Data Discovery Course

Back

Data Discovery Course

TARGIT Data Discovery enables an end-user to consume data from a wide variety of data sources and combine them with other data without knowing about data types and complicated query languages. We call these combinations for cubes.

TARGIT Data Discovery is not meant to be a replacement for a traditional BI solution, but rather be the add-on that gives business analysts and other user types with basic understanding of data structures a tool to instantly deploy company-wide solutions in a collaborative spirit.

The engine of the TARGIT Data Discovery module is a service called *TARGIT Data Service*. The TARGIT Data Service is an in memory engine designed to work with datasets of millions of rows and must be installed on the same server hosting the ANTserver component.

TARGIT Data Service will work with up to two million rows of data out of the box, but with the addition of the *Data Discovery license module* to the TARGIT solution, it can exceed the two million rows limit – limited only by memory.

Running Data Discovery the first time

Requirements

The following requirements must be present for successful installation of the TARGIT Data Discovery module:

- TARGIT Decision Suite 2015 or newer
- A newer browser (IE10 and newer)
- A valid license that includes the Data Discovery module
- .NET Framework 4.5 or newer
- Windows Server 2008 SP2 and newer
- Enabled Windows Features for:
  - WCF Activation over port and http
  - IIS ASP.NET 4.5

Data Discovery Administrator

At least one user needs to be appointed *Data Discovery Administrator*. This is done through the *Rights* settings in the TARGIT Management client.

Add single file

The “Add file” option of TARGIT Data Discovery is the option we would like to exploit as part of this lesson. This is where you can easily add ad-hoc data – typically single Excel files or single CSV files – and instantly use the TARGIT client to analyze data from those files.

Dimensions and Measures folders

When you create cubes with TARGIT Data Discovery, you will notice that all columns of the source file are categorized as either measures or dimensions or both.

Furthermore, these measures and dimensions are arranged into a number of *Display folders* where eg. a measure occur in a *sum* folder, an *avg* folder, a *max* folder etc. Each of these measure folders represent different *aggregation* types. The idea is, that you should be able to pick the proper aggregation type for each measure - eg. the appropriate aggregation type for *Contribution* could be *sum*, while for *Contribution Margin* it could be *avg*.

Updating single files

Files that have been uploaded through the *Add file* option, needs to be uploaded again to reflect updated data. You can add new data to the file or update existing data - but you should not change the structure nor the name of the file.

Data Source Column types

The Data Discovery tool will automatically detect data types of the individual columns in the source file. Only columns containing numeric data will be treated as Measures, while all other columns will be treated as dimensions. Furthermore, date columns will be detected and treated specifically as time dimensions.

However, sometimes you will want to fine tune this detection.

Attribute settings

By default, Data Discovery will produce each measure with an abundance of aggregation types – sum, avg, cnt etc. This is just to ensure that all options are available to the end-user, as Data Discovery is not able to make a qualified decision on e.g. one proper aggregation type. You may however lessen the options by making this qualification yourself.

Order and Member Property

You can apply these two settings to any attribute in your Data Discovery cube.

The *Order* setting is an option to have an attribute sorted by a different attribute. A common example is to display Item names, but to have the list sorted by Item numbers.

The *Member property* setting is an option to display one-to-one correlated data in a crosstab *without* putting additional stress on the server. A common example is to display Adress, Phone number and Email next to a Customer dimension.

Combining data sources

TARGIT Data Discovery may of course also work with multiple, related data sources. The data sources do not need to come from the same source or to be of the same type or same format – in fact, as long as TARGIT Data Discovery is able to read the data, these data can be mashed up to fulfill any analytical needs.

To perform a successful data mashup, you will still need to be able to relate data from different sources to each other. This relation requires common keys across the data sources to be mashed up.

Formats: String operations

TARGIT Data Discovery comes with an extensive library of functions for extracting, modifying, cleansing and enriching data from your Data Sources.

This is something you may often need, especially when working with data through web services, as sometimes the format of these data may initially be very different from what you require.

If you are familiar with the functions in Excel, it will not take you long to learn the more than 100 available functions in TARGIT Data Discovery. The two sets of functions are similar or almost identical.

Formats: Unpivot table

If possible, we would like to have data delivered in the "*rows equals transactions* and *columns equals dimension attributes or measures*" fashion, but sometimes this is not the case.

Especially when working with data coming from Excel sheets, you may experience data in the pivoted format, where dimension values have been added to both axes to form a *grid* of transactions rather than just a *list* of transactions.

Before such data can be useful to TARGIT Data Discovery, we will need to *unpivot* the data.

Sharing cubes and data sources

An important thing to know about the Data Sources and the Cubes you create in TARGIT Data Discovery is, that, by default, the things you create are available only to yourself!

In other words, even though many people within an organization are working on the same central TARGIT Data Discovery installation, their data are by default not shared. We think this is the best way of handling data governance: That, by default, you do not need to fear that the wrong people might accidently be able to see your uploaded data.

Only when you actively *want* to or *need* to share your data with other people within your organization, you can of course do that – and only with the individuals chosen by you.

Example: Combining external data with Data Warehouse data

A very common request among TARGIT users is to be able to mash up external ad-hoc data with data in the Enterprise Data Warehouse – without having to involve the Enterprise Data Warehouse ETL and Data Modelling procedures.

While this, on paper, may seem like a simple task, it is in fact not possible to simply extract all Enterprise Data Warehouse data as a data source for TARGIT Data Discovery.

Instead, you will need to define a specific data extract from the Enterprise Data Warehouse needed for mash up with the external data.

When you create a TARGIT analysis upon data from your Enterprise Data Warehouse, you actually make small well-defined data extracts for each object in the analysis.

In other words, if you define a crosstab in a TARGIT analysis to show the necessary Enterprise Data Warehouse data for your data mash up, you can then use this TARGIT analysis as a data source in TARIT Data Discovery.

Example: Weather data

The *Weather plugin* is just one of many online data sources from where you may extract "big data". Depending on your business, weather may actually have an impact your KPIs. To examine this, to find any correlation between weather data and eg. sales data, simply add Weather data as a data source.

Example: Use R script to merge multiple files into one data source

To execute R scripts it is required that the open source ‘R’ programming language is installed. Go to the R homepage at https://www.r-project.org/ or go directly to one of the mirrored download pages, e.g. http://cran.uib.no/.

R is in fact a powerful statistical programming language that allows you to run scripts for extracting data sets for statistical purposes. The purpose of this lesson is not to teach you details on how to use R – many online resources are available to give you a head start on that.

In this lesson you will simply learn how to work with one specific and useful script that allows you to merge data of multiple files into one data source.