DATA SOLUTIONS DE SPIRLET

Data profiling extension (2.5) for SAP Sybase PowerDesigner

A DATA SOLUTIONS DE SPIRLET White Paper

Revised October 2015

All information also available on http://www.data-solutions-de-spirlet.com

DATA SOLUTIONS DE SPIRLET

Executive Summary

In September 2013, we release our first version of “Data Profiling Extension for SAP Sybase PowerDesigner ”. The first version was a showcase of services we can provide and not really a product to sell on the market.

Beginning January 2014, we released our second version with

- More comfort in parameters settings
- Extraction in excel
- Capability to populate/transfer retrieved data into logical and conceptual models

We are happy releasing new edition extending capabilities and this white paper goes a bit deeper and discusses the new capabilities of the latest edition of “Data Profiling Extension for SAP Sybase PowerDesigner ”:

We release and validate a new version each time SAP release PowerDesigner, at least.

Why such solution

Initial idea of this product was suggested by several usual questions when analyzing source systems like ERP, CRM or any custom database:

- What is data quality of my sources
- How can I make control on all data profiling parameters
- How can I add control on vertical and horizontal constraints
- How to make analysis results shareable with other teams
- How can I automate analysis
- How to improve data knowledge
- How to determine appropriate data quality rules and cleasing in earlier stages of a project
- How to easily share information with people involved in a project

Making data profiling is essential before working on it in earlier stages of a project. It will help determining data quality and data cleansing rules.

Unfortunately companies continue to process such subject using hours after hours, days or weeks with scrapped or incomplete reports

Data Solutions de Spirlet

DATA SOLUTIONS DE SPIRLET

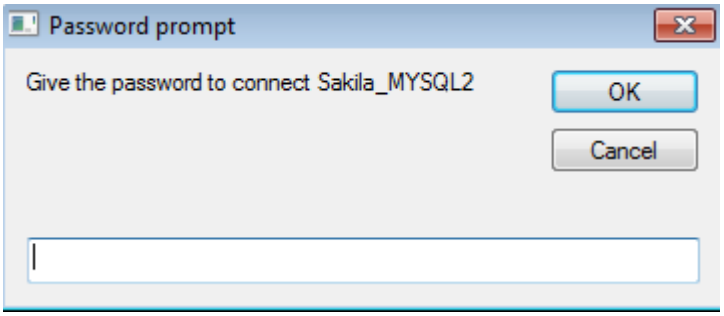
Our solution

« Data Profiling Extension for SAP Sybase PowerDesigner » is an extension model created within PowerDesigner Physical Model Execution space. The product will permit to define:

- (NEW 2.5) Capability to data profile a unique table column or view column.

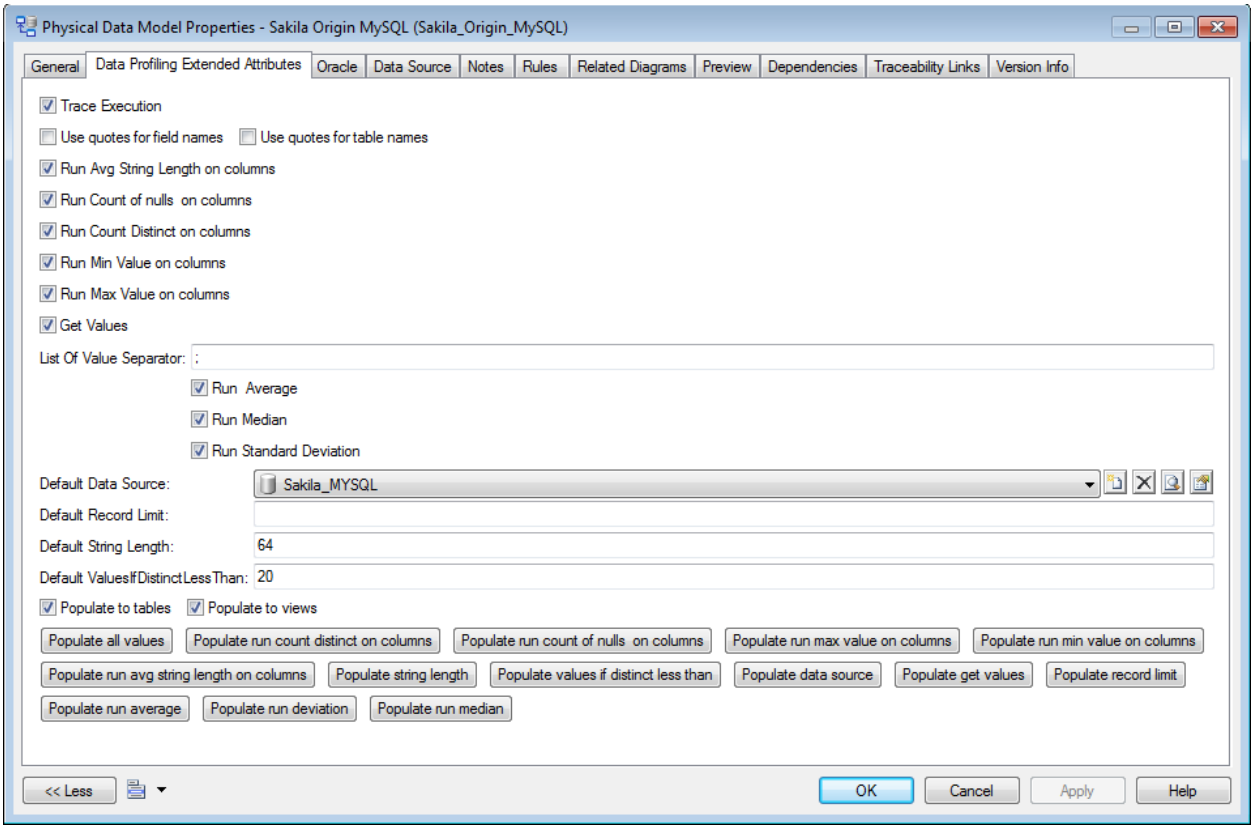
- (NEW 2.5) Average, Standard deviation and Median calculation is possible on database supporting these instructions.
- We added two extended models for logical and conceptual models which will permit to populate information directly in other models. This will permit to share information to all level of your organization.

- Data source identification without security breach (if not stored in ODBC definition)

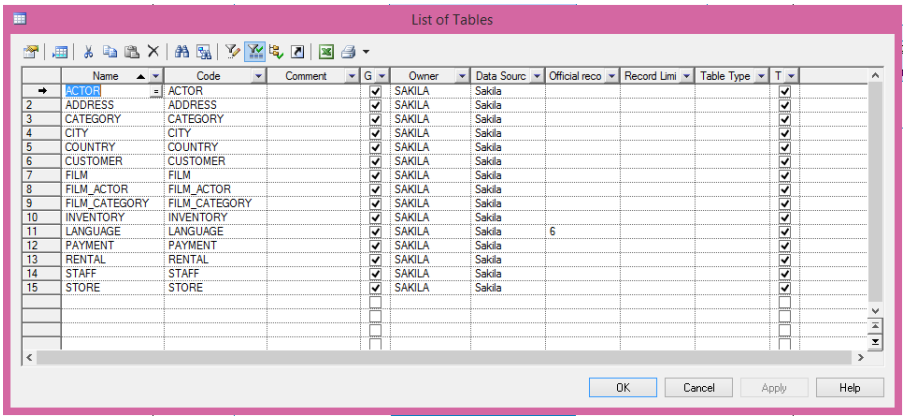


DATA SOLUTIONS DE SPIRLET

- Default parameters for all data model elements and capacity to populate them in data model

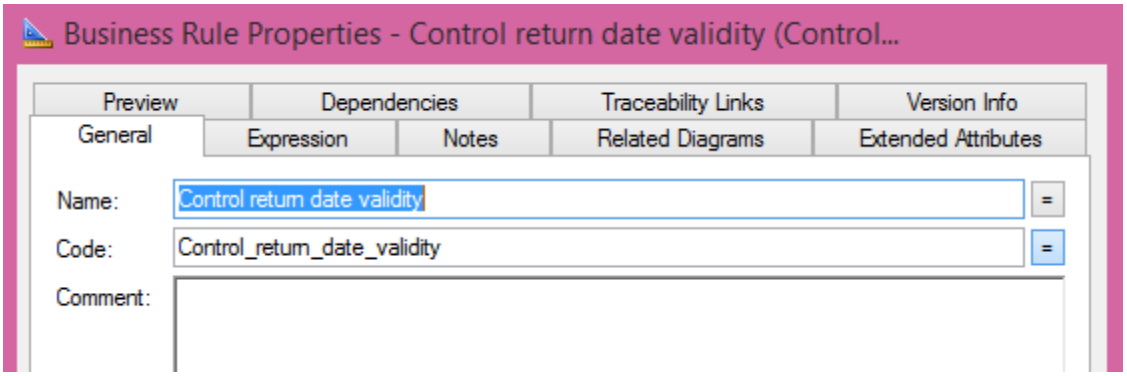


- Selection of tables and views and associated columns to be data profiled



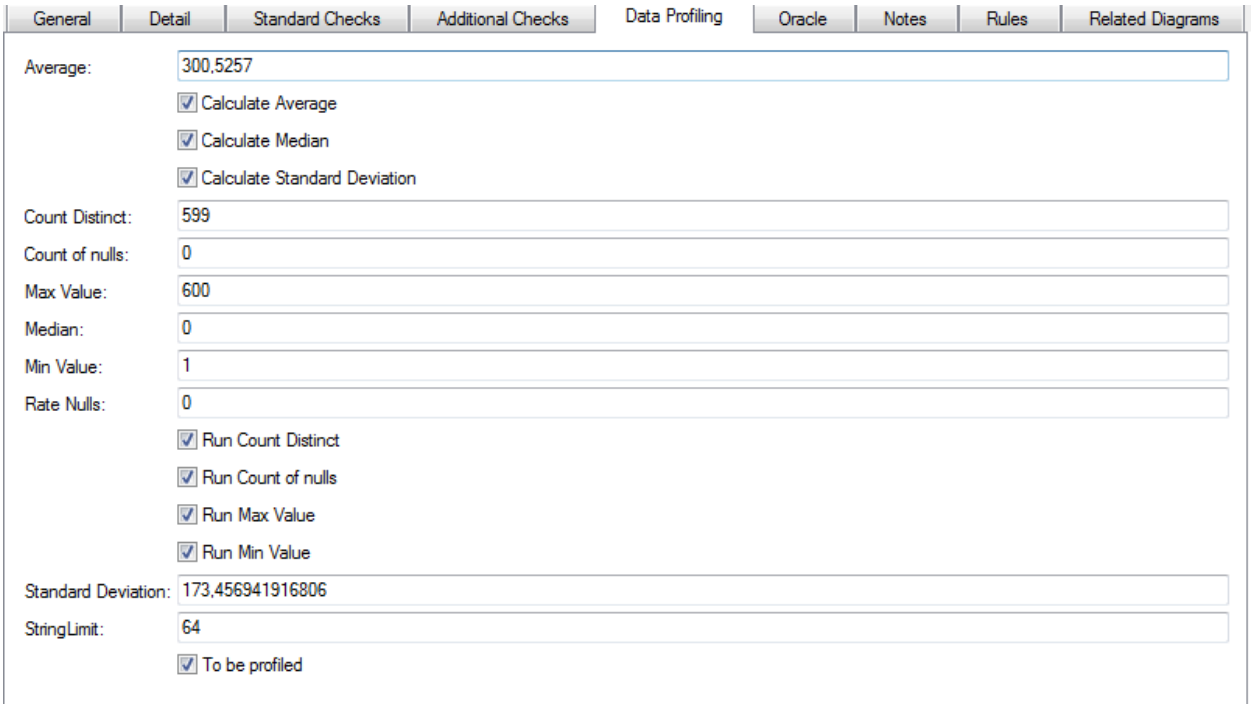
Data Solutions de Spirlet

- Control of horizontal and vertical constraints

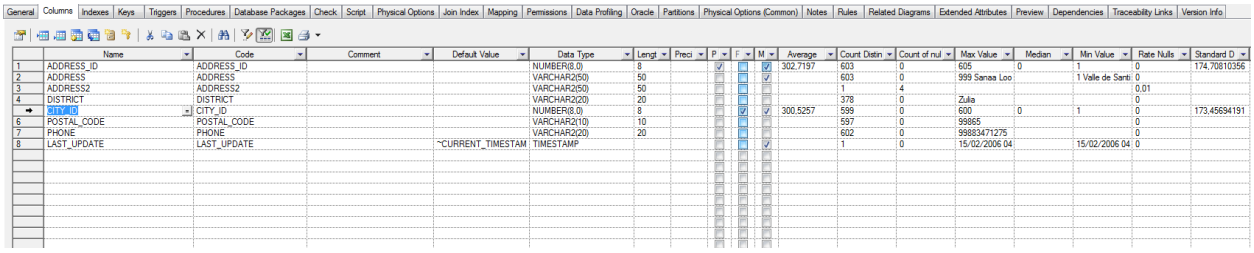


DATA SOLUTIONS DE SPIRLET

- How each column will be profiled:
  - Min value
  - Max value
  - Distinct values
  - Null values (and implicitly rate null values)
  - Limit the number of rows to be processed (sampling)
  - Limit string size to be processed
  - Obtain complete list of values if the count is less than a specified value
  - NEW (2.5):
    - Average
    - Median
    - Standard Deviation



- Final results can be directly stored with the physical data model



(NEW 2.1) Business Rules check are now also exported
Export is enhanced with new attributes

Data Solutions de Spirlet

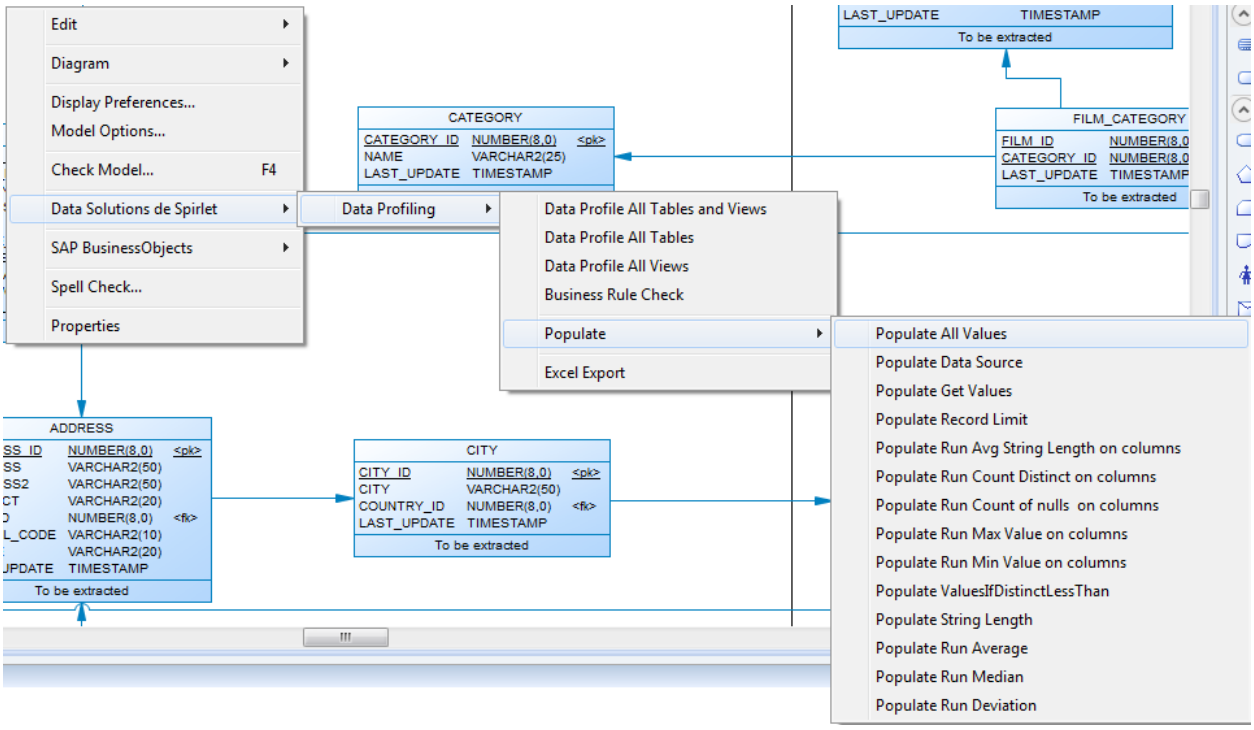
DATA SOLUTIONS DE SPIRLET

- Physical data model, and the extension, can be stored and shared into PowerDesigner Repository
- Impact and Linage capabilities of PowerDesigner are also fully operational and completed by detailed data profiling information

- Details can be automatically exported to prepare report and summary in a more succinct way

Data Source :	Sakila_MYSQL		Column Name	Column Code	Data Type	Length	Precision	String Limit (if applicable)	Primary Key	Foreign Key	Mandatory	Count Distinct	Count Nulls	Min Value	Max Value	Rate null values	Values Average	Median	Standard Deviation
Record Count:	603		ADDRESS_ID	ADDRESS_ID	NUMBER(8,0)	8	0	64	TRUE	FALSE	TRUE	603	0	1	605	0,00%	3.027.197	0	174.708.103.565.173
Record Limit:			ADDRESS	ADDRESS	VARCHAR2(50)	50	0	64	FALSE	FALSE	TRUE	603	0	1 Valle de Santiago Avenue	999 Sanaa Loop	0,00%			
			ADDRESS2	ADDRESS2	VARCHAR2(50)	50	0	64	FALSE	FALSE	FALSE	1	4			0,66%			
			DISTRICT	DISTRICT	VARCHAR2(20)	20	0	64	FALSE	FALSE	FALSE	378	0		Zulia	0,00%			
			CITY_ID	CITY_ID	NUMBER(8,0)	8	0	64	FALSE	TRUE	TRUE	599	0	1	600	0,00%	3.005.257	0	173.456.941.916.806
			POSTAL_CODE	POSTAL_CODE	VARCHAR2(10)	10	0	64	FALSE	FALSE	FALSE	597	0		99865	0,00%			
			PHONE	PHONE	VARCHAR2(20)	20	0	64	FALSE	FALSE	FALSE	602	0		9988347127 5	0,00%			
			LAST_UPDATE	LAST_UPDATE	TIMESTAMP	0	0	64	FALSE	FALSE	TRUE	1	0	15/02/2006 04:45:30	15/02/2006 04:45:30	0,00%			
List of Attached Business Rules																			
Name	Code	Percentage validati	Expression																
Check PostalCodes&Districts	Check_PostalC	0,00%	district is not null and postal_code is null																

- All actions are available by associated context and complete menu



Data Solutions de Spirlet

DATA SOLUTIONS DE SPIRLET

In summary

	Manual working	***Our automated data profiling***
Time to analyse	Slow, need to build separated SQL, measured in **days**	Fast(*), all operations are automated, measured in **minutes** or **hours**
Control of parameters	Very limited	Complete
Capacity to share	Very limited	Complete, included in data models and excel report can be directly shared
Impact and Linage	Not possible	Full, integrated with PowerDesigner
Sampling	Manual process	Integrated

(*) depends on you infrastructure parameters: cpu, memory, network and database servers

Few references

[Ralph Kimball et al. (2008), "The Data Warehouse Lifecycle Toolkit", Second Edition, Wiley Publishing, Inc., ISBN 9780470149775], (p. 297) (p. 376)

[Ralph Kimball (2004), "Kimball Design Tip #59: Surprising Value of Data Profiling", Kimball Group, Number 59, September 14, 2004, (www.rkimball.com/html/designtipsPDF/ KimballDT59 SurprisingValue.pdf)]

Data Solutions de Spirlet