KLE Society's

KLE Technological University



**An Industry Training Report**

**On**

**Customer Data Platform on GCP**

*Submitted in partial fulfillment of the requirement for the degree of*

**Bachelor of Engineering in**

**Computer Science and Engineering**

Submitted By

Name: P Suhas Rao

USN: 01FE21BCS161

**Under the guidance of**

**Dr. Manohar Madgi**

COMPUTER SCIENCE &ENGINEERING,

HUBBALLI–580 031 (India).

Academic year 2024-25



**B. V. Bhoomaraddi College Campus, Vidyanagar, Hubballi - 580031.**

**Karnataka (India)**

**COMPUTER SCIENCE AND ENGINEERING**

**CERTIFICATE**

This is to certify that Industry Training entitled “**Customer Data Platform on GCP**” is a Bonafide work carried out by the student Mr. P Suhas Rao bearing USN 01FE21BCS161 in partial fulfillment of the completion of 8th semester B.E. course during the year 2024 – 25 at **Cognizant**. The Industry Training report has been approved as it satisfies the academic requirement with respect to the training work prescribed for the above said course.

**Name of the Guide HoD**

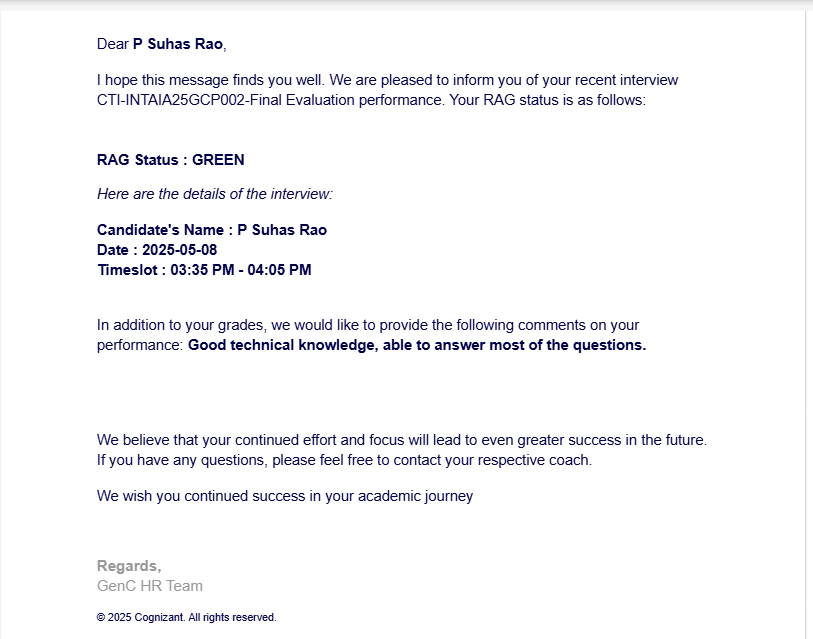
Dr. Manohar Madgi

**Name of the examiners Signature with date**

1 -------------------------- ------------------ 1 -------------------------- ---------------

2 -------------------------- ----------------- 2 -------------------------- ---------------

**CERTIFICATE**

****

**DECLARATION**

I hereby declare that the Industrial Training Report entitled “**Customer Data Platform on GCP**” is an authentic record of my own work as requirements of Industry Training during the period from 14-02-2025 to 15-05-2025 for the award of the degree of B.E. Under the guidance of Dr. Manohar Madgi.

**A black text with a white background

AI-generated content may be incorrect. P Suhas Rao**

**Date: 01-06-2025**

**01FE21BCS161**

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of a number of individuals whose professional guidance and encouragement helped me in the successful completion of this report work.

I also take this opportunity to thank Dr. Vijaylakshmi M, Professor and Head, Department of Computer Science and Engineering for having provided us academic environment which nurtured our practical skills contributing to the success of our project.

I sincerely thank our guide Dr. Manohar Madgi for his/her guidance and wholehearted co-operation during the course of completion.

I sincerely thank Geebiga , Cognizant for her support, inspiration and wholehearted co-operation during the course of completion.

My gratitude will not be complete without thanking our beloved parents, our seniors and our friends who have been a constant source of aspirations.

i

**About Company**

**Cognizant Technology Solutions Corporation**:

Cognizant Technology Solutions Corporation is a leading American multinational information technology services and consulting company. Founded in 1994 in Chennai, India, it is now headquartered in Teaneck, New Jersey, USA, and operates globally.

**Core Business**: Cognizant helps clients modernize their technology, optimize business processes, and transform customer and employee experiences. We achieve this by leveraging cutting-edge digital capabilities to enable businesses to thrive in a rapidly changing market.

**Services and Expertise:** Cognizant offers a wide range of services designed to help businesses modernize technology, reimagine processes, and transform experiences. Key service areas include:

* **Digital Transformation:** Guiding clients in their digital journeys.
* **Artificial Intelligence (AI) & Data:** Solutions leveraging AI, machine learning, and advanced data management. This includes generative AI and agentic AI systems.
* **Cloud Services:** Cloud-native transformation and hybrid cloud solutions.
* **Software Engineering:** Developing and modernizing software applications.
* **Internet of Things (IoT):** Harnessing connectivity to optimize business processes.
* **Consulting:** Strategic and operational consulting across various industries.
* **Cybersecurity:** Comprehensive solutions for risk, threat, and vulnerability management.
* **Business Process Services (BPS):** Optimizing and managing business processes.
* **Quality Engineering & Assurance:** Ensuring the quality and reliability of IT systems.

**Financial Snapshot (Recent Data):**

* **Full-Year 2024 Revenue:** Approximately $19.7 billion.
* **Q1 2025 Revenue:** $5.115 billion.

**Industry Recognition:** Cognizant is consistently recognized as a leader in the IT services industry, being a component of major indices such as the NASDAQ-100, S&P 500, Forbes Global 2000, and Fortune 500. This reflects our strong market position and consistent performance.

ii

**Table of Contents**

**Chapter No.**

**Page No.**

**Table of Contents**

**1 Introduction 1**

**2 Tools and Technology used 2**

**3 Snapshots 3**

**4 Results and Discussions 19**

**5 Conclusion and Future Scope 29**

**6 References 30**

iii

**List of Figures**

1. **Figure 1** – High Level Architecture
2. **Figure 2** – Buckets
3. **Figure 3** – Data in respective buckets
4. **Figure 4** – BigQuery Datasets
5. **Figure 5** – Tables in CRM
6. **Figure 6** – Tables in Web
7. **Figure 7** – Tables in Transaction
8. **Figure 8** – Code for Publishing Data to Pub/Sub Topic
9. **Figure 9** – Employee Streaming Dataflow Job
10. **Figure 10** – Data Streamed into BigQuery Table
11. **Figure 11** – Customer.js
12. **Figure 12** – Customer.json
13. **Figure 13** – Customer Streaming Dataflow Job
14. **Figure 14** – Customer Table Schema
15. **Figure 15** – Transformed Customer Data
16. **Figure 16** – Click\_Events.js
17. **Figure 17** – Click\_Events.json
18. **Figure 18** – Click Events Streaming Dataflow Job
19. **Figure 19** – Click Events Table Schema
20. **Figure 20** – Transformed Click Events Data
21. **Figure 21** – Key Generation in KMS
22. **Figure 22** – Python Code for Encryption
23. **Figure 23** – Bucket Containing Encrypted Data
24. **Figure 24** – Encrypted Transactional Data in BigQuery
25. **Figure 25** – Encrypted Orders Data
26. **Figure 26** – Encrypted Orders Data
27. **Figure 27** – Encrypted Order Items Data
28. **Figure 28** – Encrypted Payments Data
29. **Figure 29** – Permissions Provided for Service Account
30. **Figure 30** – Storage Admin Role
31. **Figure 31** – Bucket Accessed by User
32. **Figure 32** – KPI 1
33. **Figure 33** – KPI 2
34. **Figure 34** – KPI 3
35. **Figure 35** – KPI 4
36. **Figure 36** – KPI 5
37. **Figure 37** – KPI 6
38. **Figure 38** – KPI 7
39. **Figure 39** – KPI 8
40. **Figure 40** – Customer Overview Dashboard
41. **Figure 41** – Web Analytics Dashboard
42. **Figure 42** – Transaction Insights Dashboard
43. **Figure 43** – Monitoring Dashboard
44. **Figure 44** – Cloud Storage Monitoring
45. **Figure 45** – BigQuery Error Alerting
46. **Figure 46** – Logging Dashboard

iv

Chapter 1:

# Introduction

In today’s data-driven business landscape, understanding customer behavior and preferences is crucial for delivering personalized experiences and driving strategic decisions. The **Customer Data Platform (CDP)** developed on **Google Cloud Platform (GCP)** is a comprehensive solution designed to centralize, process, and analyze customer-related data from multiple sources in real time. This platform empowers organizations to unify customer data, ensure data security, and derive actionable insights through advanced analytics and visualization tools.

The primary objective of this project is to build a scalable and secure data pipeline that ingests real-time data from CRM systems, web analytics platforms, and transactional databases. The data is processed using GCP-native services and stored in **BigQuery**, a powerful data warehouse that supports complex querying and analytics. Sensitive information is encrypted using **Cloud Key Management Service (KMS)** to ensure compliance with data privacy regulations.

The CDP architecture is designed to be modular and cloud-native, leveraging services such as **Pub/Sub** for real-time data ingestion, **Dataflow** for stream and batch processing, and **Looker Studio** for interactive dashboard creation. The system also integrates **Cloud Monitoring and Logging** to ensure operational reliability and provide real-time alerts on system performance.

This platform is particularly beneficial for data analysts, engineers, and business managers. Analysts can explore customer trends and KPIs through dashboards, engineers can maintain and scale the data pipelines, and managers can make informed decisions based on real-time insights. The CDP not only enhances data governance and operational efficiency but also serves as a foundational layer for future integration with machine learning models and advanced analytics tools.

By implementing this CDP on GCP, the project demonstrates how cloud technologies can be harnessed to build robust, secure, and intelligent data systems that support modern business needs.

Chapter 2:

# Tools and Technology Used

The following tools and technologies were utilized in the development and deployment of the Customer Data Platform on GCP:

* **Google Cloud Platform (GCP)** – The core cloud infrastructure hosting all services.
* **Google Cloud Pub/Sub** – Used for real-time data ingestion and message streaming.
* **Google Cloud Dataflow** – Employed for both batch and streaming data processing.
* **Google BigQuery** – Acts as the central data warehouse for storing and querying processed data.
* **Google Cloud KMS (Key Management Service)** – Ensures encryption of sensitive fields to maintain data privacy and compliance.
* **Google IAM (Identity and Access Management)** – Manages access control and permissions across services.
* **Looker Studio** – Used to build interactive dashboards for visualizing business KPIs.
* **Google Cloud Monitoring & Logging** – Provides observability, error tracking, and alerting for system health and performance.
* **Python & Google Colab** – Used for scripting data transfers and automation tasks.
* **JavaScript & JSON Schema** – Used for defining transformation logic and data schemas in Dataflow jobs.

Chapter 3:

**Snapshots**

# Architecture Diagram

A diagram of a cloud monitoring system

AI-generated content may be incorrect.

(1) High Level Architecture

# Data Sources

* **CRM Data** (CSV format) – 9 tables -> customers, addresses, demographics, preferences, loyalty program, engagement, customer feedback, customer segments, data login activity.
* **Web Analytics Data** (JSON format) - 5 tables - > click events, conversion events, device information, page views, session data.
* **Transactional Data** (CSV format) – 15 tables -> orders, payments, invoices, shipments, returns, refunds, shipment tracking, order returns, payment transaction, order items, inventory, product catalog, promotions, supplier data, inventory logs.

**Note: All data are generated using faker library of python.**

Data Ingestion and Storage

Raw data files were initially stored in Google Drive. These files were programmatically transferred to Google Cloud Storage (GCS) Buckets namely crm\_sample, web\_analy, suhas\_transaction using a Python script executed in Google Colab. This automated approach ensured secure and efficient migration of data from Drive to the cloud.

Once the data was available in GCS, it was imported into BigQuery by creating three separate datasets corresponding to the source systems:

**CRM**

**Web**

**Transaction**

Each dataset was structured with appropriate tables reflecting the schema of the original source data. This served as the raw data layer for the analytics workflow.

A screenshot of a computer

AI-generated content may be incorrect.

(2) Buckets

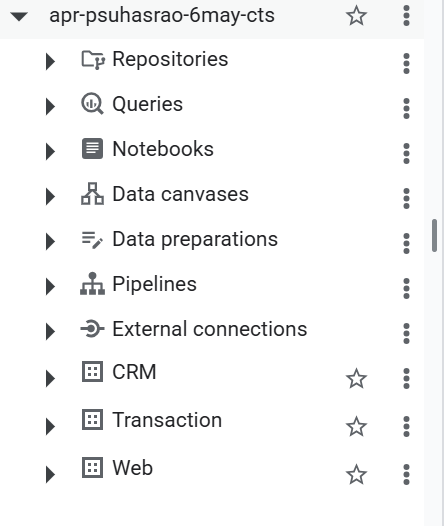
A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

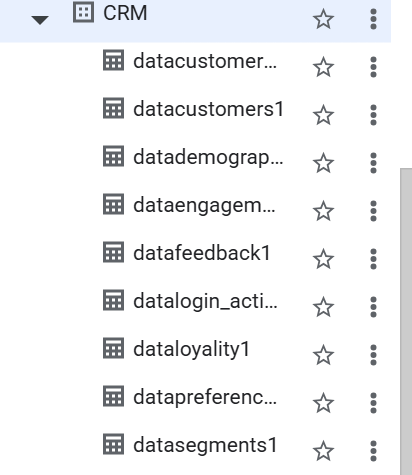
AI-generated content may be incorrect.

# A screenshot of a computer AI-generated content may be incorrect.

(3) Data in respective buckets



(4) BigQuery Datasets



(5) Tables in CRM (6) Tables in Web

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

(7) Tables in Transaction

Real-time Streaming Simulation Using Pub/Sub and Dataflow

To simulate a real-time data streaming pipeline, I utilized Google Cloud Pub/Sub and Dataflow. However, due to the high cost and time involved in streaming the full dataset of 20 lakh records, a scaled-down version was implemented using a sample CSV file (employee.csv) for demonstration purposes.

**Pub/Sub Topic Creation and Data Publishing**

* A Pub/Sub topic named emp\_ex was created.
* A Python script was developed to read the employee.csv file and publish records one by one to the emp\_ex topic.
* A delay of 10 seconds was introduced between each record to simulate real-time ingestion behavior.

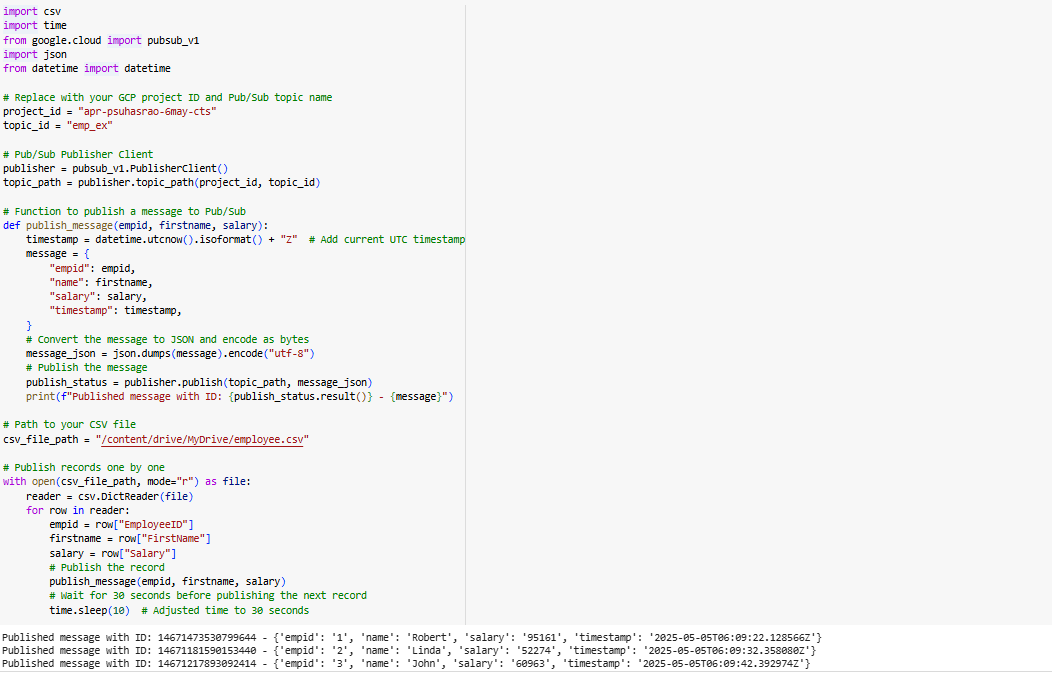
**Streaming Dataflow Job**

* An **empty table** was created in BigQuery with the schema corresponding to the structure of the employee data.
* A **Dataflow job** was launched using the **"Pub/Sub to** **BigQuery" template**.

Input: Pub/Sub topic emp\_ex

Output: BigQuery table

* As the script ran, data published to Pub/Sub **was ingested in near real-time** by Dataflow and populated into the BigQuery table.



(8) Code For Publishing Data to Pub/Sub Topic

A screenshot of a computer

AI-generated content may be incorrect.

(9) Employee Streaming Dataflow Job

A screenshot of a computer

AI-generated content may be incorrect.

(10) Data Streamed into BigQuery Table

Data Processing with Dataflow (Batch Mode)

Although the intended architecture involved real-time streaming using Pub/Sub and Dataflow, the available data was static. Hence, a batch processing approach was adopted to simulate the ingestion pipeline.

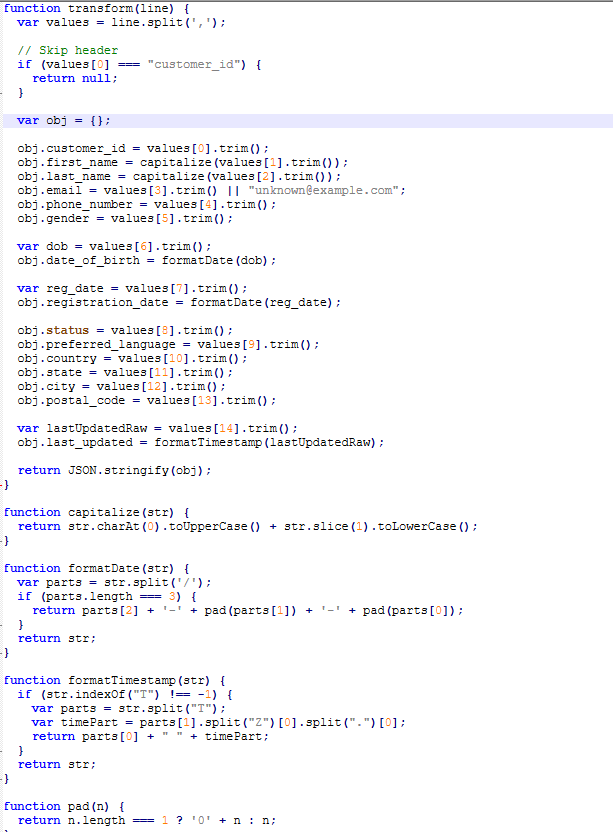
Each source (e.g., Customer, Click Events) included:

* A **JavaScript** **(.js) transformation file** defining how each record should be parsed and cleaned.
* A **JSON schema file** specifying the structure for the BigQuery table.

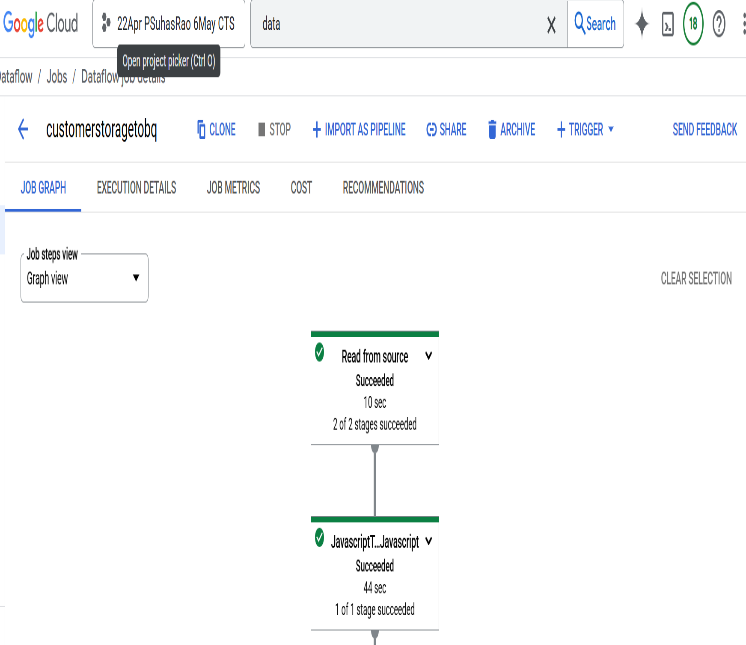
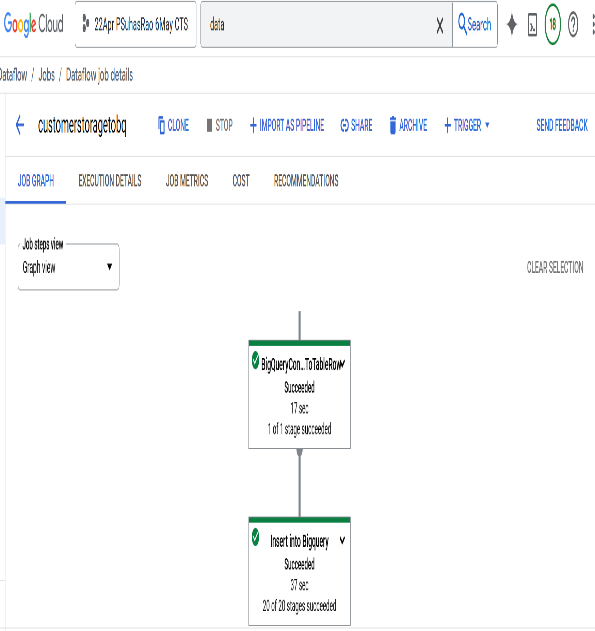
The **Cloud Dataflow "Text Files on Cloud Storage to BigQuery" template** was used to launch Dataflow jobs. These jobs:

* Pulled raw data from Cloud Storage.
* Applied the transformation logic using the .js files.
* Loaded the clean and structured data into BigQuery according to the defined .json schema.

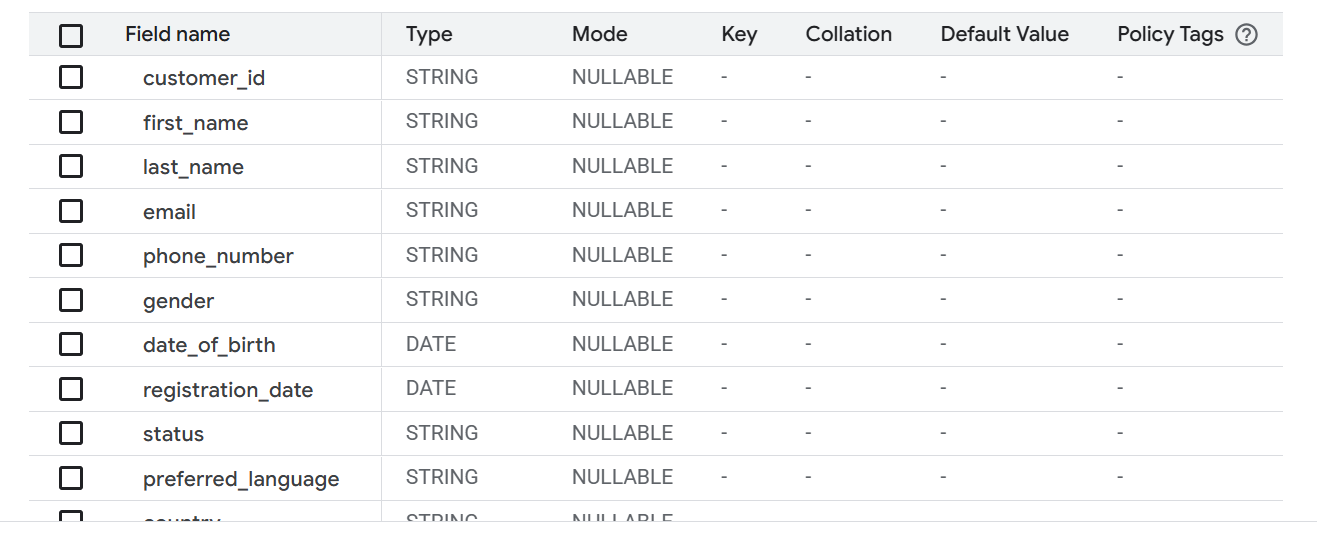
This method effectively mimicked a real-time pipeline using batch jobs, ensuring schema enforcement and preprocessing before storage in BigQuery.

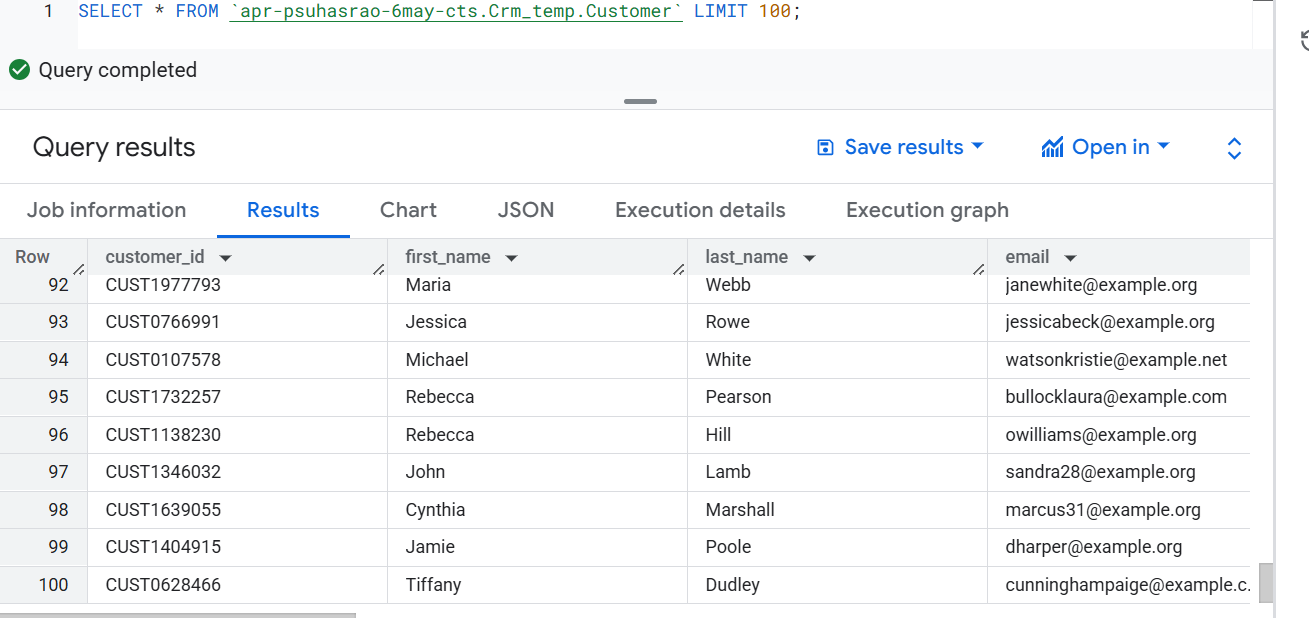
(11) Customer.js (12) Customer.json

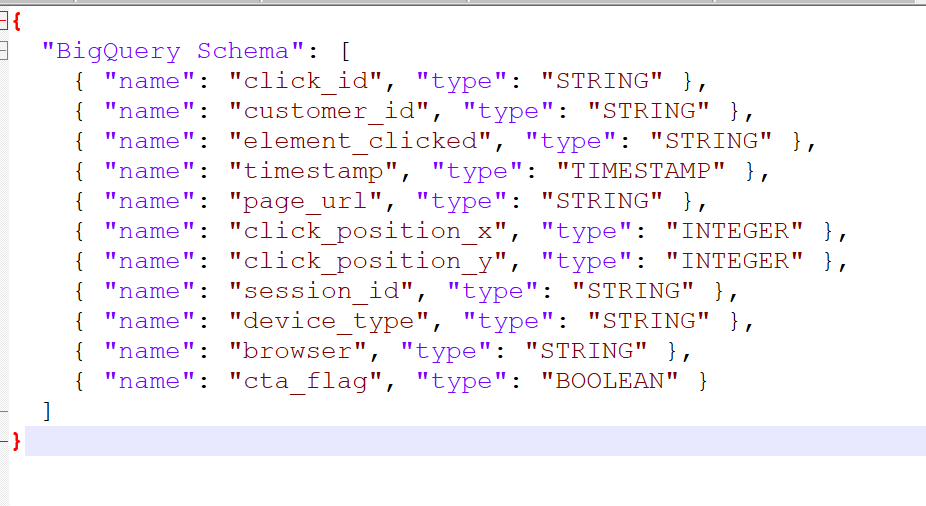
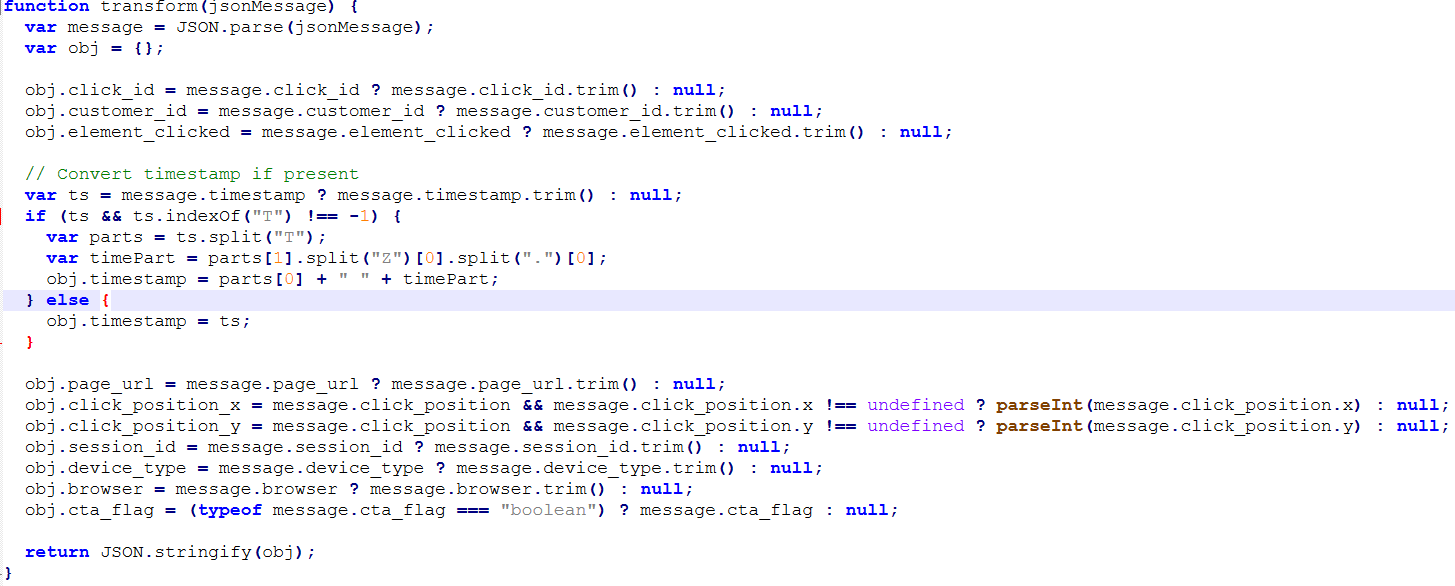
(13) Customer Streaming Dataflow Job



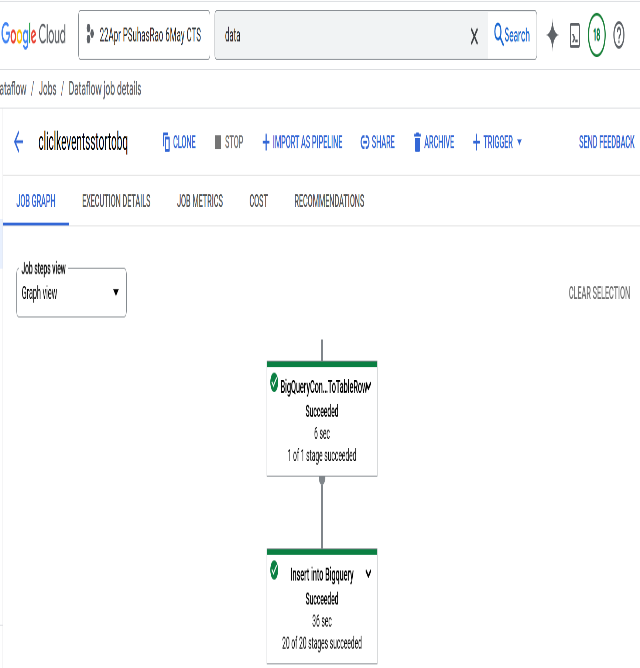
(14) Customer Table Schema



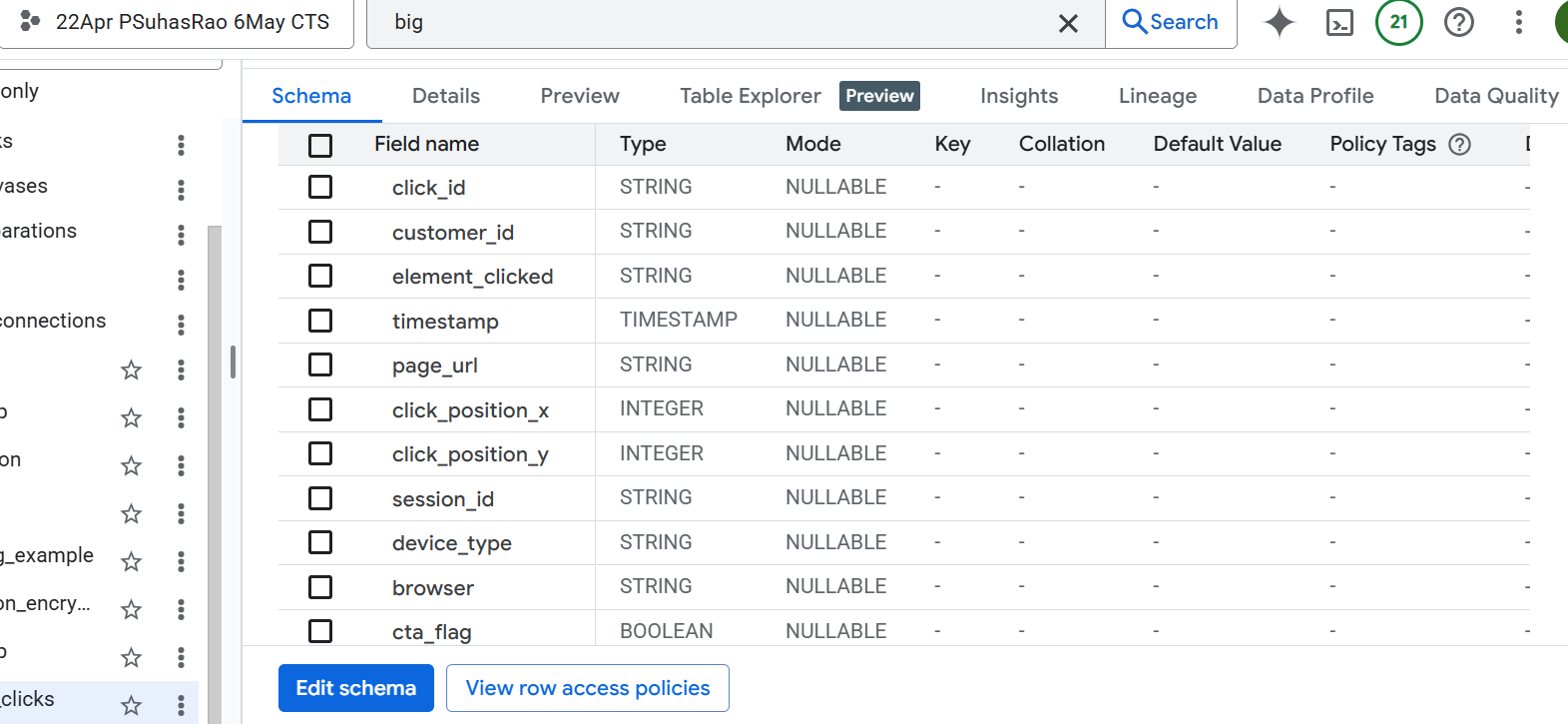
(15) Transformed Customer Data



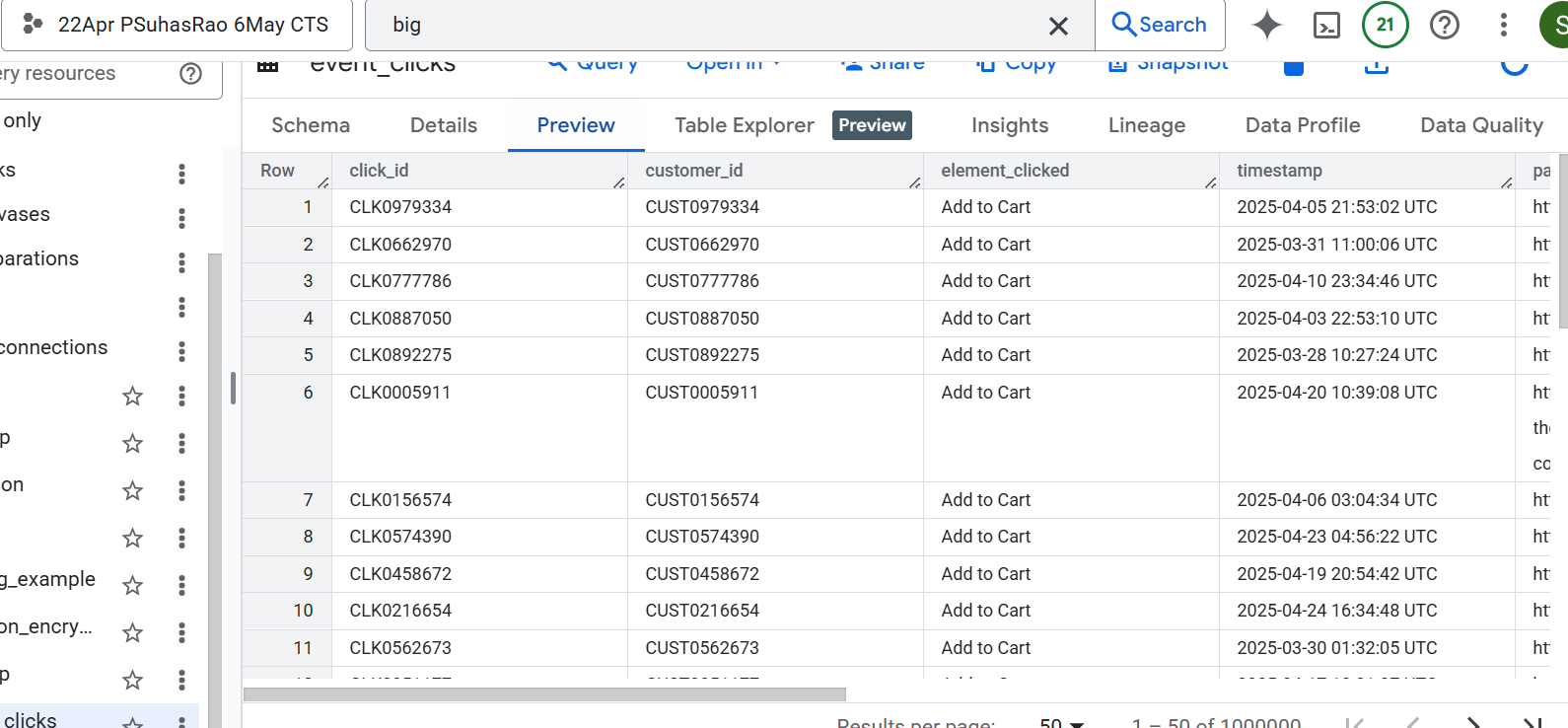
(16) Click\_Events.js (17) Click\_Events.json

(18) Click events Streaming Dataflow Job



(19) Click Events Table Schema



(20) Transformed Click Events Data

Encryption of Sensitive fields in Transactional Data

The following sensitive fields were encrypted before storage:

**orders1**: customer\_id, shipping\_address, billing\_address, payment\_method

**order\_items**: supplier\_id, supplier\_name

**payments1**: customer\_id, payment\_method, transaction\_id, authorization\_code, account\_number\_masked, payment\_ip

**invoices1**: customer\_id, billing\_address, payment\_method, bank\_name, account\_number\_masked

**shipments1**: customer\_id, tracking\_number, origin, destination

**returns1**: customer\_id, return\_reason, processed\_by

**refunds1**: customer\_id, bank\_name, account\_number\_masked, refund\_reason

**inventory**: supplier\_id, batch\_number

**product\_catalog**: supplier\_id

**promotions**: promotion\_code

**payment\_transactions1**: customer\_id, account\_number\_masked, transaction\_reference

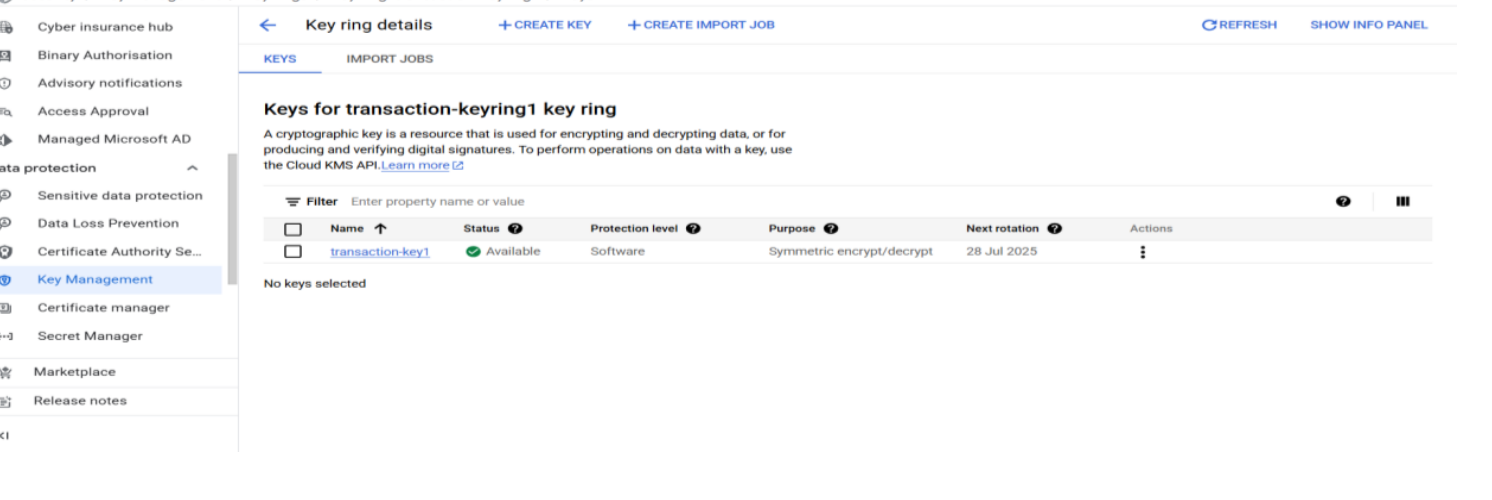
**shipment\_tracking1**: customer\_id, tracking\_number, delivery\_address

**order\_returns1**: customer\_id, pickup\_address

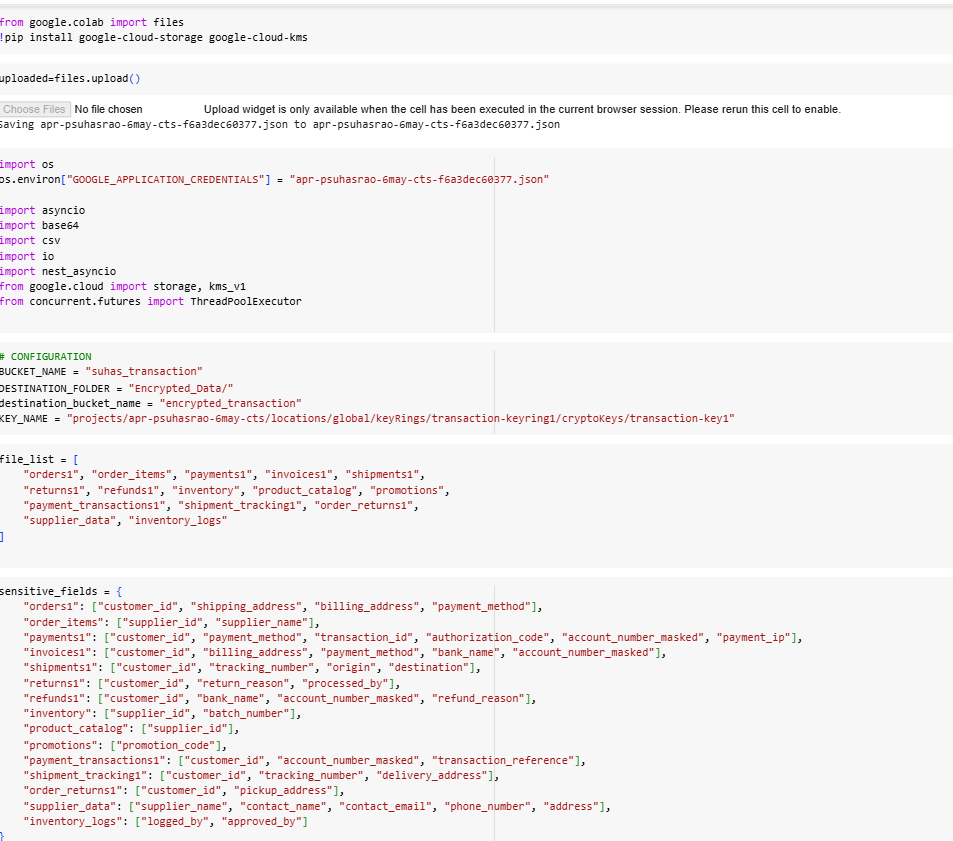
**supplier\_data**: supplier\_name, contact\_name, contact\_email, phone\_number, address

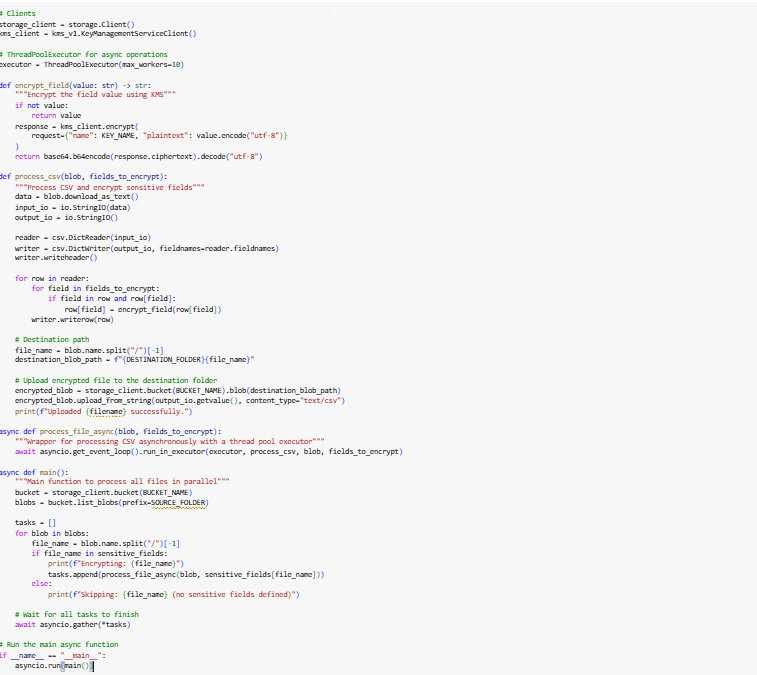
**inventory\_logs**: logged\_by, approved\_by

These fields were identified as containing personally identifiable information (PII) or confidential data and were encrypted to ensure data security during storage and processing.

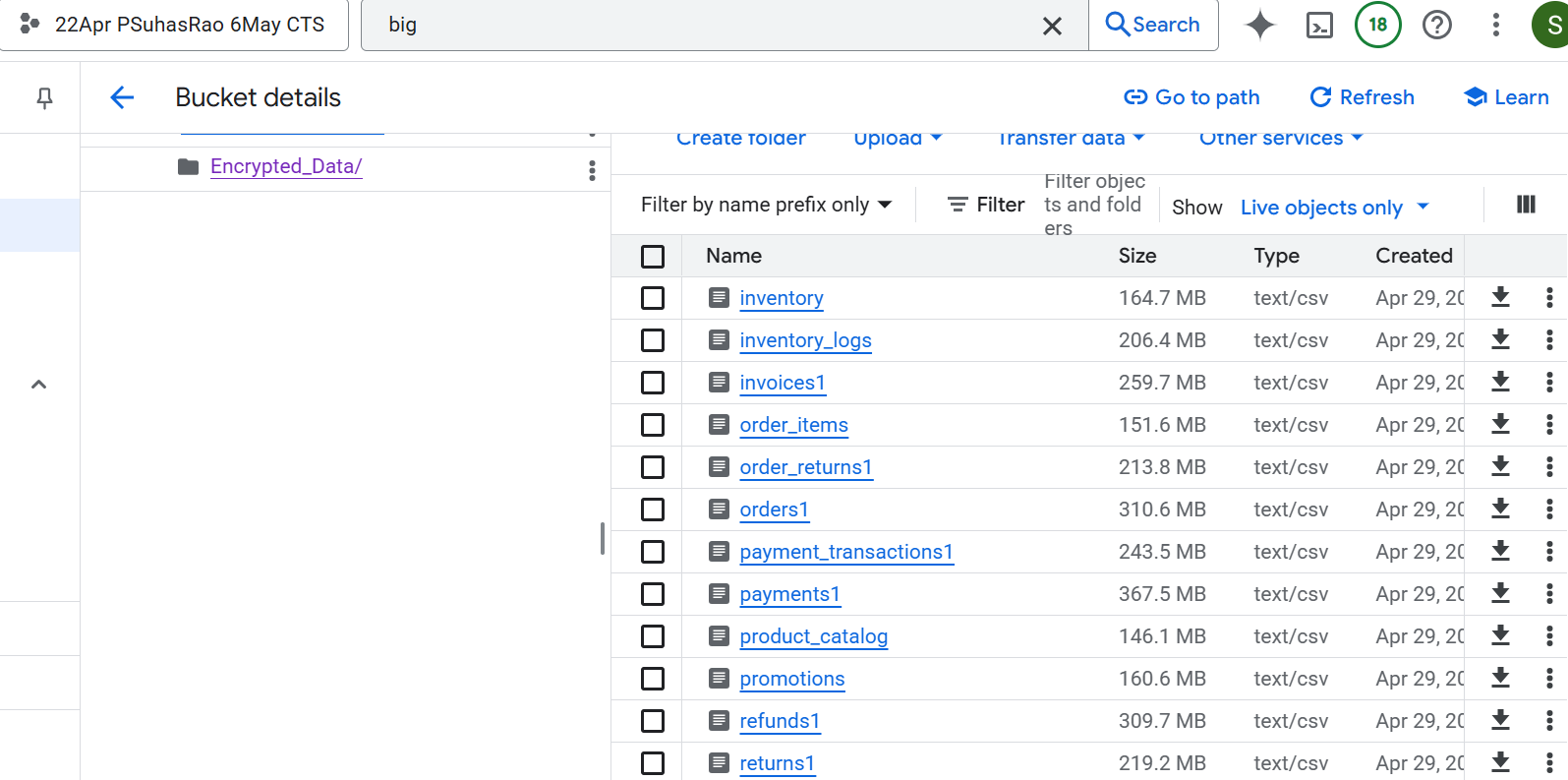


(21) Key Generation in KMS

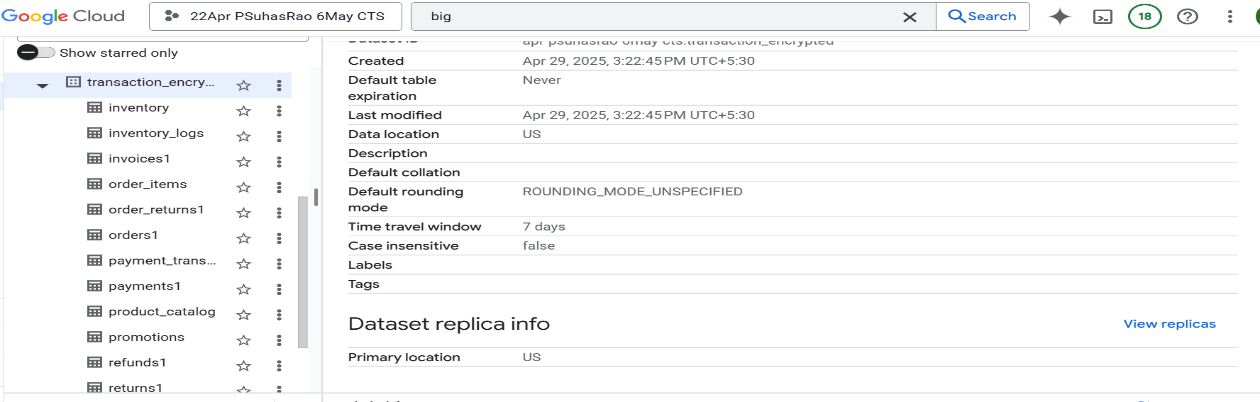




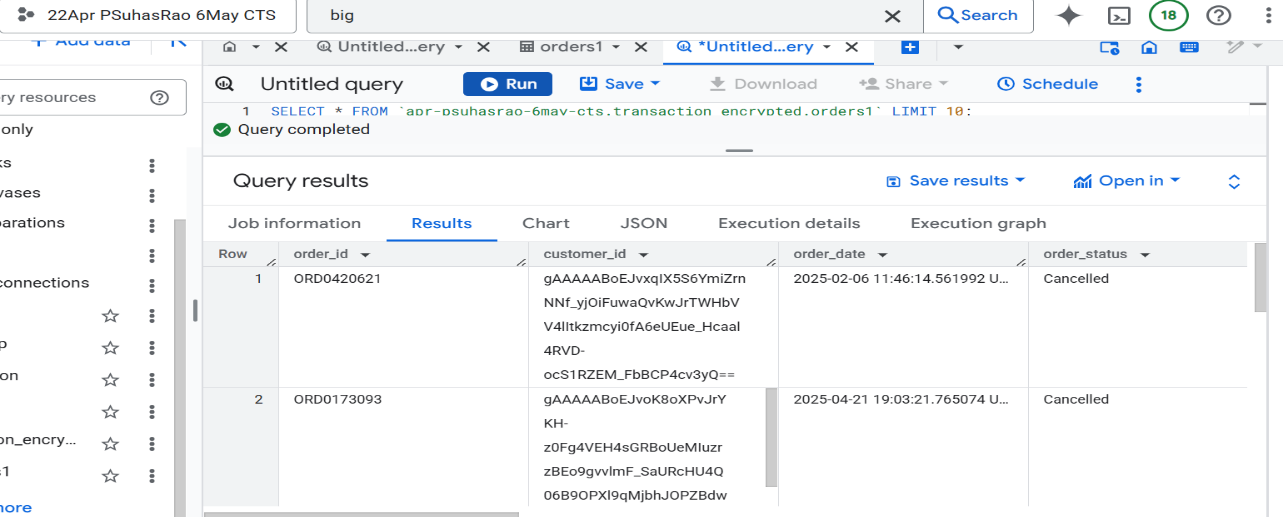
(22) Python Code for Encryption



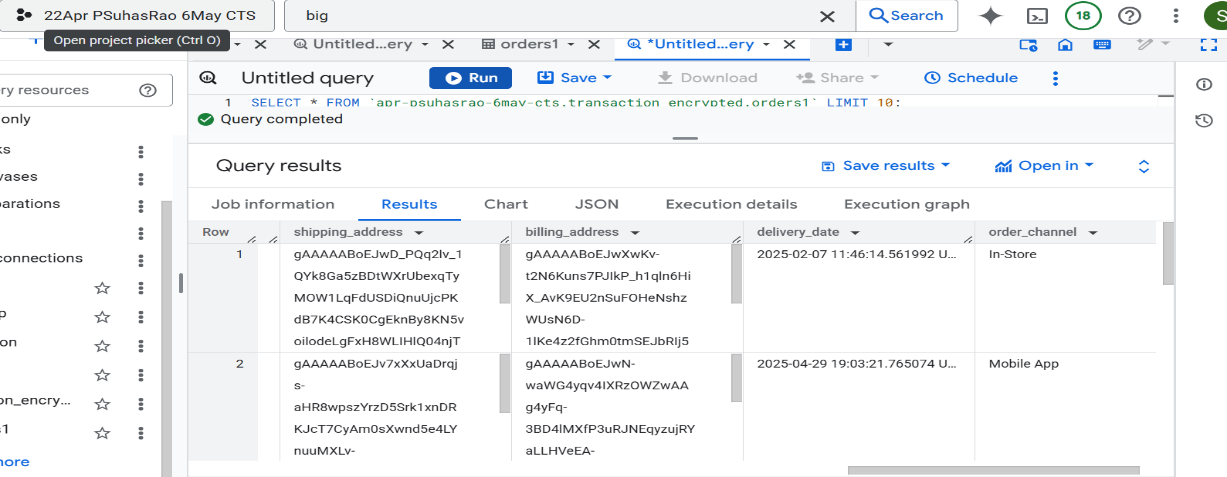
(23) Bucket in which Encrypted Data is stored



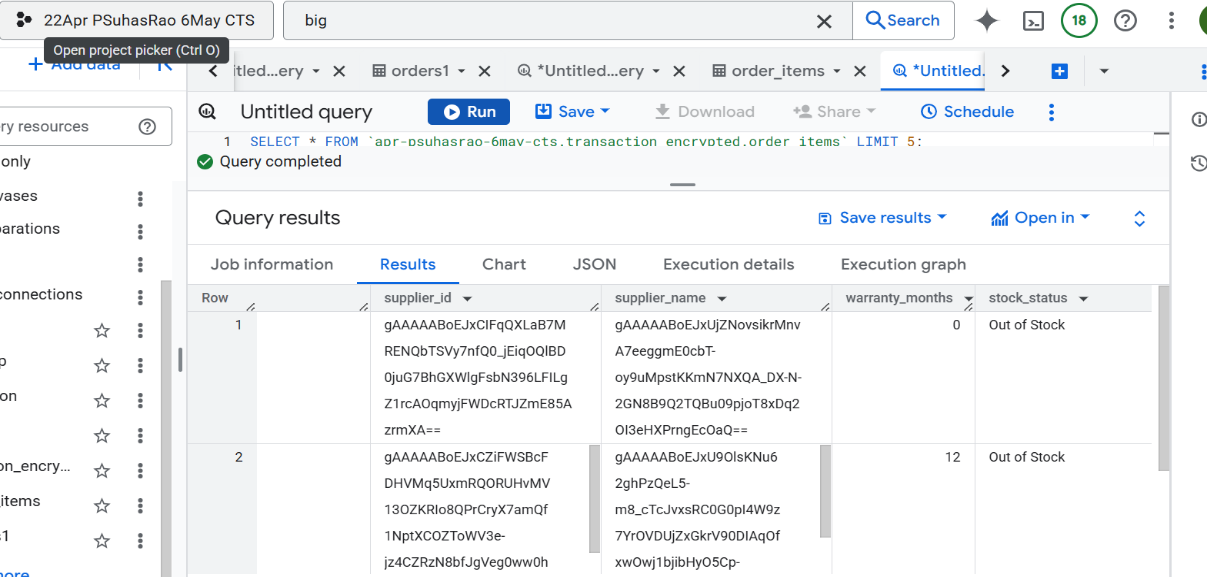
(24) Encrypted Transactional Data in BigQuery



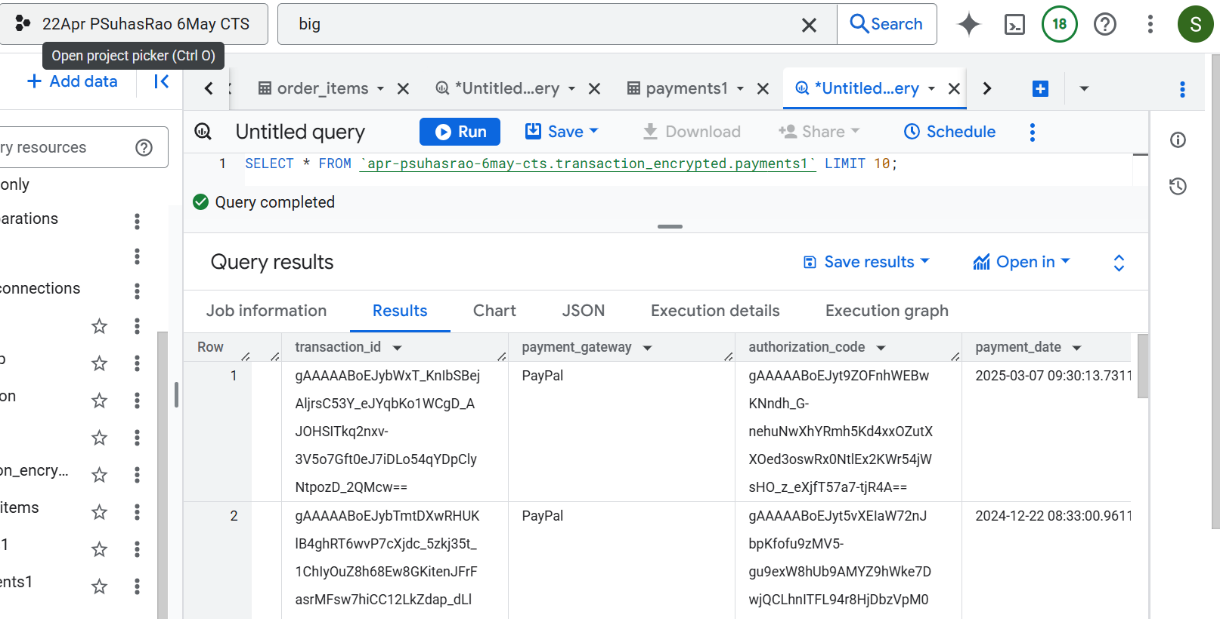
(25) Encrypted Orders Data



(26) Encrypted Orders Data



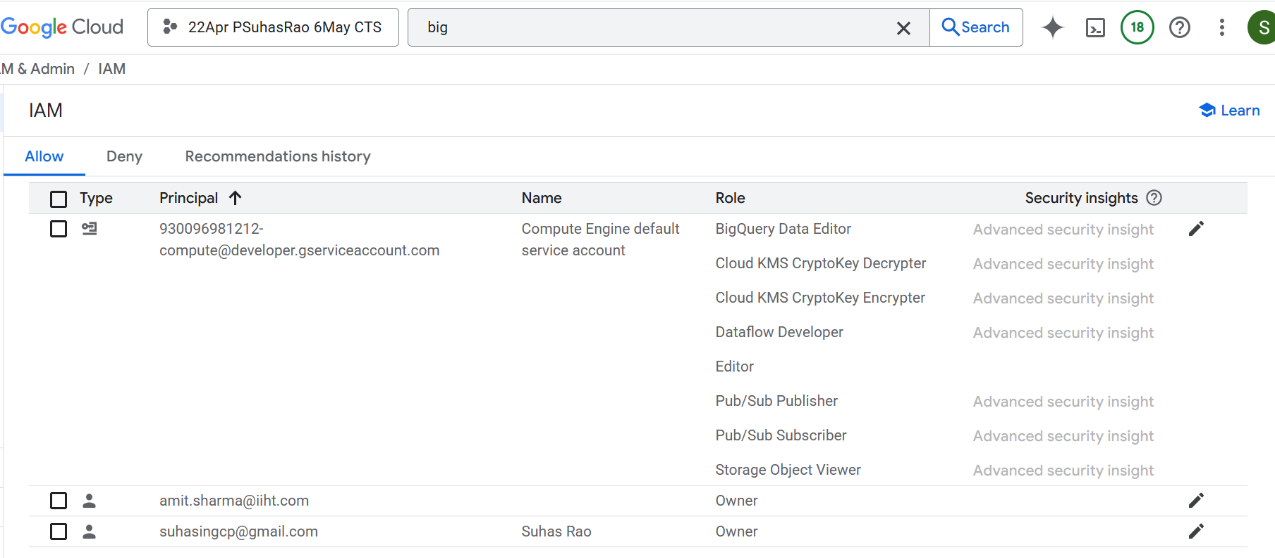
(27) Encrypted Order Items Data



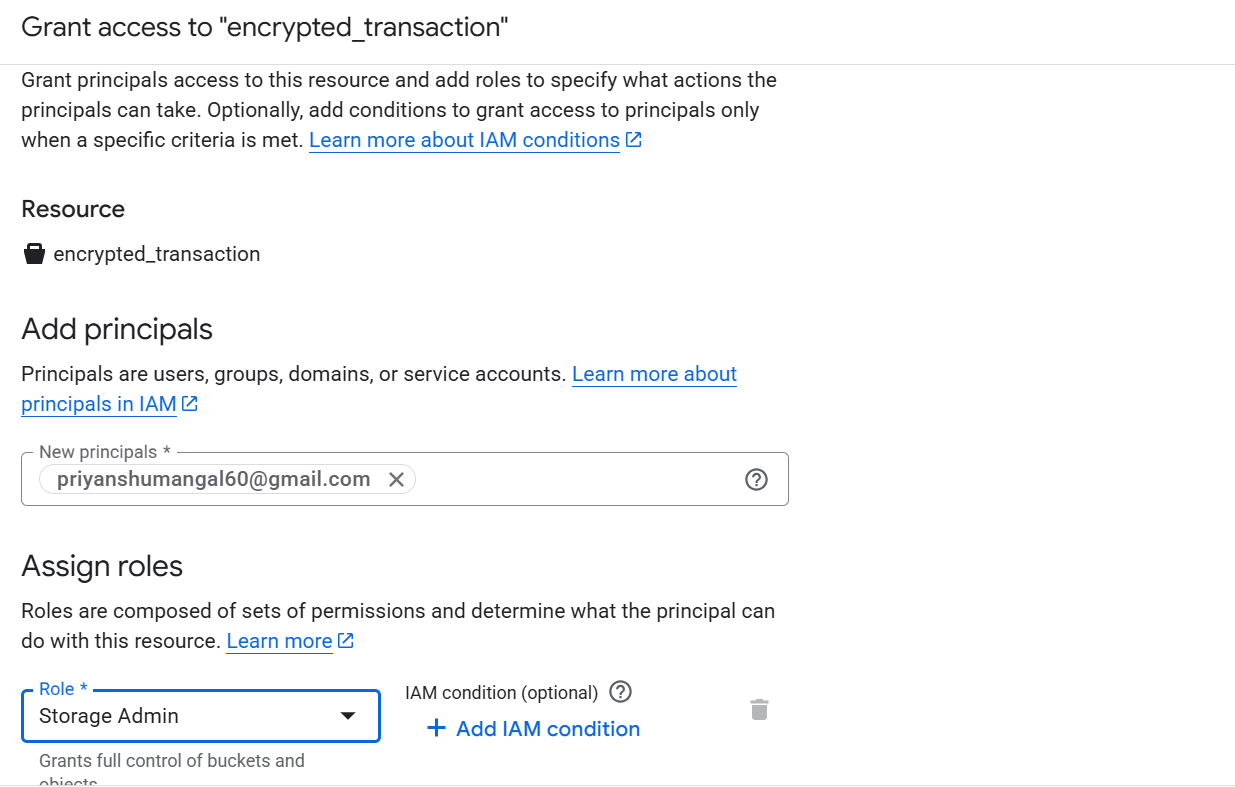
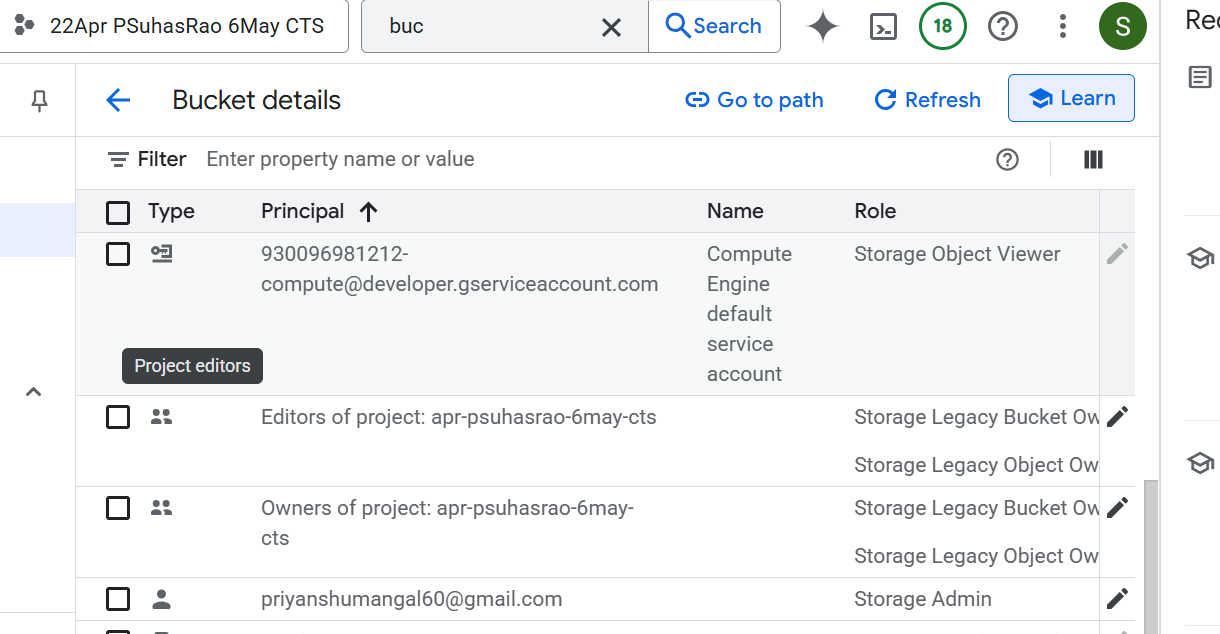
(28) Encrypted Payments Data

IAM and VPC

We were not granted permission to access the VPC at the organizational level.

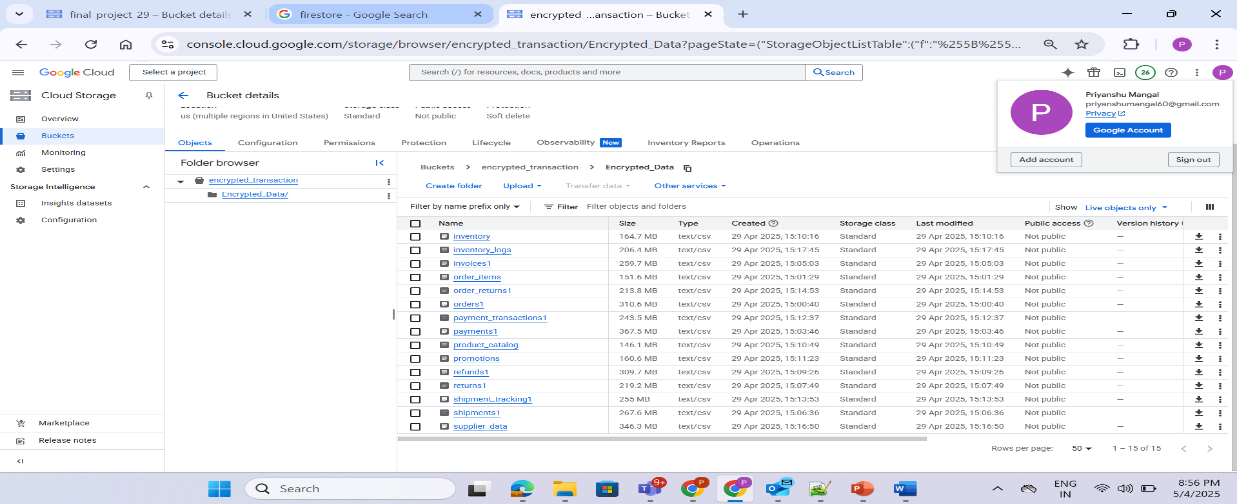


(29) Permissions Provided for Service Account

(30) Storage Admin Role

Provided [priyanshumangal60@gmail.com](mailto:priyanshumangal60@gmail.com) a role of storage admin for the bucket encrypted\_transaction so that he can access the sensitive data.



(31) Bucket Accessed By [priyanshumangal60@gmail.com](mailto:priyanshumangal60@gmail.com)

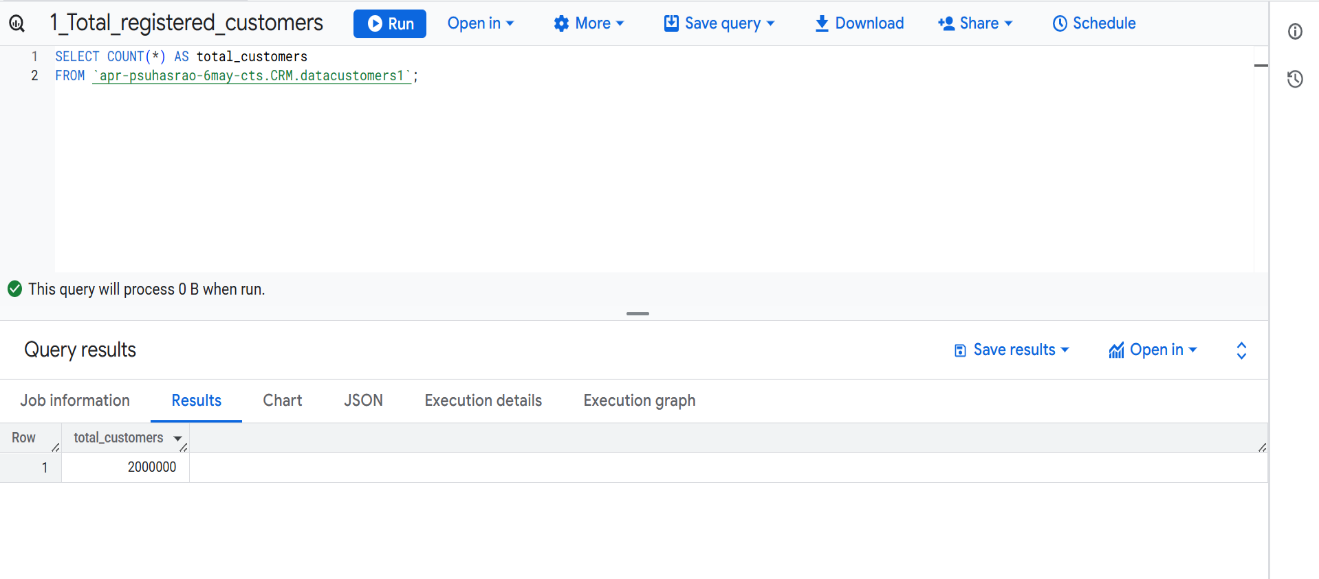
Chapter 4:

**Results and Discussions**

To evaluate the effectiveness of the implemented Customer Data Platform and to derive actionable insights from the processed data, a set of business-critical Key Performance Indicators (KPIs) were defined and visualized through interactive dashboards. These metrics and visualizations provide a comprehensive view of customer behavior, engagement trends, and transactional performance.

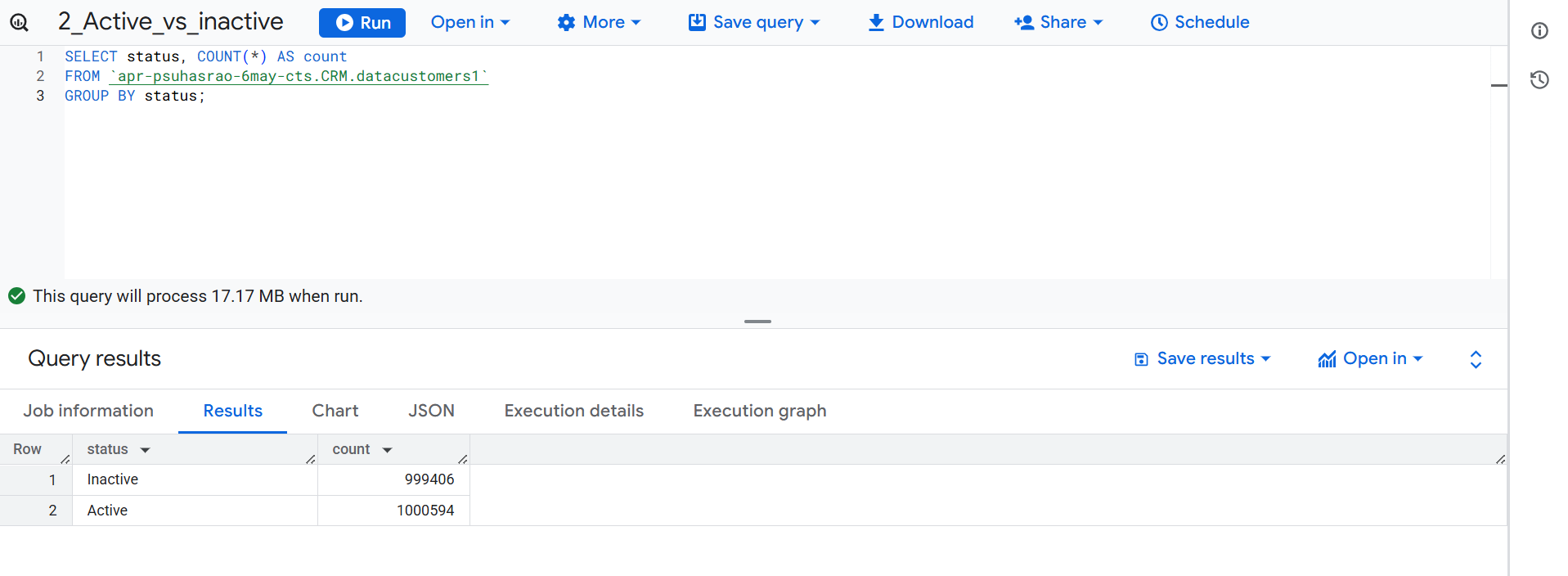
Key Performance Indicators (KPI’s)

**KPI 1: Total Registered Customers**



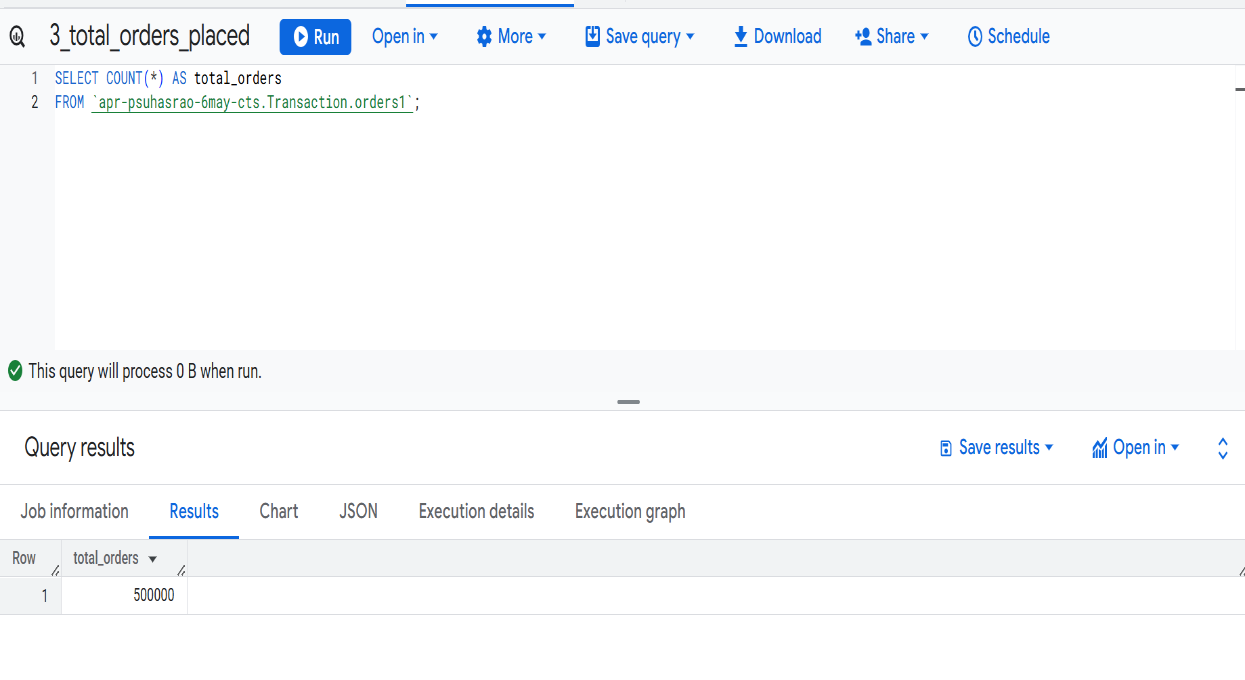
(32) KPI 1

**KPI 2: Active vs Inactive Customers**

****

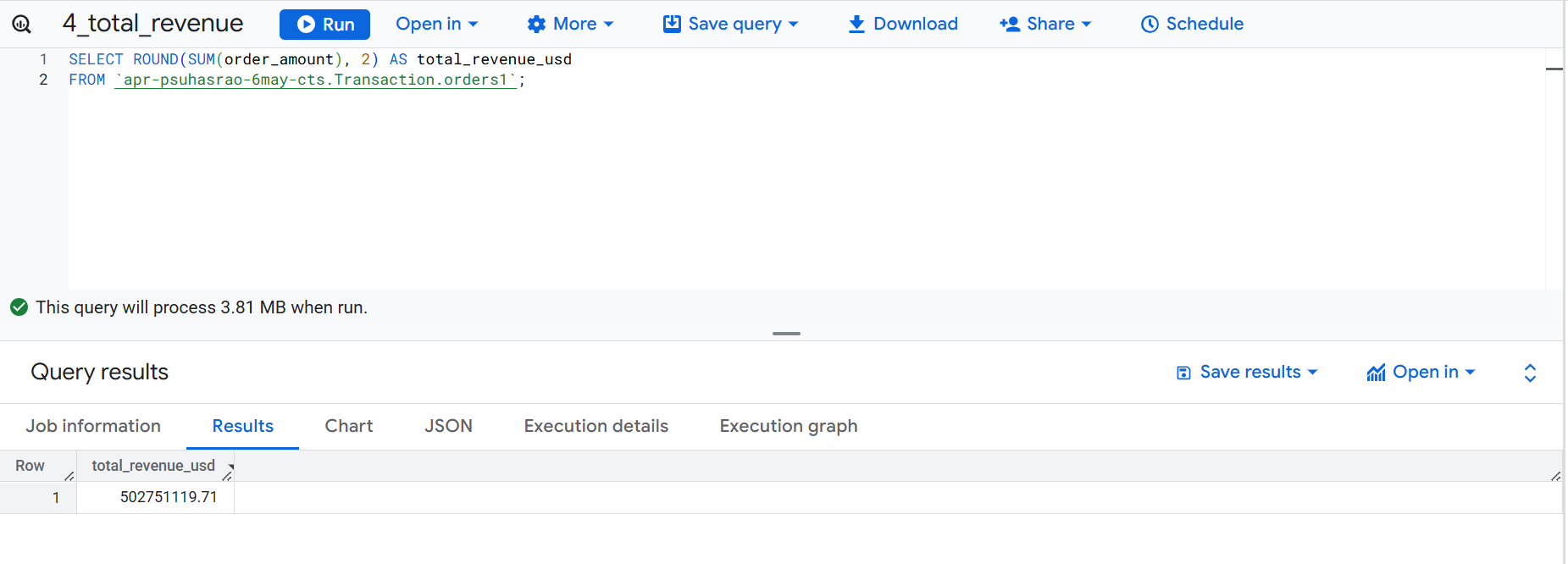
(33) KPI 2

**KPI 3: Total Orders Placed**

****

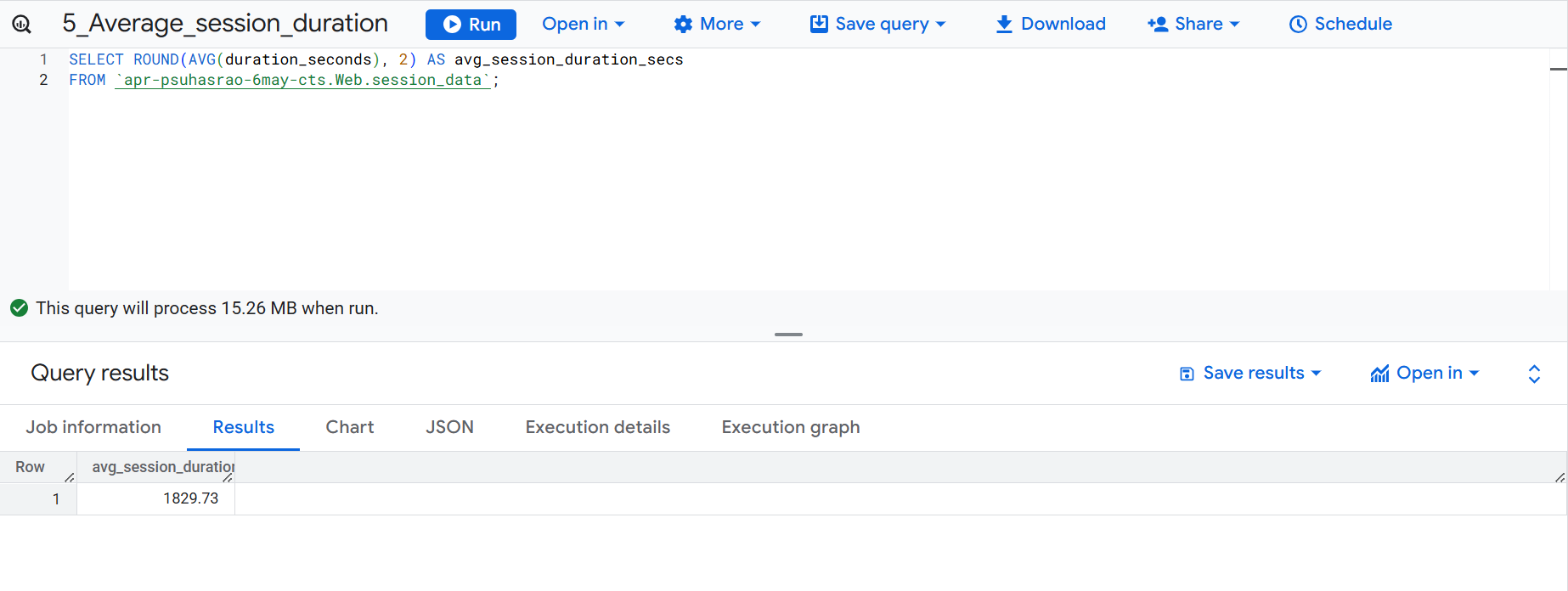
(34) KPI 3

**KPI 4: Total Revenue**

****

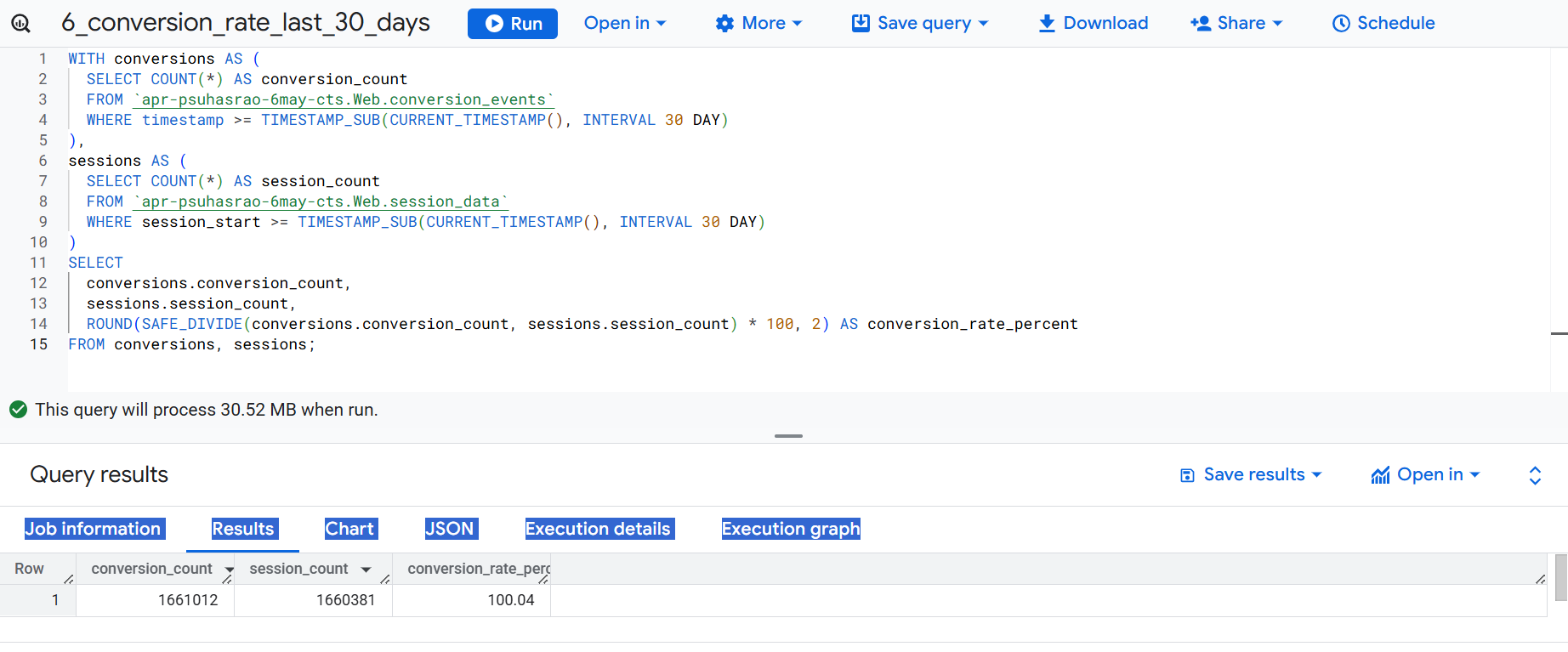
(35) KPI 4

**KPI 5: Average Session Duration**

****

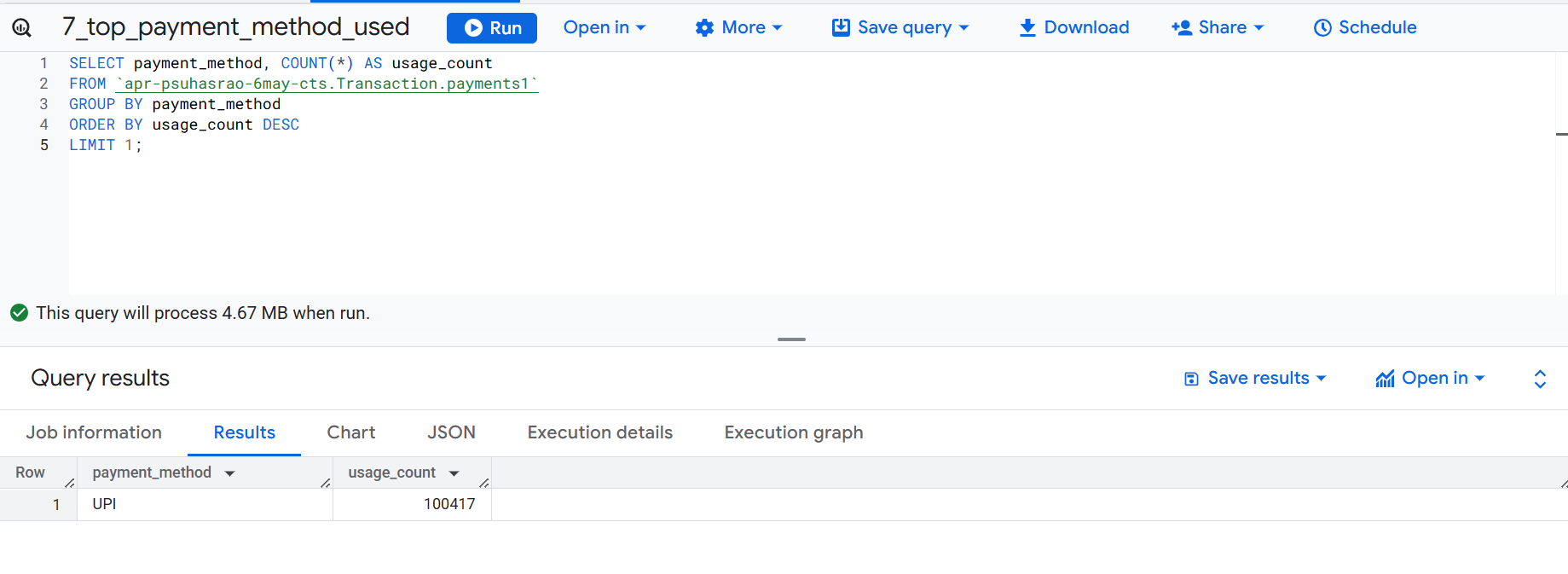
(36) KPI 5

**KPI 6: Conversion Rate (Last 30 Days)**

****

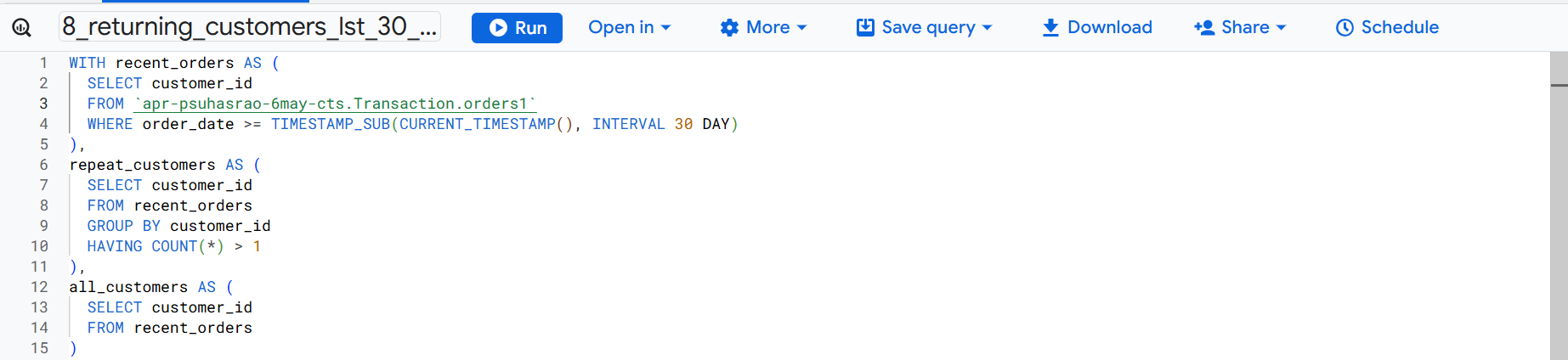
(37) KPI 6

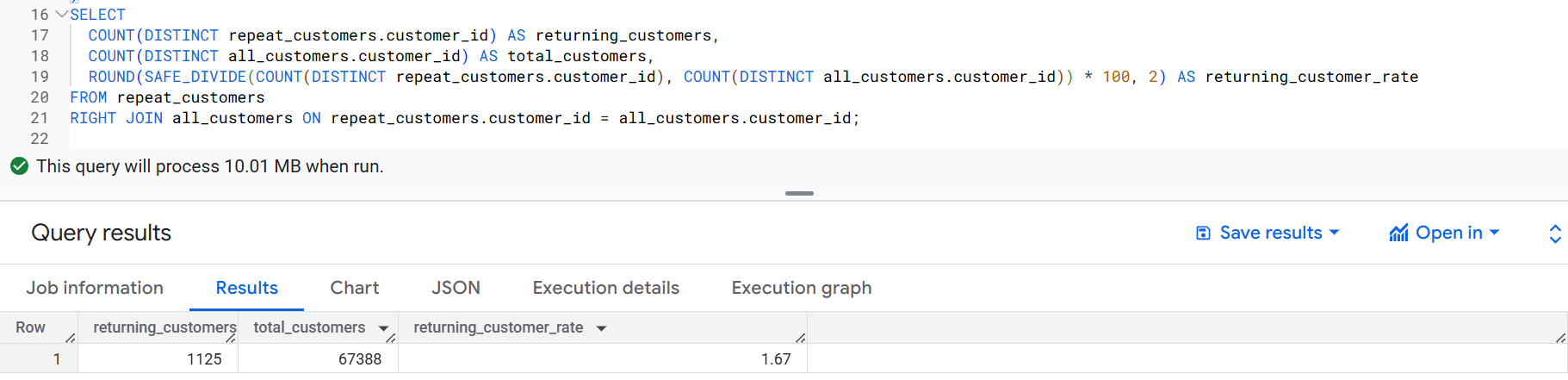
**KPI 7: Top Payment Method Used**

****

(38) KPI 7

**KPI 8: Returning Customer Rate (Last 30 Days)**

****

****

(39) KPI 8

**Dashboards**

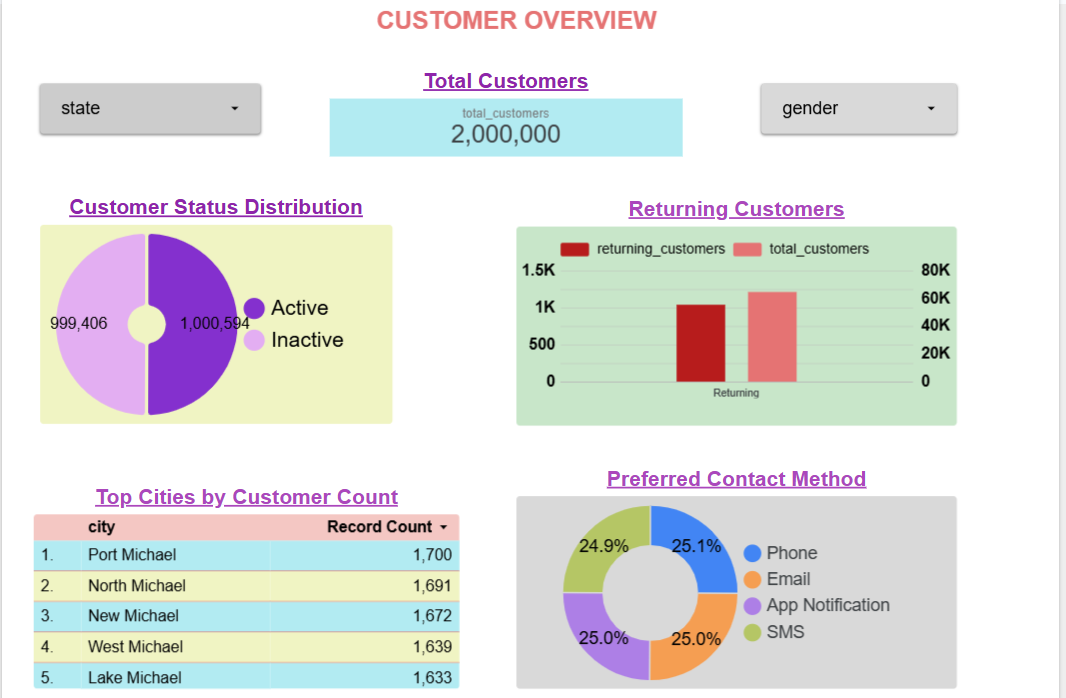
To visualize the processed data effectively, three dashboards were developed:

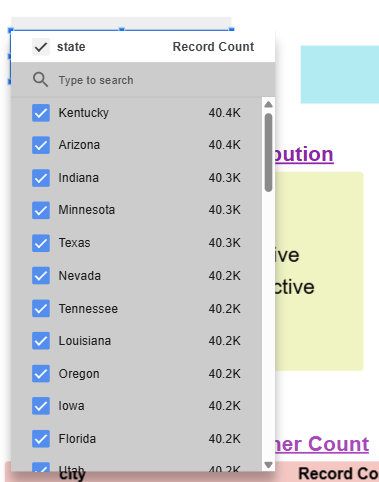
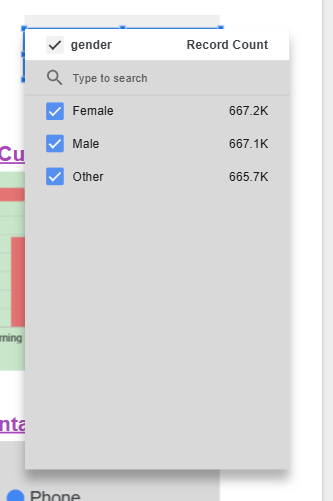
**Customer Overview**: Displays key customer metrics including total customers, active users, and customer segmentation.

**Web Engagement Analytics**: Highlights user interaction patterns such as page views, click events, browsers used etc.

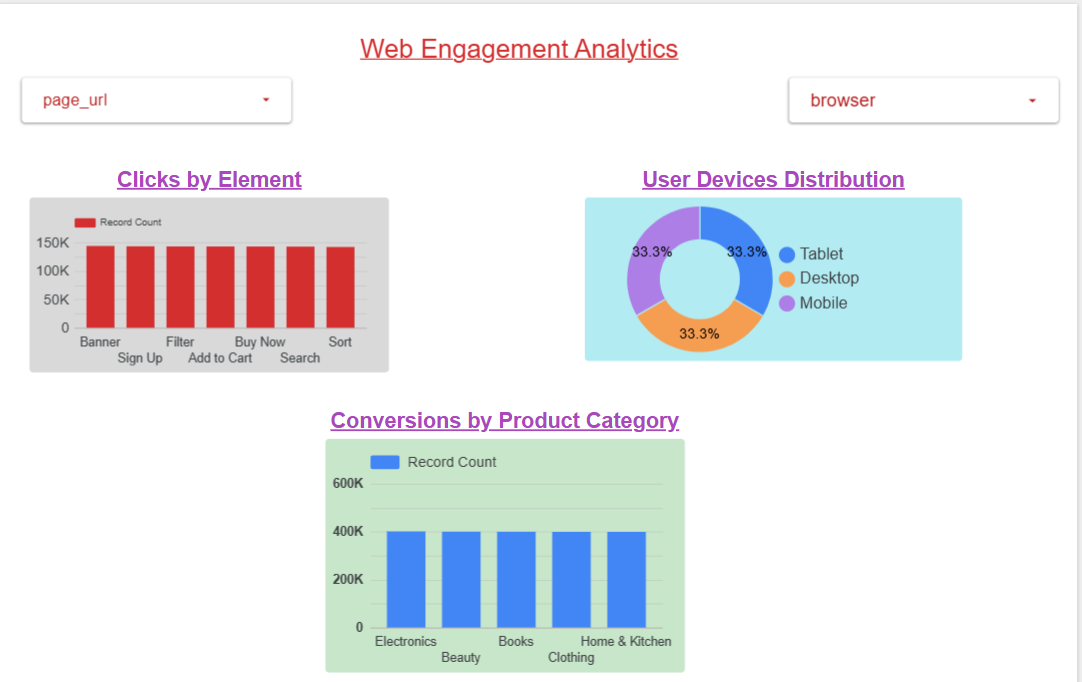
**Transaction Insights**: Provides visibility into order volumes, payment modes, and revenue trends to support financial analysis.

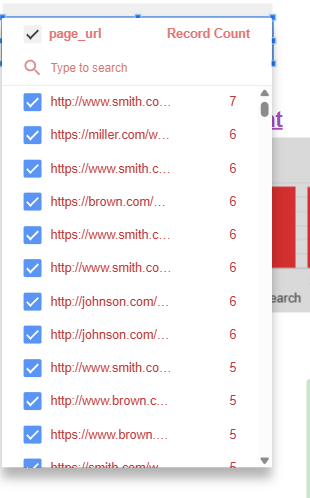
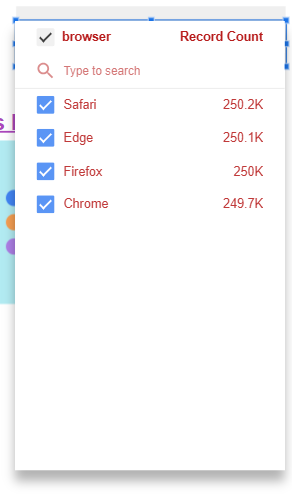
These dashboards help monitor performance and support data-driven decisions.



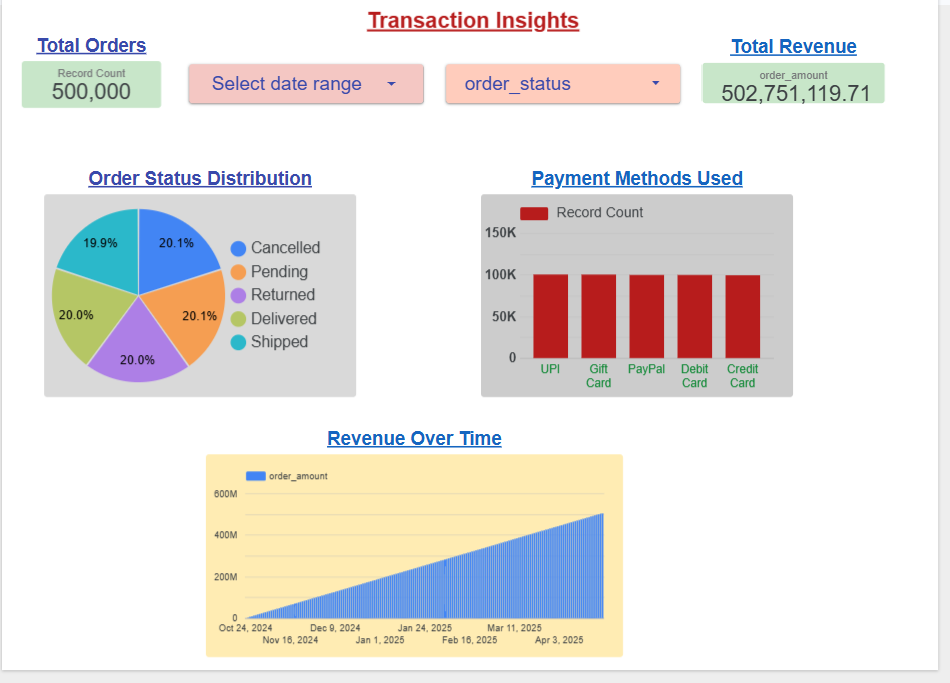
 

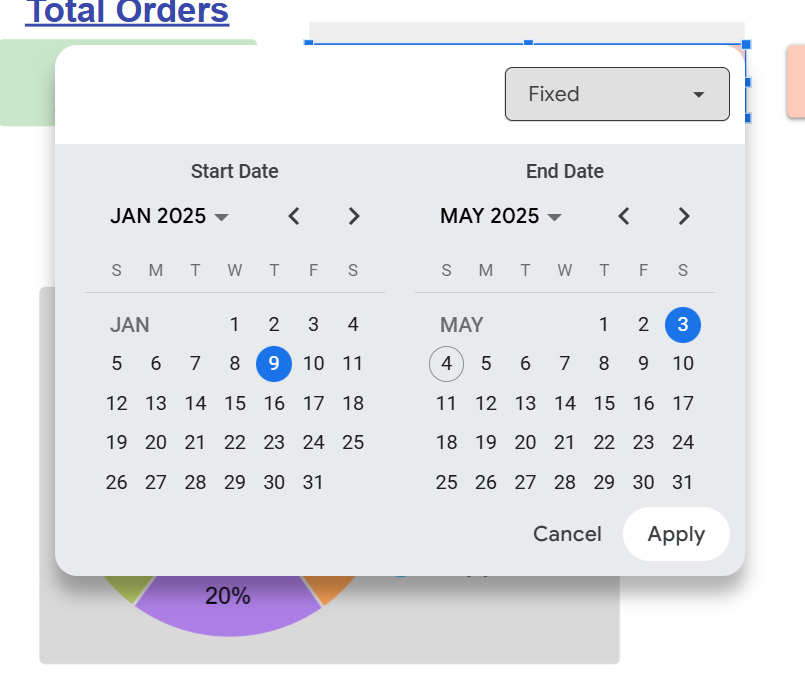
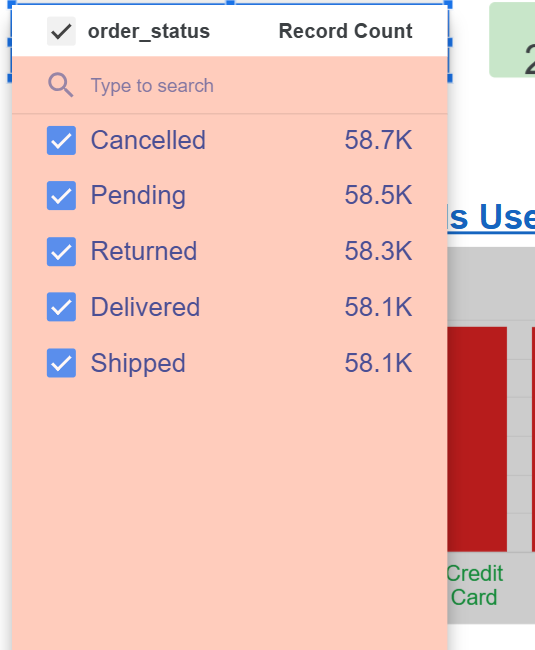
(40) Customer Overview Dashboard



(41) Web Analytics Dashboard

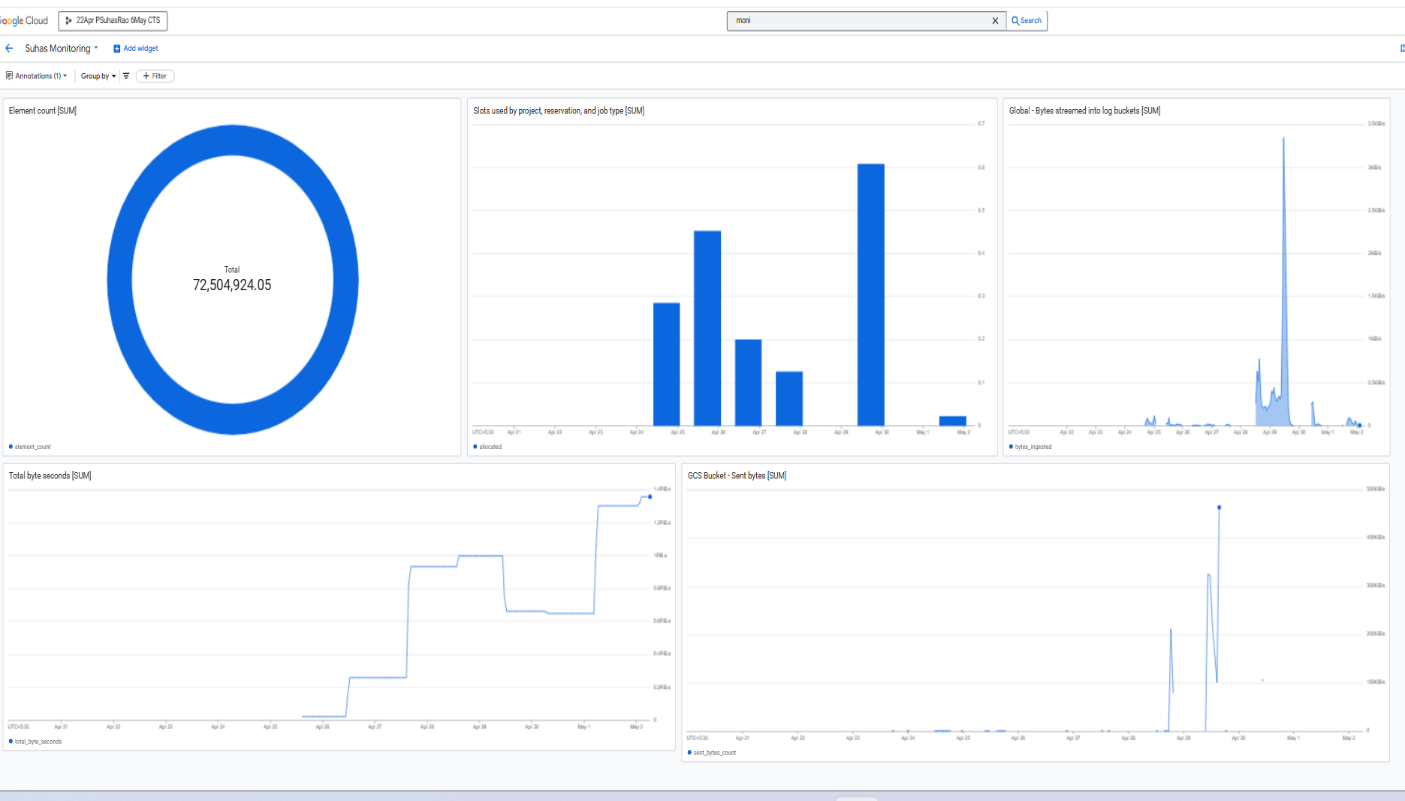


(42) Transaction Insights Dashboard

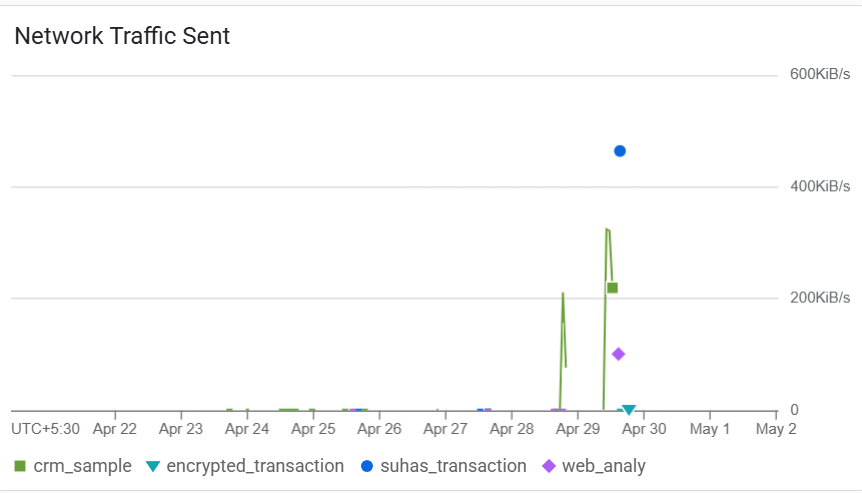
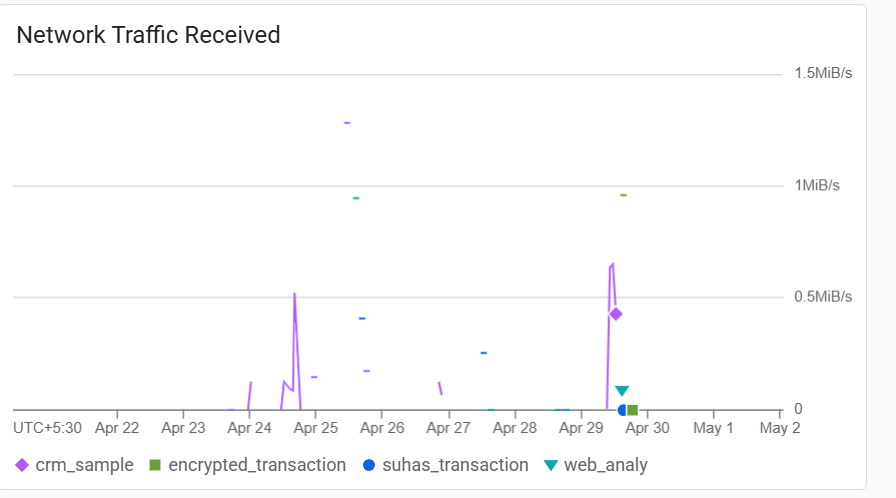
**Monitoring and Logging**

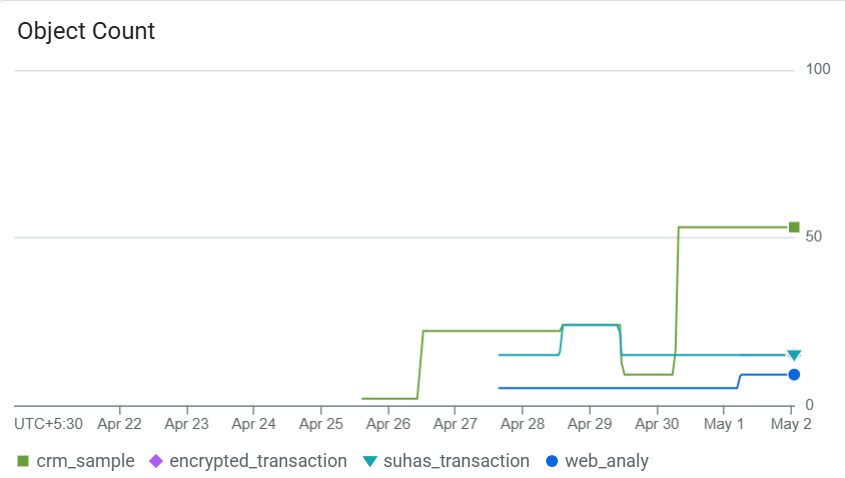
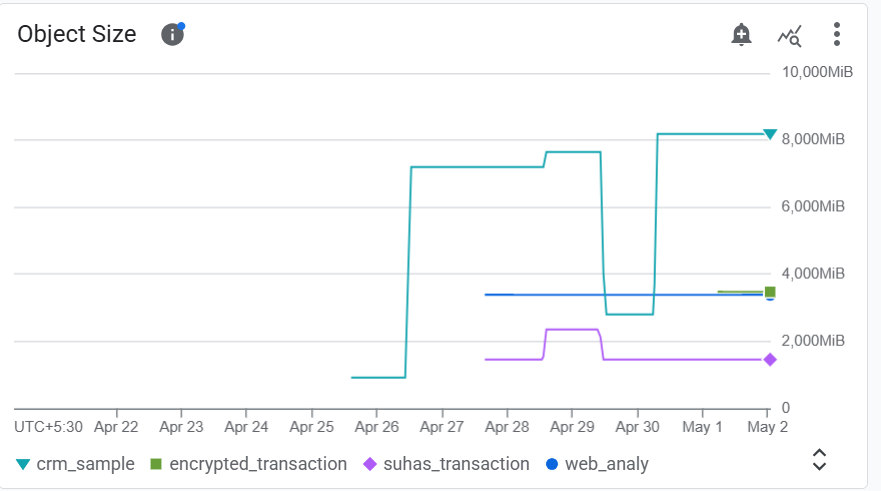
* Use of Google Cloud Monitoring to track BigQuery job errors.
* Alerting mechanisms (e.g., email alerts when errors occur).
* Logging dashboards to monitor data ingestion and processing pipelines.



(43) Monitoring Dashboard

**Cloud Storage Monitoring**

**** 

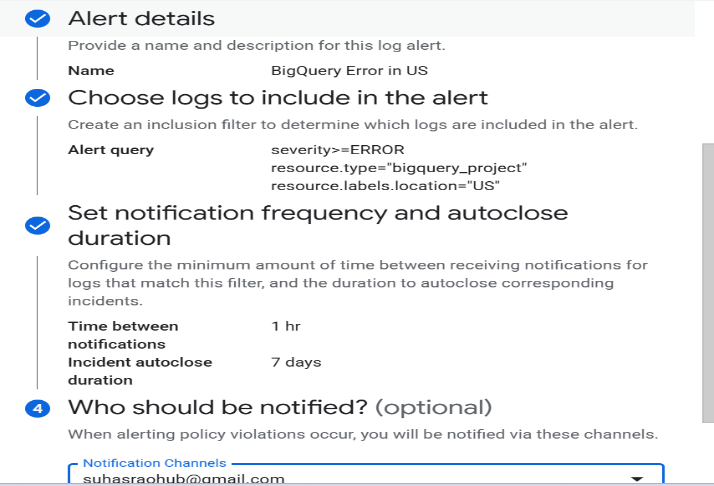
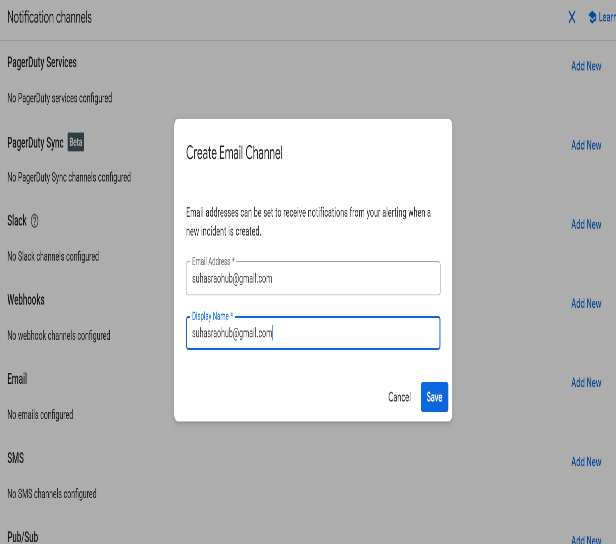
****

(44) Cloud Storage Monitoring

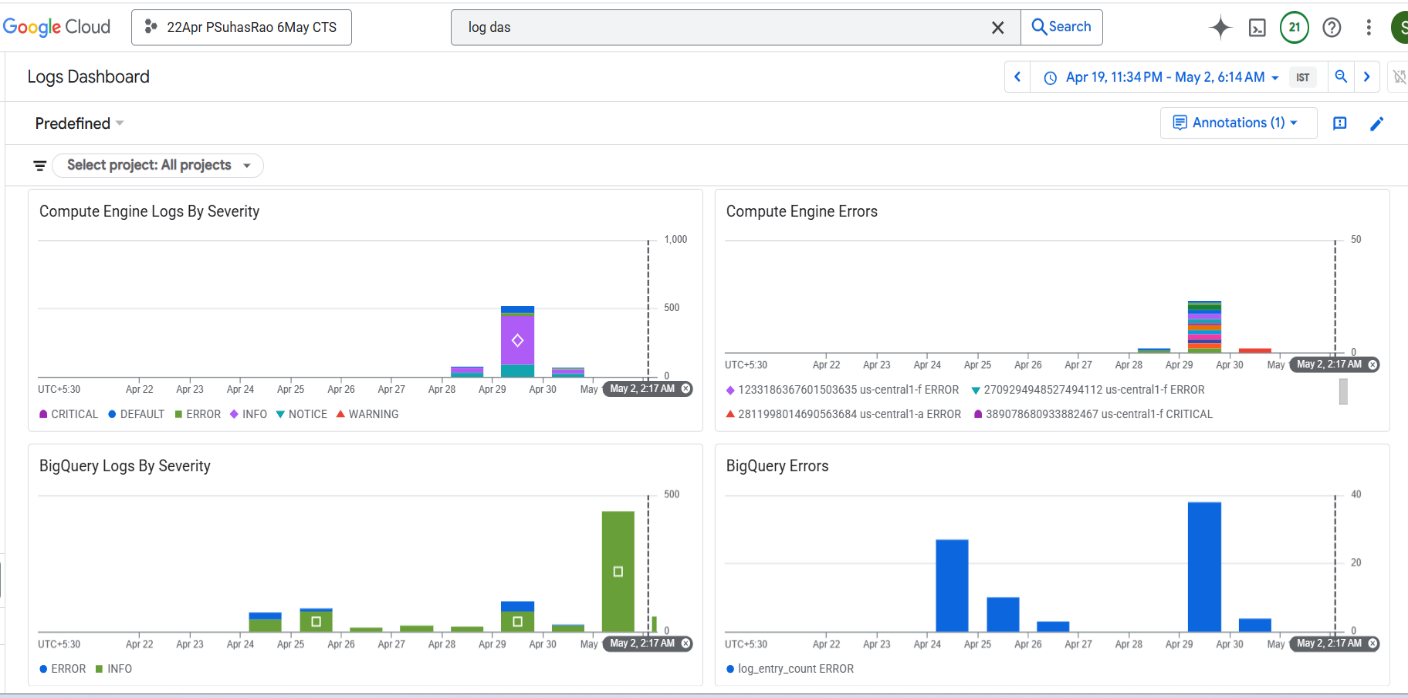
An alert will be sent to my email address [suhasraohub@gmail.com](mailto:suhasraohub@gmail.com) whenever error is triggered in bigquery project.

A screenshot of a computer

AI-generated content may be incorrect.

(45) BigQuery Error Alerting



(46) Logging Dashboard

Chapter 5:

# Conclusion and Future Scope

**Conclusion**

The implementation of the Customer Data Platform (CDP) on Google Cloud Platform (GCP) successfully demonstrated the ability to ingest, process, secure, and visualize large volumes of customer and web analytics data in real time. By leveraging GCP-native services such as Pub/Sub, Dataflow, BigQuery, Cloud KMS, IAM, and Looker Studio, the platform ensures scalability, security, and operational efficiency.

The project achieved its core objectives, including:

* Real-time and batch data ingestion from CRM, web, and transactional sources.
* Secure storage and encryption of sensitive data fields.
* Computation of key business metrics (KPIs).
* Visualization of insights through interactive dashboards.
* Implementation of monitoring and alerting mechanisms for system reliability.

This platform provides a strong foundation for data-driven decision-making and enhances data governance across business functions.

**Future Scope**

While the current implementation meets essential business requirements, there are several opportunities to enhance the platform further:

* **Integration of Machine Learning Models**: Incorporating predictive analytics for customer churn, segmentation, and recommendation systems.
* **Third-Party Tool Integration**: Connecting with external marketing platforms like Salesforce, HubSpot, or Google Ads for campaign automation.
* **Real-Time Personalization**: Using real-time data to personalize user experiences across digital platforms.
* **Scalability Improvements**: Optimizing pipelines for larger datasets and higher throughput.

Chapter 5:

**References**

* Google Cloud Platform Documentation – <https://cloud.google.com/docs>
* Pub/Sub Documentation – <https://cloud.google.com/pubsub/docs>
* Dataflow Documentation – <https://cloud.google.com/dataflow/docs>
* BigQuery Documentation – <https://cloud.google.com/bigquery/docs>
* Cloud KMS Documentation – <https://cloud.google.com/kms/docs>
* IAM Documentation – <https://cloud.google.com/iam/docs>
* Looker Studio Documentation – <https://cloud.google.com/looker/docs>
* Cloud Monitoring and Logging – <https://cloud.google.com/monitoring/docs>
* Faker Library for Python – <https://faker.readthedocs.io>