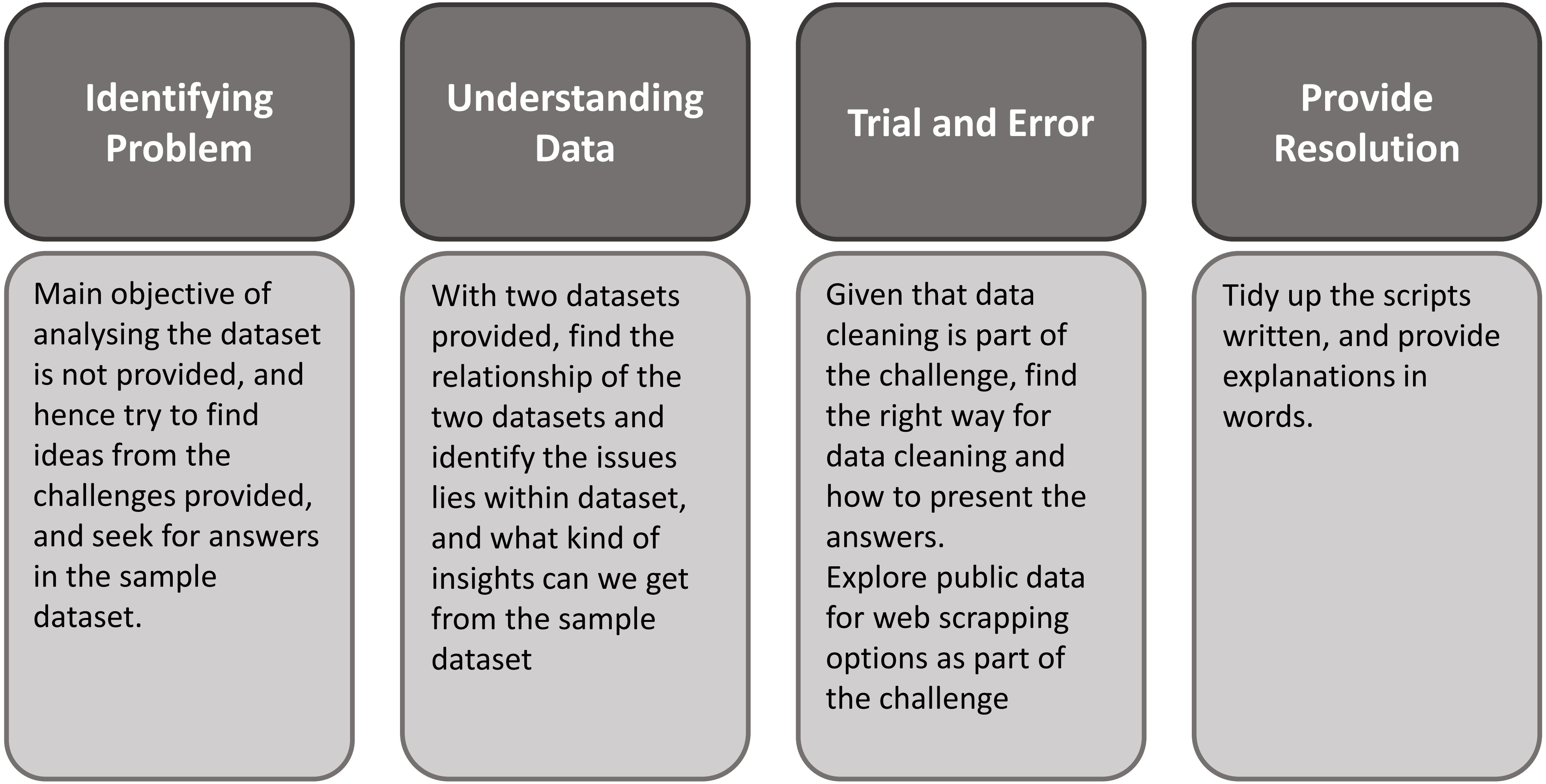
Take Home Test

Data science / analytics / engineering

Chan Xin Jo
25 August 2022

Thought Process



Identifying Problem

Main objective of analysing the dataset is not provided, and hence try to find ideas from the challenges provided, and seek for answers in the sample dataset.

Understanding Data

With two datasets provided, find the relationship of the two datasets and identify the issues lies within dataset, and what kind of insights can we get from the sample dataset

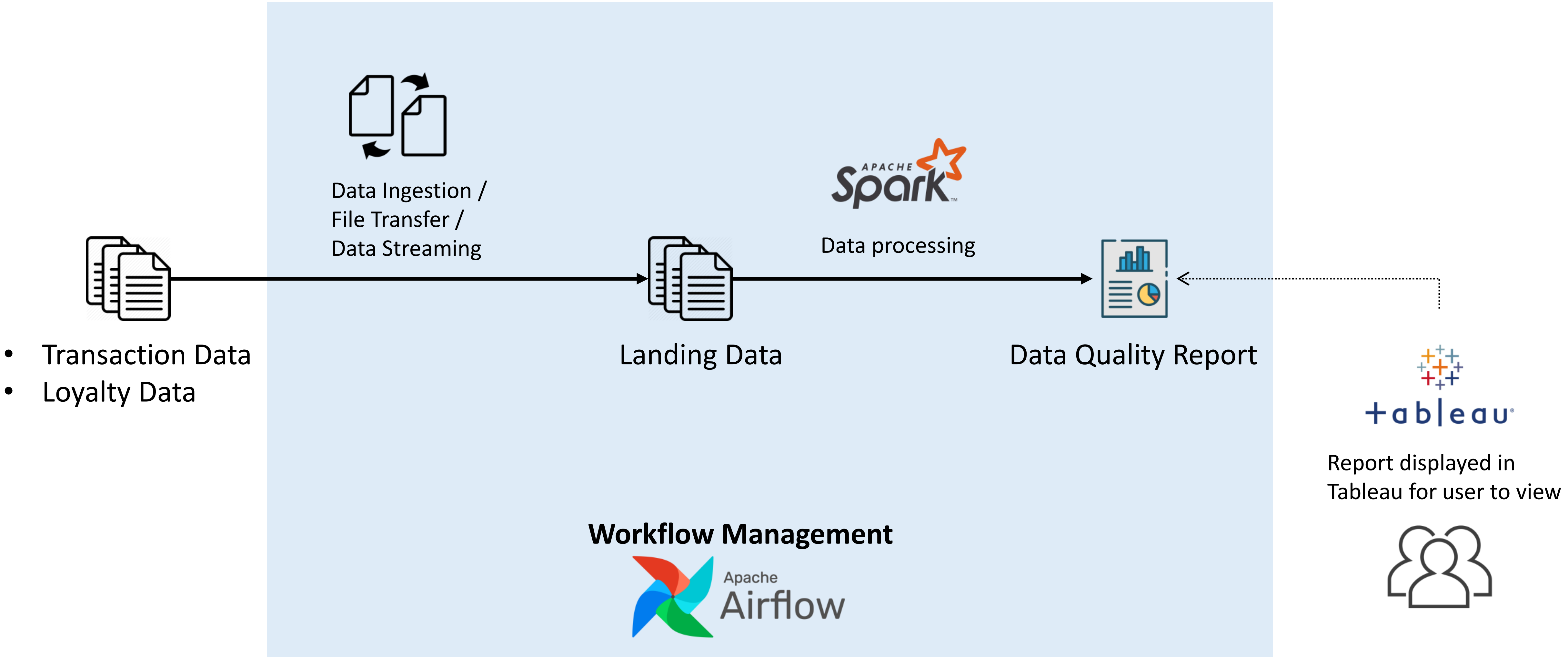
Trial and Error

Given that data cleaning is part of the challenge, find the right way for data cleaning and how to present the answers.
Explore public data for web scrapping options as part of the challenge

Provide Resolution

Tidy up the scripts written, and provide explanations in words.

Pipeline Architecture



- Transaction Data
- Loyalty Data

Data Ingestion / File Transfer / Data Streaming

Landing Data

Data processing

Data Quality Report

Report displayed in Tableau for user to view

Workflow Management

PySpark Scripts Written

test_data_exploration.py

- Explore datasets provided with some data profiling / study
- Join both dataset and check join result
- Highlight errors in the dataset and provide summary

scrape_us_census_city.py

- Read CSV data saved in US Census
- Uses Spark only to save data in local directory

test_data_enrichment.py

- Using data scraped in previous script, join with the transaction (clean) dataset to get state information
- Note: Not able to verify/check my result due to timeout issues that I couldn’t resolve in my local machine