# **My Dataflow Pipeline Documentation**

## **Pipeline Overview**

* **Purpose:** To ingest, clean, and load historical sales and store data from CSV files into BigQuery for analysis and reporting.
* **Source:** CSV files stored in a Google Cloud Storage bucket (gs://mgmt599-assignment2-bucket/kaggle-store-sales/).
* **Transformations:**
  + **Stores Data:** CSV parsing, data type casting, and filtering out malformed rows.
  + **Sales Data:** CSV parsing, data type casting (e.g., date conversion), and filtering out malformed rows.
* **Destination:** BigQuery tables in the dataset assignment\_dataflow under project focal-furnace-465023-g3.

## **Pipeline Configuration**

* Job name: **stores-data-pipeline**, **sales-data-pipeline**
* Region: **us-central1**
* Machine type: **default (DataflowRunner-managed)**
* Max workers: **auto-scaled**

## **Data Flow**

### **Stores Data Pipeline**

1. **Read from:** gs://mgmt599-assignment2-bucket/kaggle-store-sales/stores.csv
2. **Transform:**
   1. Parse each line of CSV.
   2. Convert data types: store\_nbr and cluster as integers.
   3. Filter out rows with missing or invalid values.
3. **Write to:** focal-furnace-465023-g3.assignment\_dataflow.stores\_data

### **Sales Data Pipeline**

1. **Read from:** gs://mgmt599-assignment2-bucket/kaggle-store-sales/train.csv
2. **Transform:**
   1. Parse each line of CSV.
   2. Convert date field to DATE type.
   3. Convert sales to float and onpromotion to integer.
   4. Filter out rows with missing or invalid values.
3. **Write to:** focal-furnace-465023-g3.assignment\_dataflow.sales\_data

## **Lessons Learned**

* **What was challenging?**
  + Managing type conversions, especially ensuring the date field in the sales data was properly parsed.
  + Handling and logging malformed records to avoid pipeline failures.
* **What would you do differently?**
  + Implement more robust data validation with clearer error logging to identify problematic rows.
  + Parameterize file paths and schema definitions to make the pipeline reusable for similar datasets.