ONE Desktop Workshop

Data Quality Indicator

Prepared for: v15.4.x

Prepared by: Ataccama

Dated: October 2024

Contents of the Document

Introduction

The goal of this lab is to measure the quality of data in the input file by testing their conformance to the business rules defined in the ***Data Quality Indicator (DQI)*** step. This step allows for the running of multiple evaluation rules, delivering records enriched with the results of the evaluation rules. It can also provide a summary of the evaluation in a separate data stream.

Tasks

Create a Plan with DQI Step

Within the following task, you are required to create a plan that will read an input file, evaluate the data from it across several business rules and deliver the data to a specified output. There will be a separate summary report in the additional data stream.

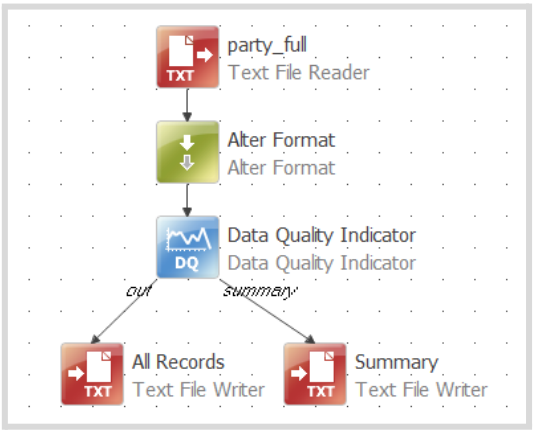
- Create a new plan **data_quality_evaluation.plan** in your **\plans** folder.
- Obtain the **party_full_1.csv** source file and place a **Text File Reader** for it into the plan. Make sure you configure it to read the data properly.
- Add a new **Alter Format** step and use it to create a new column **exp** of type ***STRING***. This attribute will be used to store the explanation codes for the applied business rules.
- Add a **DQI** step into your plan after the **Alter Format**. Configure it as follows:
  - Use the new **exp** column to be the ***Explained Column***.
  - Set the **ALL** for the ***Out Records Filter Type*** option.
  - Create the evaluation ***Rules*** as per the following criteria:
    - **src_birth_date** should not be empty.
    - **src_sin** should be 9 digits long.
    - **meta_last_update** should not be in the future.
  - Create an ***Expression*** and a relevant ***Code*** for each Rule definition.

Try to put together the correct definitions yourself now! Use common sense and construct the logic as per the criteria list. Refer to the end of this document for correct answers if in doubt.

HINT: Use functions **trashNonDigits()** and **now()** for some of the logic.

- Configure the **DQI** step output – create two **Text File Writer** steps and name the output files with corresponding names - **DQ_all_records.csv** and **DQ_summary.csv**.

When finished, your plan should look like this:



- **Run** the plan and observe the results in both files.

The following two outputs will be created:

- The first file **DQ_all_records.csv** contains the original source data and an additional explanation column **exp** corresponding to the violated business rule(s):

4190	Kukla Caron	N/A	12721161			4190	2008-08-04	SIN_LENGTH_INVALID
4191	Tenhoff Doloris		422308932			4191	2008-03-11	BIRTH_DATE_NULL
4192	Cainion Bess	1947-06-24	538563990	4141334191910715		4192	2004-04-12	
4193	Bearded Talia	2003-11-28	42892729			4193	2008-03-07	SIN_LENGTH_INVALID
4194	Thiengtham Adalbe...	1947-12-26	355680109			4194	2008-02-02	
4195	Nga Hornbuckle	1962-04-10	849979000	4141334191957245		4195	2004-03-03	
4196	Simich Sulema	1976-04-24	9482801			4196	2008-11-11	SIN_LENGTH_INVALID
4197	Carley Yeargin	1945-03-12	527770127	4141334191910665		4197	2004-10-10	
4198	Raymundo Fink		861079994	4141334192363955		4198	2004-05-01	BIRTH_DATE_NULL
4199	Terrilyn Pillard	1955-05-28	97523393			4199	2008-07-03	SIN_LENGTH_INVALID
4200	Karma Daub		92009109			4200	2008-06-02	BIRTH_DATE_NULL SIN_LENGTH_INVALID
4201	Andria Lindstedt	1981-11-22	468234919	5542334201926745		4201	2004-08-04	

- The second file **DQ_summary.csv** contains a summary of the information entered in the DQI step. Notice that the logic here is reversed as to what's defined in the ***Expression*** field:

	Name	Code	Short_Descri...	Descrip...	Expression	Success_Count	Total
1	Empty Birth Date	BIRTH_DATE_NULL			src_birth_date is no...	124341	139317
2	SIN Length	SIN_LENGTH_INV...			length(trashNonDigi...	125844	139317
3	Meta Update In Fu...	META_IN_FUTURE			meta_last_update...	139317	139317

Conclusion

We have come to the end of this workshop. We have used a Data Quality Indicator step to evaluate the data and create output with both results and a summary of the applied rules.

Continue to experiment with more rules applied or with new data samples that are provided in the sample source files.

Correct answers, hints, and useful tips

Here is how the **Data Quality Indicator** step could be configured.

Notice that the **meta_last_update** column values are processed as **DATE** format so it can be compared to the result of the **now()** or **today()** function. Make sure you read the values from it in a matching format or perform a data type conversion within the rule definition.