TECH
CONSULTING

Walmart sales Analysis

Subtitle: Leveraging Big Data for Insights using cloud technologies

By David Ishimwe Ruberamitwe

TECH
CONSULTING

Overview

Business Problem	01
Objectives	02
Dimensional Modeling diagram	03
Data Pipeline	04
Full and Incremental load	06
Data Cleaning	07
Testing data quality	08
Dimensional modeling	09
Data Analysis	10
Kinesis	11
Conclusion	12

BUSINESS PROBLEM

Challenges:

- Managing inventory across Walmart's extensive network(above **10500** stores).
- Fluctuating consumer demand due to seasonal trends and economic factors.

Impact:

- Overstock leads to high costs, while understock causes lost sales.

Goal:

- Analyze historical sales data to uncover insights that can guide inventory management.

OBJECTIVES

Sales Trends:

- What are the overall sales trends over time?

Seasonal Impacts:

- How do holidays and promotions affect sales?

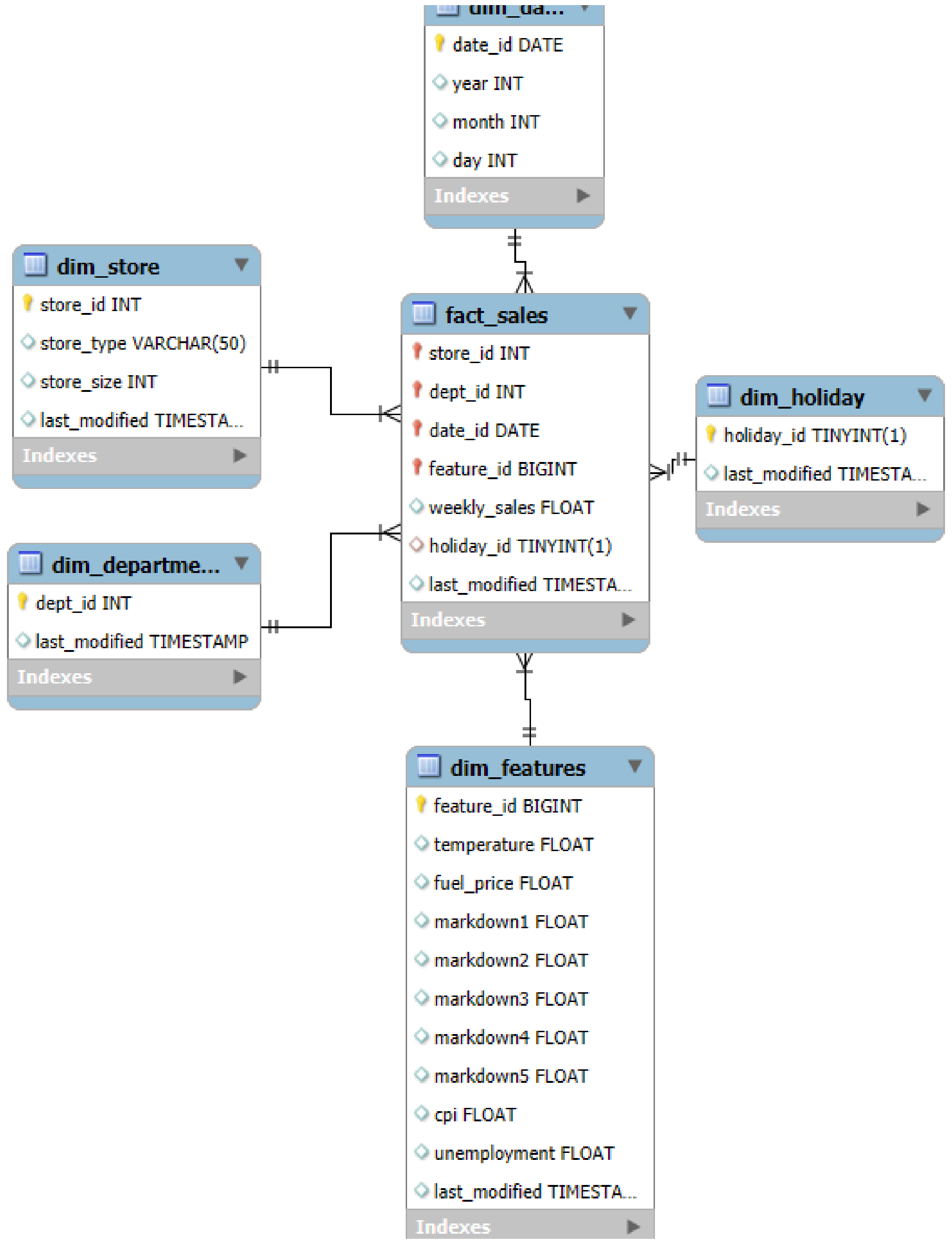
External Impact:

- Would fuel price and CPI affect sales?

Leveraging Big Data Technologies:

- Explore how big data tools and technologies can enhance the depth and accuracy of sales data analysis for better decision-making.

DIMENTIONAL MODEL DIAGRAM



dim_da...
- date_id DATE
- year INT
- month INT
- day INT

Indexes

dim_store
- store_id INT
- store_type VARCHAR(50)
- store_size INT
- last_modified TIMESTA...

Indexes

fact_sales
- store_id INT
- dept_id INT
- date_id DATE
- feature_id BIGINT
- weekly_sales FLOAT
- holiday_id TINYINT(1)
- last_modified TIMESTA...

Indexes

dim_holiday
- holiday_id TINYINT(1)
- last_modified TIMESTA...

Indexes

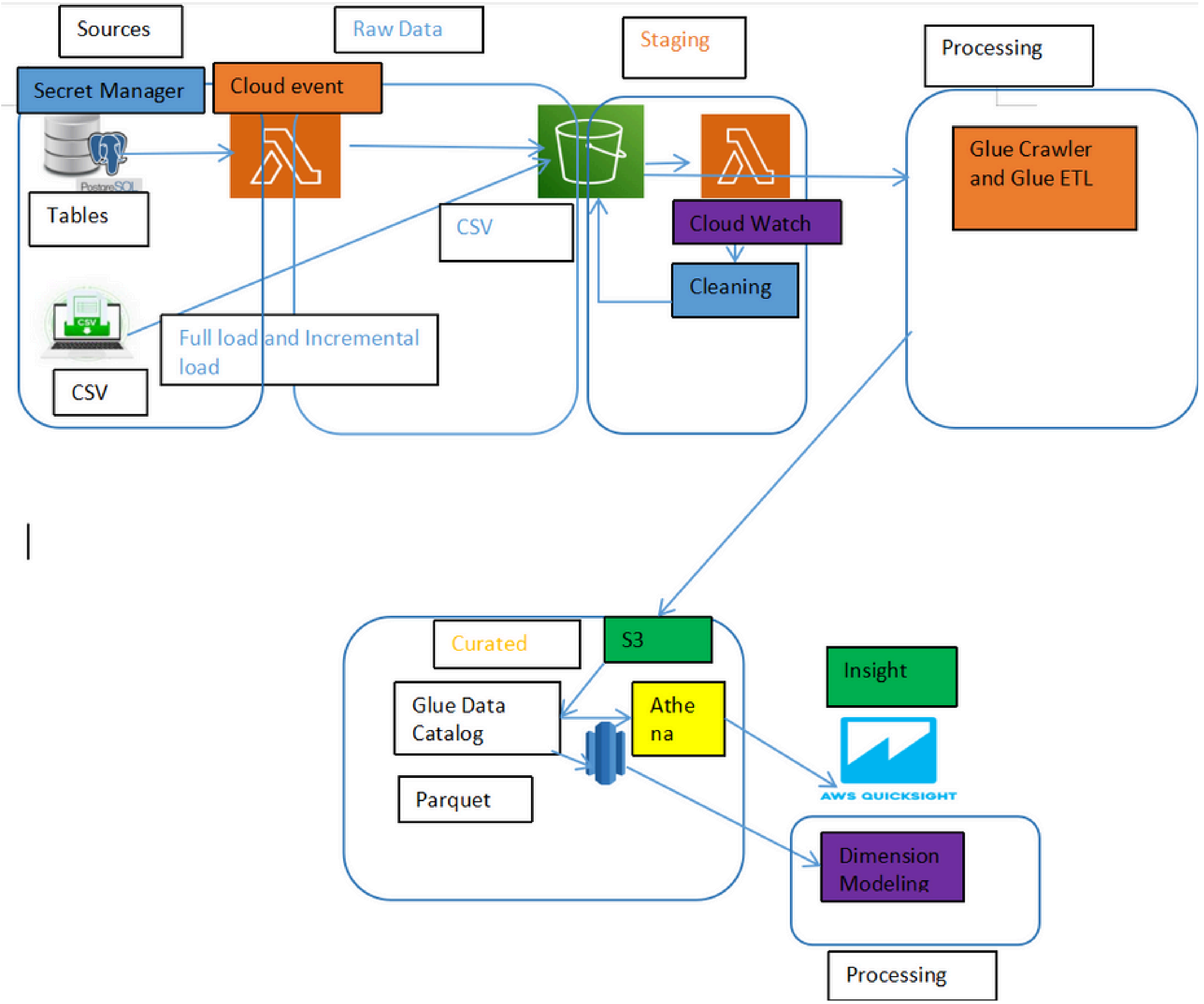
dim_departme...
- dept_id INT
- last_modified TIMESTAMP

Indexes

dim_features
- feature_id BIGINT
- temperature FLOAT
- fuel_price FLOAT
- markdown1 FLOAT
- markdown2 FLOAT
- markdown3 FLOAT
- markdown4 FLOAT
- markdown5 FLOAT
- cpi FLOAT
- unemployment FLOAT
- last_modified TIMESTA...

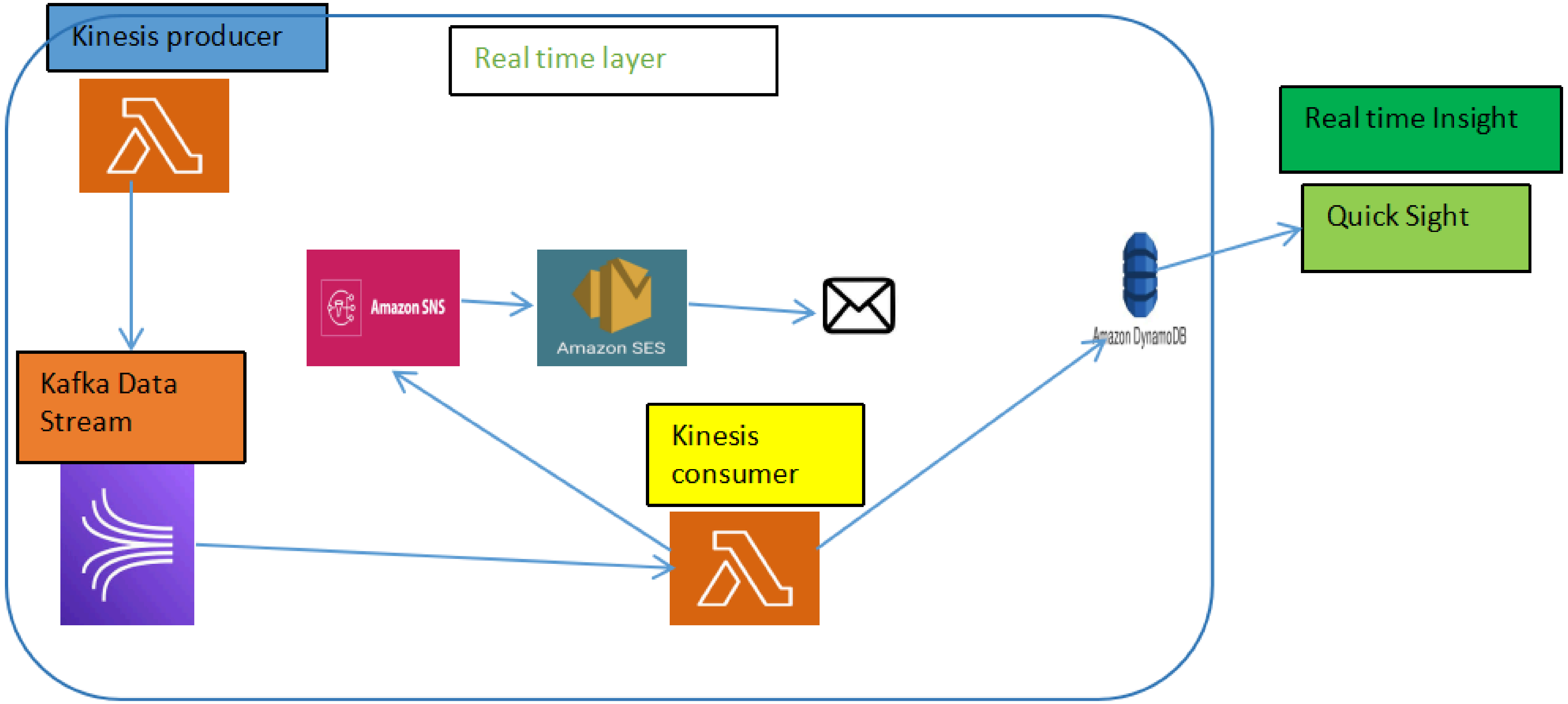
Indexes

PIPELINE DESIGN(BATCH PROCESSING)



TECH
CONSULTING

PIPELINE DESIGN(REAL-TIME PROCESSING PART)



FULL LOAD AND INCREMENTAL LOADING

Full Load Setup:

- Performed complete data loading of 1000 records.

Incremental Load Logic:

- Filtered new/updated records based on timestamp (last_modified).

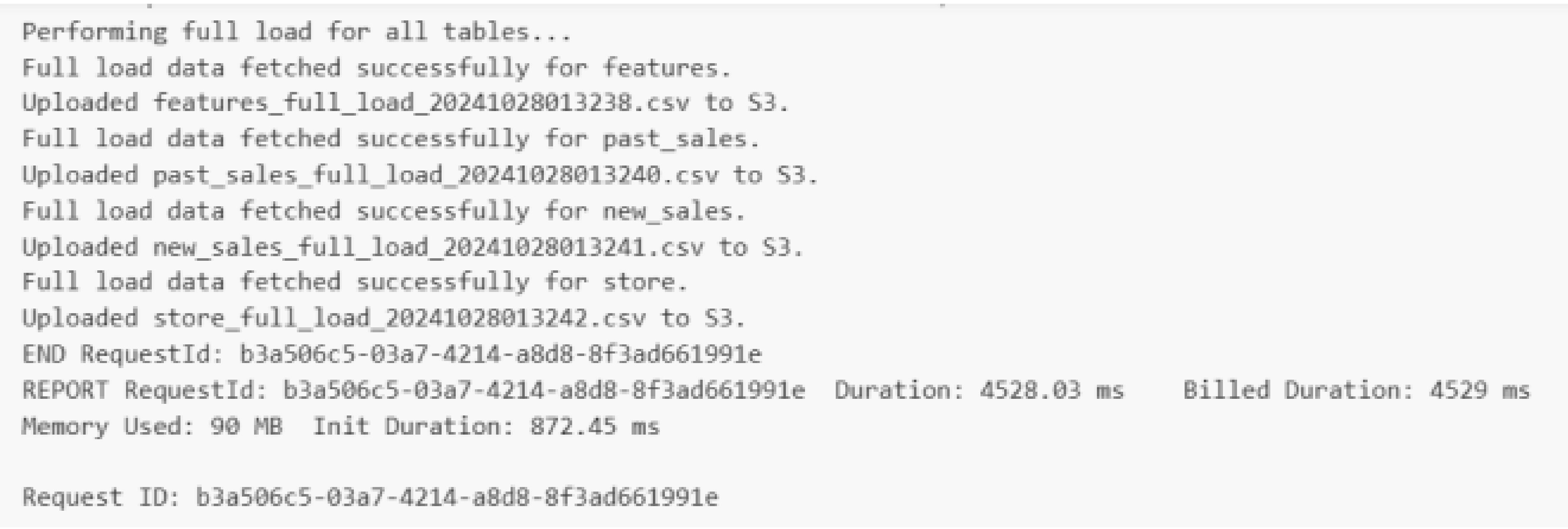
Optimized Data Handling:

- Processed only new data to reduce resource usage.

Seamless S3 Integration:

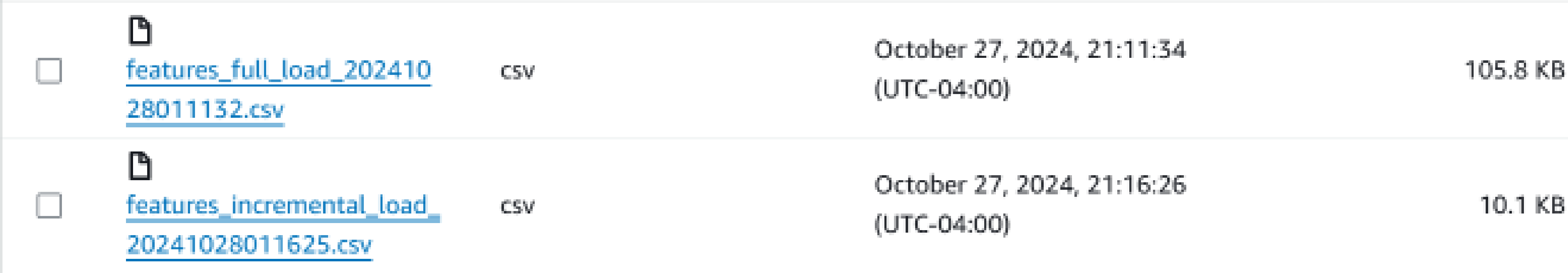
- Consistent data format for smooth appends and updates.

Result



```
Performing full load for all tables...
Full load data fetched successfully for features.
Uploaded features_full_load_20241028013238.csv to S3.
Full load data fetched successfully for past_sales.
Uploaded past_sales_full_load_20241028013240.csv to S3.
Full load data fetched successfully for new_sales.
Uploaded new_sales_full_load_20241028013241.csv to S3.
Full load data fetched successfully for store.
Uploaded store_full_load_20241028013242.csv to S3.
END RequestId: b3a506c5-03a7-4214-a8d8-8f3ad661991e
REPORT RequestId: b3a506c5-03a7-4214-a8d8-8f3ad661991e  Duration: 4528.03 ms    Billed Duration: 4529 ms
Memory Used: 90 MB  Init Duration: 872.45 ms

Request ID: b3a506c5-03a7-4214-a8d8-8f3ad661991e
```



□	features_full_load_20241028011132.csv	csv	October 27, 2024, 21:11:34 (UTC-04:00)	105.8 KB
□	features_incremental_load_20241028011625.csv	csv	October 27, 2024, 21:16:26 (UTC-04:00)	10.1 KB

CLEANING DATA ON S3

1.Remove Deduplication

- Loaded incremental and full datasets from S3 into Lambda function.
- Removed duplicate records to ensure data integrity.

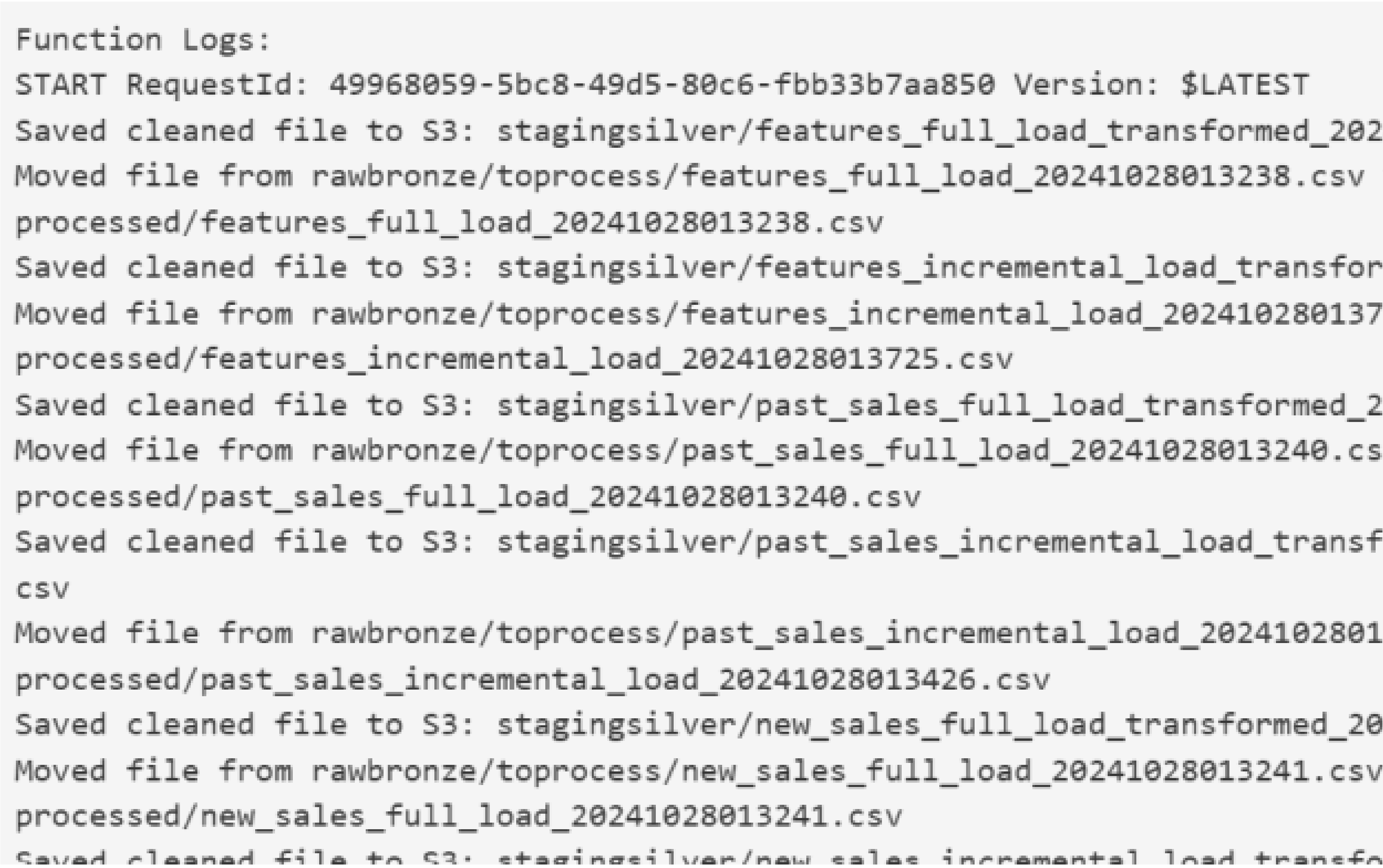
2.Data Normalization

- Standardized text fields (trimming spaces, converting to lowercase).
- Converted date columns to a consistent format.

3.Outlier Handling

- Applied Z-score method for outlier detection.
- Filtered out records with |Z| > 3 for cleaner analysis.

Result after cleaning the data



```
Function Logs:
START RequestId: 49968059-5bc8-49d5-80c6-fbb33b7aa850 Version: $LATEST
Saved cleaned file to S3: stagingsilver/features_full_load_transformed_202
Moved file from rawbronze/toprocess/features_full_load_20241028013238.csv
processed/features_full_load_20241028013238.csv
Saved cleaned file to S3: stagingsilver/features_incremental_load_transfor
Moved file from rawbronze/toprocess/features_incremental_load_202410280137
processed/features_incremental_load_20241028013725.csv
Saved cleaned file to S3: stagingsilver/past_sales_full_load_transformed_2
Moved file from rawbronze/toprocess/past_sales_full_load_20241028013240.cs
processed/past_sales_full_load_20241028013240.csv
Saved cleaned file to S3: stagingsilver/past_sales_incremental_load_transf
csv
Moved file from rawbronze/toprocess/past_sales_incremental_load_2024102801
processed/past_sales_incremental_load_20241028013426.csv
Saved cleaned file to S3: stagingsilver/new_sales_full_load_transformed_20
Moved file from rawbronze/toprocess/new_sales_full_load_20241028013241.csv
processed/new_sales_full_load_20241028013241.csv
Saved cleaned file to S3: stagingsilver/new_sales_incremental_load_transfo
```

TECH CONSULTING

DATA QUALITY CHECK

1. **Check Deduplication**

- Loaded incremental and full datasets from S3 into Lambda function.
- checked if duplicate records were removed.

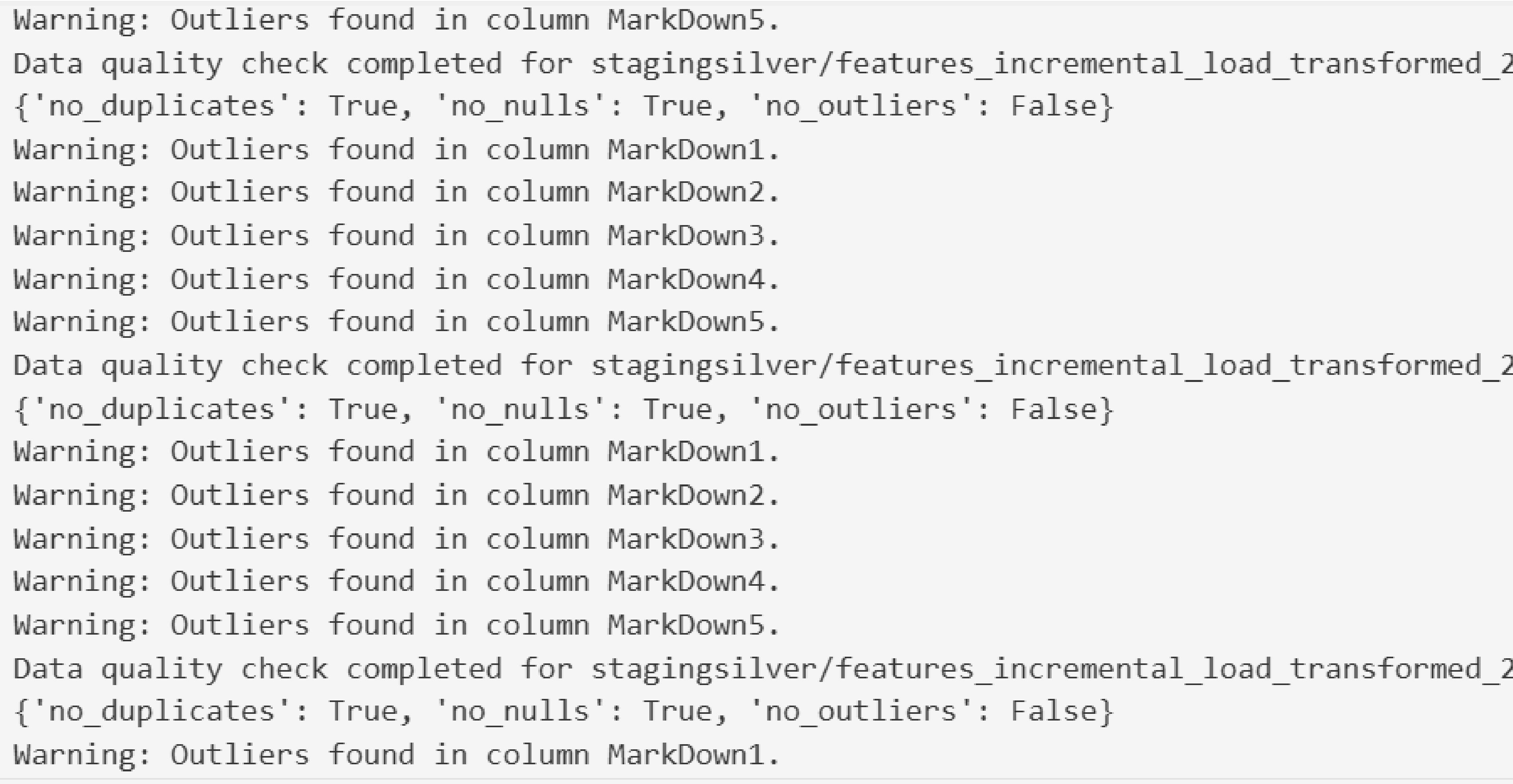
2.Check null values

- check if there is any null values in files.

3.Outlier Checking

- check if there is any outlier in the file

Result after checking data quality



```
Warning: Outliers found in column MarkDown5.
Data quality check completed for stagingsilver/features_incremental_load_transformed_2
{'no_duplicates': True, 'no_nulls': True, 'no_outliers': False}
Warning: Outliers found in column MarkDown1.
Warning: Outliers found in column MarkDown2.
Warning: Outliers found in column MarkDown3.
Warning: Outliers found in column MarkDown4.
Warning: Outliers found in column MarkDown5.
Data quality check completed for stagingsilver/features_incremental_load_transformed_2
{'no_duplicates': True, 'no_nulls': True, 'no_outliers': False}
Warning: Outliers found in column MarkDown1.
Warning: Outliers found in column MarkDown2.
Warning: Outliers found in column MarkDown3.
Warning: Outliers found in column MarkDown4.
Warning: Outliers found in column MarkDown5.
Data quality check completed for stagingsilver/features_incremental_load_transformed_2
{'no_duplicates': True, 'no_nulls': True, 'no_outliers': False}
Warning: Outliers found in column MarkDown1.
```

TECH CONSULTING

CREATING TABLES IN THE STAGING DIRECTORY USING GLUE CRAWLER TO DATA CATLOG

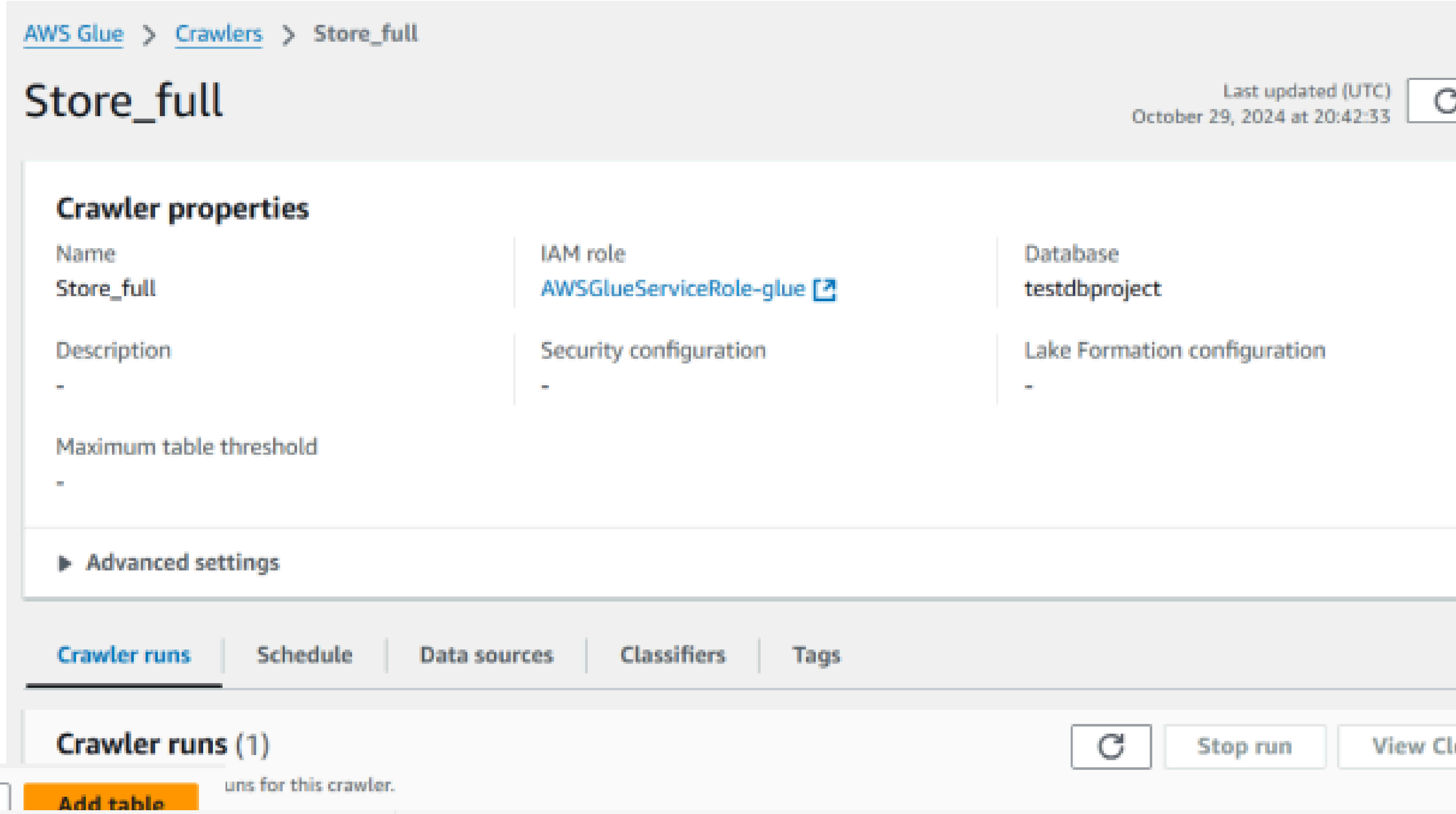
1.Created Glue crawler for getting schemas.

2.Created Data Catalog with database called testdbproject.

3.Run crawlers in order for it to get schema of tables.

4. Finally checking tables in Data Catalog.

Crawler Execution



Store_full

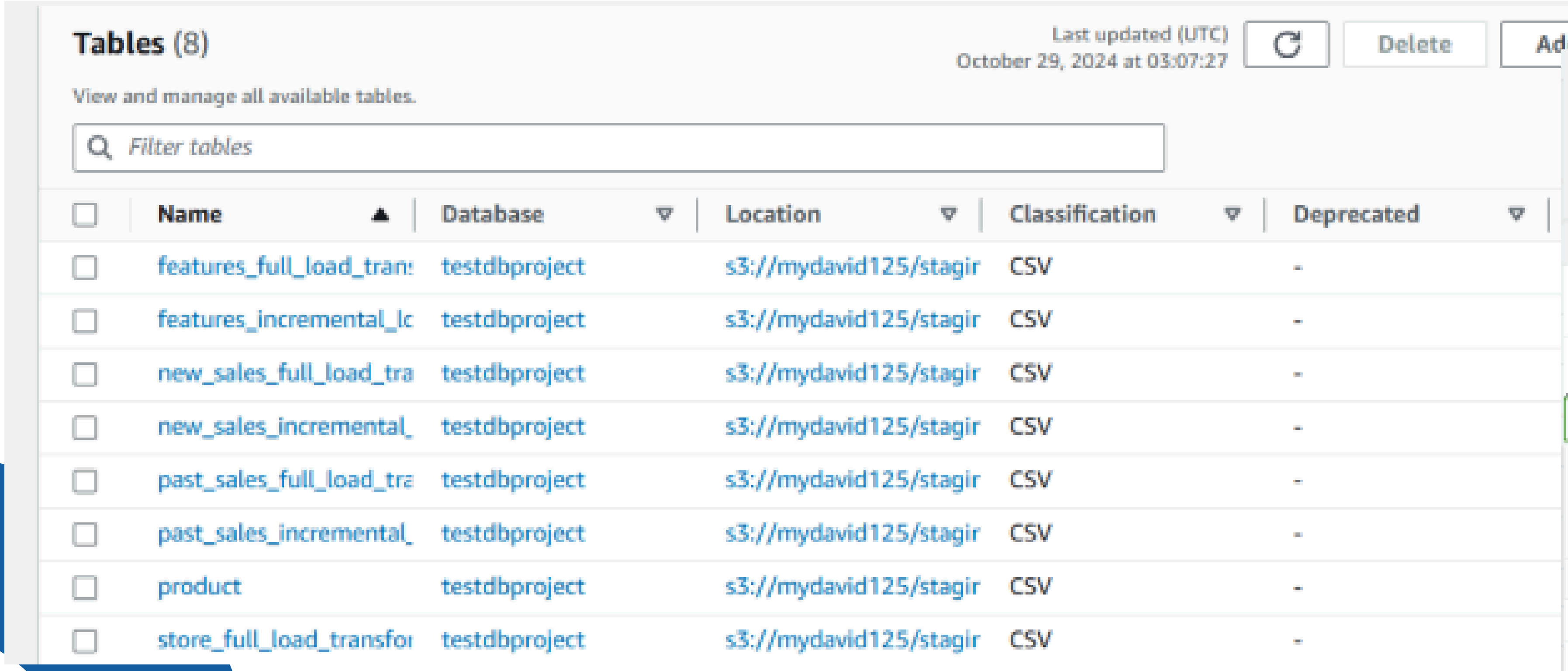
Last updated (UTC)
October 29, 2024 at 20:42:33

Crawler properties

Name: Store_full
IAM role: AWSGlueServiceRole-glue
Database: testdbproject
Description: -
Security configuration: -
Lake Formation configuration: -
Maximum table threshold: -

Crawler runs (1)

uns for this crawler.

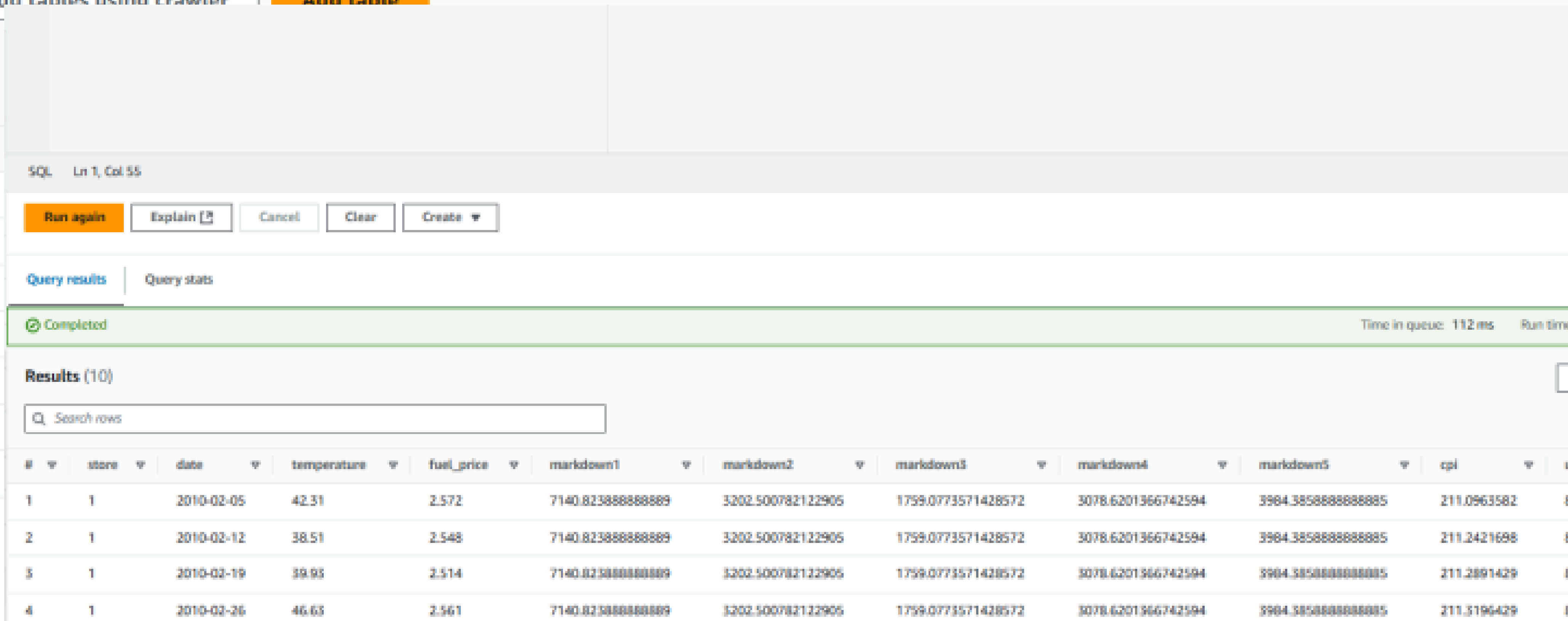


Tables (8)

Last updated (UTC)
October 29, 2024 at 03:07:27

View and manage all available tables.

□	Name	Database	Location	Classification	Deprecated
□	features_full_load_tran:	testdbproject	s3://mydavid125/stagir	CSV	-
□	features_incremental_lc	testdbproject	s3://mydavid125/stagir	CSV	-
□	new_sales_full_load_tra	testdbproject	s3://mydavid125/stagir	CSV	-
□	new_sales_incremental_	testdbproject	s3://mydavid125/stagir	CSV	-
□	past_sales_full_load_trz	testdbproject	s3://mydavid125/stagir	CSV	-
□	past_sales_incremental_	testdbproject	s3://mydavid125/stagir	CSV	-
□	product	testdbproject	s3://mydavid125/stagir	CSV	-
□	store_full_load_transfoi	testdbproject	s3://mydavid125/stagir	CSV	-



Completed

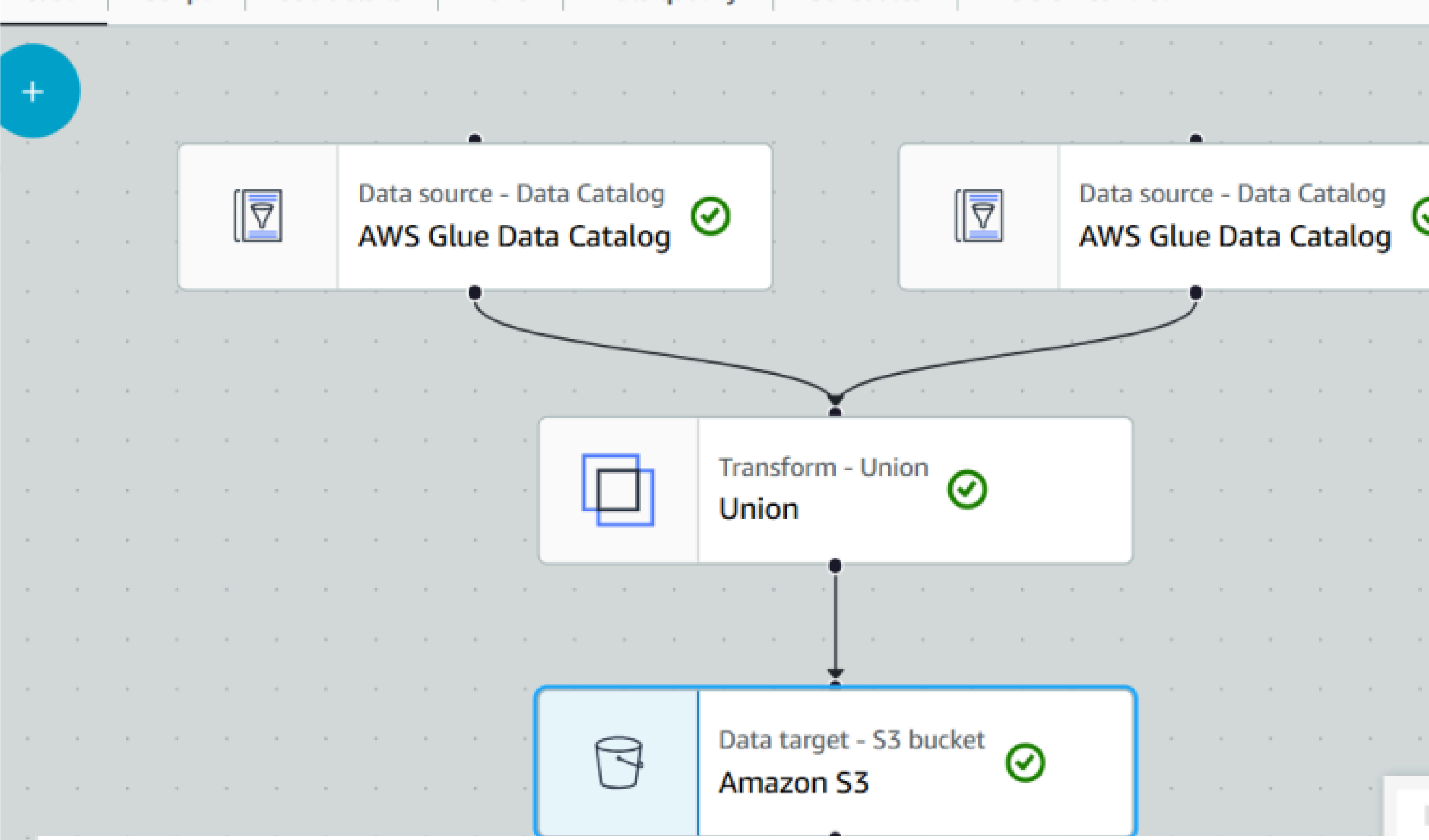
Time in queue: 112 ms

Results (10)

#	store	date	temperature	fuel_price	markdown1	markdown2	markdown3	markdown4	markdown5	cpi
1	1	2010-02-05	42.31	2.572	7140.823888888889	3202.500782122905	1759.0773571428572	3078.6201366742594	3984.3858888888885	211.0963582
2	1	2010-02-12	38.51	2.548	7140.823888888889	3202.500782122905	1759.0773571428572	3078.6201366742594	3984.3858888888885	211.2421698
3	1	2010-02-19	39.93	2.514	7140.823888888889	3202.500782122905	1759.0773571428572	3078.6201366742594	3984.3858888888885	211.2891429
4	1	2010-02-26	46.63	2.561	7140.823888888889	3202.500782122905	1759.0773571428572	3078.6201366742594	3984.3858888888885	211.3196429

TECH CONSULTING

ETL PROCESS USING GLUE

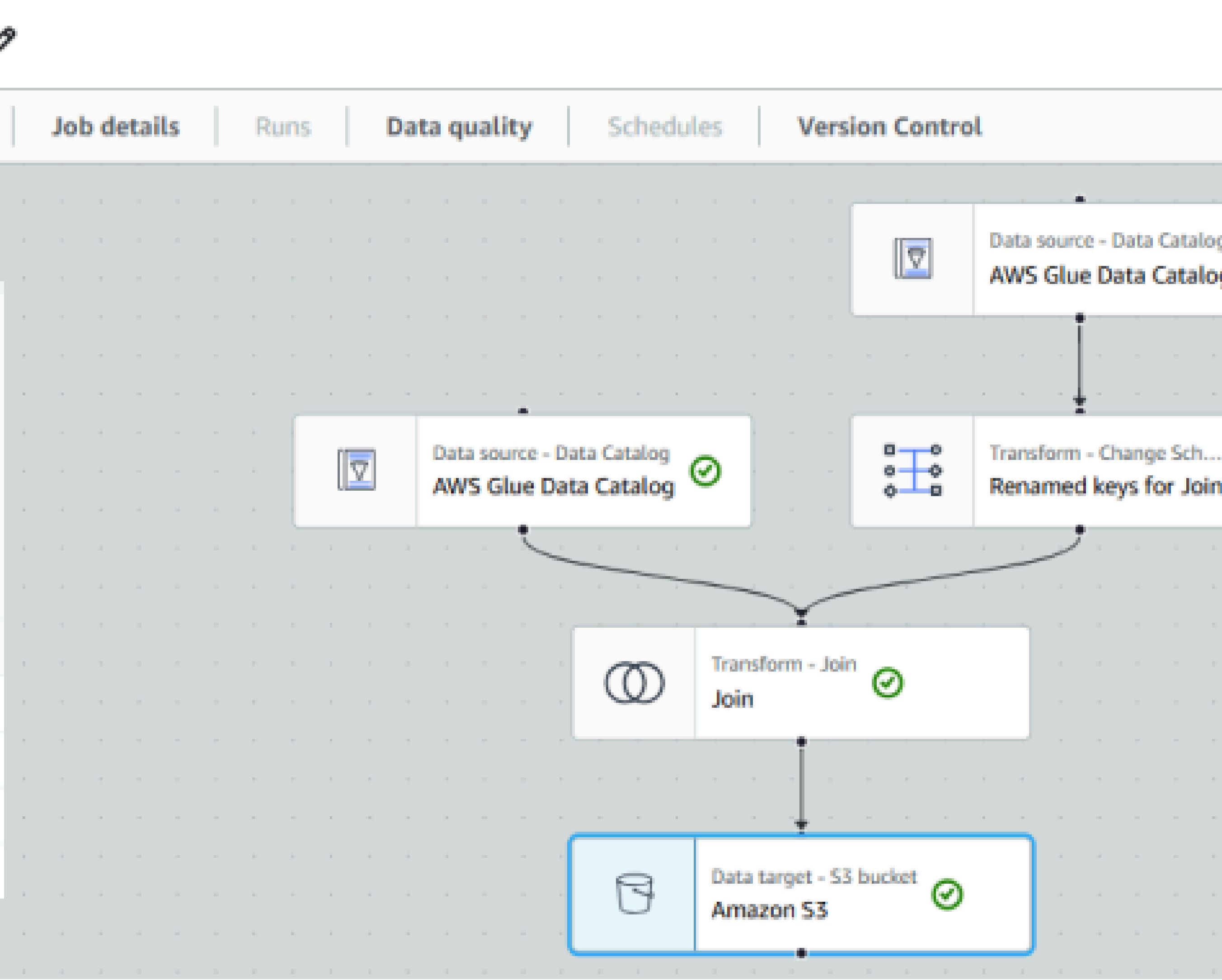


1.Created ETL job to append incremental and full load tables on Glue.

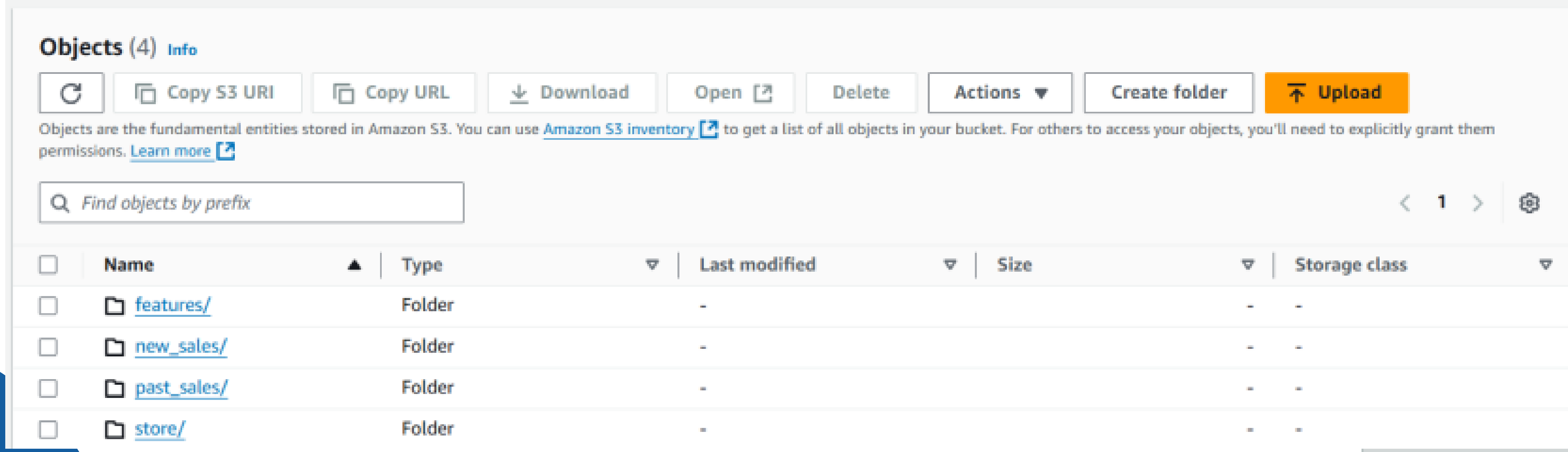
2.Created ETL job to join all files into one that I used in my analysis using Quick Sight.

Untitled job

Visual | Script | Job details | Runs | Data quality | Schedules | Version Control



Result After first job



Objects (4) Info

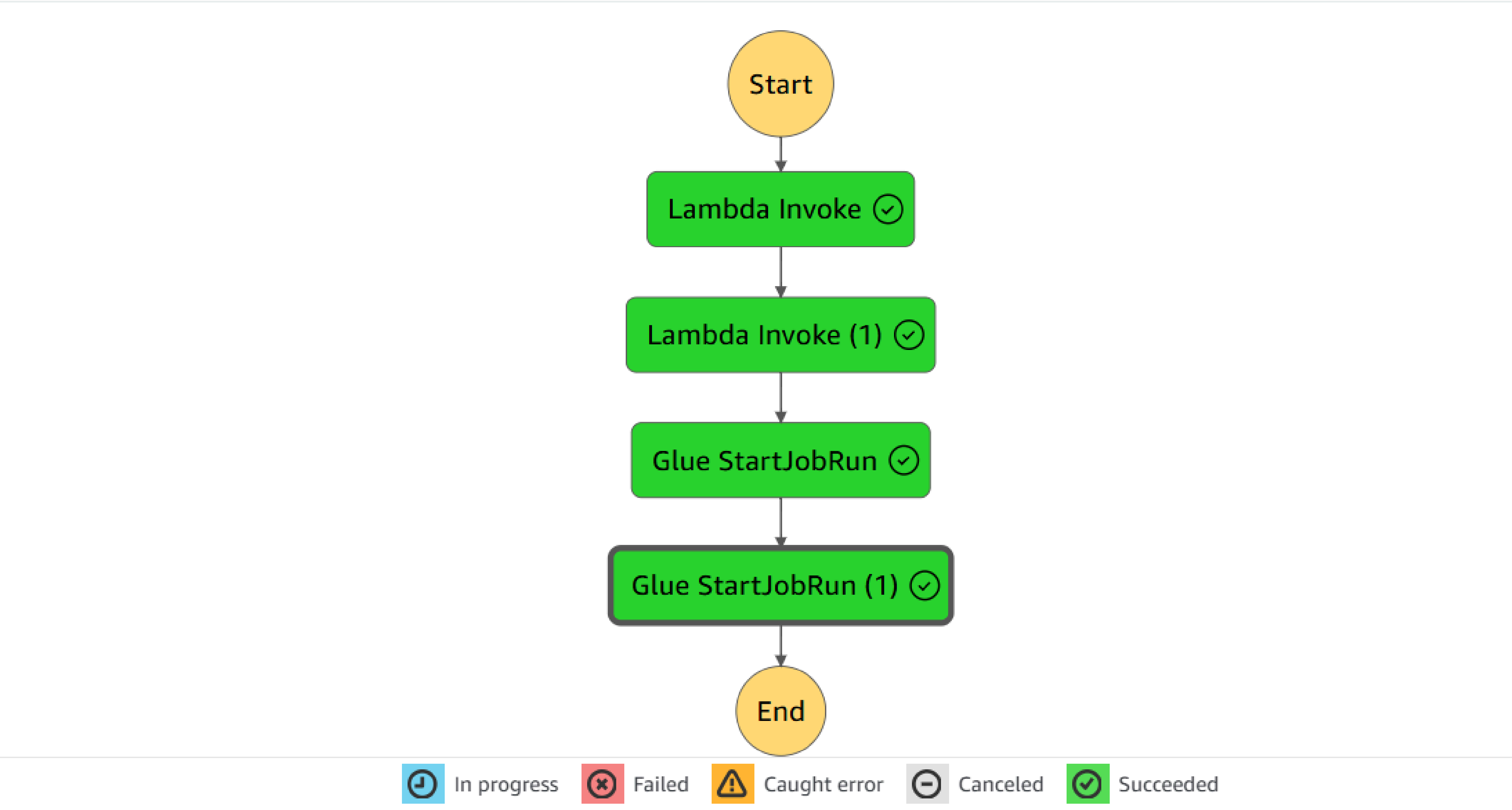
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	features/	Folder	-	-	-
□	new_sales/	Folder	-	-	-
□	past_sales/	Folder	-	-	-
□	store/	Folder	-	-	-

TECH CONSULTING

STEP FUNCTION

- The first lambda is the ingestion stage.
- The second Lambda is the cleaning stage.
- The first Glue job is for joining full and incremental files.
- The second Glue job is for joining files from the first job into one to be used for data analysis.



TECH CONSULTING

WALMART SALES DIMENSIONAL MODELING ON REDSHIFT

1.Star Schema:

- Central fact table with multiple dimension tables as external table or redshift spectrum on S3.

2.Dimension Tables:

- Key attributes for Store, Date, Department, Holiday, Features.

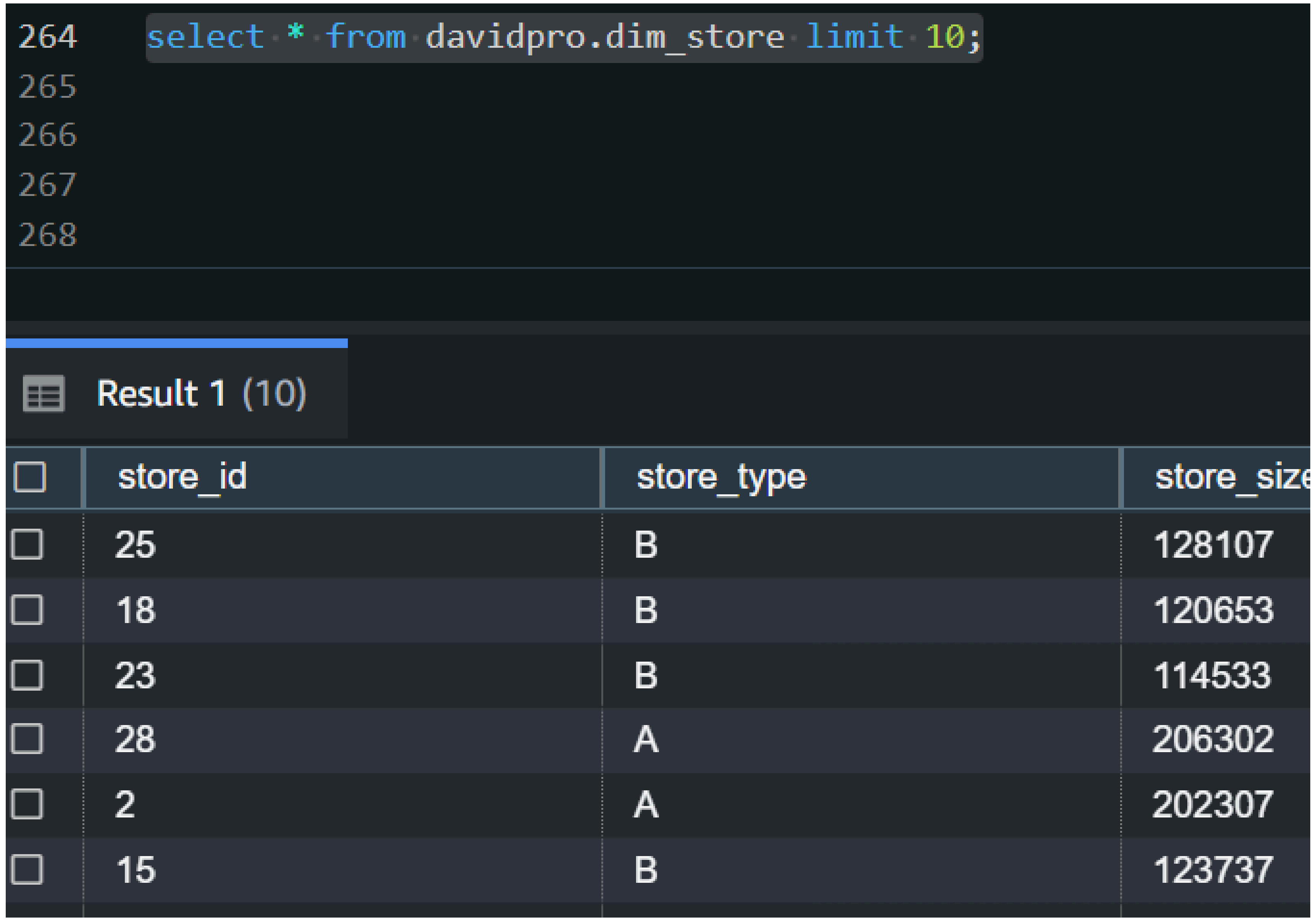
3.Efficient Loading:

- parquet format for optimized storage and performance.

4.Fact Table Joins:

- Linked dimensions for advanced sales analysis.

dim_store table



```
select * from davidpro.dim_store limit 10;
```

□	store_id	store_type	store_size
□	25	B	128107
□	18	B	120653
□	23	B	114533
□	28	A	206302
□	2	A	202307
□	15	B	123737

DATA ANALYSIS

TOTAL SALES OVER TIME ANALYSIS

Peak Sales: Significant spike in sales during late 2011, likely due to the holiday season.
Sharp Decline: Sales dropped drastically in early 2012.

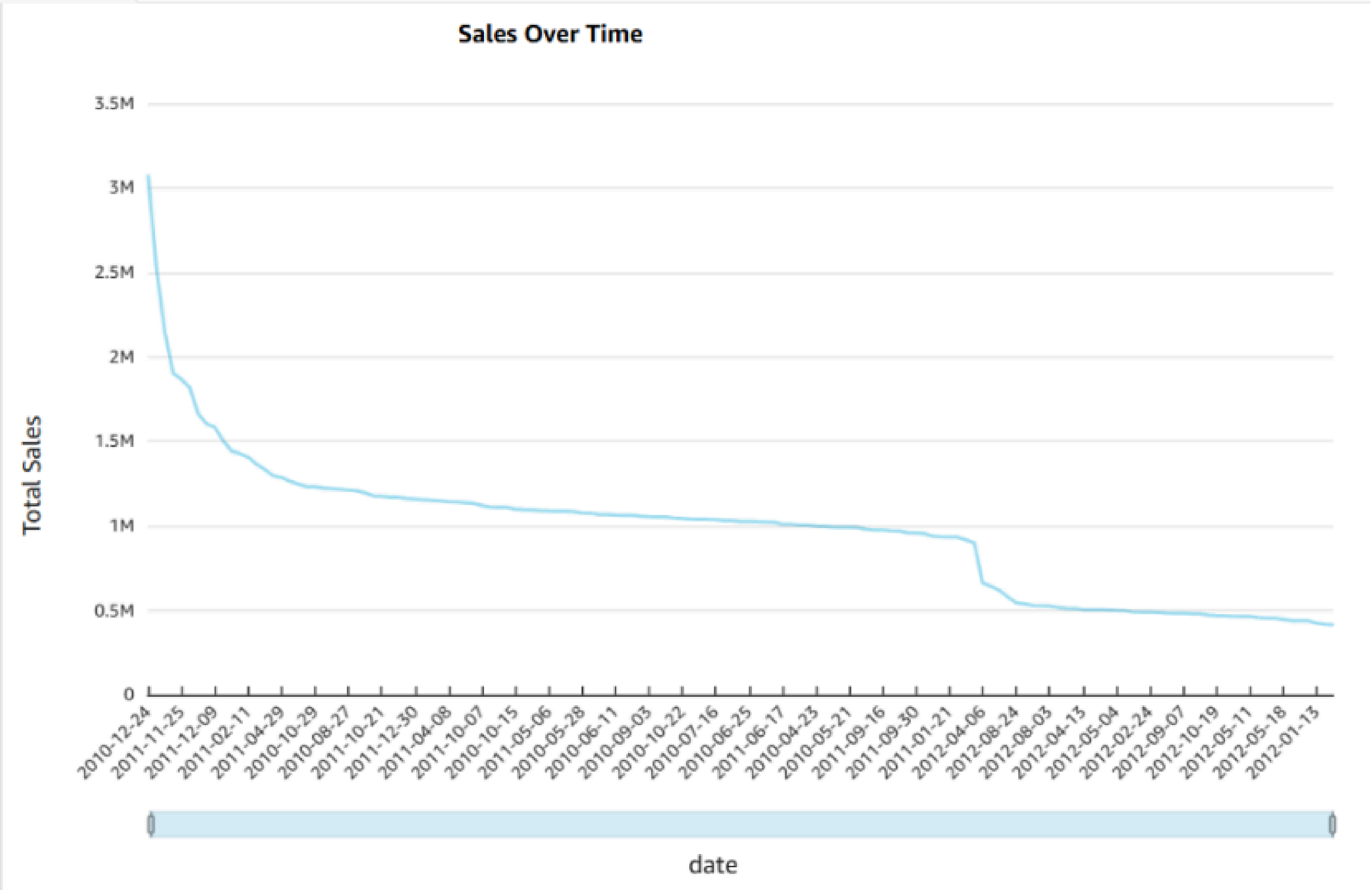
Stabilization: Sales remained relatively stable at lower levels throughout 2012

TECH CONSULTING

DATA ANALYSIS

SALES BY MONTH AND YEAR ANALYSIS

2011 Spike: Sharp rise in December 2011, reflecting a significant holiday sales surge.
2012 Trends: Fluctuations throughout 2012 with peaks around March and gradual decline towards year-end.
Holiday Impact: December sales tend to show a strong performance compared to other months across both years.

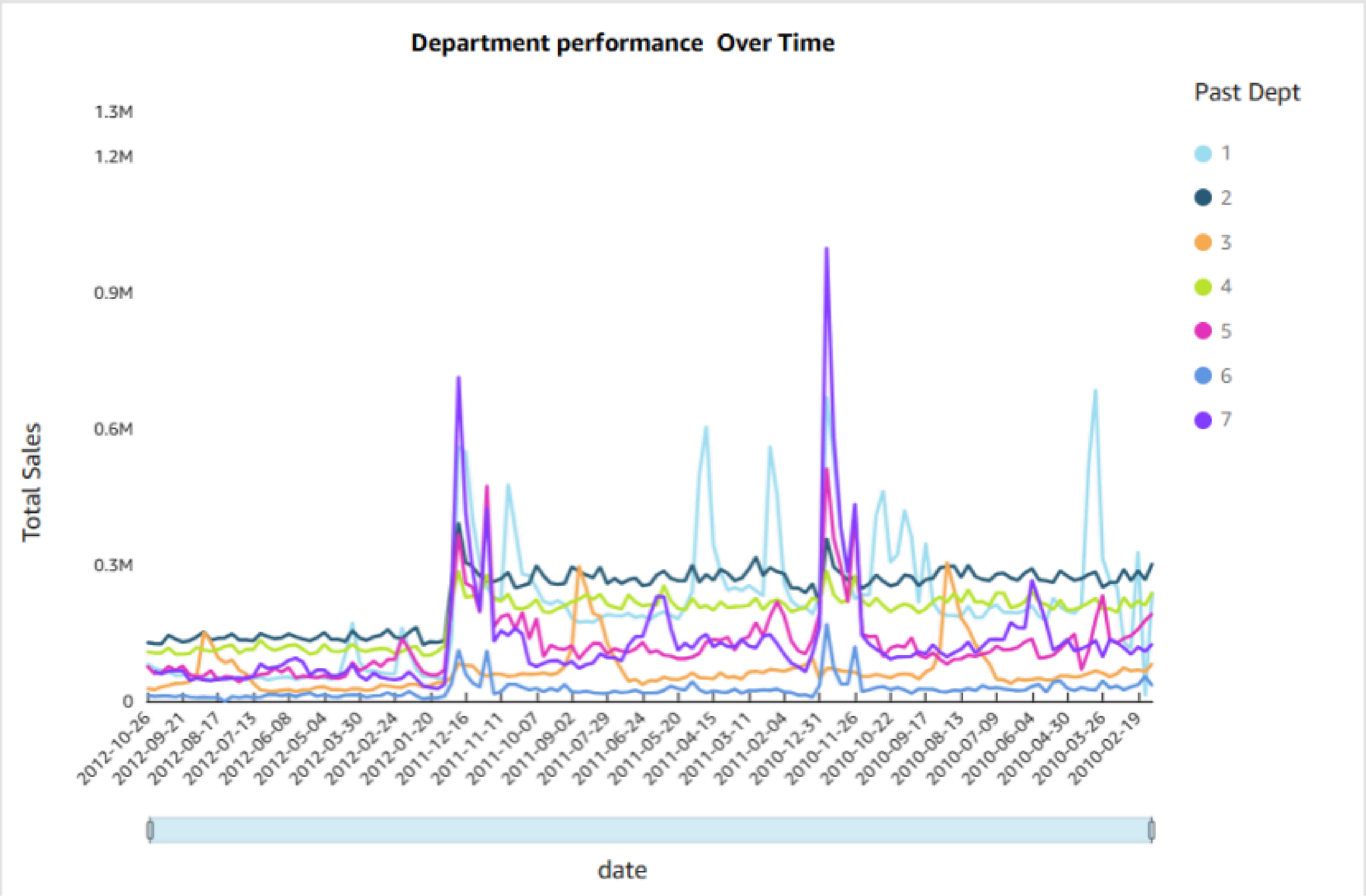


TECH CONSULTING

DATA ANALYSIS

DEPARTMENT PERFORMANCE OVER TIME ANALYSIS

Insights: Among top ten store , the first department that performed better than any others is store with ID=7 over time

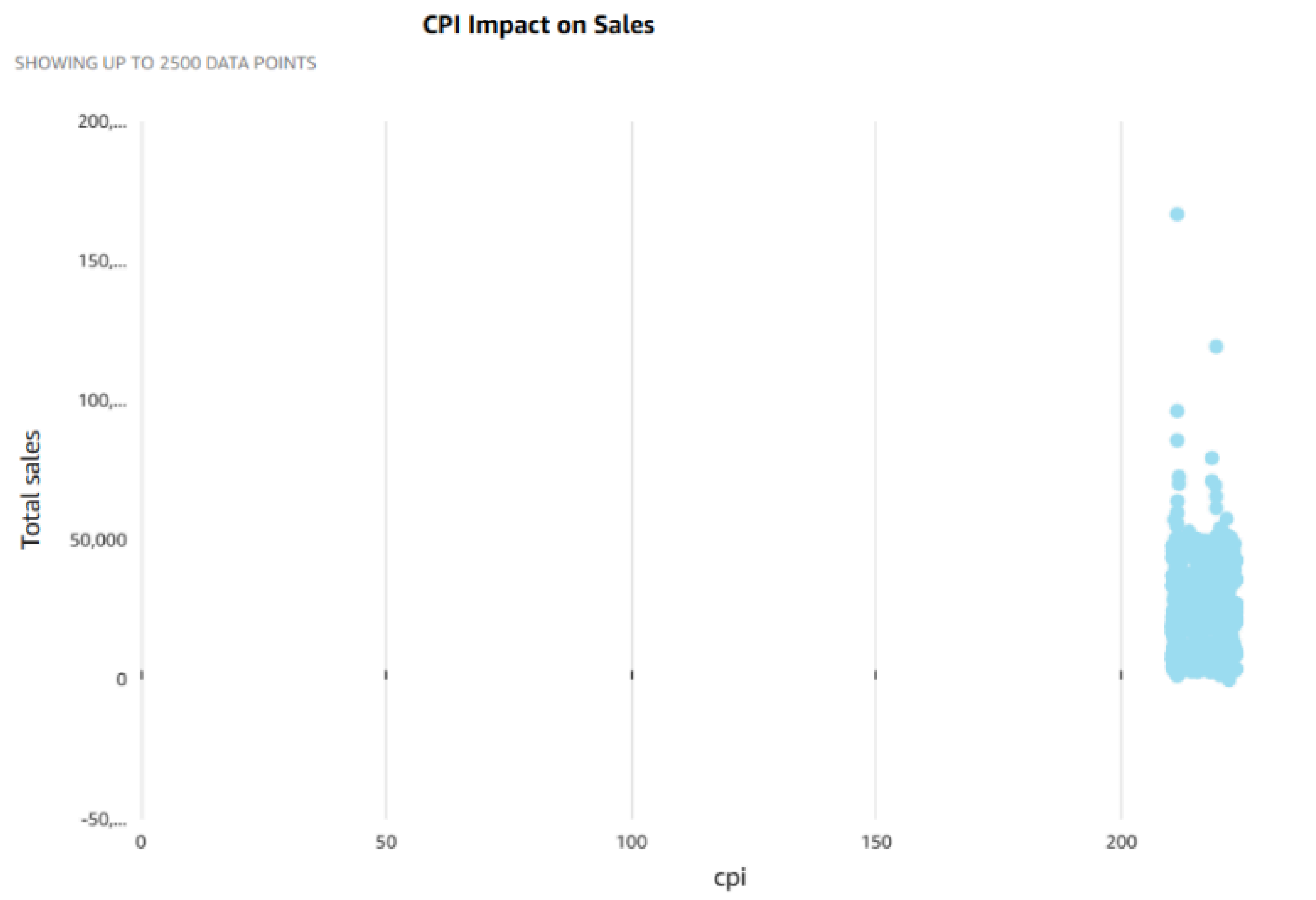


TECH CONSULTING

DATA ANALYSIS

CPI VS. TOTAL SALES ANALYSIS

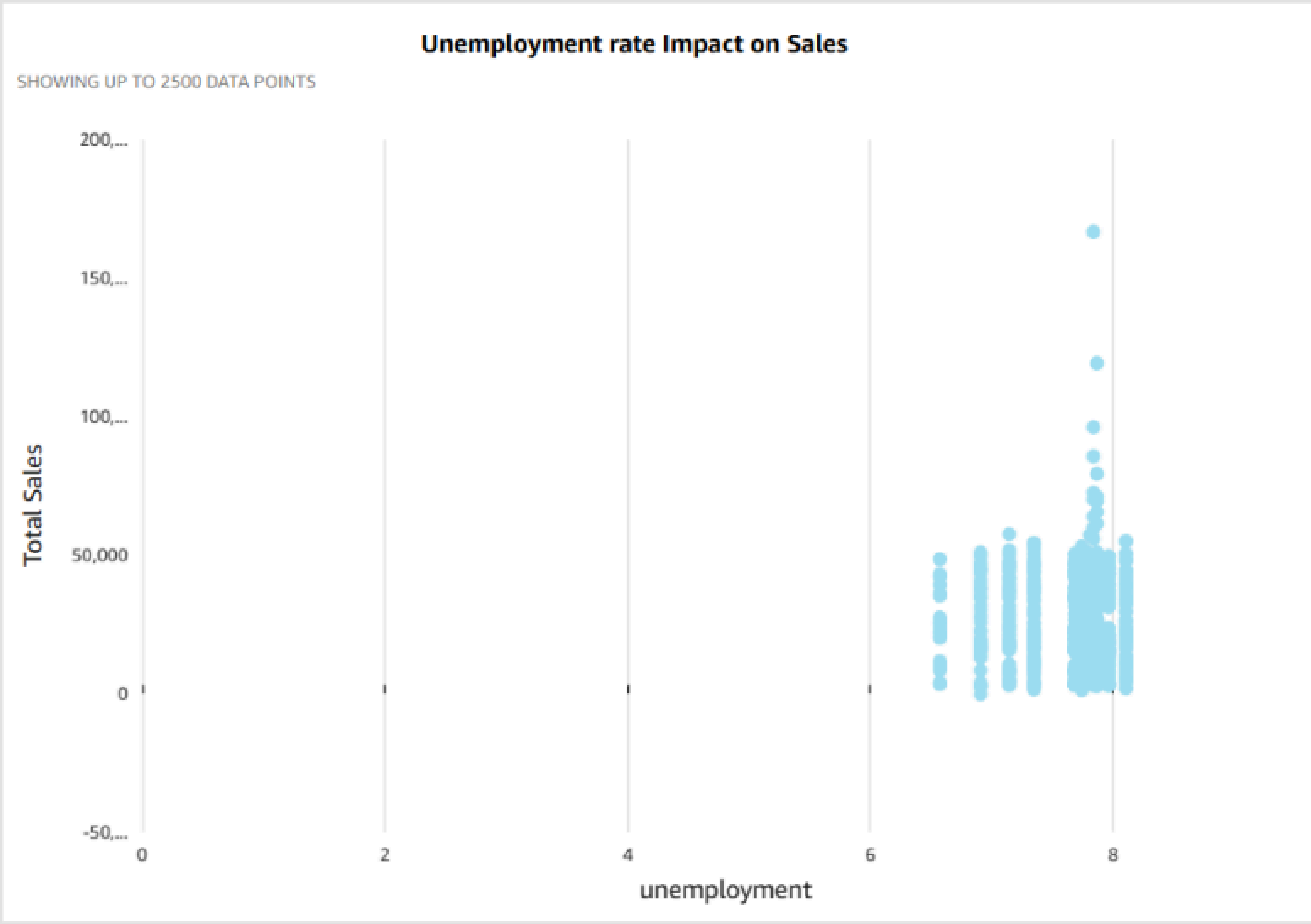
No Clear Trend: The scatter plot indicates **no strong or visible correlation** between the Consumer Price Index (CPI) and total sales.



DATA ANALYSIS

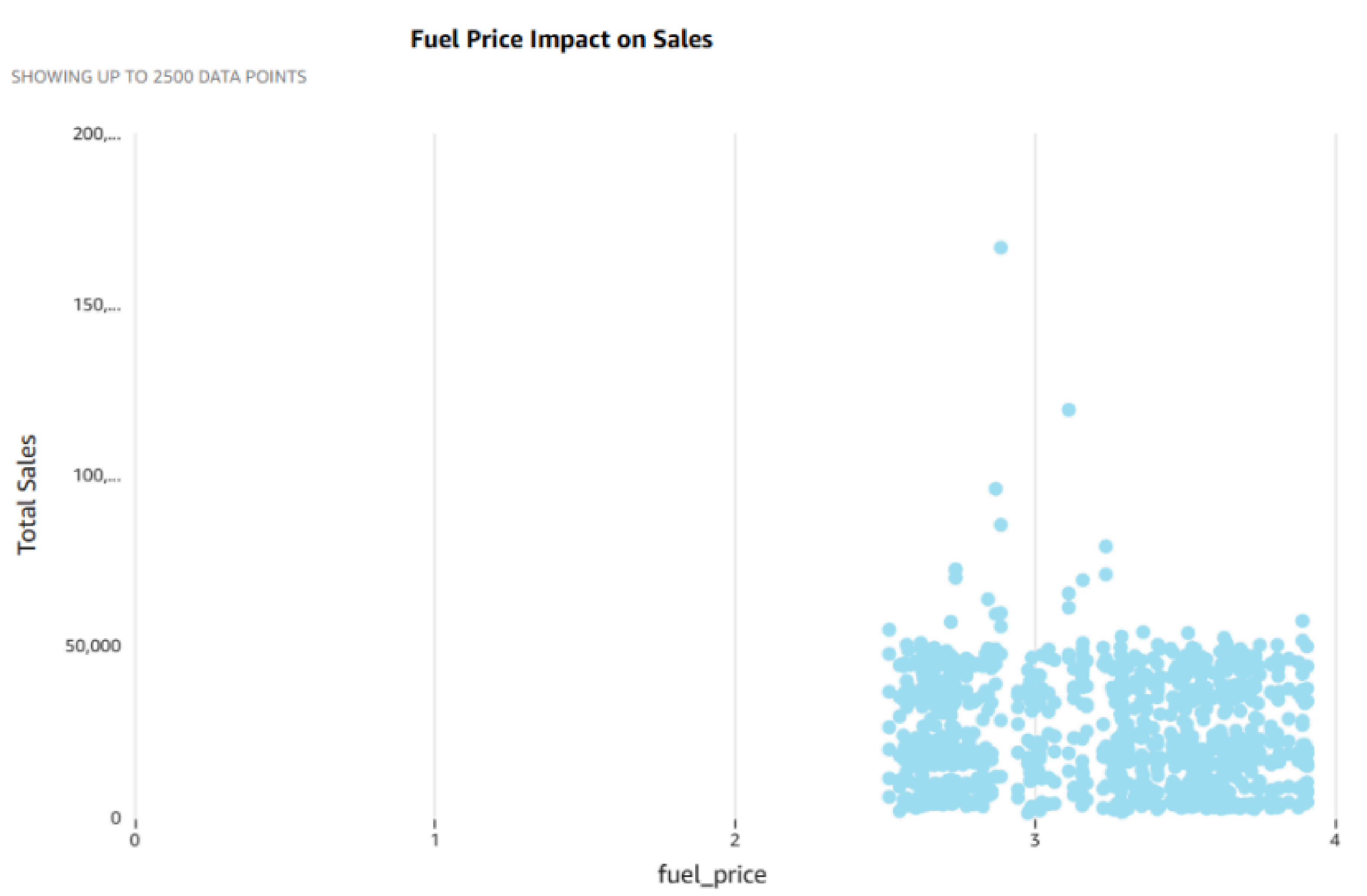
ENEMPLOYMENT RATE VS. TOTAL SALES ANALYSIS

No Clear Trend: The scatter plot indicates **no strong or visible correlation** between the enemployment rate and total sales.



FUEL VS. TOTAL SALES ANALYSIS

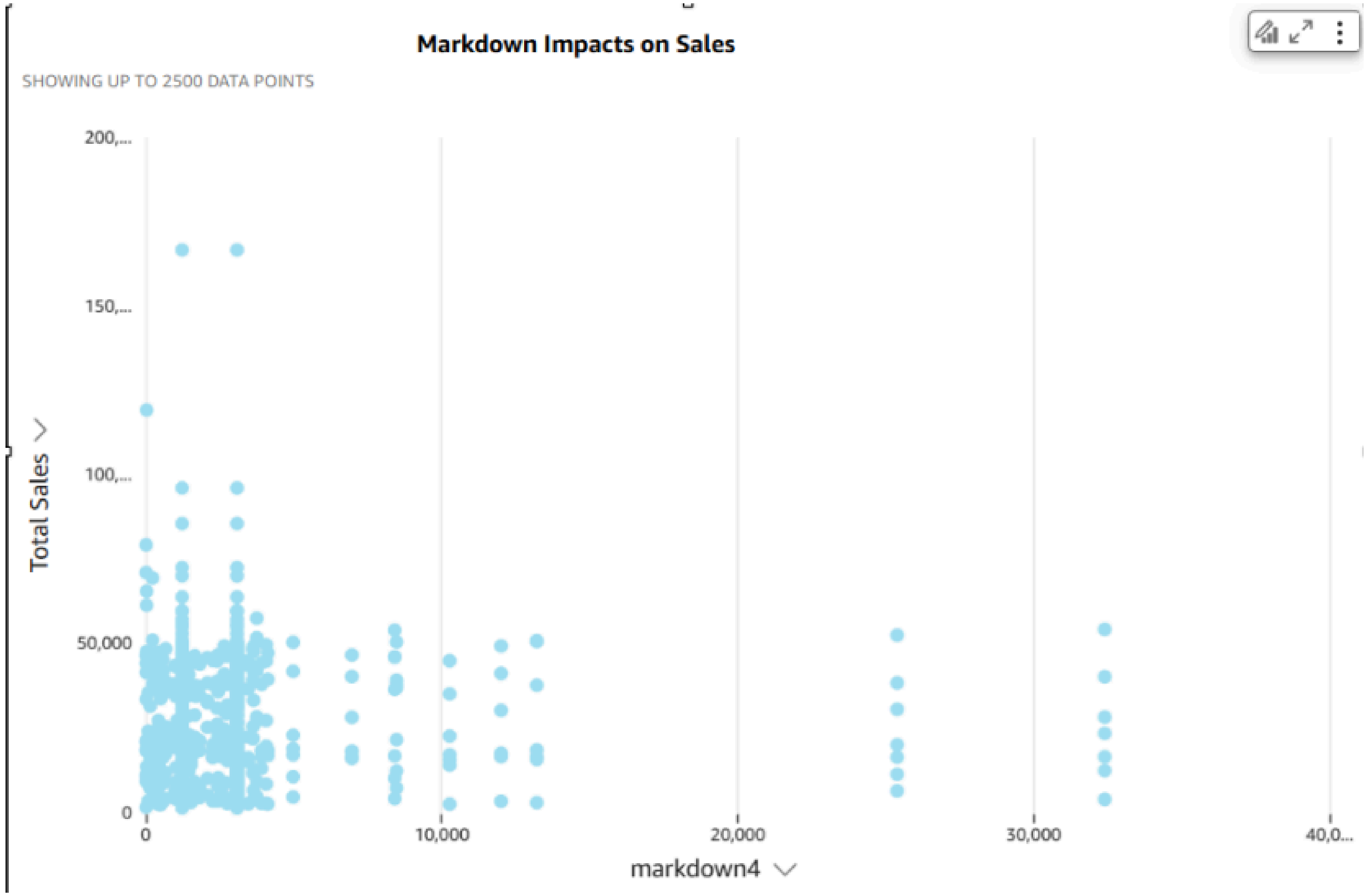
No Strong Correlation: The scatter plot shows **no clear relationship** between fuel prices and total sales, as data points are dispersed across various price levels without a discernible pattern.



DATA ANALYSIS

MARKDOWN VS. TOTAL SALES ANALYSIS

No Strong Correlation: The scatter plot shows **no clear relationship** between markdown and total sales, as data points are dispersed across various price levels without a discernible pattern.



KINESIS STREAMING

USE CASE

- Kinesis streams **real-time sales data** from all Walmart stores, enabling instant updates and analysis of store performance.
- The system identifies sudden **sales spikes or drops**, alerting Walmart to potential inventory shortages or operational issues.
- Real-time tracking of **holiday sales trends**, allowing Walmart to assess the impact of promotions and adjust strategies instantly.
- Kinesis-powered alerts notify Walmart teams of **critical changes in sales via SNS subscriber**, enabling rapid decision-making and stock adjustments.

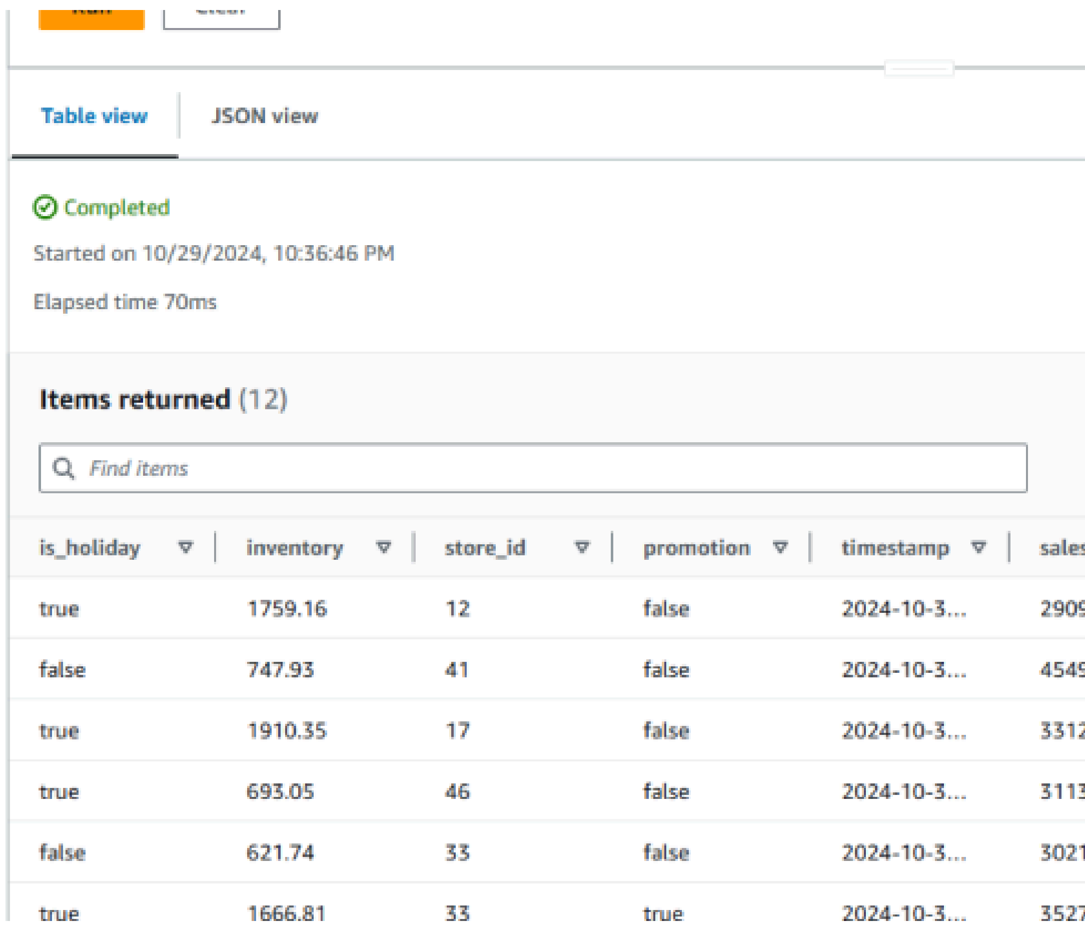
TECH CONSULTING

KINESIS STREAMING

IMPLEMENTATION

- Utilized Kinesis Producer lambda function to generate hourly sales to send to Kinesis Data Stream.
- Created Kinesis Consumer to Consume the data from producer with attached Kinesis trigger to it so that it can can capture data as soon as they are generated.
- Programmed consumer to store generated data into DynamoDB
- Integrated SNS in Consumer for me to get acknowledgment via Email subscriber.

Kinesis consumer result data in DynamoDB with SNS subscriber email



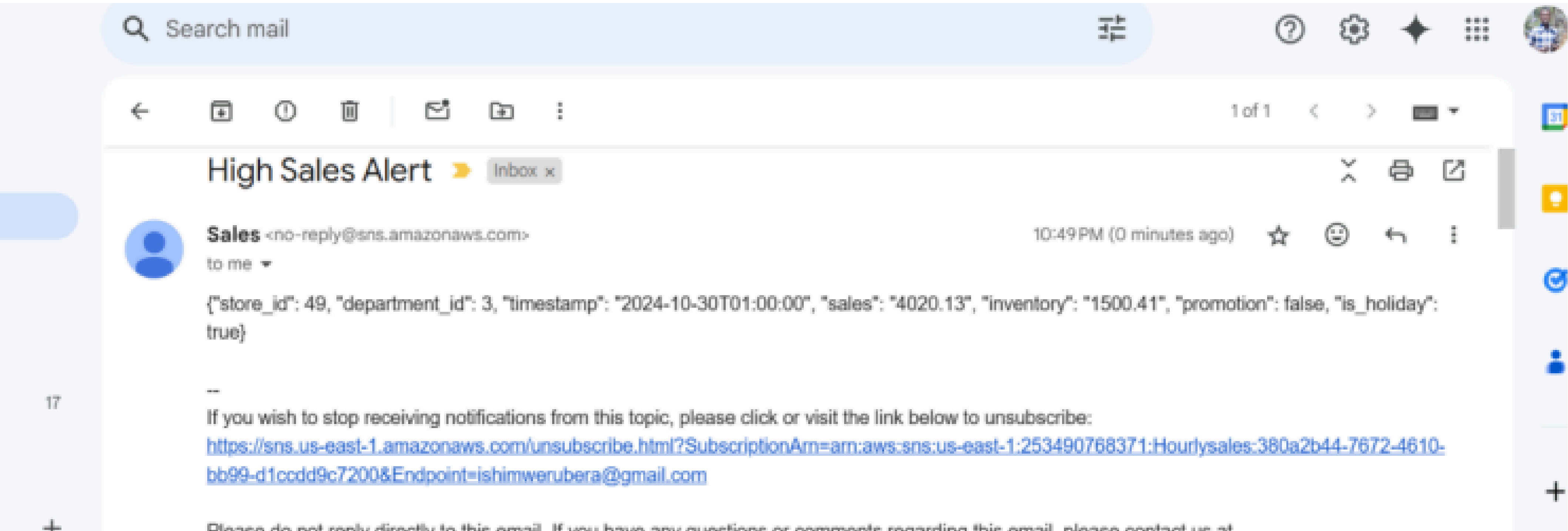
✓ Completed

Started on 10/29/2024, 10:36:46 PM

Elapsed time 70ms

Items returned (12)

is_holiday	inventory	store_id	promotion	timestamp	sales
true	1759.16	12	false	2024-10-3...	2909
false	747.93	41	false	2024-10-3...	4549
true	1910.35	17	false	2024-10-3...	3312
true	693.05	46	false	2024-10-3...	3113
false	621.74	33	false	2024-10-3...	3021
true	1666.81	33	true	2024-10-3...	3527



High Sales Alert

Sales <no-reply@sns.amazonaws.com>

to me

10:49 PM (0 minutes ago)

{"store_id": 49, "department_id": 3, "timestamp": "2024-10-30T01:00:00", "sales": "4020.13", "inventory": "1500.41", "promotion": false, "is_holiday": true}

--

If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://sns.us-east-1.amazonaws.com/unsubscribe.html?SubscriptionArn=arn:aws:sns:us-east-1:253490768371:Hourlysales:380a2b44-7672-4610-bb99-d1ccdd9c7200&Endpoint=ishimwerubera@gmail.com

Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at

TECH CONSULTING

CONCLUSION

In conclusion, big data technologies provide powerful tools for gaining real-time insights efficiently. By leveraging Kinesis Streaming, we can process sales data in real time, enabling timely decision-making. Additionally, using **Parquet** for data storage proves far more efficient than CSV, offering better memory usage and faster queries.

From the insights gained, it's clear that **holidays have a significant impact** on Walmart's sales, with a notable spike in December. To capitalize on this, I recommend **ensuring ample inventory** during this period to meet increased demand and maximize revenue.

CHALLENGES

1.Layer Dependency Issues like Frequent access restrictions hindered smooth operations in Lambda functions.

2. Hard to use Quick sight due to that it took time to generate a graph and it is costly.

3. Remember to give access roles using IAM.

FUTURE IMPROVEMENT

1. **Enhanced Real-Time Analytics:** Incorporating advanced machine learning
2. **Optimized Data Storage:** Further optimization of data storage by implementing a hybrid system of Parquet and ORC files based on usage patterns.
3. **Improved Alert System**: Expanding the alerting mechanism to include predictive alerts based on sales patterns, weather forecasts, or regional events could further enhance inventory management and promotional planning.
4. **Implementing CDC** using kinesis streaming
5. **Implementing:** code pipeline for CI/CD.
6. **Implementing:** Implementing data analysis for data from dynamoDB.

TECH
CONSULTING

THANK YOU!

Q&A