

Case Study: Optimized Real-Time Data Processing with Caching, Persisting, and Broadcasting

Problem Statement


























You are tasked with building a high-performance data pipeline for processing, aggregating, and storing sales data using the Walmart Recruiting Dataset. The goal is to focus on optimizing the performance of Spark jobs by leveraging caching, persisting, and broadcasting while handling large-scale data efficiently. The pipeline should support data validation, enrichment with metadata, and efficient storage in Google Cloud Storage (GCS) using JSON and Parquet formats. Additionally, you must strategically identify scenarios where caching can improve performance, especially when data is accessed repeatedly across different stages of the pipeline.

Case Study Tasks 1. Data Preparation

Load and preprocess the Walmart Recruiting Dataset. The dataset includes:

- train.csv: Contains historical weekly sales data for various stores and departments.
- features.csv: Contains additional data about stores and regions.
- stores.csv: Contains store metadata.

File Structure:

- ✓  CaseStudy_4 [CaseStudy4] ~/Desktop/Scala_DE_Training/CaseStudy_3
 - >  .bsp
 - >  .idea
 - >  project [CaseStudy4-build] sources root
 - ✓  src
 - ✓  main
 - >  resources
 - ✓  scala
 - ✓  casestudy.operations
 -  AggregationMetrics
 -  StorageOptimisation
 - ✓  descriptor
 -  descriptor.desc
 - ✓  Kafka
 -  SalesConsumer
 -  SalesProducer
 - ✓  protos
 -  SalesReport.proto
 - >  SalesReport
 -  CaseStudy4
 - >  test
 - >  target
 -  build.sbt
 - >  External Libraries
 - >  Scratches and Consoles

casestudy_datasets

Location

us (multiple regions in United States)

Storage class

Standard

Public access

Not public

Protection

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

Buckets > casestudy_datasets > Datasets

CREATE FOLDER UPLOAD TRANSFER DATA OTHER SERVICES

Filter by name prefix only Filter Filter objects and folders Show Live objects only

	Name	Size	Type	Created	Storage class	Last modified	Public access	
	features.csv	578.4 KB	text/csv	Dec 12, 2024, 1:11:04 AM	Standard	Dec 12, 2024, 1:11:04 AM	Not public	
	stores.csv	532 B	text/csv	Dec 12, 2024, 1:11:02 AM	Standard	Dec 12, 2024, 1:11:02 AM	Not public	
	train.csv	12.2 MB	text/csv	Dec 12, 2024, 1:11:05 AM	Standard	Dec 12, 2024, 1:11:05 AM	Not public	

Load all the datasets into dataframes

Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	2010-02-05 00:00:00	42.31	2.572	NA	NA	NA	NA	NA	211.0963582	8.106	false
1	2010-02-12 00:00:00	38.51	2.548	NA	NA	NA	NA	NA	211.2421698	8.106	true
1	2010-02-19 00:00:00	39.93	2.514	NA	NA	NA	NA	NA	211.2891429	8.106	false
1	2010-02-26 00:00:00	46.63	2.561	NA	NA	NA	NA	NA	211.3196429	8.106	false
1	2010-03-05 00:00:00	46.5	2.625	NA	NA	NA	NA	NA	211.3501429	8.106	false
1	2010-03-12 00:00:00	57.79	2.667	NA	NA	NA	NA	NA	211.3806429	8.106	false
1	2010-03-19 00:00:00	54.58	2.72	NA	NA	NA	NA	NA	211.215635	8.106	false
1	2010-03-26 00:00:00	51.45	2.732	NA	NA	NA	NA	NA	211.0180424	8.106	false
1	2010-04-02 00:00:00	62.27	2.719	NA	NA	NA	NA	NA	210.8204499	7.808	false
1	2010-04-09 00:00:00	65.86	2.77	NA	NA	NA	NA	NA	210.6228574	7.808	false
1	2010-04-16 00:00:00	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	false
1	2010-04-23 00:00:00	64.84	2.795	NA	NA	NA	NA	NA	210.4391228	7.808	false
1	2010-04-30 00:00:00	67.41	2.78	NA	NA	NA	NA	NA	210.3895456	7.808	false
1	2010-05-07 00:00:00	72.55	2.835	NA	NA	NA	NA	NA	210.3399684	7.808	false
1	2010-05-14 00:00:00	74.78	2.854	NA	NA	NA	NA	NA	210.3374261	7.808	false
1	2010-05-21 00:00:00	76.44	2.826	NA	NA	NA	NA	NA	210.6170934	7.808	false
1	2010-05-28 00:00:00	80.44	2.759	NA	NA	NA	NA	NA	210.8967606	7.808	false
1	2010-06-04 00:00:00	80.69	2.705	NA	NA	NA	NA	NA	211.1764278	7.808	false
1	2010-06-11 00:00:00	80.43	2.668	NA	NA	NA	NA	NA	211.4560951	7.808	false
1	2010-06-18 00:00:00	84.11	2.637	NA	NA	NA	NA	NA	211.4537719	7.808	false

only showing top 20 rows

Store	Dept	Date	Weekly_Sales	IsHoliday
1	1	2010-02-05 00:00:00	24924.5	false
1	1	2010-02-12 00:00:00	46039.49	true
1	1	2010-02-19 00:00:00	41595.55	false
1	1	2010-02-26 00:00:00	19403.54	false
1	1	2010-03-05 00:00:00	21827.9	false
1	1	2010-03-12 00:00:00	21043.39	false
1	1	2010-03-19 00:00:00	22136.64	false
1	1	2010-03-26 00:00:00	26229.21	false
1	1	2010-04-02 00:00:00	57258.43	false
1	1	2010-04-09 00:00:00	42960.91	false
1	1	2010-04-16 00:00:00	17596.96	false
1	1	2010-04-23 00:00:00	16145.35	false
1	1	2010-04-30 00:00:00	16555.11	false
1	1	2010-05-07 00:00:00	17413.94	false
1	1	2010-05-14 00:00:00	18926.74	false
1	1	2010-05-21 00:00:00	14773.04	false
1	1	2010-05-28 00:00:00	15580.43	false
1	1	2010-06-04 00:00:00	17558.09	false
1	1	2010-06-11 00:00:00	16637.62	false
1	1	2010-06-18 00:00:00	16216.27	false

only showing top 20 rows

```

+-----+-----+-----+
|Store|Type|  Size|
+-----+-----+-----+
|   1|   A|151315|
|   2|   A|202307|
|   3|   B| 37392|
|   4|   A|205863|
|   5|   B| 34875|
|   6|   A|202505|
|   7|   B| 70713|
|   8|   A|155078|
|   9|   B|125833|
|  10|   B|126512|
|  11|   A|207499|
|  12|   B|112238|
|  13|   A|219622|
|  14|   A|200898|
|  15|   B|123737|
|  16|   B| 57197|
|  17|   B| 93188|
|  18|   B|120653|
|  19|   A|203819|
|  20|   A|203742|
+-----+-----+-----+
only showing top 20 rows

```

2. Data Validation and Enrichment

Validate the sales data and enrich it with metadata by performing the following steps:

- Ensure no missing or invalid values in critical columns.

```

+-----+-----+-----+-----+
|Store_missing|Dept_missing|Date_missing|Weekly_Sales_missing|
+-----+-----+-----+-----+
|           0|           0|           0|           0|
+-----+-----+-----+-----+

```

- Perform joins with features.csv and stores.csv on relevant keys.

Data has been saved in Parquet format at: gs://casestudy_datasets/result/enrichedDF

[Store]	Date	Dept	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Feature_IsHoliday	Type	Size
1	1/2010-02-05 00:00:00	1	24924.5	false	42.31	2.572	NA	NA	NA	NA	NA	211.0963582	8.106	false	A	151315
1	1/2010-02-12 00:00:00	1	46039.49	true	38.51	2.548	NA	NA	NA	NA	NA	211.2421698	8.106	true	A	151315
1	1/2010-02-19 00:00:00	1	41595.55	false	39.93	2.514	NA	NA	NA	NA	NA	211.2891429	8.106	false	A	151315
1	1/2010-02-26 00:00:00	1	19403.54	false	46.63	2.561	NA	NA	NA	NA	NA	211.3196429	8.106	false	A	151315
1	1/2010-03-05 00:00:00	1	21827.9	false	46.5	2.625	NA	NA	NA	NA	NA	211.3501429	8.106	false	A	151315
1	1/2010-03-12 00:00:00	1	21043.39	false	57.79	2.667	NA	NA	NA	NA	NA	211.3806429	8.106	false	A	151315
1	1/2010-03-19 00:00:00	1	22136.64	false	54.58	2.72	NA	NA	NA	NA	NA	211.215635	8.106	false	A	151315
1	1/2010-03-26 00:00:00	1	26229.21	false	51.45	2.732	NA	NA	NA	NA	NA	211.0180424	8.106	false	A	151315
1	1/2010-04-02 00:00:00	1	57258.43	false	62.27	2.719	NA	NA	NA	NA	NA	210.8204499	7.808	false	A	151315
1	1/2010-04-09 00:00:00	1	42960.91	false	65.86	2.77	NA	NA	NA	NA	NA	210.6228574	7.808	false	A	151315
1	1/2010-04-16 00:00:00	1	17596.96	false	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	false	A	151315
1	1/2010-04-23 00:00:00	1	16145.35	false	64.84	2.795	NA	NA	NA	NA	NA	210.4391228	7.808	false	A	151315
1	1/2010-04-30 00:00:00	1	16555.11	false	67.41	2.78	NA	NA	NA	NA	NA	210.3895456	7.808	false	A	151315
1	1/2010-05-07 00:00:00	1	17413.94	false	72.55	2.835	NA	NA	NA	NA	NA	210.3399684	7.808	false	A	151315
1	1/2010-05-14 00:00:00	1	18926.74	false	74.78	2.854	NA	NA	NA	NA	NA	210.3374261	7.808	false	A	151315
1	1/2010-05-21 00:00:00	1	14773.04	false	76.44	2.826	NA	NA	NA	NA	NA	210.6170934	7.808	false	A	151315
1	1/2010-05-28 00:00:00	1	15580.43	false	80.44	2.759	NA	NA	NA	NA	NA	210.8967606	7.808	false	A	151315
1	1/2010-06-04 00:00:00	1	17558.09	false	80.69	2.705	NA	NA	NA	NA	NA	211.1764278	7.808	false	A	151315
1	1/2010-06-11 00:00:00	1	16637.62	false	80.43	2.668	NA	NA	NA	NA	NA	211.4560951	7.808	false	A	151315
1	1/2010-06-18 00:00:00	1	16216.27	false	84.11	2.637	NA	NA	NA	NA	NA	211.4537719	7.808	false	A	151315

only showing top 20 rows

3. Aggregation

Compute key metrics to analyze sales performance:

Store-Level Metrics: Total weekly sales, average weekly sales, and top-performing stores.

```
+-----+-----+
|Store|  Total_Weekly_Sales|
+-----+-----+
|   12| 1.442911465599992E8|
|    1| 2.224067667699993E8|
|   13| 2.865179503599986E8|
|    6|2.2375858833999985E8|
|    3|5.7587925869999826E7|
|    5| 4.547611971000015E7|
|   15| 8.913486826000041E7|
|    9| 7.779116531999972E7|
|    4|2.9954526929999995E8|
|    8|1.2995150966999948E8|
+-----+-----+
only showing top 10 rows
```

Data has been saved in JSON format at: gs://casestudy_datasets/result/totalWeeklySalesByStore

```
+-----+-----+
|Store|Average_Weekly_Sales|
+-----+-----+
|   12| 14924.611766652792|
|    1| 21742.767305699414|
|   13| 27394.392423749745|
|    6| 21958.644586849838|
|    3| 6379.5198703888145|
|    5| 5065.2839953219145|
|   15| 9043.716341314976|
|    9| 8804.885718166353|
|    4| 29175.54001168793|
|    8| 13147.663867867208|
+-----+-----+
only showing top 10 rows
```

Data has been saved in JSON format at: gs://casestudy_datasets/result/avgWeeklySalesByStore

Store Level Metrics

45

Data has been saved in JSON format at: gs://casestudy_datasets/result/storeMetrics

Store-Level Metrics:

```
+-----+-----+-----+
|Store| Total_Weekly_Sales|Average_Weekly_Sales|
+-----+-----+-----+
| 20|3.0140138144999975E8| 29618.84644752356|
| 4|2.9954526929999995E8| 29175.54001168793|
| 14| 2.890018644399991E8| 28877.084776178966|
| 13| 2.865179503599986E8| 27394.392423749745|
| 2|2.7538715550000036E8| 26956.456098277245|
| 10| 2.716241093899999E8| 26461.189419386257|
| 27|2.5385718997000015E8| 24892.8407501471|
| 6|2.2375858833999985E8| 21958.644586849838|
| 1| 2.224067667699993E8| 21742.767305699414|
| 39|2.0744649423000023E8| 21081.96079573173|
+-----+-----+-----+
```

only showing top 10 rows

Department-Level Metrics: Total sales, weekly trends, and holiday vs. non-holiday sales.

Data has been saved in JSON format at: gs://casestudy_datasets/result/totalSalesByDept

```
+-----+-----+
|Dept|Week| Weekly_Sales|
+-----+-----+
| 3| 22| 978958.7699999998|
| 8| 52| 2431078.54|
| 28| 16| 85671.78|
| 32| 26| 799303.57|
| 33| 37| 596466.0599999999|
| 35| 45|215307.93999999997|
| 47| 32| 757.47|
| 49| 49| 447444.82|
| 56| 26| 538773.2699999999|
| 56| 30| 164233.09|
| 58| 32| 356324.74|
| 58| 34| 384105.88|
| 71| 4| 342943.51|
| 79| 10|3165061.6599999988|
| 83| 11| 482608.3899999999|
| 92| 8| 9811172.65|
| 92| 2| 7189642.949999999|
| 39| 11| 0.78|
| 47| 24| 194.0|
| 65| 50| 79848.31|
+-----+-----+
```

only showing top 20 rows

Data has been saved in JSON format at: gs://casestudy_datasets/result/weeklyTrendsByDept

Store	Date Dept Weekly_Sales IsHoliday Temperature Fuel_Price MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5	CPI Unemployment Feature_IsHoliday Type	Size
1 2010-02-05 00:00:00 1	24924.5 false 42.31 2.572 NA NA NA NA NA NA 211.0963582	8.106 false A 151315	
1 2010-02-12 00:00:00 1	46839.49 true 38.51 2.548 NA NA NA NA NA NA 211.2421698	8.106 true A 151315	
1 2010-02-19 00:00:00 1	41595.55 false 39.93 2.514 NA NA NA NA NA NA 211.2891429	8.106 false A 151315	
1 2010-02-26 00:00:00 1	19403.54 false 46.63 2.561 NA NA NA NA NA NA 211.3196429	8.106 false A 151315	
1 2010-03-05 00:00:00 1	21827.9 false 46.5 2.625 NA NA NA NA NA NA 211.3501429	8.106 false A 151315	
1 2010-03-12 00:00:00 1	21843.39 false 57.79 2.667 NA NA NA NA NA NA 211.3806429	8.106 false A 151315	
1 2010-03-19 00:00:00 1	22136.64 false 54.58 2.72 NA NA NA NA NA NA 211.215635	8.106 false A 151315	
1 2010-03-26 00:00:00 1	26229.21 false 51.45 2.732 NA NA NA NA NA NA 211.0180424	8.106 false A 151315	
1 2010-04-02 00:00:00 1	57258.43 false 62.27 2.719 NA NA NA NA NA NA 210.8204499	7.808 false A 151315	
1 2010-04-09 00:00:00 1	42960.91 false 65.86 2.77 NA NA NA NA NA NA 210.6228574	7.808 false A 151315	
1 2010-04-16 00:00:00 1	17596.96 false 66.32 2.808 NA NA NA NA NA NA 210.4887	7.808 false A 151315	
1 2010-04-23 00:00:00 1	16145.35 false 64.84 2.795 NA NA NA NA NA NA 210.4391228	7.808 false A 151315	
1 2010-04-30 00:00:00 1	16555.11 false 67.41 2.78 NA NA NA NA NA NA 210.3895456	7.808 false A 151315	
1 2010-05-07 00:00:00 1	17413.94 false 72.55 2.835 NA NA NA NA NA NA 210.3399684	7.808 false A 151315	
1 2010-05-14 00:00:00 1	18926.74 false 74.78 2.854 NA NA NA NA NA NA 210.3374261	7.808 false A 151315	
1 2010-05-21 00:00:00 1	14773.04 false 76.44 2.826 NA NA NA NA NA NA 210.6170934	7.808 false A 151315	
1 2010-05-28 00:00:00 1	15580.43 false 80.44 2.759 NA NA NA NA NA NA 210.8967606	7.808 false A 151315	
1 2010-06-04 00:00:00 1	17558.09 false 80.69 2.705 NA NA NA NA NA NA 211.1764278	7.808 false A 151315	
1 2010-06-11 00:00:00 1	16637.62 false 80.43 2.668 NA NA NA NA NA NA 211.4560951	7.808 false A 151315	
1 2010-06-18 00:00:00 1	16216.27 false 84.11 2.637 NA NA NA NA NA NA 211.4537719	7.808 false A 151315	

only showing top 20 rows

Data has been saved in JSON format at: gs://casestudy_datasets/result/totalSalesByDept

Dept Week	Weekly_Sales
3 22	978958.7699999998
8 52	2431078.54
28 16	85671.78
32 26	799303.57
33 37	596466.0599999999
35 45	215307.93999999997
47 32	757.47
49 49	447444.82
56 26	538773.2699999999
56 30	164233.09
58 32	356324.74
58 34	384105.88
71 4	342943.51
79 10	3165061.6599999988
83 11	482608.3899999999
92 8	9811172.65
92 2	7189642.949999999
39 11	0.78
47 24	194.0
65 50	79848.31

only showing top 20 rows

4. Storage Optimization

Optimize the storage of processed and aggregated data by:

Storing enriched datasets in Parquet format for compact storage and efficient querying.

console.cloud.google.com/storage/browser/casestudy_datasets/result/enrichedDF/_temporary/0/task_20241214120136789729485537633413_0015_m_000000?pageSta...

Google Cloud My First Project Search (/) for resources, docs, products, and more Search

Cloud Storage Bucket details GO TO PATH REFRESH LEARN

casestudy_datasets

Location: us (multiple regions in United States) Storage class: Standard Public access: Not public Protection: Soft Delete

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser

casestudy_datasets

Datasets/

result/

enrichedDF/

_temporary/

0/

task_2024121412013678972...

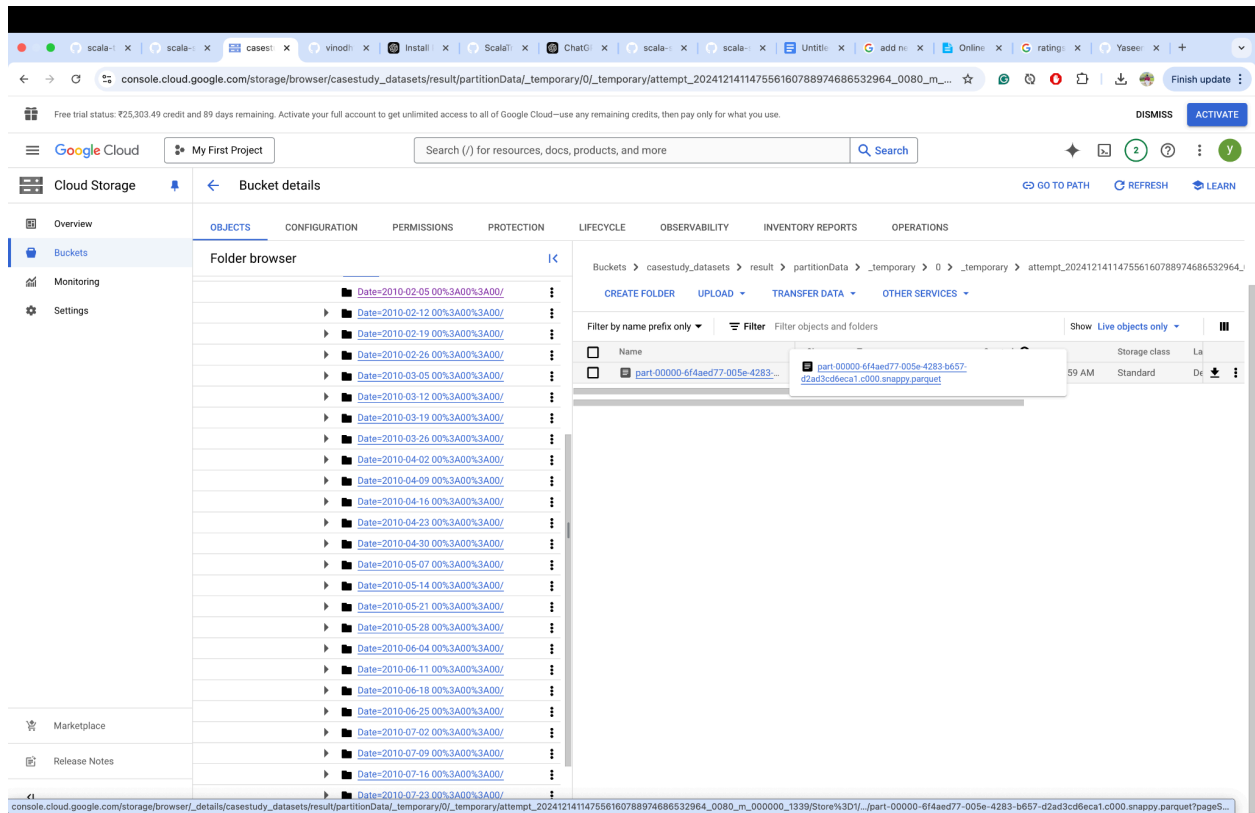
part-00000-97a147a0-c56a-4f89-b1cb-001b2808b55a-c000.snappy.parquet

part-00000-97a147a0-c56a-4f89-b1cb-001b2808b55a-c000.snappy.parquet

Storage class: Standard Last modified: Dec 14, 2024, 12:02:33 PM Pub: Not

https://console.cloud.google.com/storage/browser/_details/casestudy_datasets/result/enrichedDF/_temporary/0/task_20241214120136789729485537633413_0015_m_000000/part-00000-97a147a0-c56a-4f89-b1cb-001b2808b55a-c000.snappy.parquet?pageState={"StorageObjectListTable":...

Partitioning the data by Store and Date.



- Saving aggregated metrics in JSON format for lightweight storage.

cloud

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

GO TO PATH

REFRESH

LEARN

Cloud Storage

Bucket details

Overview

Buckets

Monitoring

Settings

casestudy_datasets

Location

us (multiple regions in United States)

Storage class

Standard

Public access

Not public

Protection

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

casestudy_datasets

Datasets/

result/

avgWeeklySalesByStore/

holidayAndNonHolidayPath/

partitionData/

totalSalesByDept/

totalSalesByStore/

weeklyTrendsByDept/

Buckets > casestudy_datasets > result > totalSalesByStore

CREATE FOLDER

UPLOAD

TRANSFER DATA

OTHER SERVICES

Filter by name prefix only

Filter

Filter objects and folders

Show

Live objects only

Name	Size	Type	Created	Storage class	Last modified	P
_SUCCESS	0 B	application/octet-stream	Dec 14, 2024, 5:05:01 PM	Standard	Dec 14, 2024, 5:05:01 PM	N
part-00000-192cbe53-e2e9-4474-...	2.4 KB	application/octet-stream	Dec 14, 2024, 5:04:55 PM	Standard	Dec 14, 2024, 5:04:55 PM	N

cloud

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

GO TO PATH

REFRESH

LEARN

Cloud Storage

Bucket details

Overview

Buckets

Monitoring

Settings

casestudy_datasets

Location

us (multiple regions in United States)

Storage class

Standard

Public access

Not public

Protection

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

casestudy_datasets

Datasets/

result/

avgWeeklySalesByStore/

holidayAndNonHolidayPath/

partitionData/

totalSalesByDept/

totalSalesByStore/

weeklyTrendsByDept/

Buckets > casestudy_datasets > result > totalSalesByStore

CREATE FOLDER

UPLOAD

TRANSFER DATA

OTHER SERVICES

Filter by name prefix only

Filter

Filter objects and folders

Show

Live objects only

Name	Size	Type	Created	Storage class	Last modified	P
_SUCCESS			Dec 14, 2024, 5:05:01 PM	Standard	Dec 14, 2024, 5:05:01 PM	N
part-00000-192cbe53-e2e9-4474-9e21-d1af12a43987-c000.json			Dec 14, 2024, 5:04:55 PM	Standard	Dec 14, 2024, 5:04:55 PM	N

https://console.cloud.google.com/storage/browser/_details/casestudy_datasets/result/totalSalesByStore/part-00000-192cbe53-e2e9-4474-9e21-d1af12a43987-c000.json?pageState={"StorageObjectListTable":{"f":"","%255B%255D"}}&authuser=1&hl=en&inv=A...

After Updating the Data Using Kafka Stream

```
+-----+-----+-----+-----+
|Store|  Total_Weekly_Sales|Average_Weekly_Sales|Data_Count|
+-----+-----+-----+-----+
|    6|2.2376112132095712E8|  21954.584117048384|    10192|
|    1|2.2240856540525395E8|  21740.817732673895|    10230|
|    8|1.2995324295149413E8|  13146.509150378768|     9885|
|    2|2.7538903110717773E8|  26954.001282879293|    10217|
+-----+-----+-----+-----+
```

```
+-----+-----+-----+-----+
|Store|Dept|Total_Weekly_Sales|Average_Weekly_Sales|Data_Count|
+-----+-----+-----+-----+
|    1|   3| 1882316.995253907|  13071.645800374354|    144|
|    6|  17| 2331574.693681641|  16191.490928344729|    144|
|    2|  17| 1973967.757177735|  13708.109424845381|    144|
|    6|   5| 4929967.197275389|   34235.88331441242|    144|
|    8|   4|3017504.5114941397|  20954.892440931526|    144|
+-----+-----+-----+-----+
```

Store Dept		Holiday_Sales	NonHoliday_Sales
1	3	115125.37000000001	1767191.6252539067
6	17	176511.87	2155062.8236816404
2	17	164592.81999999998	1809374.9371777351
6	5	504211.35727539053	4425755.84
8	4	225412.6914941406	2792091.8199999994