Case_Study_4: Data Processing Optimising

Bhargav Jangam

Detailed Code Workflow

First, I created the **enriched DataFrame**, which combines data from train.csv, stores.csv, and features.csv. This DataFrame is stored in GCS in Parquet format, partitioned by store and date. The cached data from stores.csv and features.csv is also broadcasted for later use.

Parallelly, I created dataframes for various metrics:

1. **Store Metrics**, which includes total_sales and average_weekly_sales.
2. **Top-Performing Stores Metrics**.
3. **Department Metrics**, which contains total_sales.
4. **Department Weekly Trend Metrics**, which track weekly trends, such as "increase by 100", "decrease by 100", or "no change".

These metrics are stored in JSON format in GCS and are also cached for future use.

Later, messages are consumed from Kafka and enriched using the previously broadcasted data from stores.csv and features.csv. This enriched data is then merged with the earlier created enriched DataFrame.

The cached metrics are used to update the stored JSON files in GCP. Specifically:

- For **Store Metrics**, the previous count and total sales (from cache) are used to compute the new average, total sales, and the sales report count. From this, the top-performing stores are identified. These updated metrics are then cached for future data.
- For **Department Metrics**, the current totals are added to the previous totals(from cache).
- For **Weekly Trends**, I first created empty columns matching the structure of the previous weekly trends DataFrame. The data is then sorted, and new weekly trends are computed by comparing the weekly sales of the current and previous dates. These updated trends are cached for future use.

Below is the proof of execution and the resulting images showcasing the outcomes of the process.

Execution Proof and Resulting Images

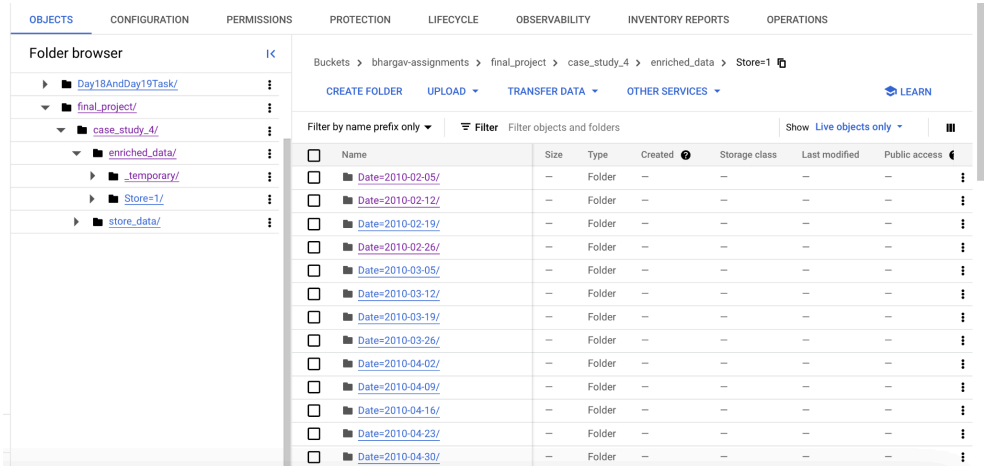
Kafka - Sales Produced messages

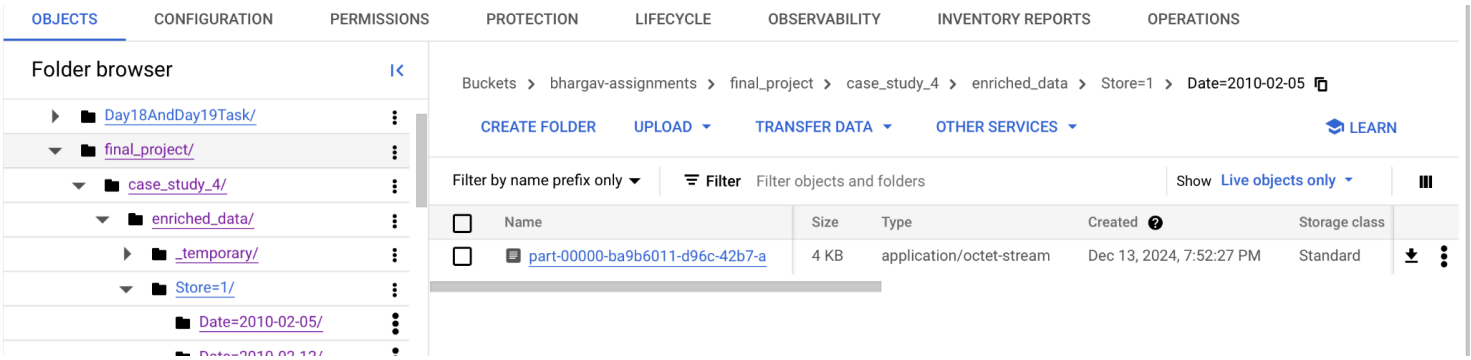
```
bhargavjangam@apple-ka-MacBook-Pro config % kafka-console-consumer.sh --bootstrap-server localhost:9092 -topic sales-topic --group consumer-group-2  --from-beginning

11
2013-01-06!???@

55
2013-02-03!??@

44
?@(3-01-27!

22
2013-01-13!???@(

33
2013-01-20!???@(

66
2013-02-10!j?@

77
2013-02-17!???@

88
2013-02-24!???@

99
2013-03-03!?F?@

1010
2013-03-10!???@

1111
2013-03-17!???@

1212
2013-03-24!?y?@
```

Initial Enriched DataFrame

```
Initial Enriched DataFrame:
+-----+----------+---------+----+------------+----+------+-----------+----------+---------+---------+---------+---------+---------+-----------+------------+
|Store|      Date|IsHoliday|Dept|Weekly_Sales|Type|  Size|Temperature|Fuel_Price|MarkDown1|MarkDown2|MarkDown3|MarkDown4|MarkDown5|        CPI|Unemployment|
+-----+----------+---------+----+------------+----+------+-----------+----------+---------+---------+---------+---------+---------+-----------+------------+
|    1|2010-02-05|    false|   1|     24924.5|   A|151315|      42.31|     2.572|       NA|       NA|       NA|       NA|       NA|211.0963582|       8.106|
|    1|2010-02-12|     true|   1|    46039.49|   A|151315|      38.51|     2.548|       NA|       NA|       NA|       NA|       NA|211.2421698|       8.106|
|    1|2010-02-19|    false|   1|    41595.55|   A|151315|      39.93|     2.514|       NA|       NA|       NA|       NA|       NA|211.2891429|       8.106|
|    1|2010-02-26|    false|   1|    19403.54|   A|151315|      46.63|     2.561|       NA|       NA|       NA|       NA|       NA|211.3196429|       8.106|
|    1|2010-03-05|    false|   1|     21827.9|   A|151315|       46.5|     2.625|       NA|       NA|       NA|       NA|       NA|211.3501429|       8.106|
+-----+----------+---------+----+------------+----+------+-----------+----------+---------+---------+---------+---------+---------+-----------+------------+
only showing top 5 rows
```

GCS uploaded Proof for Initial Enriched Dataframe

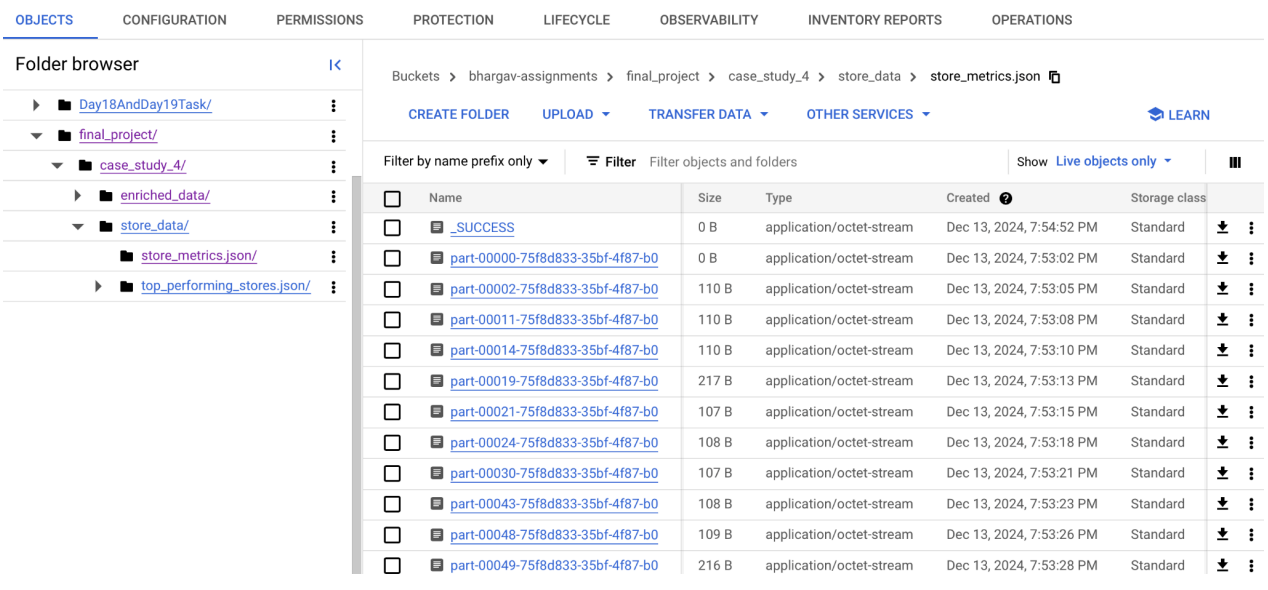




Initial Store Metric DataFrame

```
Initial Store Metrics DataFrame:
+-----+--------------------+----------------------+------------------+
|Store|Total_Sales         |Number_of_Sales_Report|Avg_Weekly_Sales  |
+-----+--------------------+----------------------+------------------+
|1    |2.2240676676999993E8|10229                 |21742.767305699414|
|5    |4.547611971000015E7 |8978                  |5065.2839953219145|
|8    |1.2995150966999948E8|9884                  |13147.663867867208|
|7    |8.1599495809999956E7|9740                  |8377.77164373712  |
|11   |1.9396439198999974E8|10034                 |19330.7147687861  |
+-----+--------------------+----------------------+------------------+
```

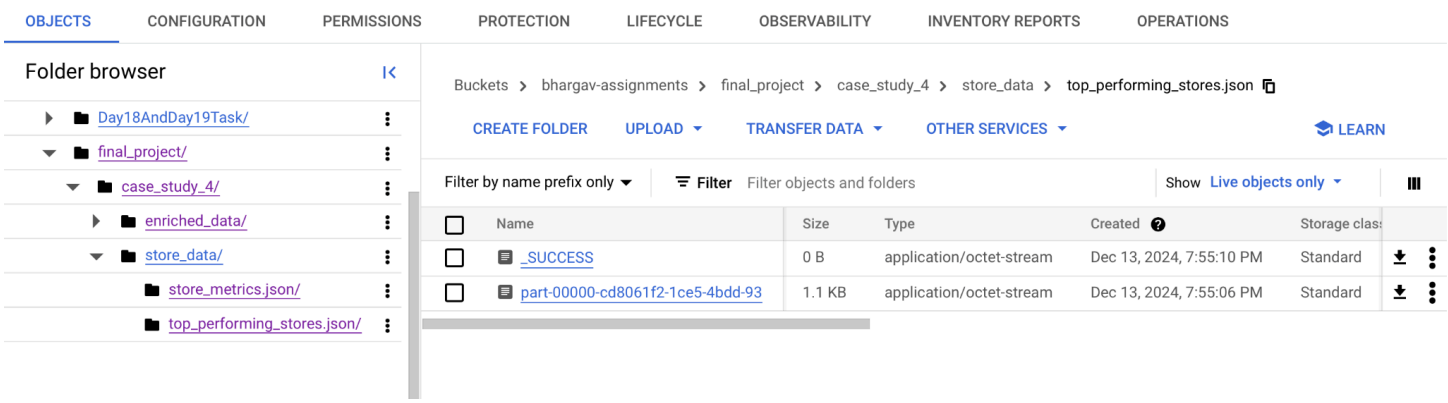
GCS uploaded Proof for Store Metric DataFrame



Initial Top Performing Stores DataFrame

```
Initial Top Performing Stores DataFrame:
+-----+------------------+---------------------+-----------------+
|Store|       Total_Sales|Number_of_Sales_Report|  Avg_Weekly_Sales|
+-----+------------------+---------------------+-----------------+
|   20|3.0140138144999975E8|                10176| 29618.84644752356|
|    4|2.9954526929999995E8|                10267| 29175.54001168793|
|   14| 2.890018644399991E8|                10008|28877.084776178966|
|   13| 2.865179503599986E8|                10459|27394.392423749745|
|    2|2.7538715550000036E8|                10216|26956.456098277245|
+-----+------------------+---------------------+-----------------+
only showing top 5 rows
```

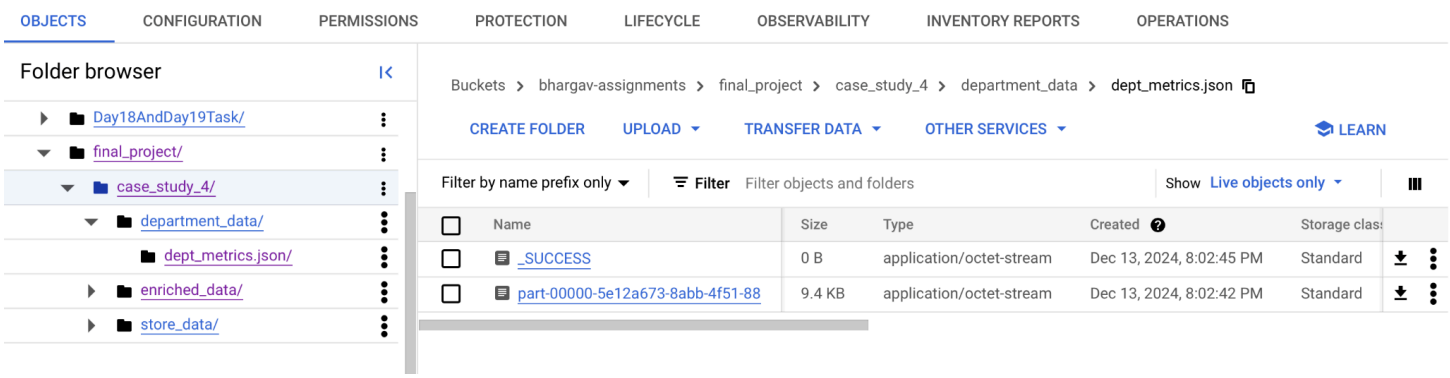
GCS uploaded Proof for Store Metric DataFrame



Initial Department Metrics Dataframe

```
Initial Department Metrics DataFrame:
+-----+----+-----------------+-------------+----------------+
|Store|Dept|      Total_Sales|Holiday_Sales|Non_Holiday_Sales|
+-----+----+-----------------+-------------+----------------+
|    2|  80|3723902.1199999996|           10|             133|
|    7|  55| 946165.2899999999|           10|             133|
|    8|  52| 278165.8300000002|           10|             133|
|   10|  85|419486.42999999993|           10|             133|
|    3|  22|443553.08999999997|           10|             133|
+-----+----+-----------------+-------------+----------------+
only showing top 5 rows
```

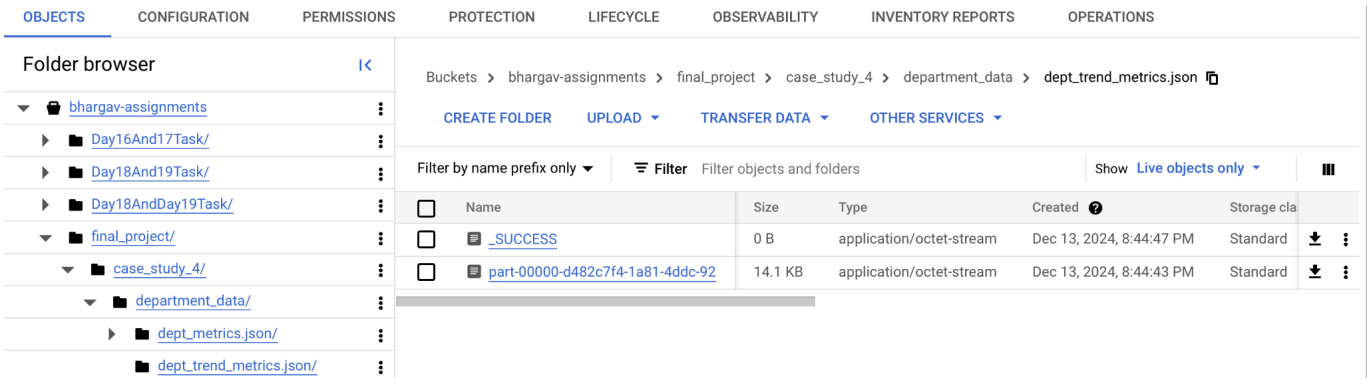
GCS uploaded Proof for Department Metrics Dataframe



Initial Department Weekly-Trends Dataframe

```
Initial Department Weekly-Trends DataFrame:
+-----+----+----------+-----------+--------------------+--------------------+
|Store|Dept|      Date|Weekly_Sales|Previous_Weekly_Sales|        Weekly_Trend|
+-----+----+----------+-----------+--------------------+--------------------+
|    1|   2|2010-02-05|    50605.27|                NULL|           No Change|
|    1|   2|2010-02-12|    44682.74|            50605.27|Decrease by 5922....|
|    1|   2|2010-02-19|    47928.89|            44682.74|Increase by 3246....|
|    1|   2|2010-02-26|    44292.87|            47928.89|Decrease by 3636....|
|    1|   2|2010-03-05|    48397.98|            44292.87|Increase by 4105....|
+-----+----+----------+-----------+--------------------+--------------------+
only showing top 5 rows
```

GCS uploaded Proof for Department Weekly-Trends Dataframe



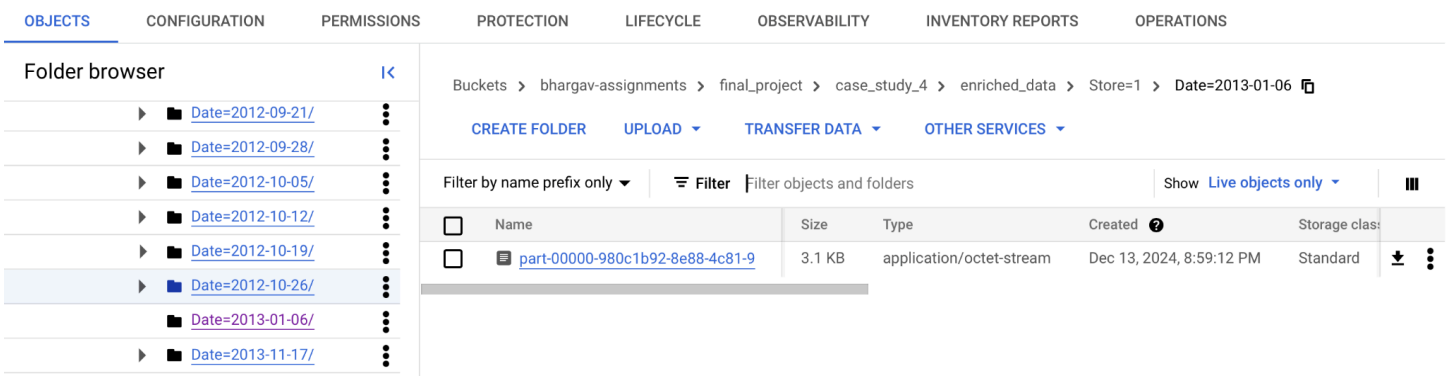
Consumed Sales Data from kafka

```
Newly Generated Data
+-----+----+----------+-----------+---------+
|Store|Dept|      Date|Weekly_Sales|IsHoliday|
+-----+----+----------+-----------+---------+
|    1|   1|2013-01-06|     14241.0|     NULL|
|    5|   5|2013-02-03|     29460.0|     NULL|
|    7|   7|2013-02-17|     29431.0|     NULL|
|    8|   8|2013-02-24|     30482.0|     NULL|
|   11|  11|2013-03-17|     42012.0|     NULL|
+-----+----+----------+-----------+---------+
only showing top 5 rows
```

New Enriched DataFrame

```
New Enriched DataFrame:
+-----+----------+---------+----+-----------+-----------+----------+---------+---------+---------+---------+---------+----+------------+----+------+
|Store|      Date|IsHoliday|Dept|Weekly_Sales|Temperature|Fuel_Price|MarkDown1|MarkDown2|MarkDown3|MarkDown4|MarkDown5| CPI|Unemployment|Type|  Size|
+-----+----------+---------+----+-----------+-----------+----------+---------+---------+---------+---------+---------+----+------------+----+------+
|    1|2013-01-06|     NULL|   1|     14241.0|       NULL|      NULL|     NULL|     NULL|     NULL|     NULL|     NULL|NULL|        NULL|   A|151315|
|    5|2013-02-03|     NULL|   5|     29460.0|       NULL|      NULL|     NULL|     NULL|     NULL|     NULL|     NULL|NULL|        NULL|   B| 34875|
|    7|2013-02-17|     NULL|   7|     29431.0|       NULL|      NULL|     NULL|     NULL|     NULL|     NULL|     NULL|NULL|        NULL|   B| 70713|
|    8|2013-02-24|     NULL|   8|     30482.0|       NULL|      NULL|     NULL|     NULL|     NULL|     NULL|     NULL|NULL|        NULL|   A|155078|
|   11|2013-03-17|     NULL|  11|     42012.0|       NULL|      NULL|     NULL|     NULL|     NULL|     NULL|     NULL|NULL|        NULL|   A|207499|
+-----+----------+---------+----+-----------+-----------+----------+---------+---------+---------+---------+---------+----+------------+----+------+
only showing top 5 rows
```

Gcs Uploaded proof for new generated data



Updated Store Metrics Dataframe

```
Updated Store Metrics Dataframe
+-----+-------------------+---------------------+-----------------+
|Store|Total_Sales        |Number_of_Sales_Report|Avg_Weekly_Sales  |
+-----+-------------------+---------------------+-----------------+
|7    |8.950107580999956E7 |9776                 |9926.891134325271 |
|11   |2.0166743898999974E8|10070                |20717.443695824237|
|8    |1.3749202166999948E8|9920                 |14567.34970570505 |
|5    |5.203849671000015E7 |9014                 |6475.234664088414 |
|1    |2.2990231276999993E8 |10265                |23046.098317638996|
+-----+-------------------+---------------------+-----------------+
```

Updated Top Performing Store Dataframe

```
Updated Top Performing Stores DataFrame
+-----+-------------------+---------------------+-----------------+
|Store|        Total_Sales|Number_of_Sales_Report|  Avg_Weekly_Sales|
+-----+-------------------+---------------------+-----------------+
|   20|3.0823889044999975E8|                10212|30745.159977556574|
|    4|3.0706263229999995E8|                10303|30426.539829770765|
|   14| 2.967195234399991E8|                10044|30202.101432539595|
|   13| 2.933396013599986E8|                10495|28502.635301490704|
|    2|2.8241471150000036E8|                10252| 28133.96845839817|
+-----+-------------------+---------------------+-----------------+
only showing top 5 rows
```

Updated Department Metrics Dataframe

```
Updated Department Metrics DataFrame:
+-----+----+-----------------+-------------+----------------+
|Store|Dept|       Total_Sales|Holiday_Sales|Non_Holiday_Sales|
+-----+----+-----------------+-------------+----------------+
|    1|  41| 915772.0599999999|           10|             133|
|    5|  10|1578731.5399999996|           11|             133|
|   41|  31| 868089.3600000001|           13|             133|
|    2|  22|2821744.9799999995|           11|             133|
|   39|   4| 6841475.659999999|           12|             133|
+-----+----+-----------------+-------------+----------------+
only showing top 5 rows
```

Updated Department Weekly Trends Dataframe

```
Updated Department Weekly-Trends DataFrame:
+-----+----+----------+-----------+--------------------+-----------------------------+
|Store|Dept|Date      |Weekly_Sales|Previous_Weekly_Sales|Weekly_Trend                 |
+-----+----+----------+-----------+--------------------+-----------------------------+
|1    |41  |2010-02-05|1011.83     |NULL                |No Change                    |
|1    |41  |2010-02-12|924.53      |1011.83             |Decrease by 87.30000000000007 |
|1    |41  |2010-02-19|1206.88     |924.53              |Increase by 282.35000000000014|
|1    |41  |2010-02-26|885.27      |1206.88             |Decrease by 321.6100000000001 |
|1    |41  |2010-03-05|1405.96     |885.27              |Increase by 520.69            |
+-----+----+----------+-----------+--------------------+-----------------------------+
only showing top 5 rows
```