Data quality with Azure Data Factory and Azure Synapse

Christian Cote

WhoAmI

Azure Platform Solution Architect at AveHealth

On-Prem ETL development using various ETL tools: DTS / SSIS, Hummungbird Genio, Informatica, Datastage

DW Experience in various domains: Pharmaceutical, finance, insurance, manufacturing and education

Specialized in Data warehouse/BI/Big Data

Writer of several books on data integration

Microsoft Data Platform Most Valuable Professional (MVP)

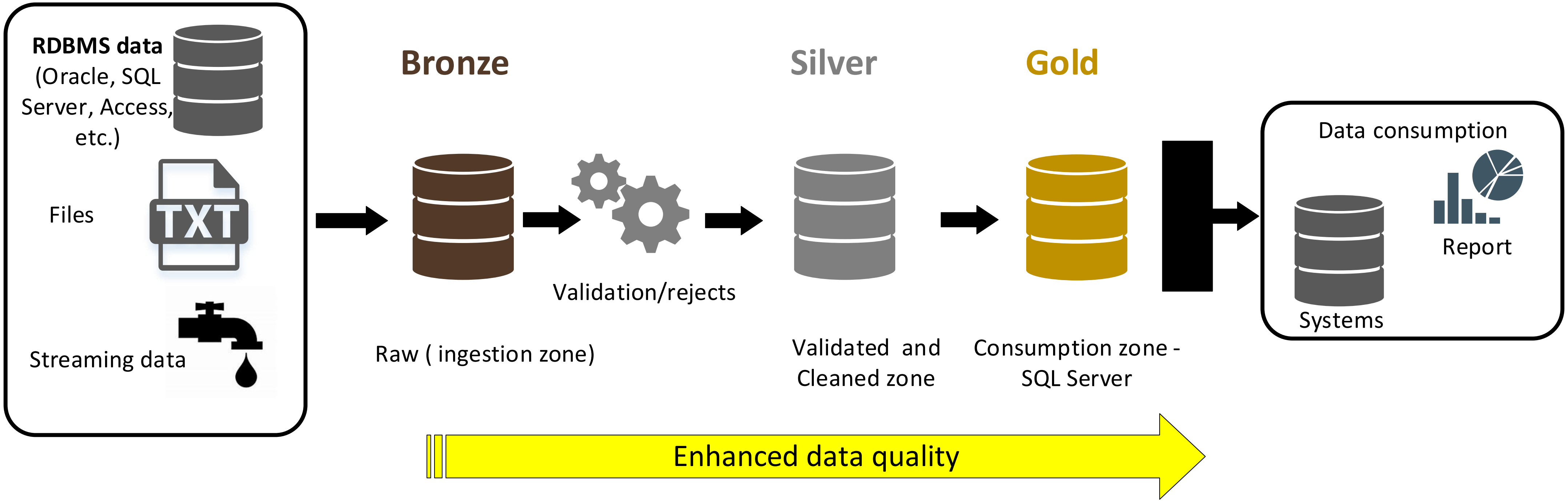
Montreal Data Platform User group leader

Agenda

- Data quality classifications
- Data Profiling

  Azure Synapse Spark

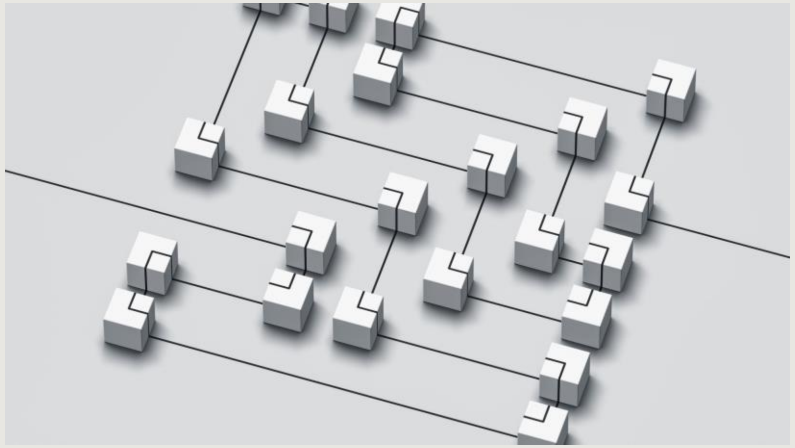
  Data pipelines tools
- Data Factory Pipelines implementation



Data integration overview

Data Quality issues

- They arise when we integrate a source system into another data structure or with other systems.
- Most organizations don’t have a data quality leadership team to tackle and correct the issues.
- They lead to data pipeline failures in production.
- Most of the time, the bad data is found out when we build the reports, way too late.



- **Consistency**: They're no contradictory information between the source systems and our reports.
- **Accuracy**: Information corresponds to the reality in all systems.
- **Completeness**: All bits of information are found in the data. E.g., We have a complete street address: street number and street name are present.
- **Auditability**: We can trace the data back to the source system.
- **Ordered**: The same format is found in every structures in the system. E.g., Date format is always YYYY/MM/DD

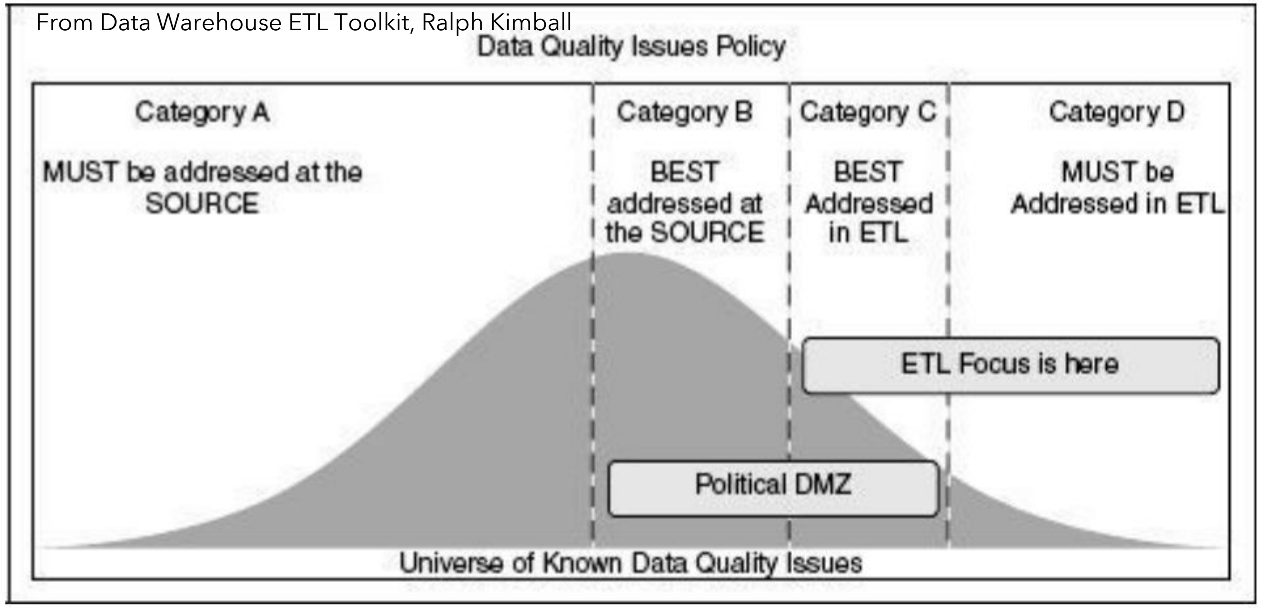
Data Quality classifications

Where should data quality issues be resolved

Zone A: Addressed in source systems Missing information that cannot be derived. Most Data Quality issues fall in this category

Zone B: Addressed in source systems Even if it might be possible to correct it technically. Technology can resolve the issues with acceptable results

Zone C: Best addressed in data integration pipelines. If the issue can be addressed by source systems, we always favor it.



From Data Warehouse ETL Toolkit, Ralph Kimball

Data Quality Issues Policy
Category A
MUST be addressed at the SOURCE
Category B
BEST addressed at the SOURCE
Category C
BEST Addressed in ETL
Category D
MUST be Addressed in ETL
ETL Focus is here
Political DMZ
Universe of Known Data Quality Issues

Zone D: Data Quality issues are better resolved while integrating the data. Third party tools can be used in the data pipelines to correct the issues.

We must be careful to not overcorrect the data as it might lead to more issues than it may resolve.

Data profiling

Data profiling categories

- **It's a mandatory activity!**
- **Find many pitfalls** in the data such as:

  NULL values

  Referential integrity

  Data type, length, supported values

  Duplication of data

  Data and Value profiling
- The data profiling exercise **assesses** the **source system data quality**.

Data profiling benefits

Prevent data pipelines to be refactored because we discovered some issues with the data.

Prevent data pipelines to fail in production because they encountered data they were not programmed for.

Help identifying data quality issues upfront.

Shorten development time for data integration developers

Null Values

They can be misleading for end user's tools.

NULL != NULL: NULL means “I don’t know what the value is”

We should fill them using values like N/A. This helps indexing in the database as by default, NULL values are not indexed.

We can present them as blank values to users using functions like NULLIF or REPLACE.

Referential Integrity (RI)

1:1, Parent child, one-to many relationships

Leads to missing and incomplete values

We need to identify the missing RI values and report it to source system administrator.

Missing RI is usually rejected by Data Integration processes.

Column values

- Dates

  Month/Day inversion like 01/02/2022. Might be Feb 1st 2022 or Jan 2nd 2022

  Oracle to SQL conversion. If we don't use format on Oracle dates, they read like 01-Feb-22
- Invalid columns length: Name with only one character, address without a street number.
- Invalid number: decimal number for an integer column.

Data and Value rules

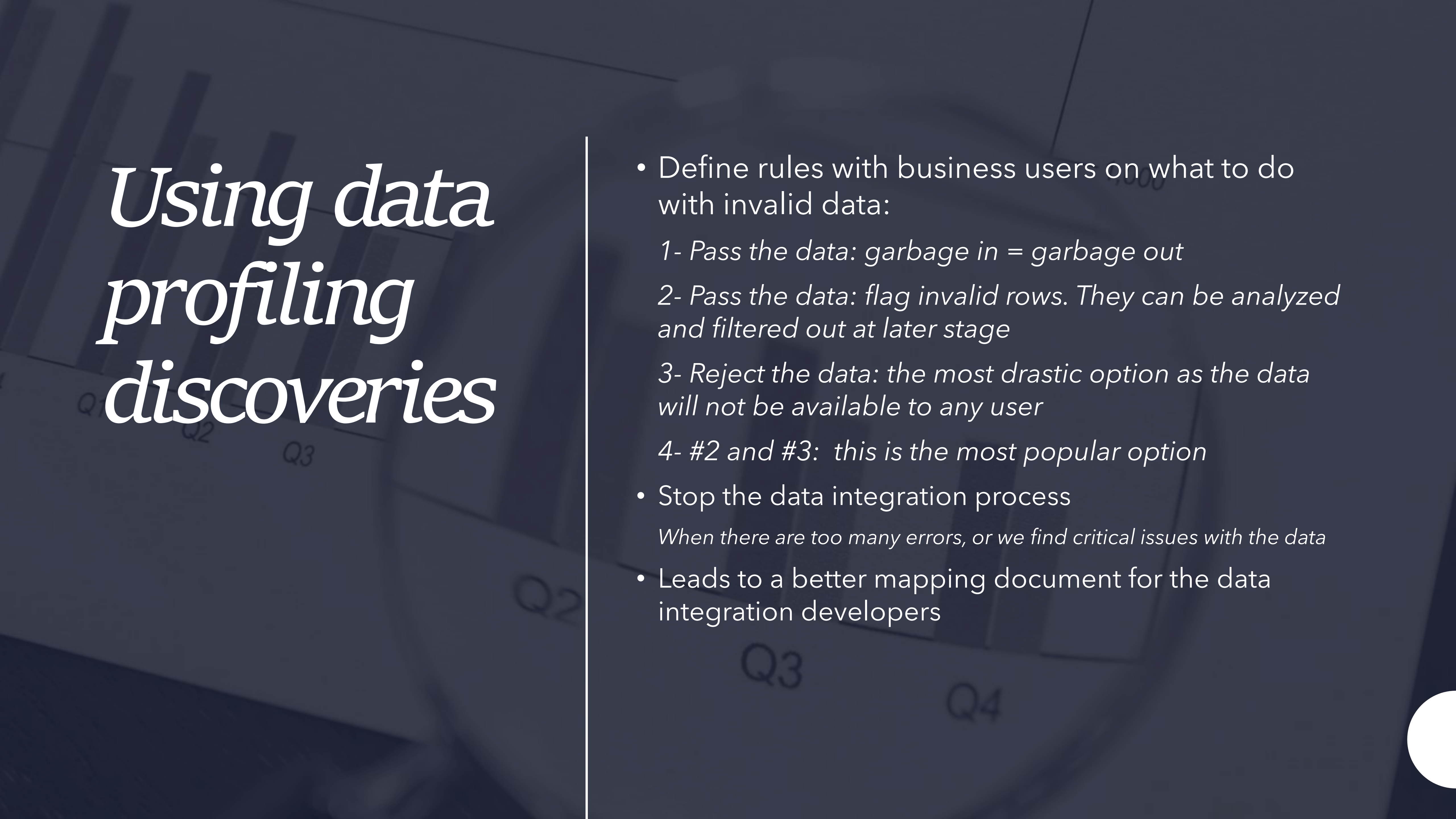
Data rule: evaluated at row level.

E.g., Invalid age: a person age < 0 or greater that 125.

Value rule: evaluated at set of values level (multiple rows).

E.g., February sales are 50% lower than January sales.

E.g., Lots of duplication leads to sales that are too high.



Using data profiling discoveries

- Define rules with business users on what to do with invalid data:

  1- Pass the data: garbage in = garbage out

  2- Pass the data: flag invalid rows. They can be analyzed and filtered out at later stage

  3- Reject the data: the most drastic option as the data will not be available to any user

  4- #2 and #3: this is the most popular option

- Stop the data integration process

  When there are too many errors, or we find critical issues with the data

- Leads to a better mapping document for the data integration developers

Profiling using Spark

- Existing libraries like Pandas Profiling can be used for small data volumes
- Great Expectations
- SQL and Python for larger volumes

Profiling large volume of data

- Most tools and open-source libraries do not support large volumes (more than 1M rows)
- We can use custom profiling methods

  E.g., Aggregate data: Count(*) distinct values per country
- Split data into manageable sets

  E.g., profile the current year of data

Demo: Profiling data with Pipelines and Spark notebook

Microsoft Excel
ro-Enabled Worksl

Demo

Sample mapping document

Data Quality assessment in pipelines

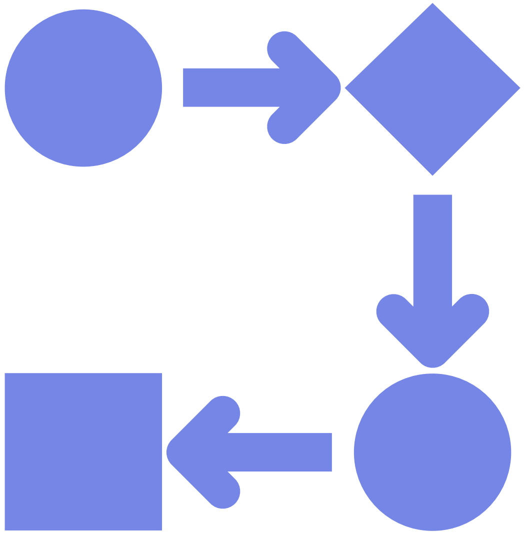
- Assert Transformation pipelines
- Reusable pipelines or flowlet
- Spark Notebook Activity

Demo

Data quality with no-code pipelines

To sum up

- Data Quality is not a one-time off activity, it's a process
- Data profiling is an essential step to assess our data sources
- The pipelines should only react to bad data, that is reject or tagging a rows as faulty.
- We do not create data, only integrate it to other systems or target model.



Useful links

- Azure Synapse documentation
  [Azure Synapse Analytics | Microsoft Azure]
- Pandas profiling
  [pandas-profiling · PyPI]
- Kimball Group dimensional modeling spreadsheet
  http://www.kimballgroup.com/wp-content/uploads/2014/03/Ch02_MDWToolkit_Datamodel_Spreadsheet_4.0_2008R2.zip

Questions?

Thank you!