ONE Desktop Workshop

Scoring & Explanations

Prepared for:	v15.4.x
Prepared by:	Ataccama
Dated:	October 2024

Contents of this Document

1. Introduction 3

2. Tasks 3

   2.1. Create a plan and add a data source 3

   2.2. Add Alter Format and Simple Scoring 3

   2.3. Configure the Simple Scoring step 4

3. Conclusion 4

Correct answers, hints, and useful tips 5

1. Introduction

Assigning a score value for a record can be one of the ways to distinguish good records from bad ones. The scoring element can be imagined as 'penalty points' assigned for failing certain rules, validity checks, or given conditions.

This workshop will explain data quality by using the **Simple scoring** step, where business rules and corresponding scores are defined.

2. Tasks

Let's revisit our business rules that were defined in our Profiling workshop (**ONE Desktop Workshop – Data Profiling**). We now have some slightly updated rules:

- **src_sin** should be exactly 9 digits, should be numbers only, and cannot start with a zero.
- **meta_last_update** must not be in the future.

We will expand on this and define these scoring rules:

- For all attributes, if they pass the business rule, give a score of 0 and an explanation of "SIN_VALID" or "META_LAST_UPDATE_VALID" as appropriate.
- if **src_sin** is null, give a score of 10,000,000 and the explanation "SIN_NULL."
- If src_sin is longer than 9 characters, but after removing non-numbers is 9 characters, give a score of 200 and the explanation "SIN_CHARS."
- if **src_sin** after converting to a number is less than 9 digits, give a score of 100,000 and the explanation "SIN_TOO_SHORT". (This includes cases with beginning zeros)
- If meta_last_update is in the future, give a score of 9,000,000 and the explanation "META_LAST_UPDATE_IN_FUTURE."

2.1. Create a plan and add a data source.

› In the **plans** folder, create a new plan called **04_party_scoring.plan**

› Add **party_full_1.csv** into the plan.

2.2. Add Alter Format and Simple Scoring

- Find the Alter **format** step and add it to the plan.
- Connect the '***out***' endpoint of the **Text File Reader** to the '***in***' endpoint of the **Alter Format** step.
- Open the **Alter Format** step and add 2 new columns: **sco_default** (INTEGER) and **exp_default** (STRING). Leave the expressions blank.
- Add a **Simple Scoring** step and connect it to the data flow

2.3. Configure the Simple Scoring step.

After having gone through the previous workshops, you should be able to work out how to fill in the **Simple Scoring** step.

*Have a try and work out how to fill in the **Simple Scoring** step based on the requirements at the beginning of the chapter.*

Consult the previous workshops for tips on expressions.

*Don't forget the **CTRL+Space** content assist and the F1 help on expressions.*

You can also use the expression debug function and templates. Good luck!

When finished, end the flow with a text file writer, name the file **party_full_1_score.csv**, and store it on **data\out**. Open the file and check the results.

3. Conclusion

We have come to the end of this workshop, where we have performed some simple scoring on our data. The lower the score, the better the quality of the data.

Looking at the data, can you think of more data quality checks to be put into place?

Correct answers, hints, and useful tips

Here is how the **Simple Scoring** step should be configured: