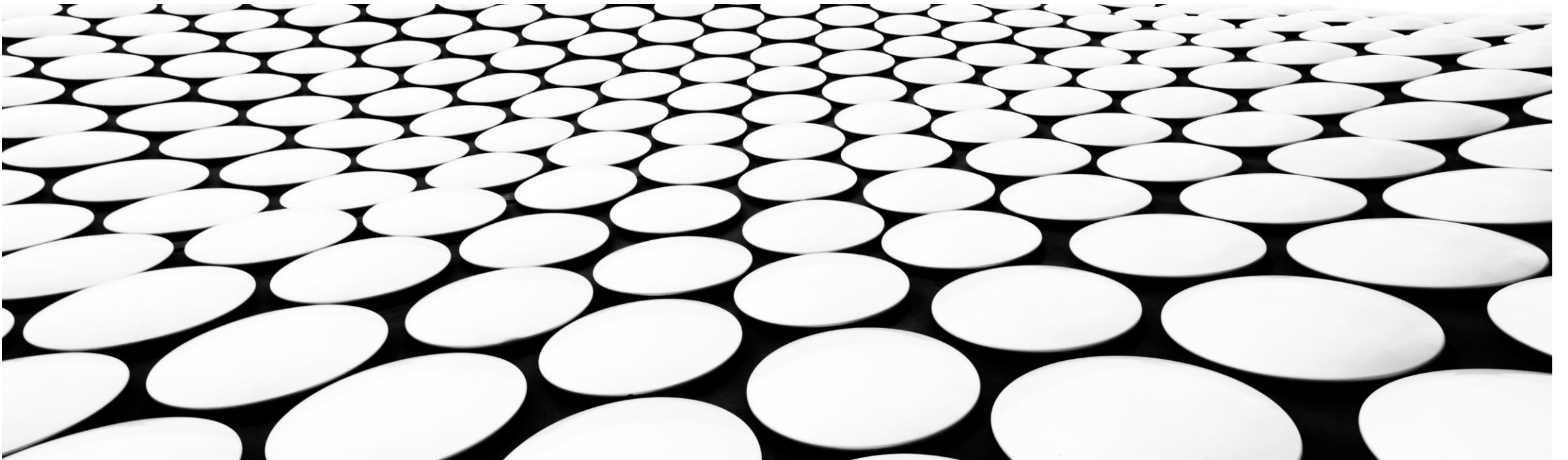


---

# INTRODUCING **DATA PROFILING** AND **RULE-BASED DATA QUALITY CHECK**

PREPARED BY HARRY WIN





# WHY DO WE NEED DATA PROFILING?

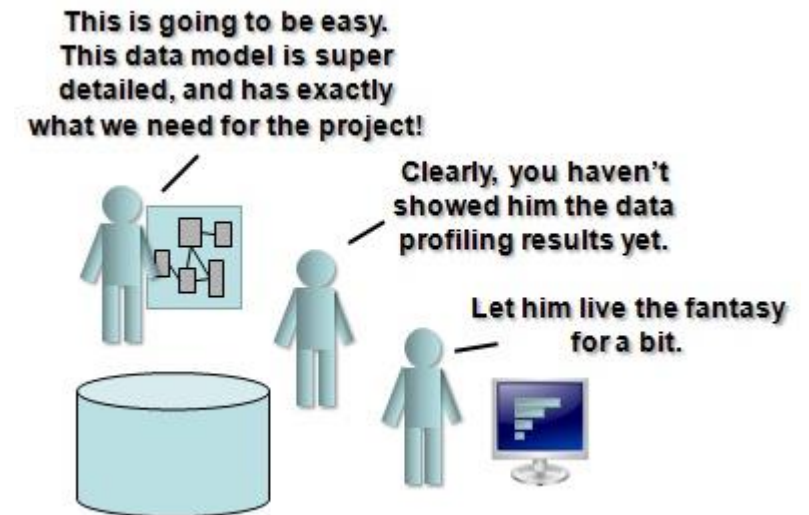
**Problem Statement:** Lack of “trust in the data” due to poor data quality leads to reduced or discontinued BI usage among Information consumers.

## How do we tackle this?

1. Data Profiling
2. Rule-based testing

# WHAT IS DATA PROFILING?

- Data Profiling helps you answer the following questions.
  - Is the data complete? Are there blank or null values?
  - Is the data unique? How many distinct values are there?
  - Is there an anomalous patterns there? What is the distribution of patterns?
  - What range of values exist? Are there any extreme values?



# SAMPLE PROFILE REPORTS

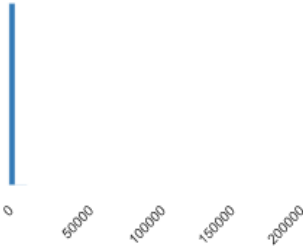
cashpaid

Real number ( $\mathbb{R}_{\geq 0}$ )

SKEWED

ZEROS

Distinct	7818	Mean	41.90946253
Distinct (%)	2.3%	Minimum	0
Missing	0	Maximum	215106.62
Missing (%)	0.0%	Zeros	216024
Infinite	0	Zeros (%)	62.8%
Infinite (%)	0.0%	Memory size	2.6 MiB



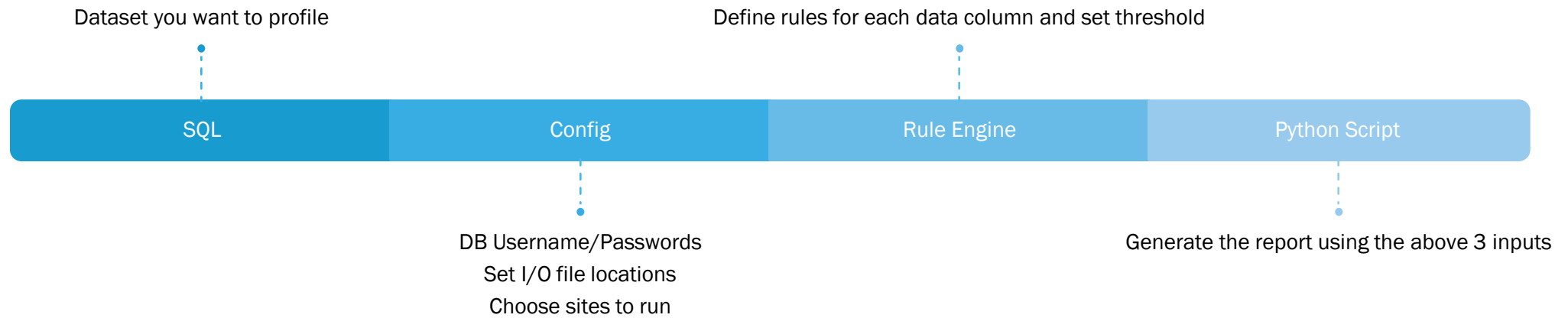
Toggle details

Individual site profile report

Master Report with  
test outcomes

DataField	Quality Check Result	Number of rows that Pass	Number of rows that Fail	Pass Rate	Pass Threshold	site	region
CASHPAID__CHECK	PASS	343704	332	99.90349847	95	BAYSTATE	New England
IS_VALID_ISNOTNULL_CHECK	PASS	344036	0	100	100	BAYSTATE	New England
RXSTATUS_ISNOTNULL_CHECK	PASS	344036	0	100	100	BAYSTATE	New England
JULIANTOCALDATE(FILLDATE::INTEGER)__CHECK	FAIL	0	16478	0	100	CNE	New England
RXSTATUS_RECOGNIZEDSTATUS_CHECK	PASS	344036	0	100	100	BAYSTATE	New England
RXNBR_ISNOTNULL_CHECK	PASS	580913	0	100	100	BERKSHIRE	New England
RFNBR_ISNOTNULL_CHECK	PASS	580913	0	100	100	BERKSHIRE	New England
QTYLEFT_GREATEROREQUAL0_CHECK	FAIL	0	11431	0	90	COOPER	Mid Atlantic
QTYLEFT_GREATEROREQUAL0_CHECK	PASS	579263	1650	99.71596435	90	BERKSHIRE	New England

# FOUR COMPONENTS



# CONFIG FILE

```
[Input]
# Please enter your username.
user = zwin

# Please enter your password.
password = |

# Please enter the sql file path. (For example, B:\BA Shared Docs\Utility Tool\multi sites SQL run selfserve\test.sql
input_sql = B:\BA Repository\HW\EDW - Data Quality Check\prescription fact\prescription_fact_ttryerx.sql

# Please enter the folder you would like to write an output to. (For example, C:\Users\zwin\Desktop)
out_folder = B:\BA Repository\HW\EDW - Data Quality Check\prescription fact\results

# Please enter the output file name. DON'T include .xlsx or .csv.
out_file = ttry_erx_QA

# If there is any date/time field in your dataset, please enter them here with comma separated. Date/Time fields mess up with report generation process and require special handling.
date_fields = rxdate, status_time, discontinued_date, status_update_date

# Are you running this for all sites? Answer Y/N only.
all_sites = N

# If you answer Y to above question, leave it blank. Otherwise, Please enter site ID separated by space. You can enter more than one site.
site_lists = 201601

# Do you want to create an individual report for each site: Answer Y/N only.
individual = Y

# If you are using custom rule engine, please enter the folder where your custom rules are saved at. Make sure "Custom Rule Engine.xlsx" file is located in that folder.
rule_file = B:\BA Repository\HW\EDW - Data Quality Check\prescription fact\Custom Rule Engine.xlsx
```

# RULE ENGINE TEMPLATE

What you enter in the template:							
RuleName (Optional)	DataFields (Required)	Condition1 (Required)	operator1 (Optional)	Condition2 (Required)	operator2 (Optional)	Condition3 (Optional)	Threshold to Pass (Required)
isnotnull	rxstatus	is not null					100%
validstatus	rxstatus	in (0,1,2,4,8,17,900,999,998)					95%
	upi_length_check						100%
Parsing your input							
,Case When (rxstatus is not null) Then 'Pass' Else 'Fail' End "RXSTATUS_ISNOTNULL_CHECK"							
,Case When (rxstatus in (0,1,2,4,8,17,900,999,998)) Then 'Pass' Else 'Fail' End "RXSTATUS_VALIDSTATUS_CHECK"							

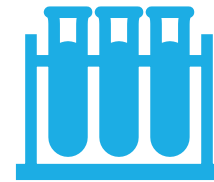
# SETTING CUSTOM RULES FOR ENGINE



1. Make sure you are using the same name as in your query. If you are using alias in your query, use that same name in the template.



2. Make sure you are using sql like syntax for your condition statement



3. You can test more than one rule for the same field. You just need to define the rule name.




4. Set business thresholds for test cases. You can leave condition and operator columns Blank for custom test cases in SQL.




# PRODUCT DEMO



# VISION

- 
- Central Business Rule Engine library – Shared ownership with John/Chris Team
    - Fills, Erx, PA/ FA, Clinical, etc

- 
- Central Profile Repository to share with Stakeholders (Strat-Ops?)
    - I:\Business Analytics\2 - SHS Partner Site Information\4 - Data Profiles

- 
- Cadence
    - Run this tool on incremental dataset from partner hospitals on a weekly/scheduled interval

---

## HIGHLIGHTS

01

Any dataset can be profiled

02

Individual site gets its own detailed profile report.

03

You get one master file to see the outcome of your test cases for all sites.

## QUESTIONS / COMMENTS

