**Challenge 1: Data Cleaning, Transformations and ETL pipeline architecture**

Data Exploration and Cleaning

Two datasets are provided. Observing based on the CSV file name, one dataset consists of loyalty data, and another dataset consists of transactions data.

Both datasets consist of:

- Name, city, phone number, and email (user information)
- UUID assigned to each record

Loyalty dataset has an additional information license plate, and transactions dataset has information of transaction ID and amount.

At the first glance, the user information in loyalty dataset is somehow "dirty" and require further cleaning. The information in transactions dataset is cleaner than loyalty dataset, presumably can be used to "clean" data for loyalty dataset.

Upon data exploration,

- The UUID assigned is unique to record i.e., no UUID duplicates within dataset
- Two datasets can be joined with UUID – if UUID is randomly generated this wouldn't be possible
- The user information of two datasets is matched after data cleaning

Business Use Case

One business use case that can be provided is the data quality issue of the dataset saved. Data quality issue often arises when the application frontend does not implement data quality checking when submitting forms. Another possible chance are dirty data is saved during data bulk loading.

When data quality issue is resolved, business could utilize the dataset for further use. For example, accurate email and phone number allows business to contact customer for product upselling.

To resolve data quality issue, a data quality report can be presented to business for business to understand how "dirty" the data is, and hence able to identify the root cause of the data error. Once the root cause is identified, developer could fix the bugs or add enhancement in the data loading script to eliminate the data issue.

Pipeline Architecture

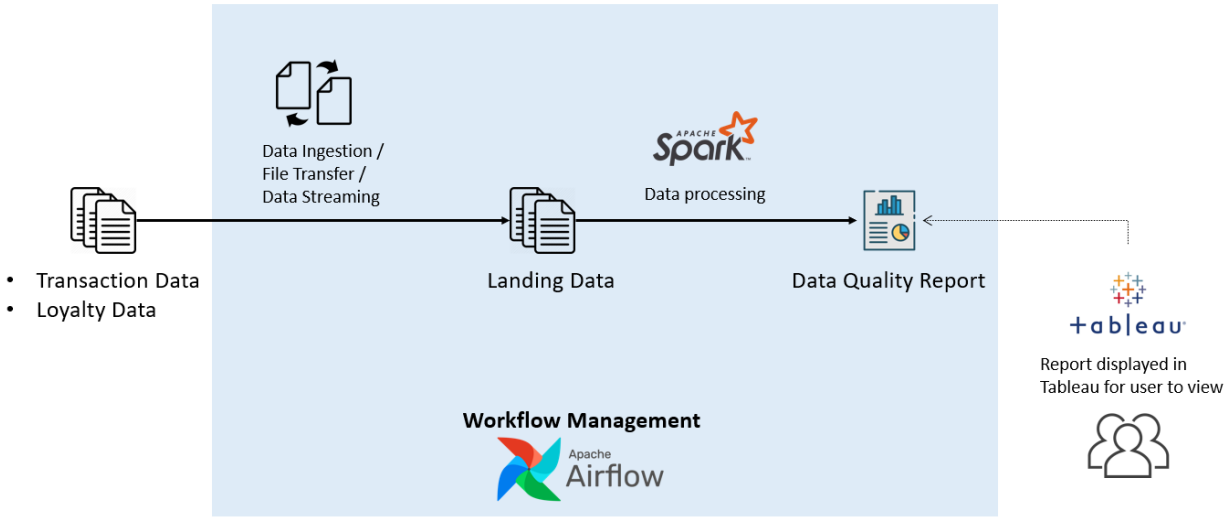


Diagram above shows the high-level architecture of the pipeline proposed:

1. Data is provided by source. Depending on security requirements, we can use external connector to connect to source database directly for data ingestion and data streaming, or source could provide data in the form of raw files for us to process.
2. Spark program is executed to check the data quality of loyalty record based on the latest transaction record. Then the information is populated into a report.
3. Report is then displayed in Tableau for business users to view the data quality report of loyalty data.
4. Process is scheduled using Apache Airflow for data loading and data processing – used for scheduled data quality report refresh.

Validation

Based on the sample data provided, data is “contaminated” with additional strings appended to random part of the original string. If the loyalty data matches with transaction data with additional strings removed, we could say that the data quality issue raised is valid.

**Challenge 2: Customer engagement**

Upon exploring the sample dataset provided, we have identified data quality issues in the loyalty dataset provided. In 10,000 records, there are only 1,152 records without data quality issue. The data quality issue identified are due to additional strings being inserted into loyalty data. For example, the name “ESTHER LEE” in transaction data is saved as “EST121HER LEE” in loyalty data.

To resolve the data quality issues, we will need to identify the root cause where the dirty data is being inserted. Identifying and fixing the root cause would help to reduce the data patching work to resolve the data quality issue “manually”, and preventing the data quality issue from happening again.

Data quality report shall be provided from time to time to ensure data quality.

**Challenge 3: Scrapping, Datasourcing and Enriching data**

Data enrichment is refining the existing data with additional data from different sources, be it from internal or external, for the purpose of using the additional data to derive more insights.

By enriching the cleaned data with US Census public data for city population, we could derive the state (or possible states) from city data, and get insights on the customer base population.