result records, select Just generate file.
   - To load campaign result records in your database, select Generate file and load it on history table.
   - To delete cached result records after loading them in the history table, select Generate file, load it and remove it.
3. To generate response records in a file, select Campaign actions and choose Response.

How to review campaign results

1. To explore a column in CWorkflow, drag the column from My Data, or a segment from Discrete Values and drop it in Data Explorer.
2. To view summary of history records for this campaign in Data Explorer, choose Summary.
3. To view details about the response records for this campaign in Data Explorer, choose Record view.

Chapter

8

Scheduling tasks

This chapter contains:

- Automating a task
- Managing scheduled tasks

Automating a task

BIRT Analytics supports automating tasks that initiate specific actions. A task initiates an action based on a trigger of one of the following types:

- Time trigger
  A task having a time trigger runs at a set time.
- Event trigger
  A task having an event trigger runs when a selected event occurs.

To automate a task, create a scheduled task to execute when the set trigger occurs.

About event types

Available events vary based on plug-ins installed. For each event, you specify a condition when you configure the event, including an attribute, an operation, and a value.

For example, you can create the following event condition:

```
Full Name = Michael
```

In this event, a task triggers each time the value Michael appears in the field Full Name.

Available events appear in the following groups:

- BIRT Analytics
  - Engineering
  - Modified definition
  - Delete repository item
  - Rename repository item
- Data Mining
  - Apply model
    - Clustering
    - Decistion tree
    - Linear regression
    - Logistic regression
    - Naive Bayes algorithm
- Campaign Workflow
  - Campaign executed
  - Cell executed
  - Step change

Typical event attributes are:

- User: User name
- Date CCYYMMDD
- Hour HHMMSS
- Engine data format
  Date in analytical engine format. This attribute has to be used in SQL queries that invoke date or date-time fields.

About action types

BIRT Analytics supports communication, query, program execution, data analysis and data modeling actions.

Available actions appear in the following groups:

- Sending
  - Send email
    Sends a message via email to one or multiple recipients. To create a Send email action, you configure the following details: e-mail contents, event trigger, or time trigger attributes. Right-click in the Content editor to display available content options. Content options vary, based on the task trigger. Common examples of email task triggers include:
  - Send execution results
    Sends an email with campaign execution results. This mail can be delivered to each user who has been involved in a campaign execution process.
  - Notify users via email
    Sends an email to all the users related to a campaign triggering a task. This email will be sent to users having permissions for the current campaign stage and a valid email account on the database.
- Conditional
  - Query. A query action evaluates whether data matches a condition. Correct query syntax follows: *action, filter, operator, value*
    - *action* is a function such as count, sum, or mean
    - *filter* is a database domain or SQL statement
    - *operator* is >, <, or =, for example
- Execute
  - Execute campaign
    Runs a campaign execution.
  - Execute response load
    Loads all campaign responses associated to an event in a new database table.
  - Export campaign cell
    Exports all records from a selected cell in a campaign. Sets the event or campaign date of execution, the records to export (all the records, all the records except control, only control) and an exportation report that will be used as a template for exporting.
- Data Model
  - Delete column
    Deletes a column. User must indicate entire column name [database].[table].[column].
  - Delete table
    Deletes a table. User must type the full table name, for example, [database].[table].
  - Delete database
    Deletes a database.
  - .Rename
    Renames a repository item. User must type full name of the repository item, for example, [database].[table].[column].

- Apply
  - Apply Model
    Applies the Clustering, Decision Tree, Linear regression, Logistic regression or Naive Bayes model. These analyses must be properly configured, trained and saved.
  - Apply model online
    Set an action to apply the model in a deferred way.

  In either case, the result is a column containing results of the applied model.

Creating a scheduled task

A scheduled task includes a trigger, task details, and an assigned action. The following procedure summarizes high-level steps for creating a scheduled task.

1. Create a new scheduled task.
2. Configure conditions that trigger the task.
3. Assign an action to the scheduled task.
4. Save the scheduled task.

How to create a task triggered by time

1. Choose Workflow➛Tasks and events➛New Scheduled Task.
2. In Scheduled task—Name, type characters that name the task.
3. Optionally, describe and activate the task, using the following options:
   - In Description, type characters that describe the task.
   - Select Starting date and Ending date. Then, use the date and time selector to specify a range of dates during which this task trigger is active.
4. In Trigger for task, expand Scheduled task on time.
5. Drag Time to the right and drop it in the box.
6. In Time, select values for each time parameter using the selector, as shown in Figure 8-1.

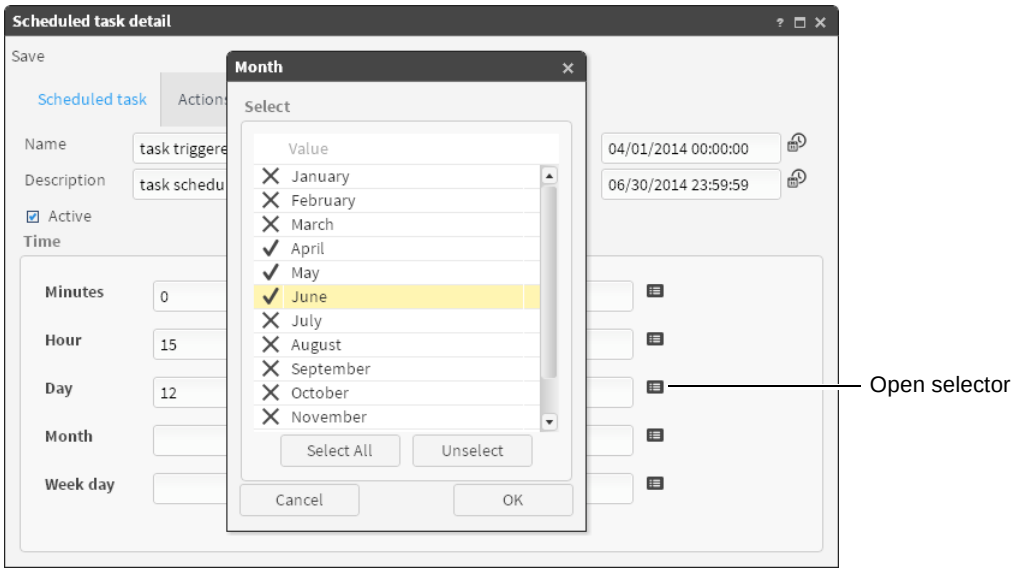


Figure 8-1 Setting time properties for a scheduled task

For example, double-click the selector icon for Month. In Month, select values for the minute, hour, day, month, and day-of-week that specify when the task triggers. Choose OK. Not every field requires a value.

Default values are: minute = 0, hour = 1, and current day. Default settings trigger a task on this day of each month at the hour and minute 0.

7 Choose Save.

How to create a condition for an event trigger

1 In Scheduled task detail, in Trigger for task, expand Scheduled task on event.

2 Drag an event and drop it as shown in Figure 8-2.

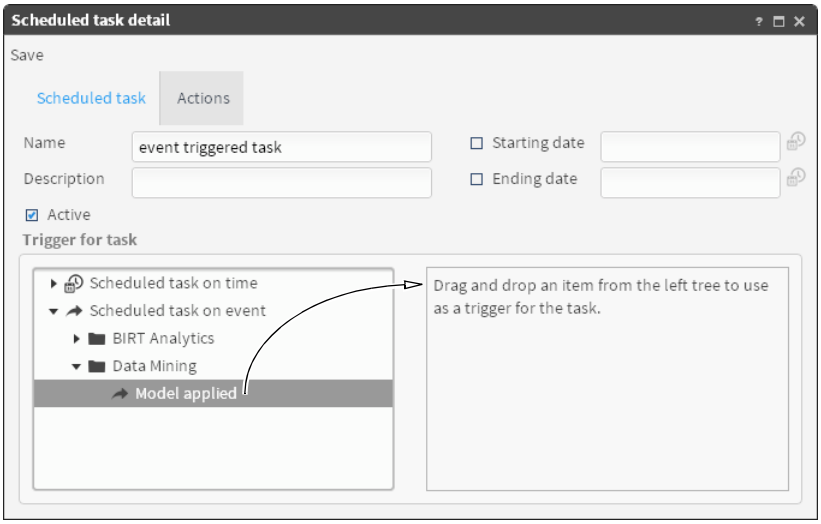


Figure 8-2 Creating a trigger for an event-driven task

3 Drag a listed attribute and drop it in Conditions, as shown in Figure 8-3.

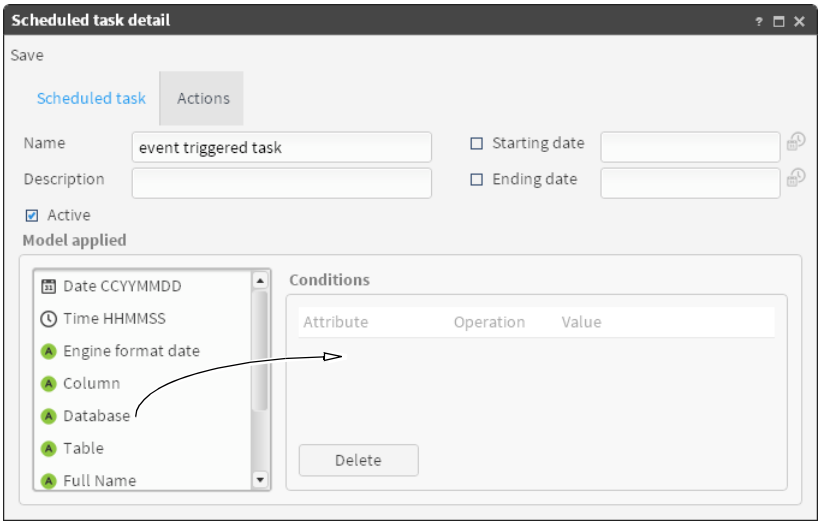


Figure 8-3 Creating a condition that triggers an event-driven task

4 In Conditions detail, select an operation and type a value for the selected attribute, as shown in Figure 8-4.

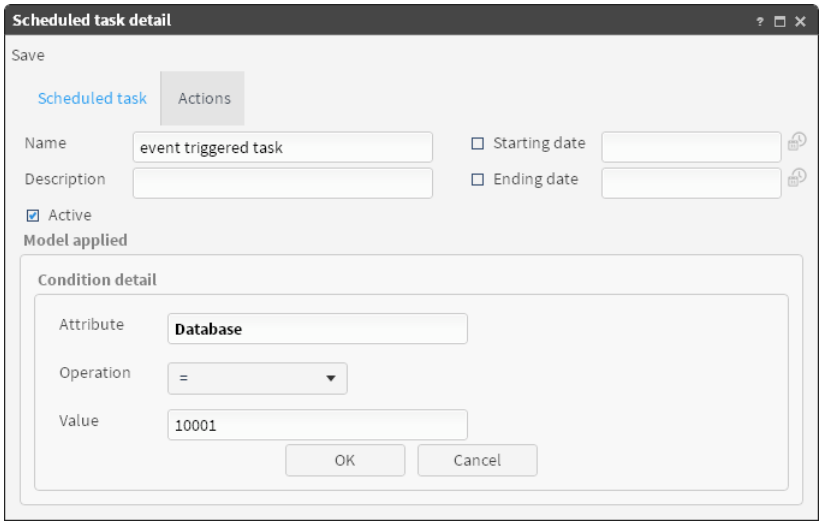


Figure 8-4 Configuring condition details for an event trigger

5 Choose OK.

How to assign an action to a scheduled task

1 In Actions, expand an action group.

2 Drag an action name and drop it in the Actions visual editor, as shown in Figure 8-5.

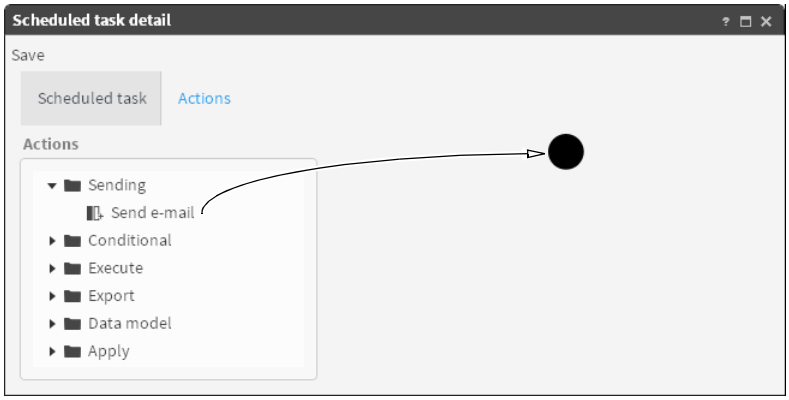


Figure 8-5 Dropping an action in Actions visual editor

3 In Action detail, define action properties using tools provided for a selected action type.

For example, to create a Send email action, provide values for the following properties, using the tool-set provided in Action detail editor, as shown in Figure 8-6.

- In To..., type a valid email address.
- Choose the check mark to validate an email address.
- In Subject, type a subject for an email message.
- In Format, select HTML or Plain Text.
- In Content, type characters that comprise an email message.

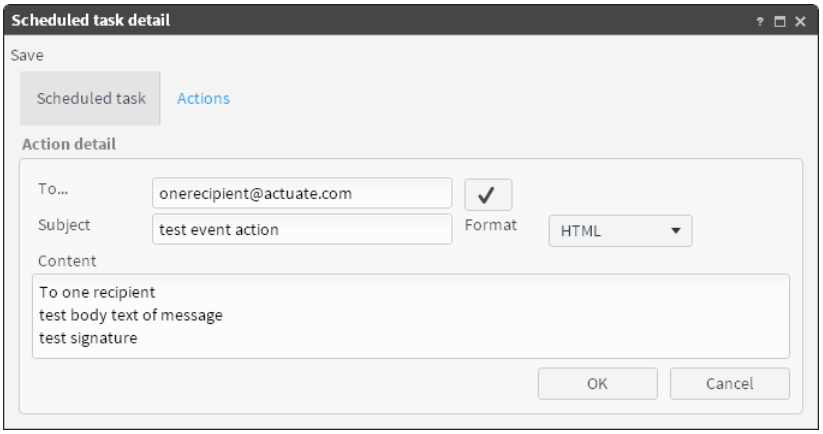


Figure 8-6 Defining properties for a Send email action

4 Choose OK. An action appears in the Actions visual editor, as shown in Figure 8-7.

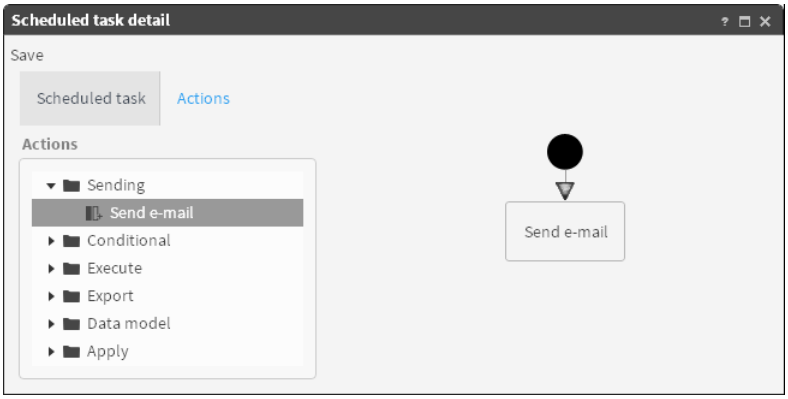


Figure 8-7 Examining the visual actions editor

How to edit action details

1 In Actions visual editor, double-click an action.

2 Action detail opens editing tools appropriate for the selected action.

How to remove an action from a scheduled task

1 In Actions, right-click an action in the visual editor.

2 Choose Delete, as shown in Figure 8-8.

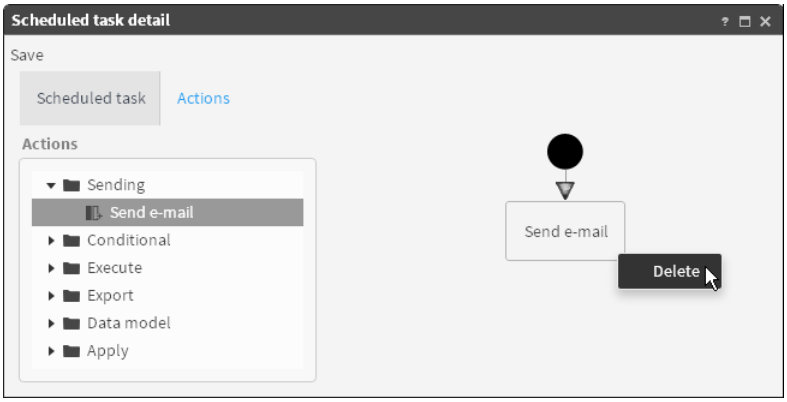


Figure 8-8 Removing an action from a scheduled task

Managing scheduled tasks

To view all scheduled tasks in BIRT Analytics, choose Workflow—Tasks and events—Scheduled tasks. Scheduled tasks appear, sorted by creation date and time in ascending order. A selected task appears highlighted, as shown in Figure 8-9.

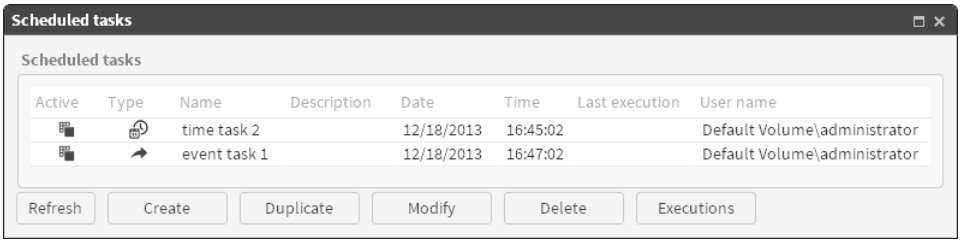


Figure 8-9 Selecting one scheduled task

Use Scheduled tasks to complete any of the following activities:

- To view current status of all scheduled tasks, choose Refresh.
- To create a new scheduled task having no properties, choose Create.
- To create a new scheduled task having the same properties as an existing one, select a task and choose Duplicate.
- To edit an existing scheduled task, select a scheduled task and choose Modify.
- To remove a scheduled task from BIRT Analytics, choose Delete.
- To view a list of scheduled tasks already completed, choose Executions.

Duplicating a scheduled task

Duplicate creates a new scheduled task having all properties of an existing, selected task. Duplicating a scheduled task, then modifying one or multiple properties in the duplicated task, saves time spent recreating identical properties for a new task. Only the time attribute, indicating the time at which a task is created, is unique for a new, duplicate task.

How to duplicate a scheduled task

1. In Scheduled tasks, select a task.
2. Choose Duplicate.
3. In Scheduled task detail, choose Save.

   A new task having all properties of the original task appears in Scheduled tasks. The new task has a unique value in Time, as shown in Figure 8-10.

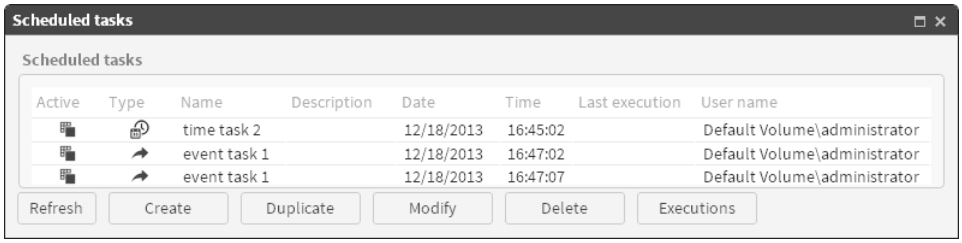


Figure 8-10 Examining a duplicated scheduled task

Modifying a scheduled task

You can modify properties of an existing, scheduled task before the time or event set as a trigger for that task occurs. Scheduled task details supports editing and saving changes to task properties.

How to modify a scheduled task

1. In Scheduled tasks, select a task from the list.
2. Choose Modify.
3. Using Scheduled task details, change at least one property of the task. For example, edit the name property.
4. Choose Save.

Using a conditional query to automate actions

Action properties support associating specific actions with defined trigger event conditions. Automating a notification process provides a basic example for associating event-triggered, scheduled tasks and unique trigger events. A basic example demonstrating how to define unique actions that result from evaluating a conditional query follows.

The following example task generates one of two unique email message actions based on evaluating an order quantity in a customer database.

How to specify actions based on query results

1. In a scheduled task, create an event trigger having conditions that monitor the customer order table:
   1. In Trigger for task, expand BIRT Analytics. Then drag and drop an Engineering event type in the scheduled task editor.
   2. In Scheduled task detail—Scheduled task, define conditions and attributes that identify a specific column, as shown in Figure 8-11.

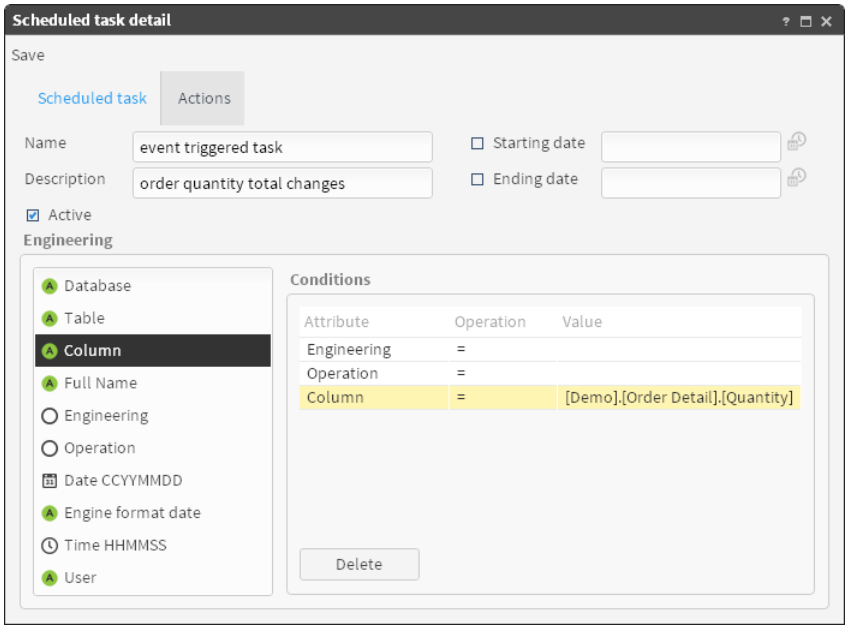


Figure 8-11 Defining specific conditions for a scheduled task

2. In Actions for the event-triggered task, expand Conditional. Drag Query and drop it in the visual editor.

3 Specify query conditions for an action using Action details. For example, the selections shown in Figure 8-12 compare a quantity in a specific field with a defined value. The result returned by the conditional query is Yes or No. Choose OK.

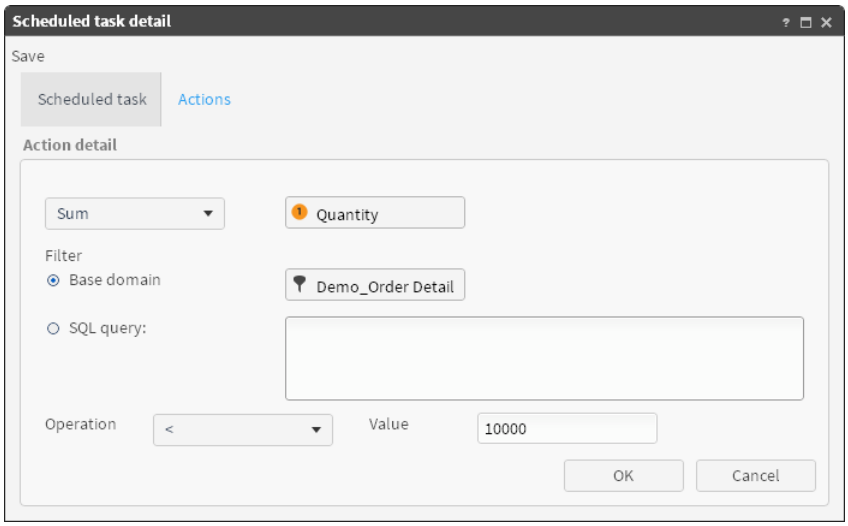


Figure 8-12 Specifying query conditions for an action

4 In Actions, add two Send email actions, using the visual editor, as follows:

1 Expand Sending. Drag and drop a Send email action on Query. Choose Yes.

2 Configure details for an email message action, appropriate for a Yes result.

3 Drag and drop another Send email action on Query.

4 Configure details for an email message action, appropriate for a No result.

5 Two Send email actions appear, one for each query condition, as shown in Figure 8-13.

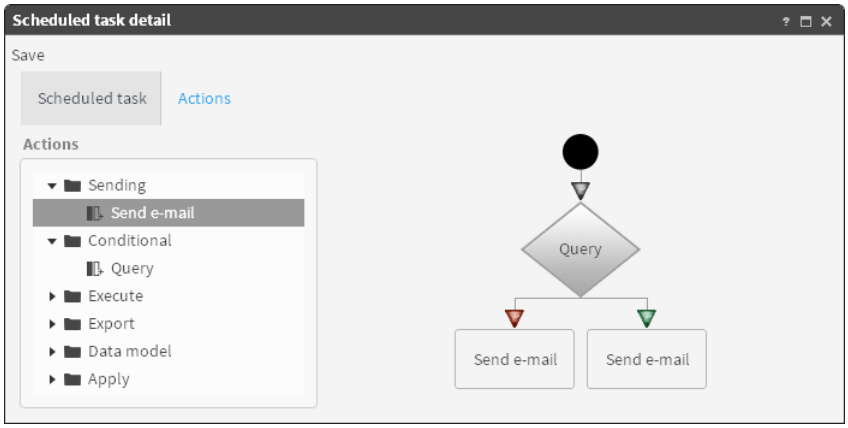


Figure 8-13 Examining a conditional query and associated actions

6 Choose Save.

A

access control list (ACL)

A group or set of users with access to a database object. Using the BIRT Analytics Administration tool, the administrator creates a security group or ACL that manages privileges for a database object.

Related terms
BIRT Analytics Administration, column, database, group, security role, table

action

An action is an event executed by a manual or task trigger. Example actions include send e-mail, query action, delete column, and apply model.

Related terms
scheduled task, trigger

Aggregates

A tool that supports grouping data from multiple tables in one table. Aggregates supports defining a function and filter as properties.

Related terms
filter, table

analysis

A tool that provides a specific view of data stored in FastDB. BIRT Analytics supports multiple analyses.

Related terms
Bubble analysis, Crosstab analysis, Evolution analysis, FastDB, Map analysis, Pareto analysis, Profile analysis, Venn analysis

antecedent

Terms representing the left-hand, or If… clause of an association rule. The antecedent clause of an association rule contains discrete data items.

Related terms
association rules, consequent

Association Rules

A predictive analytics tool that uses association rules to identify an If…Then relationship between data values stored in an information repository. For example, an association rule may show the following relationship: If a customer buys products A and B, then the customer also buys product C.

Related terms
association rules, predictive analytics

association rules

A predictive analytics technique that analyzes data for frequent If…Then patterns and calculates support and confidence criteria that identify the most important relationships. Support indicates how frequently the items appear in the database. Confidence indicates the number of times the If…Then relationships evaluate true.

An association rule has two parts, an antecedent and a consequent. The antecedent represents one or multiple data items. The consequent represents an item found in combination with the antecedent. An association rule returns a lift and a leverage value that measure how well the rule predicts the consequent.

Related terms
antecedent, Association Rules, confidence, consequent, lift, leverage, predictive analytics, support

Association Rules

A predictive analytics tool that uses association rules to identify an If…Then relationship between data values stored in an information repository. For example, an association rule may show the following relationship: If a customer buys products A and B, then the customer also buys product C.

Related terms
association rules, predictive analytics

B

baseline filter

A filter that returns a group of records to serve as a basis for comparison. For example, use the year 2012 as a baseline filter for profit, to compare profit earned in another year with profit earned in 2012.

Related terms
filter, record

big data analysis

The practice of analyzing, exploring, filtering, loading, segmenting, and studying massive quantities of data. Big data analysis uses statistics to describe qualities and predict trends in these data repositories.

Related terms
analysis, BIRT Analytics, data repository

BIRT Analytics

An application, including a data repository, data loader, and web service, that supports big data analysis.

Related terms
big data analysis, BIRT Analytics Administration

BIRT Analytics Administration

A BIRT Analytics system administration tool that runs as a browser-based application. The administrative user has full permission to modify all configurable features of the BIRT Analytics system.

Related term
BIRT Analytics

BIRT Analytics Loader module

A tool that extracts, transforms, and loads records from an external data source to FastDB.

Related terms
BIRT Analytics, BIRT Analytics Administration, FastDB

Bubble analysis

A tool that supports viewing a spatial distribution of data with respect to two axes.

Related terms
analysis, Crosstab analysis, Evolution analysis, Map analysis, Pareto analysis, Profile analysis, Venn analysis

C

calculated field

A data field that displays the result of an expression.

campaign

A set of tasks, defined for specific population segment. A campaign is completed during a defined time period to accomplish a specific goal.

Related term
segment

Canvas

A workspace for data analysis gadgets. Canvas supports arranging, assembling, and saving a collection of data visualization gadgets.

Related term
gadget

cell

A set of properties that defines campaign actions to be performed for all records in a segment.

Related terms
action, campaign, record, segment

Clustering

A predictive analytics tool that uses k-means cluster analysis. Clustering identifies groups of similar data values in large segments stored in a big data repository.

Related terms
k-means, cluster analysis, predictive analytics

cluster analysis

A data analysis task that iterates estimating of values assigned to common data attributes. Common attributes identify groups of similar items, called clusters. Comparing clusters highlights similar and different groups in big data.

Related terms
analysis, big data analysis, Clustering

column

1. A named field in a database table or query. For each data row, the column can have a different value, called the column value. The term column refers to the definition of the column, not to any particular value.
2. A vertical sequence of cells in a crosstab, grid element, or table element.

Related terms
column-oriented DBMS, database, data field, query, table

column-oriented DBMS

A column-oriented DBMS is a database management system (DBMS) that stores data tables as sections of columns of data rather than as rows of data. A column-oriented DBMS serializes all of the values of a column together, then the values of the next column, and so on.

Related terms
database, column

confidence

An expression used to identify an association rule. Confidence compares how often the consequent appears when the antecedent is met. The confidence expression has the following syntax:

```
Confidence (A,B-> C) = Support (A,B,C)/Support (A,B)
```

Related terms
association rules, support

consequent

Terms representing the left-hand, or …Then clause of an association rule. The consequent clause of an association rule contains items found in combination with items in the antecedent.

Related terms
antecedent, association rules

Convert A BIRT Analytics option that displays results from one data analysis using a different type of data analysis. For example, an analysis created using Crosstab converts to a Bubble, Evolution, or Map analysis.

Related term
analysis

count The total number of records in a field.

Related terms
field, record

Crosstab analysis

A tool that supports analyzing data using cross-tabulation, or pivoting of different fields.

Related terms
analysis, Bubble analysis, Evolution analysis, Map analysis, Pareto analysis, Profile analysis, Venn analysis

Cylinder A data visualization gadget that displays numeric values and boundaries in ranges. A Cylinder displays defined data measures as colored slices that comprise one cylinder shape.

Related terms
Dial, Funnel, gadget, Gallery, Label, Meter, Sphere

D

data analysis A process including acquiring, organizing, transforming, and modeling data to support decision-making.

Data Explorer

A tool that displays records from a database stored in FastDB. Data Explorer provides a summary view for a table and a detail view for records, tables, selections, and segments.

Related terms
Data Tree, FastDB, record, table

data field A location storing data having a specific type. A data field typically contains data from a database or other data source. A data field appears as a column when viewing a table in Data Explorer. For example, the BIRT Data Analytics Demo database includes the data field types listed in Table G-1.

Table G-1 Data field types

Icon	Field type	Description
	Calculated	Displays a value result from an expression
	Date	Contains numbers that represent day, month, and year
	Date and time	Contains numbers that represent day, month, year, and time of day
	Full numeric	Contains whole, or integer numbers, such as 1 or 1000
	Real numeric	Contains real, or partial numbers such as 1.05 or 0.003
	Time	Contains a value representing time of day
	Text	Contains a string of alphabetic characters

Related terms
record, Data Explorer, Data Tree, data types

data integration

A process through which data in varied sources is combined.

data mining A computational process used to extract and transform data to prepare it for analysis.

Related term
analysis

data repository

A physical or virtual location for storage and retrieval of data.

Related term
FastDB

Data Tree A tool that supports viewing and working with databases, tables, and records stored in FastDB. Data Tree includes Discrete Values, My Data, and My Folders viewers.

Related terms
database, Discrete Values Viewer, My Data Viewer, My Folders Viewer, record, table

data types A data type defines the limits of a data field in a BIRT Analytics database. For example, the BIRT Data Analytics demo database includes the data types listed in Table G-2.

Table G-2 Data types in BIRT Analytics Loader

Data type	Description
Date	Contains numbers that represent day, month, and year. The default format is mm_dd_yyyy.
Datetime	Date and time data from January 1, 1753, through December 31, 9999, providing accuracy to three-hundredths of a second, or 3.33 milliseconds. The default format is yyyy_mm_dd_hh_MM_ss.
Integer	Integer data from -2^31 (-2,147,483,648) through 2^31-1 (2,147,483,647).
Longint	Integer data from -2^63 (-9,223,372,036,854,775,808) through 2^63-1 (9,223,372,036,854,775,807).
Real	Floating precision number data with the following valid values: -3.40E+38 through -1.18E-38, 0 and 1.18E-38 through 3.40E+38.
String	A sequence of ASCII characters.
Time	Contains a value representing time of day. The default format is hh_MM_ss.
Unicode	A sequence of characters based on consistent encoding, representation, and handling of text as expressed in global writing systems.

Related terms
Data Explorer, data field, Data Tree, record

database

1. An integrated collection of logically related records that provides data for information application platforms, such as BIRT. The database model most commonly used is the relational model. Other typical models are entity-relationship, hierarchical, network, object, and object-relational.
2. An integrated set of logically related records stored in FastDB.

Related terms
record, table

decision tree A predictive analytics technique that predicts the value of a target variable, based on values of multiple input variables. For example, use a decision tree to predict a survival rate, based on characteristics of the population that may survive.

Related terms
Decision Tree, predictive analytics

Decision Tree

A predictive analytics tool that uses the decision tree technique to predict an outcome, based on values of multiple input variables. For example, use Decision Tree to predict the product a customer will purchase, based on customer, purchase, gender, occupation, and income data.

Related terms
association rules, predictive analytics

Decodes A tool that supports renaming a data field stored in FastDB.

Related terms
data analysis, data field, FastDB

Dial A data visualization gadget that uses a needle-shaped pointer to display defined measures and numeric values in a range.

Related terms
Canvas, Cylinder, Funnel, gadget, Gallery, Label, Meter, Sphere

Discrete Values Viewer

A tool that supports viewing discrete values in a data record, selection, or segment.

Related terms
My Data Viewer, My Folders Viewer, record, segment, selection

Downloads A tool that supports writing FastDB records to an external database.

Related terms
database, Export file, FastDB, record

Dubnium.exe The file that runs the BIRT Analytics data repository, FastDB.

Related term
FastDB

E

Evolution analysis

A tool that supports viewing a time-progression view of data values.

Related terms
analysis, Bubble analysis, Crosstab analysis, Map analysis, Pareto analysis, Profile analysis, Venn analysis

Export Analytic DB

A tool that supports creating a new database field based on a segment defined in the database. The new field is stored in FastDB.

Related terms
Export file, FastDB, segment

Export file A tool that supports creating a new text file based on a segment defined in the database. The file is stored in FastDB.

Related terms
Downloads, FastDB, segment

Expressions A tool that supports creating a logical relationship, using data fields, functions, and operators. Results of the relationship appear as a calculated field in FastDB.

Related terms
calculated field, data field, FastDB

F

FastDB The BIRT Analytics data repository. FastDB is a web service that caches data and supports executing data analysis and forecasting algorithms.

Related terms
Data Tree, database, data repository, record, table

field See data field.

filter A function that limits the number of records included a segment or selection. BIRT Analytics supports the following three filter types: baseline, target, and universal.

Related terms
baseline filter, target filter, universal filter

Forecasting A predictive analytics tool that uses the Holt-Winters, iterative method. Forecasting predicts a future trend in data exhibiting a seasonal pattern.

Related terms
Holt-Winters, predictive analytics

functionalities

The system privileges an administrator grants to a security role.

Related terms
BIRT Analytics Administration, security role

Funnel A data visualization gadget displaying numeric values and boundaries that represent groups in a range, using colored bands that display on a funnel shape.

Related terms
Cylinder, Dial, gadget, Gallery, Label, Meter, Sphere

G

gadget A computer program that provides services without requiring an application for each one. BIRT Analytics provides multiple gadgets that support data visualization.

Related terms
Cylinder, Dial, Funnel, Gallery, Label, Meter, Sphere

Gallery A tool that supports running multiple data visualization gadgets. Use the Gallery to assemble, arrange, and save gadgets on the Canvas.

Related terms
Canvas, Cylinder, Dial, Funnel, gadget, Label, Meter, Sphere

group A set of users belonging to the same organizational unit who share the same permissions for performing tasks. Using the BIRT Analytics Administration tool, the administrator creates a group from the list of available users on the system.

Related term
BIRT Analytics Administration

H

has seasonality

User-selected option that recognizes a seasonal trend in a data set.

Related terms
Holt-Winters, seasonal periodicity, seasonality

Holt-Winters A popular numerical estimation method used to forecast values in data that exhibit seasonal trends. The Holt-Winters method repeats and refines a time-series formula that includes a

level, trend, and seasonal component. The formula calculates forecast values valid for time t using a weighted average for all data prior to time t.

Related term
Forecasting

I

Import
A tool that supports adding a field to a database by uploading records from an external database. The field is stored in FastDB.

Related terms
database, FastDB, field

indexed field
A data field having an associated key. An indexed field appears in a summary table used for data retrieval.

Related terms
data field, field, table

J-K

k-means
An iterative method of cluster analysis that groups large data sets into clusters of similar data. A k-means method forms clusters around data values having the nearest mean.

Related terms
analysis, Clustering, cluster analysis, mean

kurtosis
A coefficient that describes the degree of concentration for a distribution of values, based on a mathematical average. The kurtosis coefficient is a value between -0.5 and 0.5. Colloquially, the Kurtosis coefficient is an average that indicates how sharp a distribution is with respect to a standard normal distribution.

Related terms
skewness, standard normal distribution

L

Label
A data visualization gadget that associates specific alphanumeric characters with a defined measure. A Label displays a text description of a measure in the BIRT Analytics Gallery.

Related terms
Canvas, Cylinder, Dial, Funnel, gadget, Meter, Sphere

leverage
A value that indicates how well an association rule predicts the consequent. The method used to calculate leverage differs from the method used to calculate lift.

Related terms
association rules, lift

lift
A value that indicates how well an association rule predicts the consequent. A lift value greater than one indicates that the items in the rule appear together more than expected. The method used to calculate lift differs from the method used to calculate leverage.

Related terms
association rules, leverage

Links
A tool that supports maintaining links binding columns and tables in a database stored in FastDB.

Related terms
column, table

M

make permanent

A field operation that creates a new data field from either a calculated field or a current segment. The data field appears in FastDB.

Related terms
calculated field, FastDB, field, segment

Map analysis A tool that supports plotting data values and regions on a geographic map. For example, a map analysis shows geographic regions and the number of high-net-worth customers in each region.

Related terms
analysis, Bubble analysis, Crosstab analysis, Evolution analysis, Pareto analysis, Profile analysis, Venn analysis

maximum The highest registered value in a set of values.

Related term
minimum

mean An arithmetic mean of all registered values in the field.

Related terms
median, mode

median A value that divides a field into two symmetrical parts.

Related terms
mean, mode

Meter A data visualization gadget that uses colored bars to display numeric values and boundaries in a range.

Related terms
Canvas, Cylinder, Dial, Funnel, Gallery, Label, Sphere

minimum The lowest registered value in a set of values.

Related term
maximum

mode The values having the most frequent number of occurrences in a field.

Related terms
mean, median

My Data Viewer

A tool that supports viewing fields and tables in multiple databases stored in FastDB.

Related terms
Data Explorer, database, Discrete Values Viewer, field, My Folders Viewer, table

My Folders Viewer

A tool that supports viewing reports, selections, and gadgets by a user or, if shared, by other users. My Folders appears as a tab in Data Tree and in the Start pane.

Related terms
Data Explorer, Discrete Values Viewer, gadget, My Data Viewer, selection

N

NetScaler Web Logging (NSWL) query

A type of SQL query that tracks HTTP data traffic and writes information to a log file in a standard format such as the following example:

```
Select * from [Demo].[Household]
where [Demo].[Household].[Town]='LONDON';
```

Related terms
BIRT Analytics Administration, query, security filter, SQL (Structured Query Language)

normal distribution

A bell-shaped, single-peaked, symmetric distribution of data. In a normal distribution, the mean, mode, and median coincide at the center.

Related term
standard normal distribution

Numeric Ranges

A tool that supports creating a calculated field that includes a a series of ranges into which data from numeric fields is grouped. For example, Numeric Ranges supports defining the following age ranges: Young - for age values less than 21, Adult - for age values 21 through 67, and Old - for age values greater than 67.

Related terms
calculated field, field

O-P

parameter

A variable expression that accepts a defined set of values.

Related term
filter

Parametric

A tool that supports creating a field based on a defined condition, for use as a filter on a measure.

Related term
field

Pareto analysis

A tool that supports comparing data using the Pareto principle, a commonly accepted rule which implies a data distribution with a numeric ratio of 80% to 20%. For example, the Pareto principle implies that 80% of sales result from 20% of customers.

Related terms
analysis, Bubble analysis, Crosstab analysis, Evolution analysis, Map analysis, Profile analysis, Venn analysis

predictive analytics

A subject encompassing a variety of techniques used to analyze current and historical facts to make predictions about future, or otherwise unknown events. Credit scoring is a well- known application that uses predictive analytics techniques to generate a score for an individual, based on credit history data for that individual.

Related terms
Association Rules, Clustering, Decision Tree, Forecasting

Profile analysis

A tool that supports identifying a set of similar characteristics in a group. A profile analysis compares z-score values calculated for each set of characteristics.

Related terms
analysis, Bubble analysis, Crosstab analysis, Evolution analysis, Map analysis, Pareto analysis, Venn analysis, z-score

profile

A set of associated security roles, groups, filters, and users. Using the BIRT Analytics Administration tool, the administrator creates a profile from the lists of roles, groups, filters, and users available on the system. From the BIRT Analytics security options list, choose Profiles, specify a profile name, provide a description, then select the roles, groups, filters, and users to include in the profile.

Related terms
BIRT Analytics Administration, group, security filter, security role

prompted filter

A data set filter that supports user entry of parameter values.

Related terms
filter, parameter

Q

Quantile

A tool that supports creating a new calculated field by grouping values in a numeric field, using multiple groups that contain an equal number of values. For example, use Quantile to group a field containing 2400 values into four quartiles having 600 values each.

Related terms
calculated field, field

query

A statement specifying the data rows to retrieve from a data source. For example, a query that retrieves data from a database typically is a SQL SELECT statement.

Related terms
database, SQL (Structured Query Language)

R

Ranking

A tool that supports ordering a table by generating a column of calculated values that correspond to a sorted column. The calculated values represent an ordered list of ranks.

Related terms
column, table

record

A set of related, indexed data fields in a database. A record often appears as a row shown in a table. For example, a customer record could include a numeric field for customerID, a character string field for customer name, and an alphanumeric field for age group.

Related terms
field, row

row

See record.

Related terms
field, record

S

scheduled task

A scheduled task includes a trigger, task details, and an assigned action.

Related terms
action, campaign, stage, trigger, workflow

Scratchpad A BIRT Analytics work area that supports temporary caching of multiple segments. Scratchpad also supports creating new fields based on segments or selections.

Related terms
Data Explorer, Data Tree, segment, selection

seasonal periodicity

A value indicating the number of periods in a cycle. Input a value for seasonal periodicity to initiate a forecast that predicts a seasonal pattern in a data set.

Related terms
Forecasting, Holt-Winters

seasonality In a data set, a periodic trend that corresponds to monthly, quarterly, or semi-annual periods such as seasons.

Related terms
Forecasting, Holt-Winters

security role A set of functionalities that an administrator uses to configure permissions in the BIRT Analytics system.

Related terms
BIRT Analytics Administration, functionalities, query, security filter

security filter A type of query that an administrator uses to limit access to data in the BIRT Analytics system.

Related terms
BIRT Analytics, BIRT Analytics Administration, group, NetScaler Web Logging (NSWL) query

segment A segment is a group of records sharing at least one common characteristic.

Related terms
record, selection

selection A selection is a user-specified request that returns a segment from a database.

Related terms
record, segment

skewness A value that reflects the distribution of values in a data set. Skewness values can be positive, zero, or negative. A positive value reflects a data set in which more values lie to the left of the mean value. A negative value reflects a data set in which more values lie to the right of the mean. A zero value indicates values distributed evenly around the mean, typically implying a symmetric distribution.

Related terms
kurtosis, mean

Sphere A data visualization gadget that uses a colored sphere shape to display numeric values and boundaries in a range.

Related terms
Canvas, Cylinder, Dial, Funnel, gadget, Label, Meter

SQL (Structured Query Language)

A language used to access and process data in a relational database.

Related term
database

stage A tool that supports defining users as task owners and assigning to each task owner the permissions required to perform tasks. Define a stage to identify part of a campaign.

Related terms
action, campaign, scheduled task, trigger, workflow

standard deviation

The value equal to the positive square root of variance calculated for a data set.

Related term
variance

standard normal distribution

The normal distribution in which the mean is zero and the standard deviation is one.

Related term
normal distribution

Standardize column

A tool for preprocessing data values having a distribution different from a standard normal distribution. Multiple options support value sets distributed closely, clustered, spread, or having many repeated values.

Related term
standard normal distribution

sum

The cumulated sum of all the values in a field.

Related term
sum-of-squares

sum-of-squares

The sum of all of the squared values in a set.

Related term
sum

support

An expression that calculates a ratio measuring how many transactions contain all items in an association rule. The support expression has the following syntax:

```
Support (A,B) = Transactions (A,B)/Total transactions
```

Related term
association rules

T

table

A named set of records in a database.

Related terms
database, record

target filter

A filter that returns a group of records for comparison with an established baseline. For example, use the year 2010 as a target filter for profit, to compare profit earned in 2010 with profit earned in another, baseline year.

Related terms
baseline filter, universal filter

temporal file

A temporary data file generated and stored in the system cache. Using the BIRT Analytics Administration tool, the administrator can remove the accumulated temporal files and records created by an application to optimize performance.

Related term
BIRT Analytics Administration

trigger

A trigger is a time or event that starts a scheduled task.

Related terms
action, campaign, scheduled task, stage, workflow

U

universal filter

A filter that is always applied at a lower resolution level, before changing resolution.

Related terms
baseline filter, target filter

V-Y

value

1 The content of a constant, parameter, symbol, or variable.

2 A specific occurrence of an attribute. For example, blue is a possible value for an attribute color.

Related term
parameter

variance

A value equal to the squared average of the distances between each value and the arithmetic mean.

Related term
mean

Venn analysis

A tool that supports data analysis based on crossing more than two fields. A Venn analysis identifies coincident values in multiple data segments. For example, use a Venn analysis to show how many customers buy the same three products.

Related terms
Bubble analysis, Crosstab analysis, Evolution analysis, Map analysis, Pareto analysis, Profile analysis, segment

W

workflow

A role responsible for completing tasks or stages in a campaign.

Related terms
action, campaign, scheduled task, stage, trigger

Z

z-score

A value describing whether a quantifiable difference between two groups is statistically significant.

Related term
Profile analysis

Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z