Kazakh-British Technical University

Big Data and Machine Learning, and Cloud Security and Compliance on Google Cloud

Assignment 4, Cloud Computing

Student: Gulshat Khamidulla
Date of submission: 05.12.24

2024

Table of contents

Introduction

In cloud environments, **Big Data** and **Machine Learning** are crucial because they allow businesses to handle vast amounts of data efficiently and make intelligent decisions. The cloud provides the necessary resources such as storage, processing power, and scalability to manage and analyze Big Data without needing physical infrastructure. It enables businesses to process data faster and at a lower cost.

Machine Learning, in the cloud, helps businesses analyze this data to make predictions, automate tasks, and improve services. Cloud platforms offer ready-to-use tools for developing, training, and deploying ML models, making it easier for companies to take advantage of this technology.

However, as organizations move more data and services to the cloud, **security** becomes a critical concern. Ensuring strong security measures, like encryption, access control, and monitoring, is essential to protect sensitive information stored and processed in the cloud.

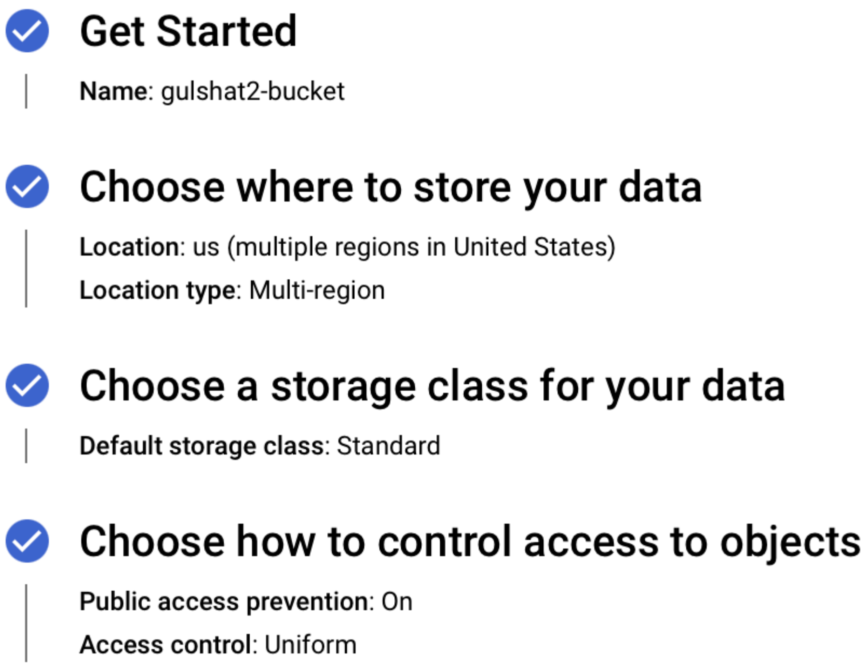
Big Data and Machine Learning on Google Cloud

Overview of the Pipeline

The pipeline described in this exercise is designed to process large datasets, train machine learning models, and deploy them for predictions. Google Cloud's ecosystem provides a comprehensive set of tools, including **BigQuery** for data processing, **Cloud Storage** for data storage, and **AI Platform** for model training and deployment. This pipeline is structured to handle data ingestion, preprocessing, model development, and monitoring.

Data Ingestion and Processing

I collected the dataset from a public source Kaggle and uploaded it to Google **Cloud Storage**. Cloud Storage provides a secure and scalable way to manage large datasets, acting as a data lake for all raw data. First of all, I needed a bucket where I uploaded the dataset. Picture 1 describes the process of creating a bucket named "gulshat2-bucket", location type is "multi-region" and default storage class and access control.



Get Started

Name: gulshat2-bucket

Choose where to store your data

Location: us (multiple regions in United States)

Location type: Multi-region

Choose a storage class for your data

Default storage class: Standard

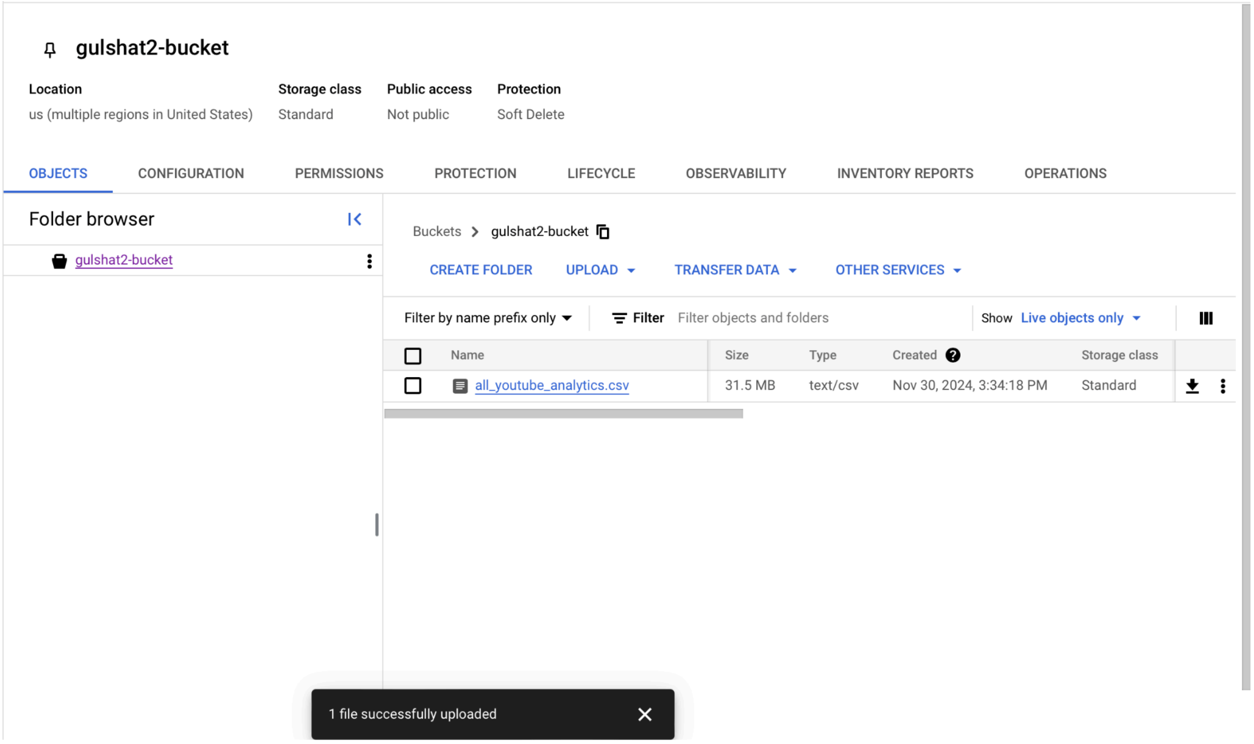
Choose how to control access to objects

Public access prevention: On

Access control: Uniform

Picture 1. Creating a bucket

Next, as shown in Picture 2, I uploaded a csv file which contains large dataset about YouTube analytics.



gulshat2-bucket

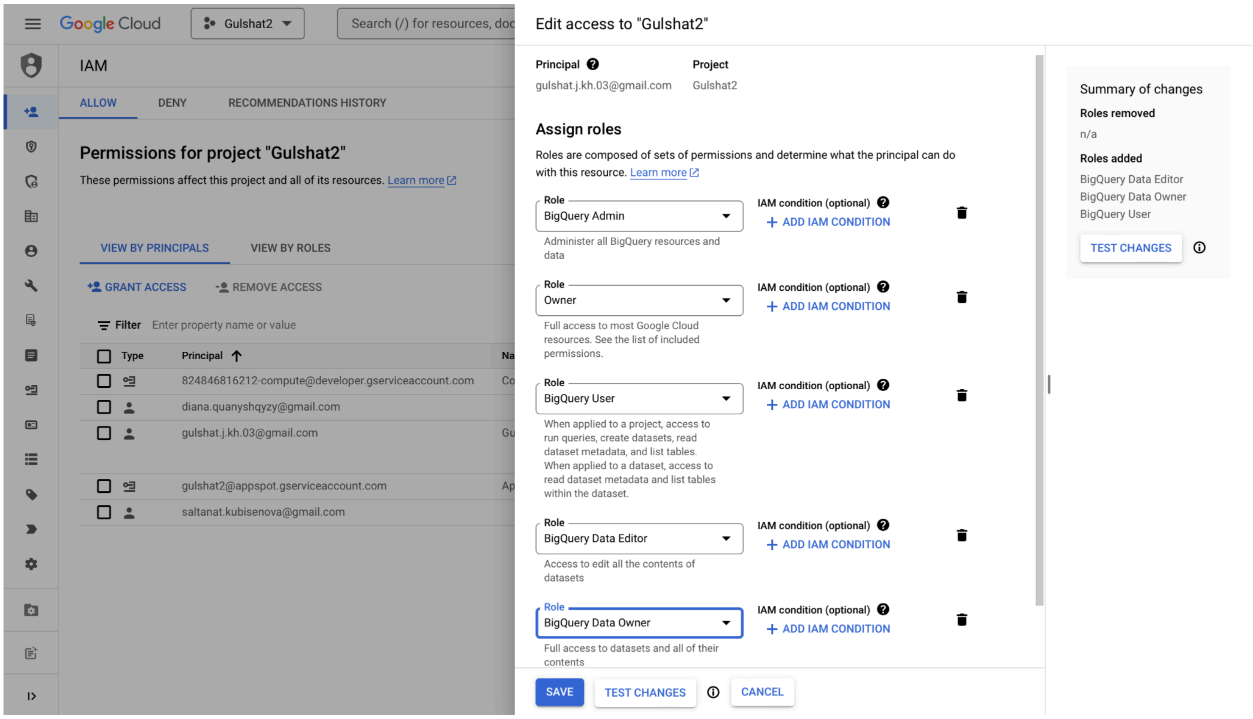
Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

□	Name	Size	Type	Created	Storage class
□	all_youtube_analytics.csv	31.5 MB	text/csv	Nov 30, 2024, 3:34:18 PM	Standard

1 file successfully uploaded

Picture 2. Uploading a dataset

Before creating a dataset in BigQuery I assigned different roles such as BigQuery Admin, BigQuery User, BigQuery Data Editor, BigQuery Data owner. I was not able to create datasets and work with them without these roles.



Google Cloud | Gulshat2 | Search (/) for resources, doc

IAM

ALLOW | DENY | RECOMMENDATIONS HISTORY

Permissions for project "Gulshat2"

These permissions affect this project and all of its resources. Learn more

VIEW BY PRINCIPALS | VIEW BY ROLES

GRANT ACCESS | REMOVE ACCESS

Filter Enter property name or value

Type	Principal ↑	Na
□	824846816212-compute@developer.gserviceaccount.com	Co
□	diana.quanyshqyzy@gmail.com	
□	gulshat.j.kh.03@gmail.com	Gu
□	gulshat2@appspot.gserviceaccount.com	Ap
□	saltanat.kubisenova@gmail.com	

Edit access to "Gulshat2"

Principal: gulshat.j.kh.03@gmail.com

Project: Gulshat2

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. Learn more

Role: BigQuery Admin — IAM condition (optional): + ADD IAM CONDITION

Administer all BigQuery resources and data

Role: Owner — IAM condition (optional): + ADD IAM CONDITION

Full access to most Google Cloud resources. See the list of included permissions.

Role: BigQuery User — IAM condition (optional): + ADD IAM CONDITION

When applied to a project, access to run queries, create datasets, read dataset metadata, and list tables. When applied to a dataset, access to read dataset metadata and list tables within the dataset.

Role: BigQuery Data Editor — IAM condition (optional): + ADD IAM CONDITION

Access to edit all the contents of datasets

Role: BigQuery Data Owner — IAM condition (optional): + ADD IAM CONDITION

Full access to datasets and all of their contents

SAVE | TEST CHANGES | CANCEL

Summary of changes

Roles removed

n/a

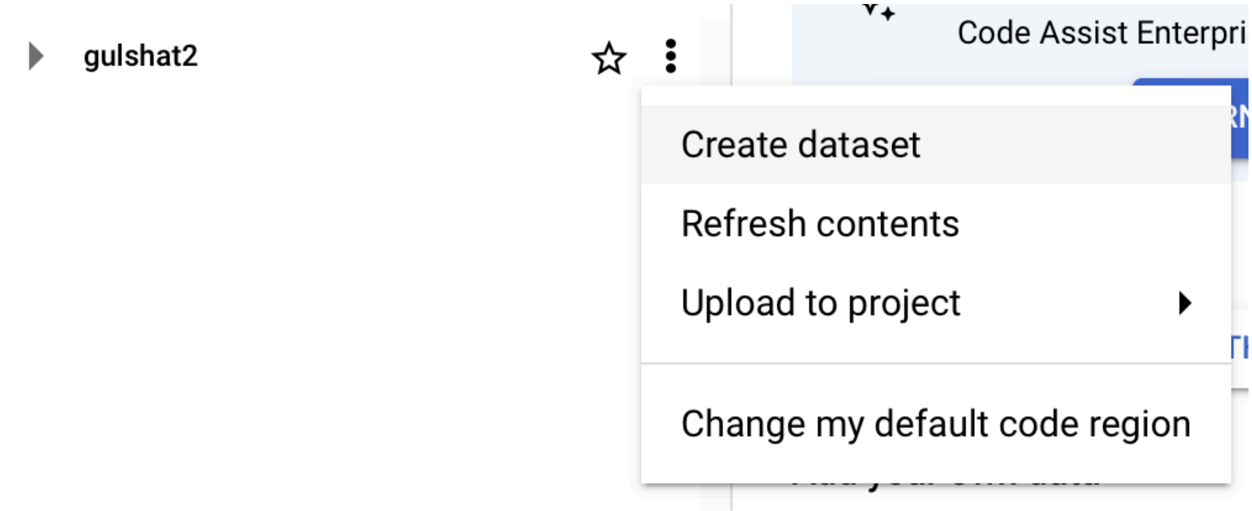
Roles added

BigQuery Data Editor
BigQuery Data Owner
BigQuery User

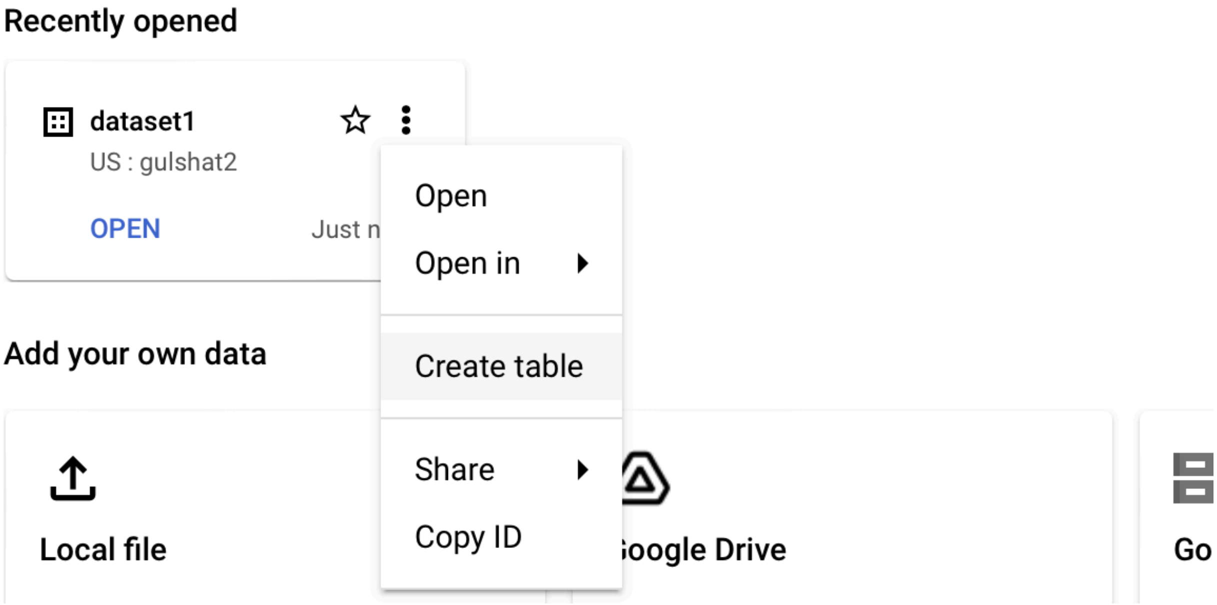
TEST CHANGES

Picture 3. IAM Permissions

Next, I proceeded to upload the **all_youtube_analytics.csv** file into BigQuery as a new table. I started by clicking on the **youtube_analytics** dataset in the resource tree. Then, I clicked on the **"Create Table"** button.

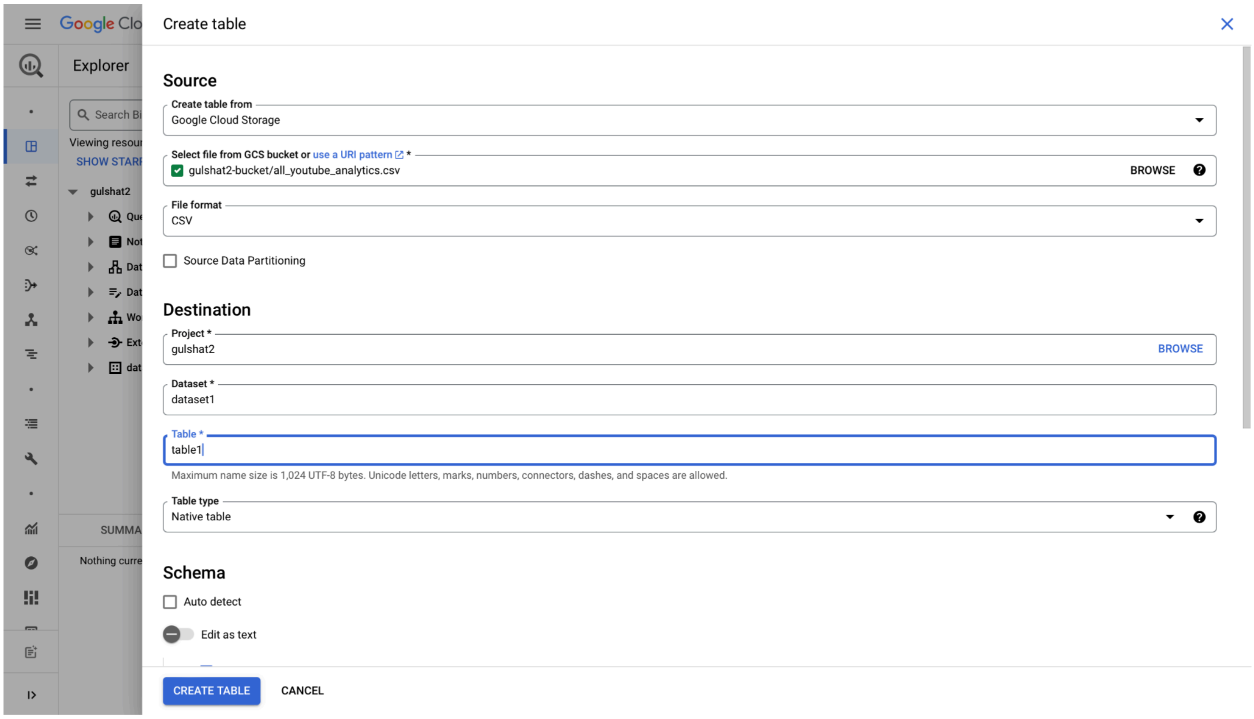


Picture 3. Creating a dataset



Picture 4. Creating a table

In the "Create Table" interface, I set the source type to **"File Upload"** and selected the **all_youtube_analytics.csv** file from my local machine. I ensured that the file format was set to **"CSV"**, as the data was stored in this format.



Create table

Source

Create table from: Google Cloud Storage

Select file from GCS bucket or use a URI pattern *: ☑ gulshat2-bucket/all_youtube_analytics.csv BROWSE

File format: CSV

□ Source Data Partitioning

Destination

Project *: gulshat2 BROWSE

Dataset *: dataset1

Table *: table1

Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type: Native table

Schema

□ Auto detect

Edit as text

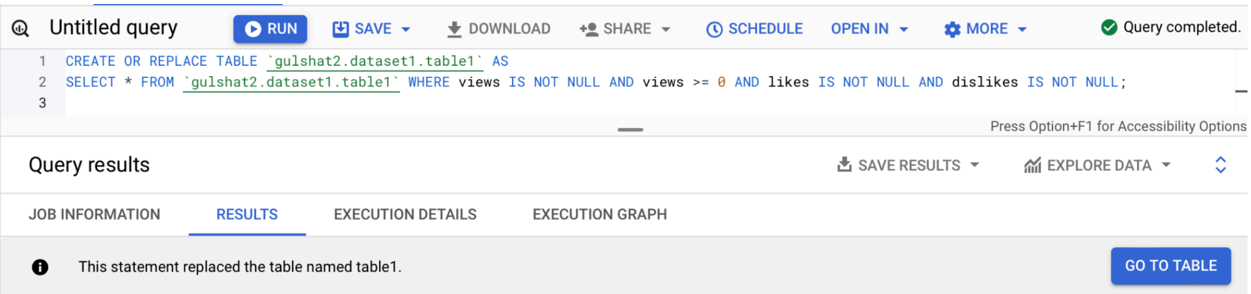
CREATE TABLE CANCEL

Picture 5. Setting a table

Once all the details were filled in, I clicked on the **"Create Table"** button, which successfully created the dataset and added it to the project's resource tree.

After creating the table, I clicked on the **table1** table to inspect its structure and data. BigQuery provided a preview of the data, and I verified that the column names in the table matched those in the original CSV file and the data types were correctly detected, with numerical fields as integer , textual fields as string, and date fields as date.

To clean the data, I opened the SQL workspace in BigQuery and wrote queries to handle potential issues in the dataset as shown in the picture 6 below..



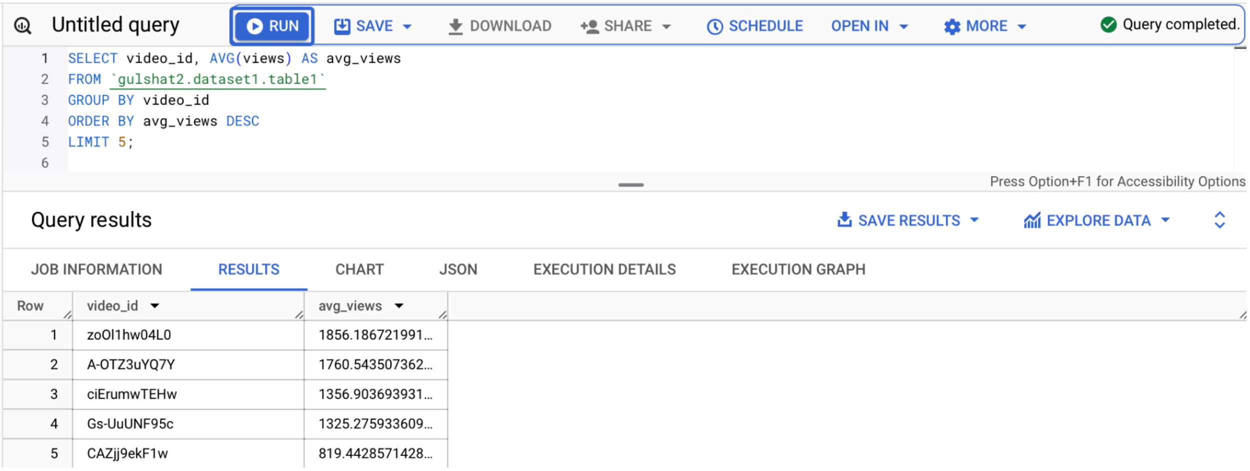
```
CREATE OR REPLACE TABLE `gulshat2.dataset1.table1` AS
SELECT * FROM `gulshat2.dataset1.table1` WHERE views IS NOT NULL AND views >= 0 AND likes IS NOT NULL AND dislikes IS NOT NULL;
```

Query results

This statement replaced the table named table1.

Picture 6. Data cleaning

The SQL query in Picture 7 and bar chart in Picture 8 retrieves the **top 5 videos with the highest average views**. The goal of this query is to analyze the dataset to find the top 5 videos with the highest average number of views. It could be useful for identifying popular videos in a dataset, understanding which videos consistently attract the most viewers, generating insights for content strategy or marketing decisions.

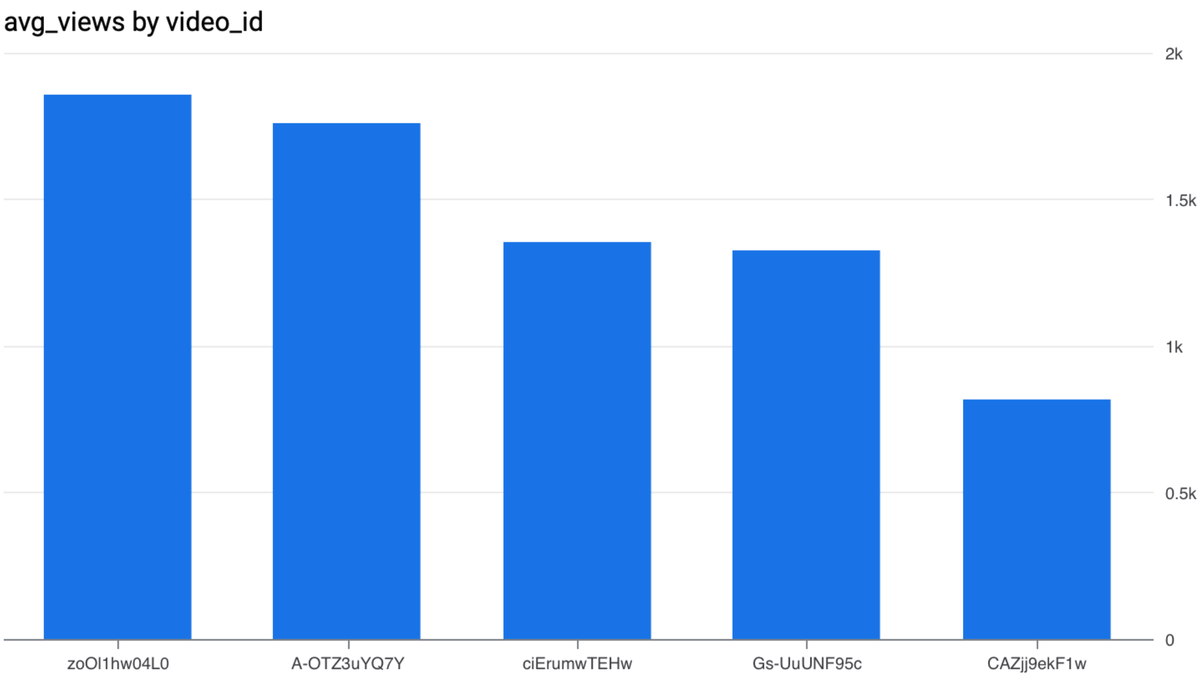


```
SELECT video_id, AVG(views) AS avg_views
FROM `gulshat2.dataset1.table1`
GROUP BY video_id
ORDER BY avg_views DESC
LIMIT 5;
```

Query results

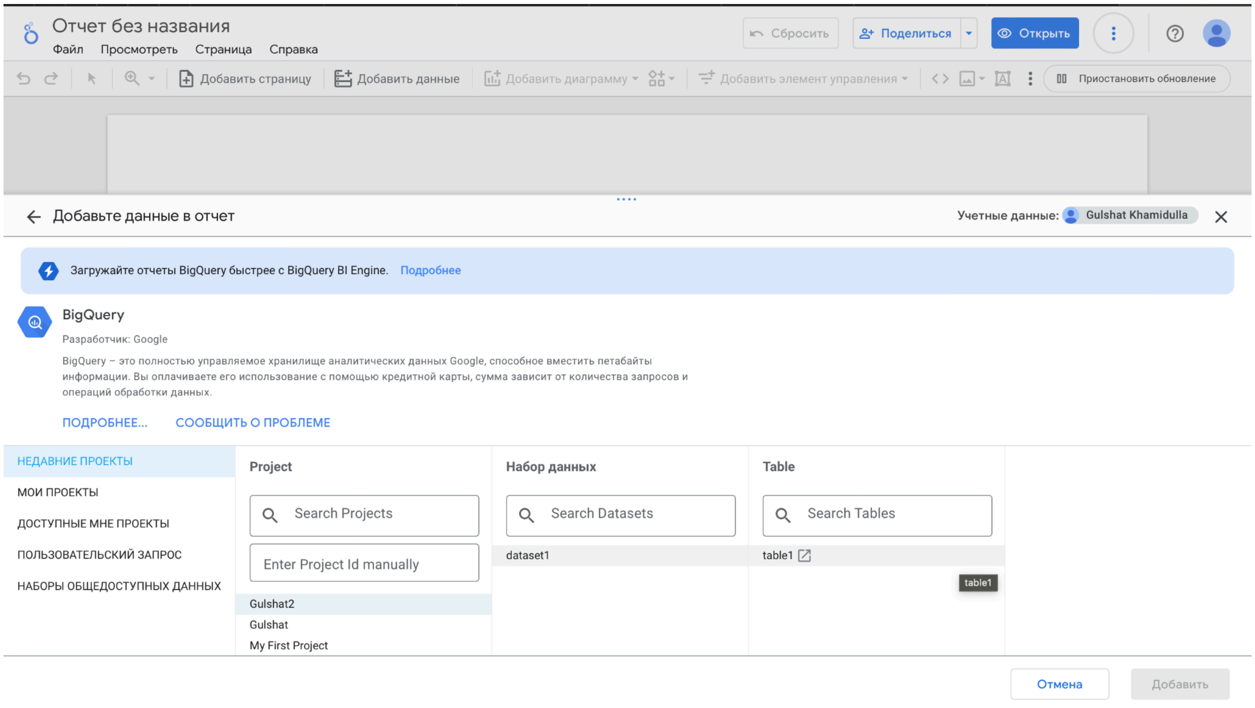
Row	video_id	avg_views
1	zoOl1hw04L0	1856.186721991…
2	A-OTZ3uYQ7Y	1760.543507362…
3	ciErumwTEHw	1356.903693931…
4	Gs-UuUNF95c	1325.275933609…
5	CAZjj9ekF1w	819.4428571428…

Picture 7. Data preprocessing



Picture 8. Data visualization

Once the cleaned and preprocessed data was ready, I connected BigQuery to **Google Data Studio** to create visualizations.



Загружайте отчеты BigQuery быстрее с BigQuery BI Engine. Подробнее

BigQuery

Разработчик: Google

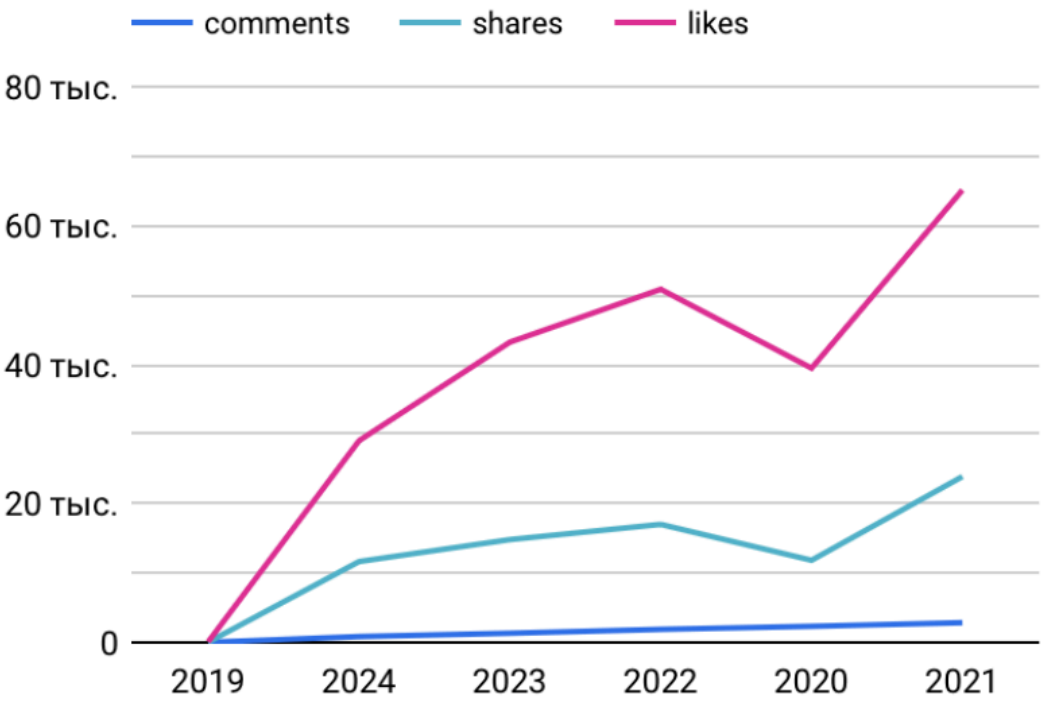
BigQuery – это полностью управляемое хранилище аналитических данных Google, способное вместить петабайты информации. Вы оплачиваете его использование с помощью кредитной карты, сумма зависит от количества запросов и операций обработки данных.

Picture 9. Connecting Google Data Studio

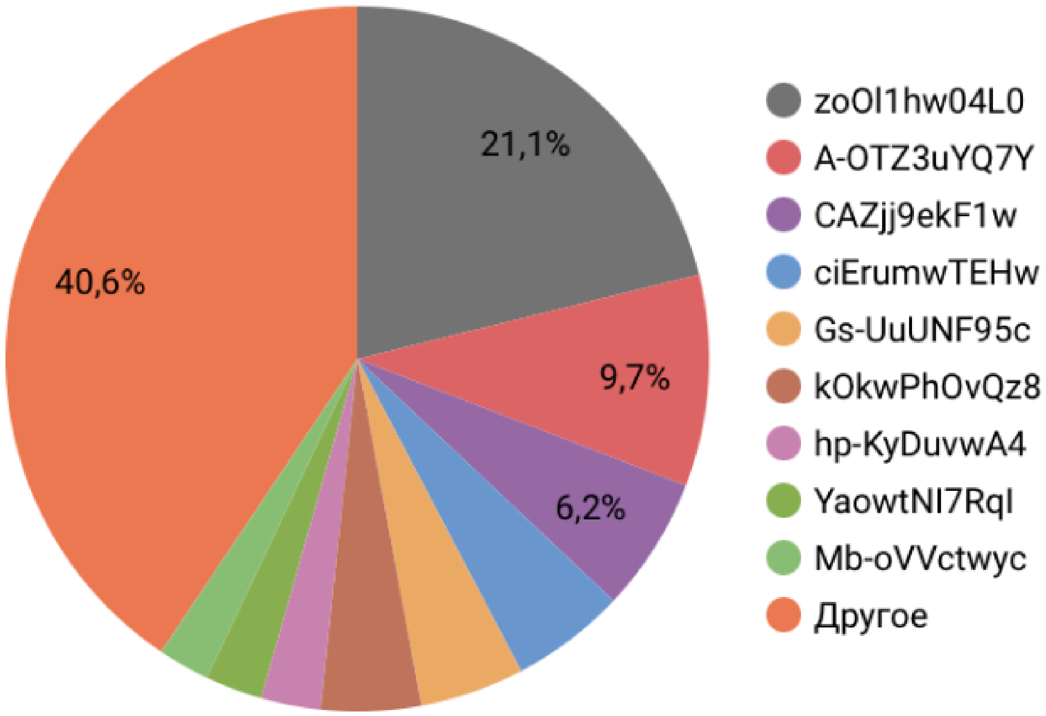
In Data Studio, I built the following charts:

- **Line Chart**: Showing the trend of likes, shares, and comments over the years.
- **Pie Chart**: Displaying the distribution of total likes, shares, and comments.
- **Table**: Provide granular data for deeper insights.

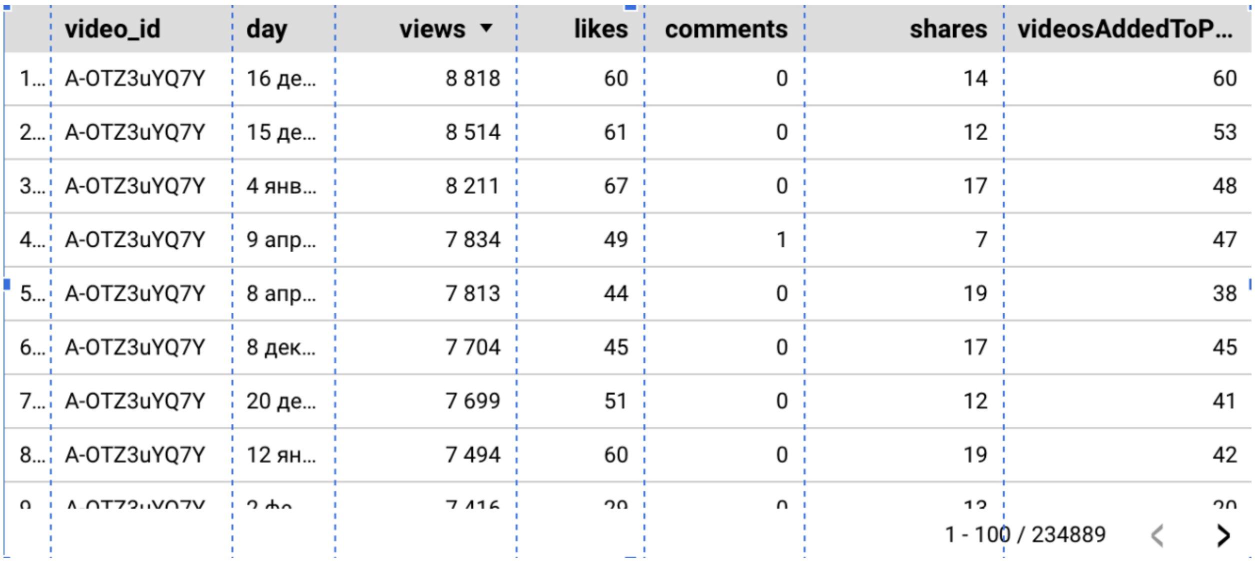
These visualizations provided a clear and concise summary of the data trends, making it easier to derive insights.



Picture 10. Line chart



Picture 11. Pie chart distribution



	video_id	day	views ▼	likes	comments	shares	videosAddedToP...
1...	A-OTZ3uYQ7Y	16 де...	8 818	60	0	14	60
2...	A-OTZ3uYQ7Y	15 де...	8 514	61	0	12	53
3...	A-OTZ3uYQ7Y	4 янв...	8 211	67	0	17	48
4...	A-OTZ3uYQ7Y	9 апр...	7 834	49	1	7	47
5...	A-OTZ3uYQ7Y	8 апр...	7 813	44	0	19	38
6...	A-OTZ3uYQ7Y	8 дек...	7 704	45	0	17	45
7...	A-OTZ3uYQ7Y	20 де...	7 699	51	0	12	41
8...	A-OTZ3uYQ7Y	12 ян...	7 494	60	0	19	42
9	A-OTZ3uYQ7Y	2 фе...	7 416	29	0	13	20

1 - 100 / 234889

Picture 12. Table

Machine Learning Model Training

First, I would prepare and preprocess the dataset, ensuring that it is clean and ready for training. Once the data is prepared, I would choose an appropriate machine learning model. After selecting the model, I would set up the training job on AI Platform by specifying the training data location, the model implementation code (either in Python scripts or a Jupyter notebook), and the hyperparameters (such as learning rate, batch size, number of epochs, etc.).

Model Deployment

Once the model is trained and evaluated, I would deploy it on AI Platform's serving capabilities. This would involve uploading the trained model to Google Cloud Storage, if it's not already there, creating a model version in AI Platform, setting up the API endpoint for making predictions. This would allow the model to receive input data and return predictions in real time.

Monitoring and Logging

I would also set up any necessary monitoring and logging for the deployed model, ensuring that it's performing as expected in production.

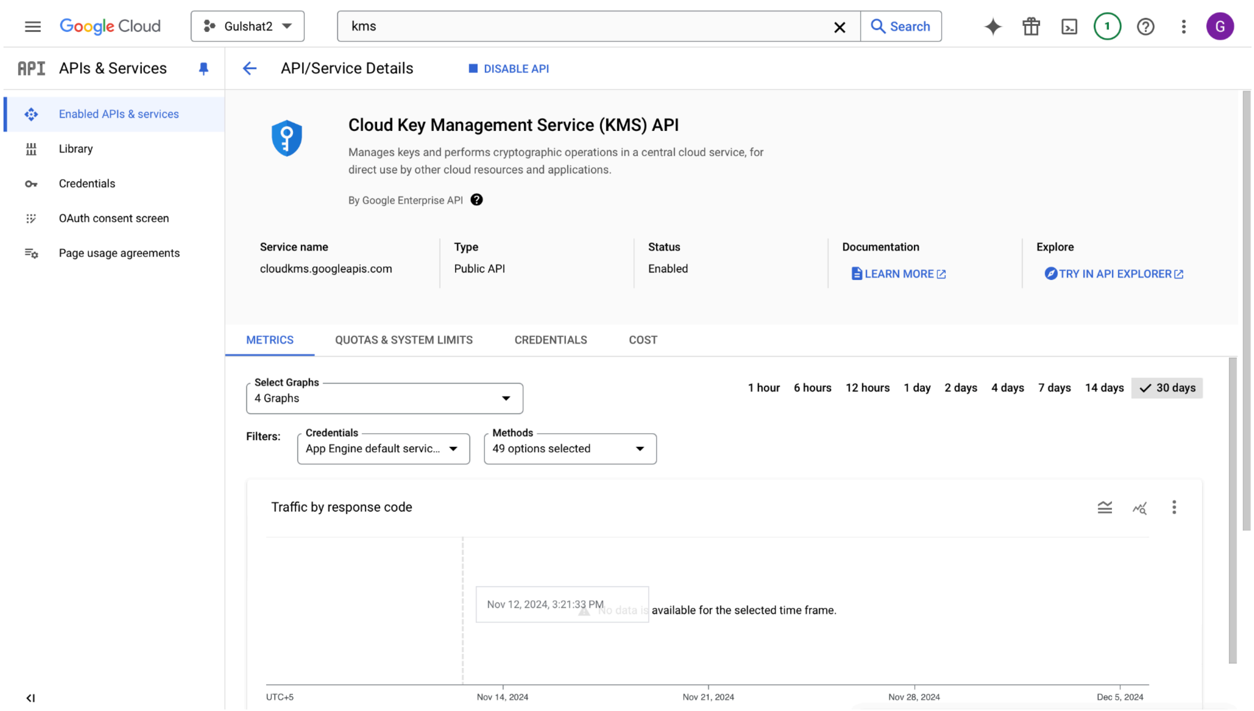
Cloud Security and Compliance

Identity and Access Management (IAM)

In terms of Identity and Access Management (IAM), I have configured a set of roles and permissions to ensure that only authorized individuals and services have access to cloud resources. I could assign various roles, such as Viewer, Editor, and Owner to different teams based on their responsibilities within the project. Additionally, custom roles have been created to grant specific permissions tailored to the needs of different users. Service accounts with precise permissions are set up for automated processes, ensuring that applications and services interact with the cloud securely.

Data Encryption

Google Cloud automatically encrypts data at rest using Google-managed encryption keys (GMEK) by default. This applies to most Google Cloud services, including Cloud Storage, Compute Engine, and BigQuery. But to take more control over the encryption process, we can use Customer-Managed Encryption Keys (CMEK) as shown in Picture 13. This gives the ability to create, manage, and control access to the encryption keys.



Cloud Key Management Service (KMS) API

Manages keys and performs cryptographic operations in a central cloud service, for direct use by other cloud resources and applications.

By Google Enterprise API

Service name	Type	Status	Documentation	Explore
cloudkms.googleapis.com	Public API	Enabled	LEARN MORE	TRY IN API EXPLORER

Select Graphs: 4 Graphs

Filters: Credentials: App Engine default servic... Methods: 49 options selected

Traffic by response code

Nov 12, 2024, 3:21:33 PM

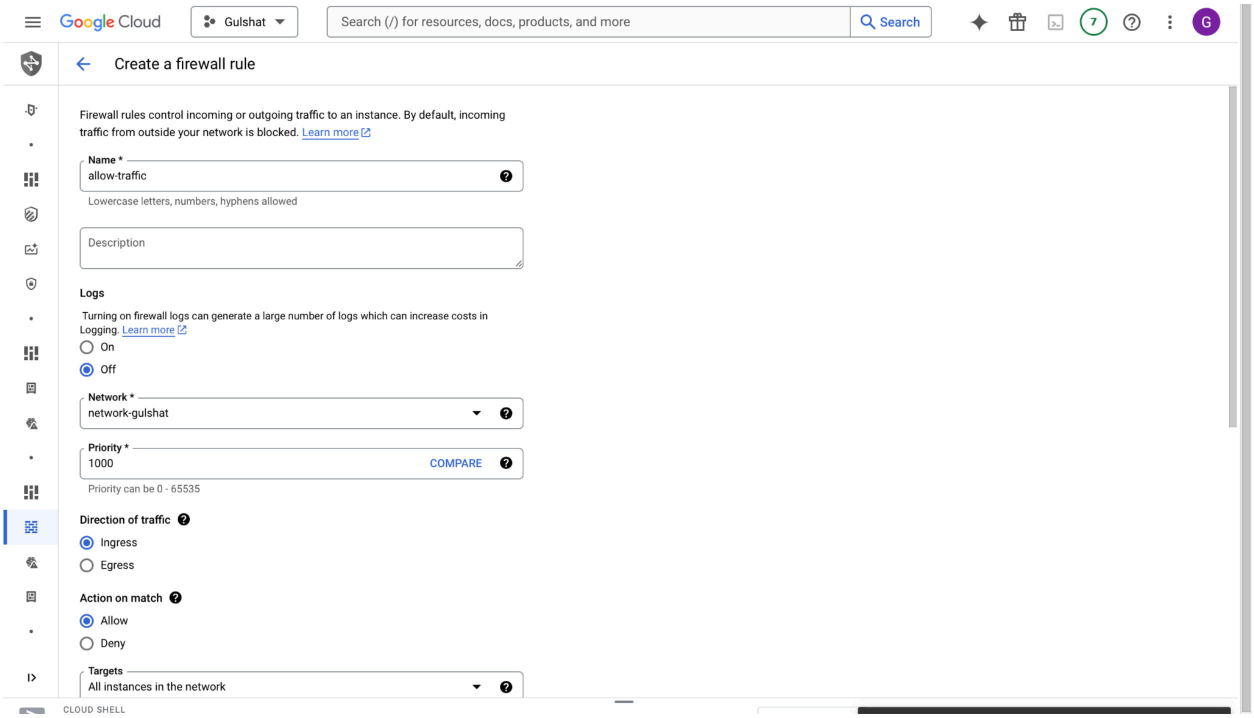
No data is available for the selected time frame.

Picture 13. Enabling KMS API

Network Security

A **Virtual Private Cloud (VPC)** is a private network within Google Cloud, where resources like virtual machines (VMs) and databases reside. **Firewall rules** control the traffic flow to and from resources within a VPC. These rules specify which traffic is allowed or denied based on **IP ranges**, **ports**, and **protocols**. They can be applied to specific resources using **tags** or **service accounts**. Google Cloud provides default rules that allow **internal communication** and **egress traffic**, while denying external ingress traffic unless specified otherwise. In picture 14 shown an example of firewall rule named **allow-traffic**.

By configuring VPCs and firewall rules, we can secure our network, control access, and isolate resources.



Create a firewall rule

Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. Learn more

Name *
allow-traffic
Lowercase letters, numbers, hyphens allowed

Description

Logs

Turning on firewall logs can generate a large number of logs which can increase costs in Logging. Learn more

○ On
● Off

Network *
network-gulshat

Priority *
1000 COMPARE
Priority can be 0 - 65535

Direction of traffic

● Ingress
○ Egress

Action on match

● Allow
○ Deny

Targets
All instances in the network

Picture 14. Example of firewall rule

Audit Logging

In Google Cloud, **Audit Logs** capture events that occur in your cloud environment, helping track access to resources and changes made. These logs are automatically enabled for all Google Cloud services. Audit logs help maintain compliance and security by providing detailed records of all actions taken within our cloud environment.

Compliance Standards

Google Cloud adheres to various **compliance standards** to ensure that your data is handled securely and in accordance with global regulations.

1) **ISO 27001**: Ensures the security of data processing.
2) **SOC 1, SOC 2, SOC 3**: Audits for controls related to security, availability, and confidentiality.
3) **GDPR**: Ensures compliance with European Union data protection regulations.
4) **HIPAA**: Supports healthcare-related compliance.
5) **PCI DSS**: Ensures secure payment card data handling.

Incident Response Planning

My **Incident Response Plan (IRP)** is designed to ensure a rapid and effective response to any security incidents that may arise in our system. The plan consists of the following steps:

1. Preparation: We define the roles and responsibilities of the incident response team and equip them with the necessary tools and resources.
2. Identification: We use monitoring tools and alerts to quickly identify potential incidents.
3. Containment: Once an incident is identified, we take immediate steps to contain the issue and prevent it from affecting other parts of the system.
4. Eradication: After containment, we remove the root cause of the incident to ensure that it cannot recur.
5. Recovery: We restore normal operations and verify that all systems are secure.
6. Post-Incident Review: We conduct a thorough review of the incident to learn from it and update our response plan based on the findings.

To ensure preparedness, I regularly run **incident response simulations** to test the effectiveness of the plan, refine the response process, and ensure readiness for any potential incidents.

Conclusion

In this report, I implemented a big data processing and machine learning pipeline using Google Cloud, while also focusing on applying security best practices and compliance measures. To ensure the project adhered to security and compliance standards, I implemented measures such as Identity and Access Management (IAM), data encryption, network security, and audit logging. These steps ensured the project remained secure and met compliance requirements. Overall, I was able to demonstrate how Google Cloud's tools and services can be used to effectively manage big data, build machine learning models, and maintain strong security and compliance.

References

https://cloud.google.com/compute/docs/access/create-enable-service-accounts-for-instances#changeserviceaccountandscopes
https://cloud.google.com/compute/docs/tutorials/basic-webserver-apache#:~:text=This%20tutorial%20shows%20you%20how%20to
https://cloud.google.com/vpc/docs/create-modify-vpc-networks#gcloud
https://cloud.google.com/logging/docs/audit
https://docs.google.com/document/d/1gZarAFyBdXm3vLIvBpW6MVGcaLQbX0sW-ufMhbYgERk/edit?tab=t.0
https://cloud.google.com/bigquery