

# 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 cleansing 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



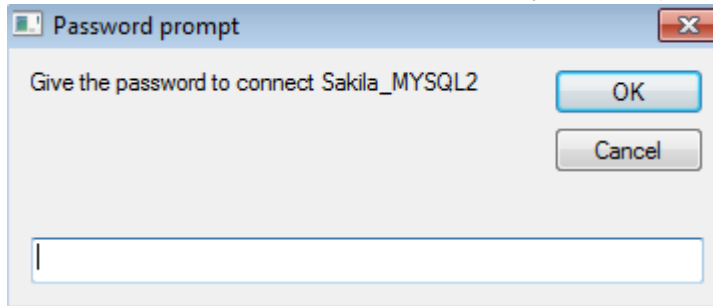
# DATA SOLUTIONS DE SPIRLET

## Our solution

---

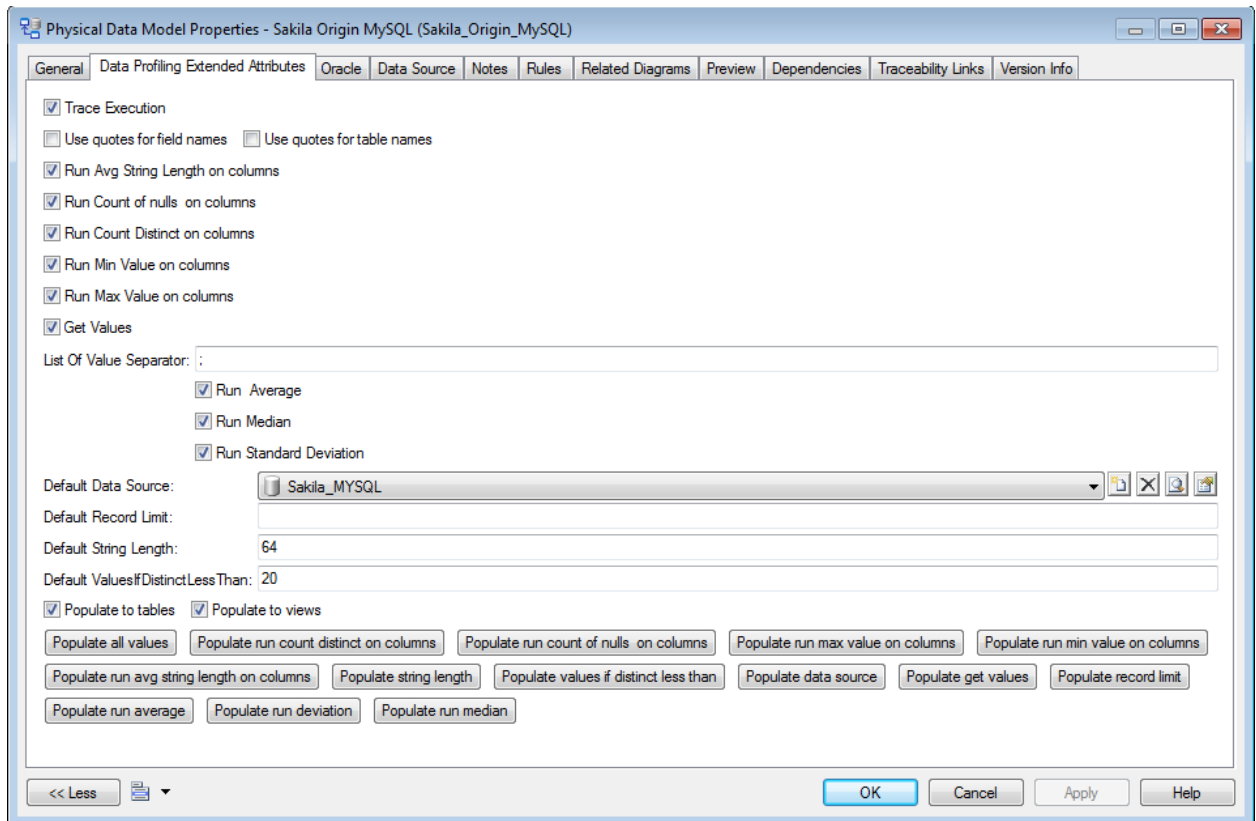
« Data Profiling Extension for SAP Sybase PowerDesigner » is an extension model created within PowerDesigner Physical Model Execution space. The product will permit to define:

- (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)

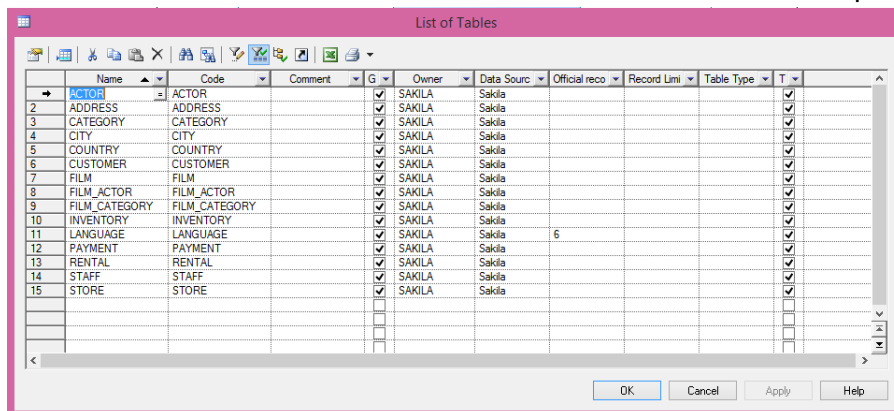


# DATA SOLUTIONS DE SPIRLET

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

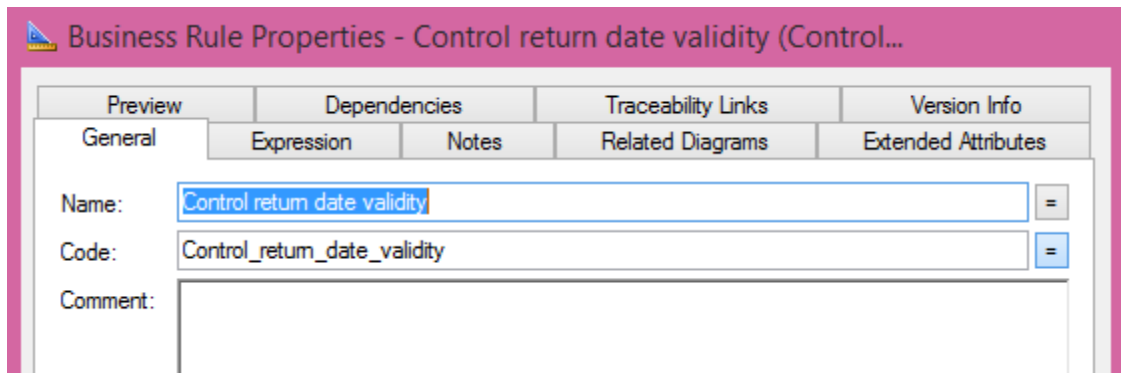


- Selection of tables and views and associated columns to be data profiled



# DATA SOLUTIONS DE SPIRLET

- Control of horizontal and vertical constraints



The image shows a screenshot of a software interface titled "Business Rule Properties - Control return date validity (Control...". The interface has a tabbed structure with the following tabs: "Preview", "Dependencies", "Traceability Links", and "Version Info". Under the "Preview" tab, there are sub-tabs: "General", "Expression", "Notes", "Related Diagrams", and "Extended Attributes". The "General" sub-tab is active, showing the following fields:

Name:	<input type="text" value="Control return date validity"/>	=
Code:	<input type="text" value="Control_return_date_validity"/>	=
Comment:	<input type="text"/>	

# 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

General	Detail	Standard Checks	Additional Checks	Data Profiling	Oracle	Notes	Rules	Related Diagrams
Average:		300,5257						
		<input checked="" type="checkbox"/> Calculate Average						
		<input checked="" type="checkbox"/> Calculate Median						
		<input checked="" type="checkbox"/> Calculate Standard Deviation						
Count Distinct:		599						
Count of nulls:		0						
Max Value:		600						
Median:		0						
Min Value:		1						
Rate Nulls:		0						
		<input checked="" type="checkbox"/> Run Count Distinct						
		<input checked="" type="checkbox"/> Run Count of nulls						
		<input checked="" type="checkbox"/> Run Max Value						
		<input checked="" type="checkbox"/> Run Min Value						
Standard Deviation:		173,456941916806						
StringLimit:		64						
		<input checked="" type="checkbox"/> To be profiled						

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

General	Columns	Indexes	Keys	Triggers	Procedures	Database Packages	Check	Script	Physical Options	Join Index	Mapping	Permissions	Data Profiling	Oracle	Partitions	Physical Options Common	Notes	Rules	Related Diagrams	Extended Attributes	Preview	Dependencies	Traceability Links	Version Info			
	1	ADDRESS_ID	ADDRESS_ID						NUMBER(8,0)	8			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						302,7197	603	0	605	0	1	0	174,70810356
	2	ADDRESS	ADDRESS						VARCHAR2(50)	50			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						603	0	999	Samae Loo	1	0	0	1
	3	ADDRESS2	ADDRESS2						VARCHAR2(50)	50			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						1	4			1	0	0,01	1
	4	DISTRICT	DISTRICT						VARCHAR2(20)	20			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						378	0	Zula	0	1	0	0	173,45694191
	5	CITY	CITY_ID						NUMBER(8,0)	8			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						599	0	600	0	1	0	0	173,45694191
	6	POSTAL_CODE	POSTAL_CODE						VARCHAR2(10)	10			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						597	0	99865	0	1	0	0	173,45694191
	7	PHONE	PHONE						VARCHAR2(20)	20			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						602	0	99883471275	0	1	0	0	173,45694191
	8	LAST_UPDATE	LAST_UPDATE					~CURRENT_TIMESTAMP	TIMESTAMP				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>						1	0	15/02/2006 04	0	1	0	0	173,45694191

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



# 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 (If applicable)	Foreign Key	Mandatory	Count	Distinct	Count Nulls	Min Value	Max Value	Rate null 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%	9.027.957	0	174.708.103.965.173
Record Limit:		ADDRESS	ADDRESS	VARCHAR2(50)	50	0		64	FALSE	FALSE	TRUE	603	0	0	1 Valle de Santiago Avenue 999 Seneca 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			9988471275	0.00%			
		LAST_UPDATE	LAST_UPDATE	TIMESTAMP	0	0		64	FALSE	FALSE	TRUE	1	1	0 15/02/2006 04:45:30	15/02/2006 04:45:30	0.00%				

- All actions are available by associated context and complete menu

The screenshot displays a PowerDesigner data model with the following tables and their attributes:

- ADDRESS**: SS\_ID (NUMBER(8,0) <pk>), SS (VARCHAR2(50)), SS2 (VARCHAR2(50)), CT (VARCHAR2(20)), D (NUMBER(8,0) <fk>), L\_CODE (VARCHAR2(10)), UPDATE (VARCHAR2(20)), UPDATE (TIMESTAMP). Marked as "To be extracted".
- CATEGORY**: CATEGORY\_ID (NUMBER(8,0) <pk>), NAME (VARCHAR2(25)), LAST\_UPDATE (TIMESTAMP). Marked as "To be extracted".
- CITY**: CITY\_ID (NUMBER(8,0) <pk>), CITY (VARCHAR2(50)), COUNTRY\_ID (NUMBER(8,0) <fk>), LAST\_UPDATE (TIMESTAMP). Marked as "To be extracted".
- FILM\_CATEGORY**: FILM\_ID (NUMBER(8,0)), CATEGORY\_ID (NUMBER(8,0)), LAST\_UPDATE (TIMESTAMP). Marked as "To be extracted".

The context menu for the CATEGORY table includes the following options:

- Edit
- Diagram
- Display Preferences...
- Model Options...
- Check Model... (F4)
- Data Solutions de Spirlet
  - Data Profiling
    - Data Profile All Tables and Views
    - Data Profile All Tables
    - Data Profile All Views
    - Business Rule Check
  - Populate
    - Populate All Values
    - Populate Data Source
    - Populate Get Values
    - Populate Record Limit
    - Populate Run Avg String Length on columns
    - Populate Run Count Distinct on columns
    - Populate Run Count of nulls on columns
    - Populate Run Max Value on columns
    - Populate Run Min Value on columns
    - Populate ValuesIfDistinctLessThan
    - Populate String Length
    - Populate Run Average
    - Populate Run Median
    - Populate Run Deviation
  - Excel Export
- SAP BusinessObjects
- Spell Check...
- Properties



# DATA SOLUTIONS DE SPIRLET

## In summary

---

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

<sup>(\*)</sup> 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](http://www.rkimball.com/html/designtipsPDF/KimballDT59_SurprisingValue.pdf))]

