Data Profiling Reports Overview

# Introduction

Detailed data analysis is a key aspect of source system analysis. When done correctly it not only ensures a good design but also well-defined business & ETL rules, data definitions, etc. It is a very time consuming and query intensive task but a lot of it can be automated giving the analyst a head start.

This project is designed to provide a way to view database, table and column statistics and also drill through to view the distribution of the values of individual columns. The reports are formatted to be printable so they can be used in a requirements/data analysis discussion.

# Technical Details

The project consists of a set of SQL Server Reporting Services 2012 reports designed to run in local mode but can also be configured to be deployed on a local or a remote server. The reports connect to tempdb but the queries run dynamically on any database selected. Only the databases the user has access to will be listed to view or analyze. The connection assumes a default SQL Server instance. For a named instance, the connection needs to be edited accordingly.

# Typical usage scenario

In a typical scenario, data from the source system is copied to a SQL Server database. The “Database Report.rdl” is run. Any of the other reports can also be run but the table, column names will have to be supplied by specifying the parameters.

The database report lists all the databases on the server with their size, creation date, owner name, physical storage location. Clicking on a database will run the “Table Report” for that database.

The table report lists all the tables in the database with number of rows, columns, indexes, approx. data size, approx. index size and also lists the existence of various keys. Clicking on a table will run the “Column Report” for that table.

The column report lists the columns with data types, nullability and more importantly the number and percentage of distinct values, Null values, blanks and zeros. This is very important information as it helps identify business and unique keys in the data. Also, the existence of blanks and zeros help identify typical ETL lookup issues and address them.

Clicking on a column name will display a column value distribution chart. The chart is mapped to a 100% scale. High numbers in the 0-10% range mean highly distinct values, while high numbers in the higher ranges mean highly repeating values. The report also lists the top 10 most frequent values in that column.