ACTUATE.
The BIRT Company™

BIRT Analytics 4.4
Technical Summary of New Features

The BIRT Analytics 4.4 release is focused on the addition of new algorithms and functionalities. With BA 4.4, Actuate has completed a tool box of data mining techniques fast and easy to use. The addition of new functions and instructions to the front-end and the loader module increases productivity for both, administrators and end-users.

Notice

The information in this white paper is proprietary to Actuate Corporation ("Actuate") and may not be used in any form without the prior consent of Actuate.

© 2014 by Actuate Corporation. All rights reserved.

Version 2 – July 2014

Table of Contents

Introduction

BIRT Analytics provides fast, free-form visual data mining and predictive analytics. The uniqueness of BIRT Analytics results from a combination of the ease of use of data discovery tools with the power and sophisticated analytic products typically reserved for data scientists.

A focus of this release has been to increase the data mining techniques pre-built into Fast DB engine. These new algorithms provide complementary ways to analyze data and create new predictive models.

Actuate has also made several other technology/architecture improvements and functional enhancements.

A summary list of new BA 4.4 improvements appears below:

- New pre-process tools and data mining algorithms
- Functional enhancements added to Loader
- Architecture and technology

New Pre-process Tools and Data Mining Algorithms

New Pre-processing Tool – DICHOTOMOUS

This new pre-processing tool allows generating Boolean columns, which only has 0 or 1 values, ("false" or "true" like *dummy* columns) from a categorical column. Only providing the column, BA 4.4 generates a column for each distinct categorical value.

For example, if we have a "Gender" column with two values: "Female" and "Male", BA 4.4 will create two new columns called "Gender EQ Female" and "Gender EQ Male". In each case that the Gender value is "Female", in the new column "Gender EQ Female" will appear a 1, the rest will be 0.

This tool is quite useful in the data preparation for a logistic regression.

New Advanced Analysis – CORRELATION

This new algorithm allows calculating the dependence relationship between two or more continuous sets of data. This calculation is done using the Pearson's correlation coefficient. It requires the following parameters:

- *Domain*: Data set used in the calculations and could be any type of segment.
- *Independent variables*: Related with the Domain, BA 4.4 provides all the continuous columns that could be added to the analysis.

BA 4.4 shows the results in a correlation matrix of the analyzed variables. Pearson correlation values goes from +1 in direct linear relationship to -1 inverse linear relationship. If this coefficient has a value close to 0 that means those variables are uncorrelated.

A quite known example of dependence is the correlation between the price of certain product and its demand.

To learn more about correlation coefficient: http://en.wikipedia.org/wiki/Correlation_and_dependence

New Advanced Analysis – LINEAR REGRESSION

This data mining technique provides a function (in this case lineal) that models the relationship between a continuous dependent variable and one or several independent variables. The pre-built algorithm in BA 4.4 to build this multivariate linear regression is ordinary least squares (OLS) estimator. This is the most common and straightforward algorithm.

It requires the following parameters:

- *Domain*: Data set used in the calculations and could be any type of segment.
- *Dependant variable*: The column that is going to be modeled.
- *Independent variables*: Related with the Domain, BA 4.4 provides all the continuous columns that could be added to the analysis.

BA 4.4 provides two tabs after calculation. In the first one, *Results*, shows the function equation of the multi-linear regression and a basic evaluation of the goodness of fit of the statistical model (based in R squared coefficient). The second one, *Statistics*, shows multiple tests calculated to evaluate in detail the goodness of fit. They are:

- R squared or coefficient of determination test.

- Adjusted R squared test.

- Standard error global or for each independent variable.

- Upper and Lower confidence level for each independent variable.

- Student's t test global and for each independent variable and its p-value.

A classic example of lineal regression is the analysis of the effect of pricing on consumer behavior.

To learn more about linear regression: http://en.wikipedia.org/wiki/Linear_regression

New Advanced Analysis – LOGISTIC REGRESSION

This data mining model predicts a binary response from a categorical dependent variable based in one or more independent variables also known as predictors. It provides a probability of the occurrence of one of the two possible outcomes from a dependent variable using a logistic function. Maximum likelihood estimation is the pre-built algorithm used in the calculations of the statistical model in BA 4.4

It requires the following parameters:

- *Domain*: Data set used in the calculations and could be any type of segment.
- *Dependant variable*: The column that is going to be modeled that must contain only Boolean values.
- *Independent variables*: Related with the Domain, BA 4.4 provides all the continuous columns that could be added to the analysis.

BA 4.4 provides two tabs after calculation as in the linear regression. In the first one, *Results*, shows the function equation of the multi-linear regression and a basic evaluation of the goodness of fit of the statistical model (in that case is used the Pearson's chi-squared test) . The second one, *Statistics*, shows multiple tests calculated to evaluate in detail the goodness of fit. They are:

- Pearson's chi-squared test and its p-value.

- Likelihood-ratio test.

- Upper and Lower confidence level for each independent variable.

- Odds ratio of each independent variable.

- Log Likelihood-ratio p-value of each independent variable.

- Significance level of each independent variable.

A common example of use of a logistic regression is filling null values of a Gender column using data as age or income. The statistical model is trained with known customer's gender and additional data (predictors), and then apply this model in the null values in gender in order to set if they are male or female.

To learn more about linear regression: http://en.wikipedia.org/wiki/Logistic_regression

Functional Enhancements Added To Loader

New Function - INDEX

The INDEX function creates the indexes of a column of a certain table. This function is useful when the project is appending in a table data coming from distinct sources or files and Loader is told to not index the columns during this process. After all data is loaded, calling this new function is it possible to index the columns. If it's not done in this way, Loader will index in each loading step, taking more time than doing it at the end. Indexing has some of the highest computational costs of all loading processes.

New Function - DROPAUTOLINKS

The DROPAUTOLINKS function delete all transient links in the repository.

New Function - DROPDERIVEDCOLUMNS

The DROPDERIVEDCOLUMNS function delete all derived columns from a table.

New Expression Function ROUND

BA 4.4 Loader had added a new ROUND function in Expression transformation that returns a number rounded to a value without decimals.

This upgrade had been also added to the Frontend expression builder in the Engineering tab.

Upgraded Expression Functions – MID, LEFT and RIGHT

The length of the strings in the column result from an expression that uses MID, LEFT or RIGHT functions will be defined by the parameter provided in the functions. Previously, the length of the column was the same of the original.

There is only one exception to this rule, in case the parameter is provided through a variable that its value is defined by a certain column, the column result for MID, LEFT or RIGHT will have a length equal to the original column.

This upgrade had been also added to the Frontend expression builder in the Engineering tab.

New Search Functionality in Transform Tab

BA 4.4 Loader had a search box in the transform tab that allows looking for specific words in the list of transform scripts. The lines that contain those words searched are shown in blue. In case we choose to look for the words in the whole project (using the options pick list at the right of the text box), results are shown in a new window, classifying the results by files and by its position in this file (line number).

This new functionality will make easier to maintain and modify any loading project.

New Test Connection Functionality

When adding a new data source in the loading process using an external database connection, it's possible to run a test of this connection in BA 4.4, with the parameters provided, clicking in the "Test Connection" button.

New Log Error Tracking Functionality

BA 4.4 Loader had improved the track of errors and the view of the loading log. In the log viewer, the text of the standard messages is printed in black, but the INFO messages are green and errors are red. Also, in the case of an error message, it shows the stack trace with the line of script where it has been generated the error.

New Sharing Loading Projects Functionality

BA 4.4 Loader had added the ability to share loading projects and data sources between users.

Architecture and Technology

A number of changes were made to update the architecture and technology:

- BA 4.4 installer had been updated and simplified
- Installation folders structure had been modified

The new BA4.4 architecture is shown below.

New installer simplified

The new BA 4.4 installation package is more simple and straight forward than previous releases. The process allows installing distributed implementations in distinct servers with ease.

Folder structure modified

Folder structure in BA 4.4 installation had changed deeply. One of the most important changes done had been the creation of a new subfolder known as "data". This folder contains the elements that could be interesting to maintain in future upgrades or releases. Those elements are repository and backend databases and loading projects.

Upgrading to BIRT Analytics 4.4

Upgrading to BA 4.4 requires a new license key specific to BA4.4.

The key steps in upgrade process include a backup of key files, uninstallation / reinstallation of the product, and a restoration of the backed up files. The user's data repositories can be preserved and will become accessible after the administrator has restored, resynchronized the repository and reassigned permissions. All the data segments created and saved by users are preserved and the loader scripts are preserved as well and will become accessible after the administrator has restored them from the initial backup process.
For more detail see the '*Upgrade from BA4.3 to BA4.4*' document.