Title: “Big Data and Machine Learning, and Cloud Security and Compliance on Google Cloud”

Kassym Orakbay

03.12.2024

Kazakh-British Technical University

1. **Executive Summary**

   This Assignment includes wide range of instruments and concepts to work with, such as data storing, machine learning, logging and security management. Moreover, it highlights key findings and implementations in the domains of Data Storage, Machine Learning, and security practices, providing a comprehensive overview of how these elements work together to create efficient, secure, and scalable systems.

Table of Contents

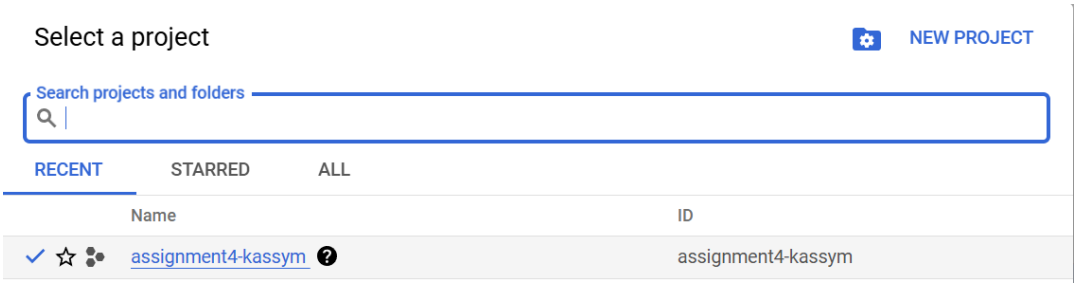
1. Introduction

The importance of processing Big Data continues to grow as humanity generates vast amounts of data every minute. Simultaneously, cloud computing and storage technologies are becoming increasingly popular worldwide. This makes it essential to integrate modern concepts to process and analyze data effectively, producing meaningful insights. Furthermore, leveraging advanced tools to implement Machine Learning (ML) models and build analytics on stored data enhances the value derived from these datasets.

In addition, platforms like Google Cloud Platform (GCP) offer comprehensive solutions for security management, ensuring the protection and integrity of data. GCP provides built-in tools for encrypting data, managing access controls, and monitoring threats, making it a reliable platform for secure data processing and analysis. Combining these capabilities allows organizations to harness the full potential of their data while maintaining robust security standards.

This assignment and report aim to teach basic concepts and techniques of using ML and Cloud storage along with tools for logging and securing data to build sustainable project.

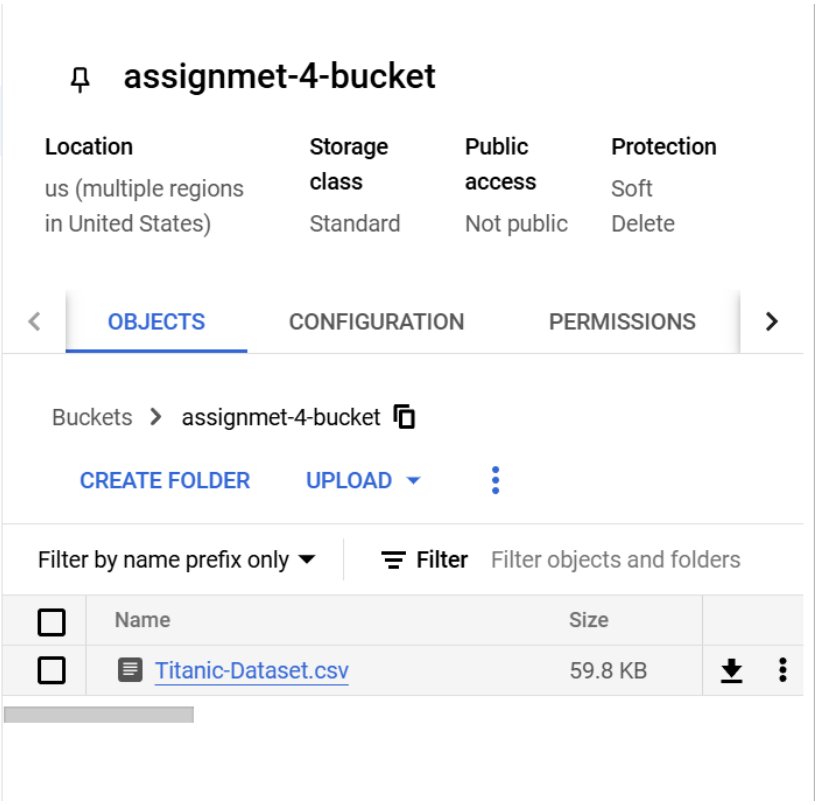
1. **Set Up a Google Cloud Project:**
   - Create a new project in the Google Cloud Console.
   - Enable necessary APIs (e.g., BigQuery, Cloud Storage, AI Platform).



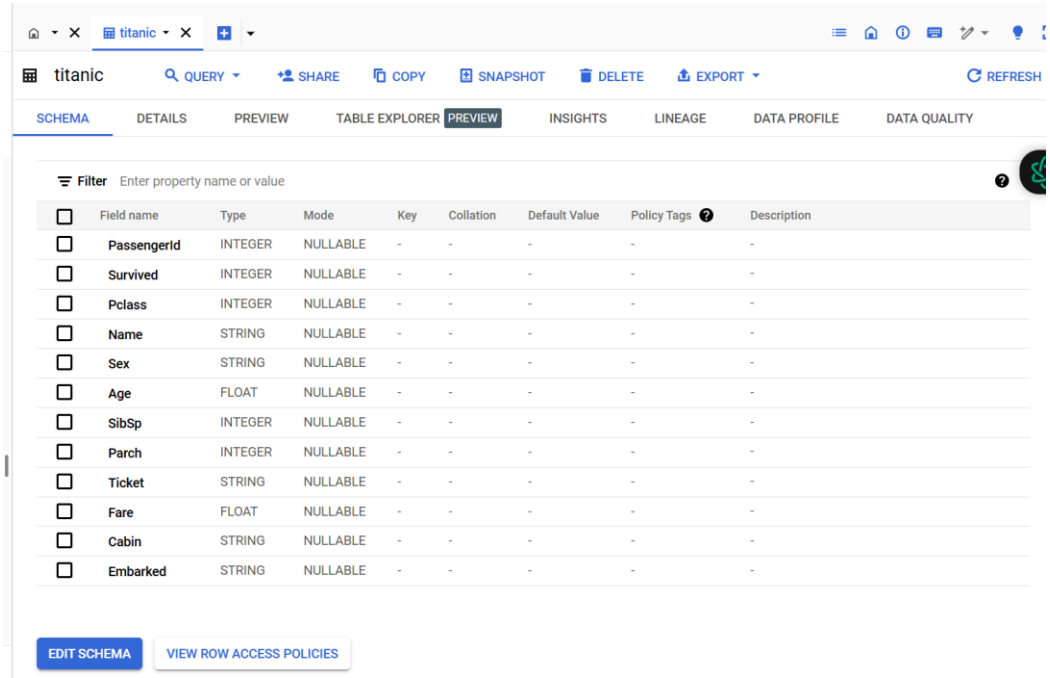
1. **Data Ingestion:**

   - Collect a large dataset relevant to your use case (e.g., public datasets from Kaggle or Google Dataset Search).

   - Upload the dataset to Google Cloud Storage.

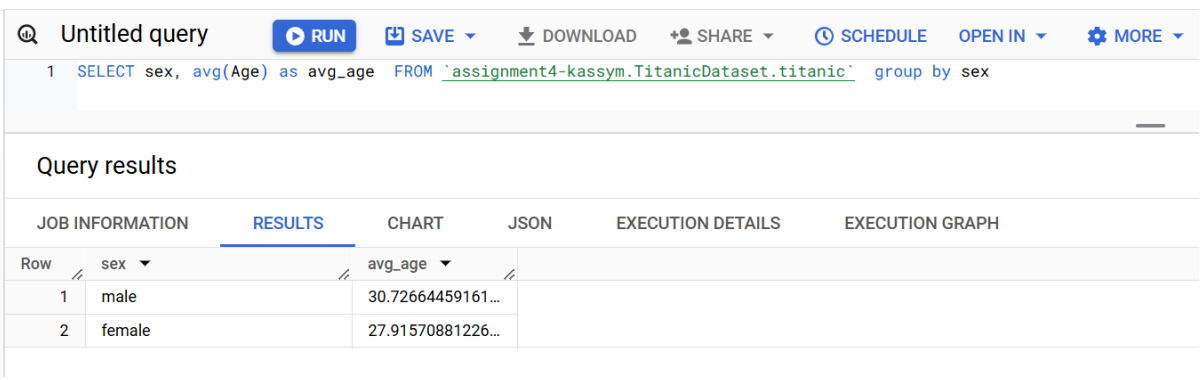


1. **Data Processing with BigQuery**
   - Use BigQuery to create a dataset and load the data from Cloud Storage.

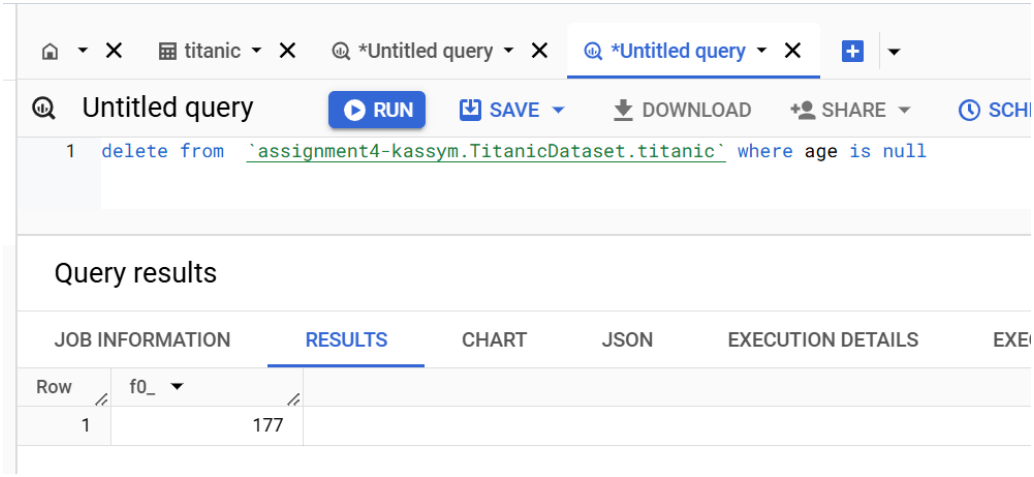


Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
PassengerId	INTEGER	NULLABLE	-	-	-	-	-
Survived	INTEGER	NULLABLE	-	-	-	-	-
Pclass	INTEGER	NULLABLE	-	-	-	-	-
Name	STRING	NULLABLE	-	-	-	-	-
Sex	STRING	NULLABLE	-	-	-	-	-
Age	FLOAT	NULLABLE	-	-	-	-	-
SibSp	INTEGER	NULLABLE	-	-	-	-	-
Parch	INTEGER	NULLABLE	-	-	-	-	-
Ticket	STRING	NULLABLE	-	-	-	-	-
Fare	FLOAT	NULLABLE	-	-	-	-	-
Cabin	STRING	NULLABLE	-	-	-	-	-
Embarked	STRING	NULLABLE	-	-	-	-	-

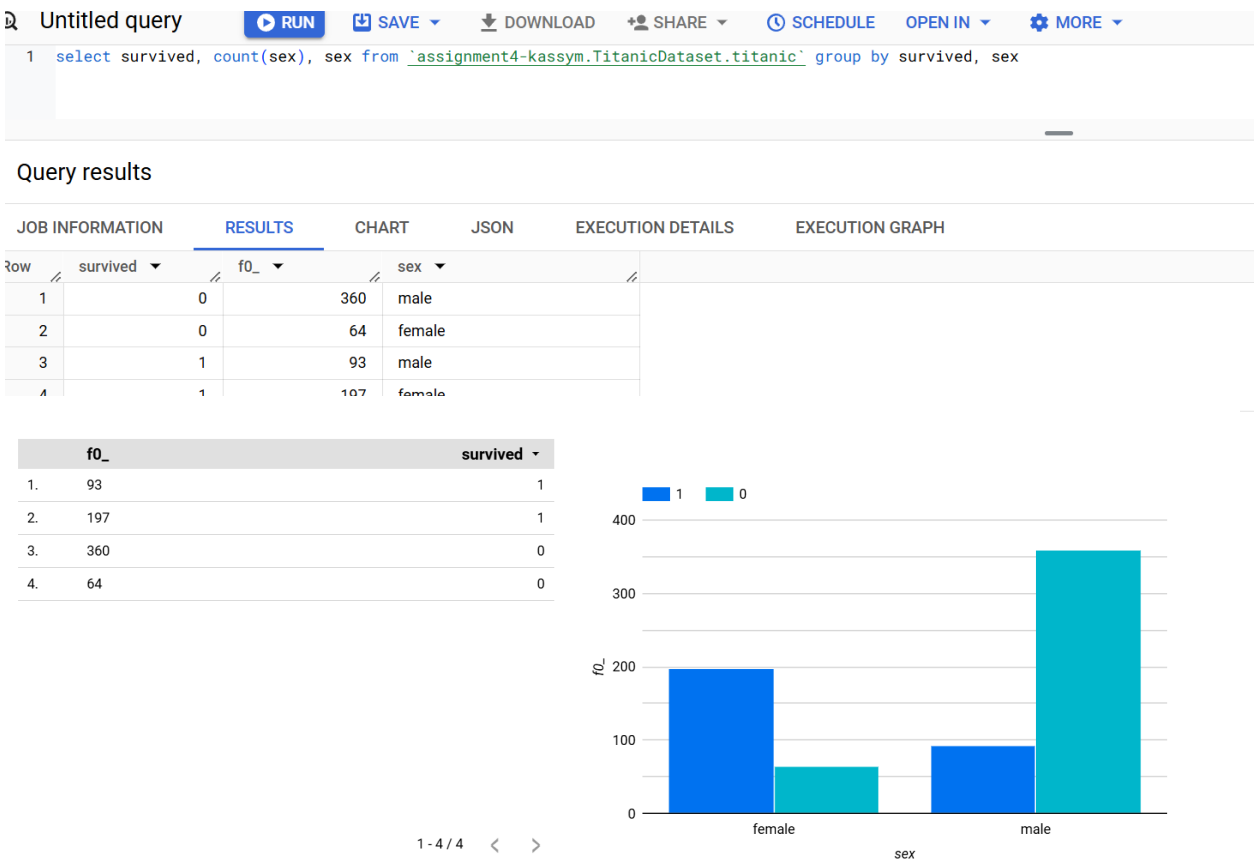
   - Perform data cleaning and preprocessing using SQL queries (e.g., filtering, aggregating).



Row	sex	avg_age
1	male	30.72664459161...
2	female	27.91570881226...



- Create summary statistics and visualize the results using Google Data Studio or similar tools.



Отчеты Looker Studio - 03.12.2024, 10:46

Файл Изменить Просмотреть Вставить Страница Упорядочить Ресурс Справка

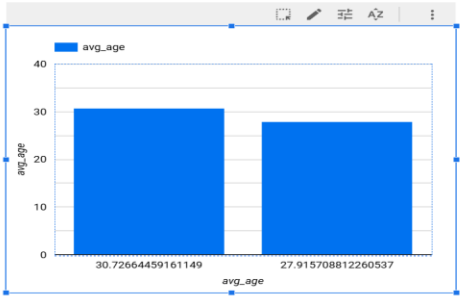
Добавить страницу | Добавить данные | Добавить диаграмму | Добавить элемент управления

+ Добавить быстрый фильтр | Сбросить

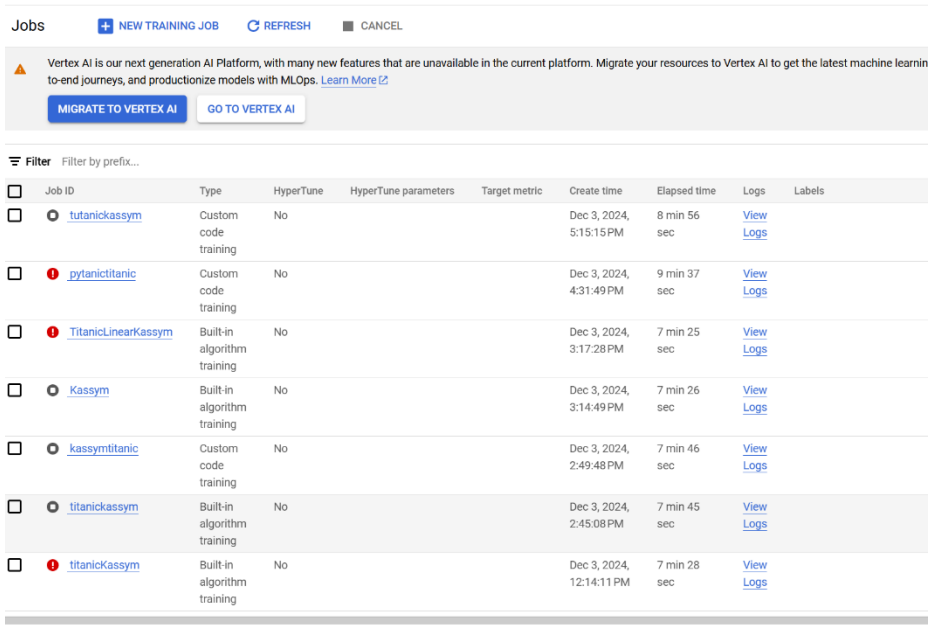
titanic

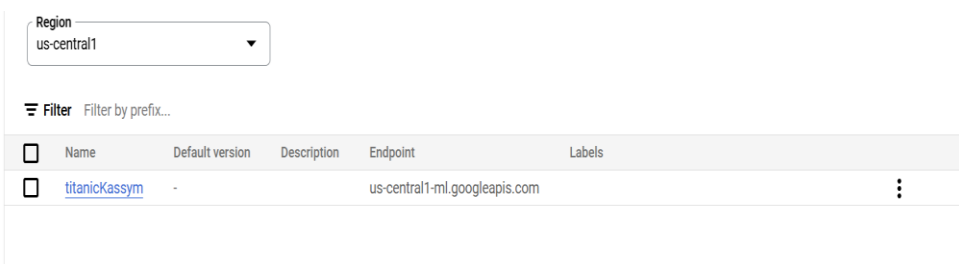
	avg_age	Record Count
1.	30.72664459161149	1
2.	27.915708812260537	1

1 - 2 / 2



1. **Machine Learning Model Training:**
    - Use the AI Platform to train a machine learning model on the processed data.
    - Choose a model suitable for the task (e.g., classification, regression) and implement it using TensorFlow or Scikit-learn.
    - Set up a training job on AI Platform, specifying the necessary configurations (e.g., training data, hyperparameters).





Pre-built container settings

Python version *
3.7
Select the Python version you used to train the model

Framework
TensorFlow

Framework version
2.11.0

ML runtime version *
2.11

Model URI *
gs:// assignmet-4-bucket BROWSE
Cloud Storage path to the entire SavedModel directory. Learn more

Online prediction deployment

Scaling
Auto scaling

Minimum number of nodes
1
Keeping a minimum number of nodes running all the time will avoid dropping requests due to nodes initialization after the service has scaled down. This setting can increase cost, as you pay for the nodes even when no predictions are served.

Machine type *
n1-standard-4, 4 vCPUs, 15 GB memory

Accelerator type
Accelerator count

Serv
Spec

Based on the model region and the machine type selected, the available accelerator types and the numbers of accelerators that can be selected may vary.

SAVE CLEAR CANCEL

Training data

Use single file stored in a GCS bucket

Data path *
gs:// assignmet-4-bucket/Titanic-Dataset.csv BROWSE
The Cloud Storage path where the training data is stored

Validation data

Use a percentage of training data

Percentage of splitting *
25 %

Test data (Optional)

Use a percentage of training data

Percentage of splitting
15 %

Training output

Output directory *
gs:// assignmet-4-bucket BROWSE
The path to the Google Cloud Storage location where you want the trained model and other training job output to be stored.

PREVIOUS NEXT CANCEL

Learn more

Training algorithm — Training data — Algorithm arguments — **Job settings**

Job ID
titanicKassym

Must start with a letter and contain only letters, numbers, and underscores. Case-sensitive. Can't be changed later. 13 / 128

Region *
us-central1

The Google Cloud Platform region where the training job runs. For efficiency, the region you select should match the region where your training data is stored in Cloud Storage.
Learn more

Scale tier *
BASIC

The resources AI Platform allocates to your training job. Learn more

PREVIOUS PROCESSING... CANCEL

Job Details CANCEL

Vertex AI is our next generation AI Platform, with many new features that are unavailable in the current platform. Migrate your res
MLOps. Learn More

MIGRATE TO VERTEX AI GO TO VERTEX AI

titanicKassym

Preparing (1 min 42 sec)

Creation time	Dec 3, 2024, 12:14:11 PM
Start time	
End time	
Logs	View Logs
Training input	

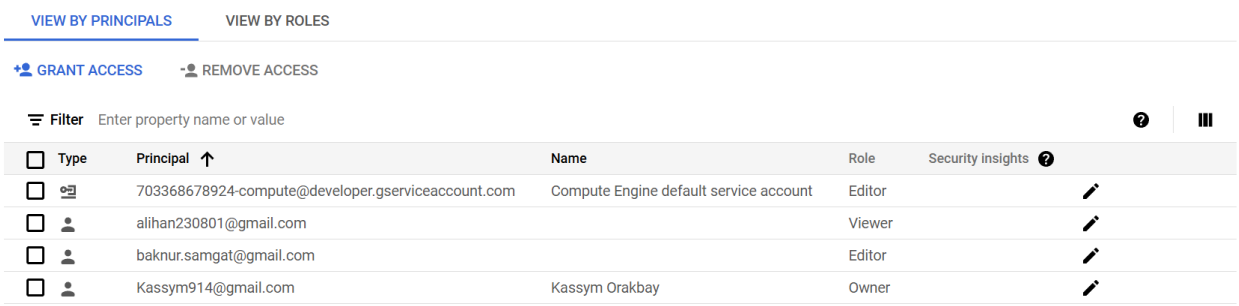

```
{
  "args": [
    "--preprocess",
    "--training_data_path=gs://assignmet-4-bucket/Titanic-Dataset.csv",
    "--validation_split=0.25",
    "--test_split=0.15",
    "--objective=reg:linear",
    "--eval_metric=eval_metric",
    "--booster=gblinear",
    "--colsample_bylevel=1",
    "--num_boost_round=8",
    "--max_depth=6",
    "--eta=0.3",
    "--csv_weight=0",
    "--base_score=0.5",
    "--updater=shotgun"
```

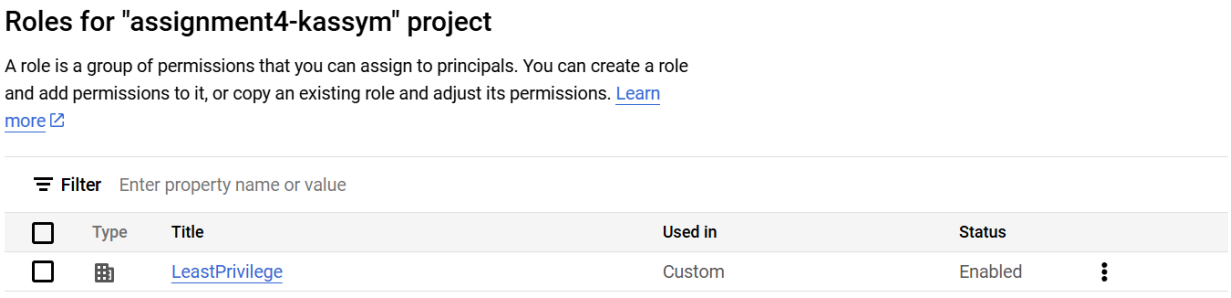
Exercise 2: Cloud Security and Compliance

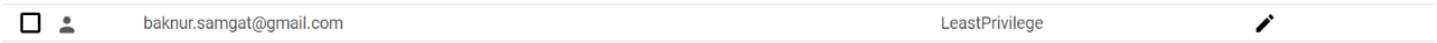
Objective: Implement security best practices and compliance measures for a Google Cloud project.

Tasks:

1. **Identity and Access Management (IAM):**
   - Configure IAM roles and permissions for different users in your project.
   - Implement the principle of least privilege for service accounts and users.





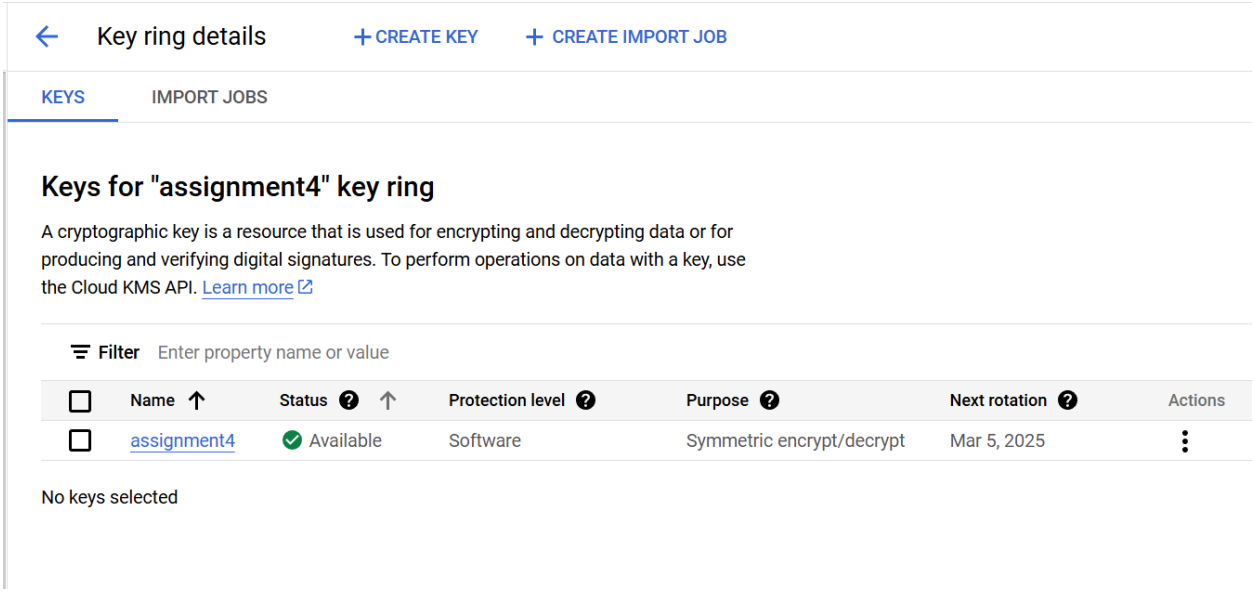


2. **Data Encryption**

- Ensure that data is encrypted at rest and in transit.

  - All data stored in Google Cloud is automatically encrypted at rest by default. No additional configuration is needed.

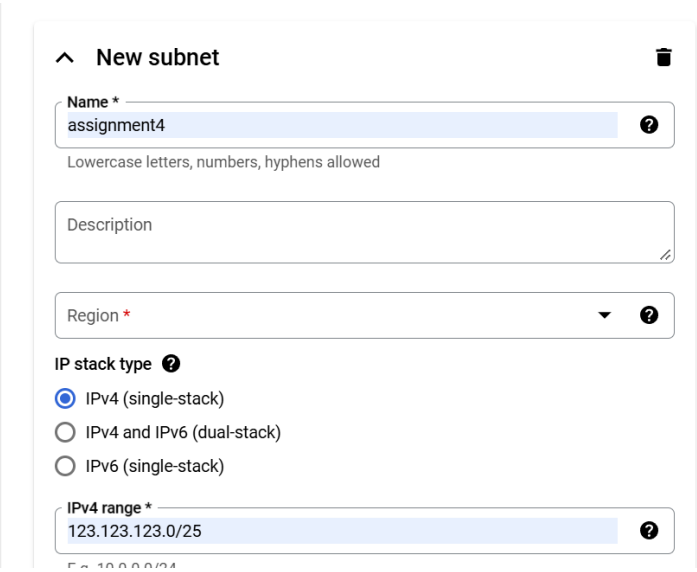
- Utilize Google Cloud KMS for managing encryption keys.

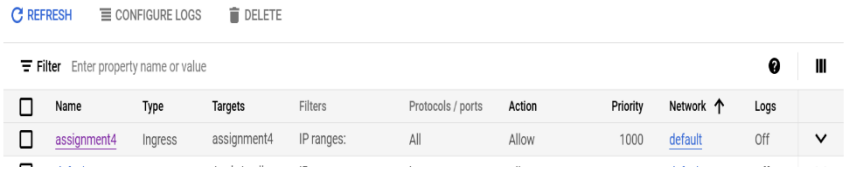


3. **Network Security**

- Set up Virtual Private Cloud (VPC) and configure firewall rules to restrict inbound and outbound traffic.

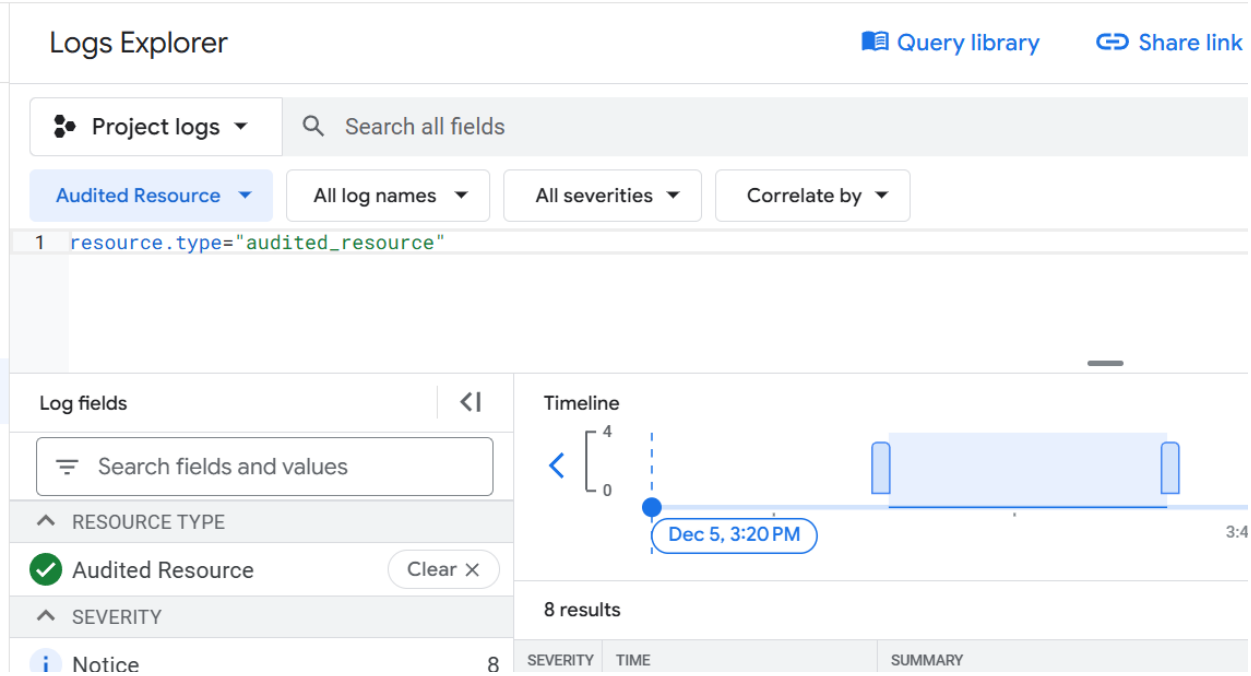
- Implement private Google access and ensure that sensitive data is not exposed to the public internet.





4. **Audit Logging**

- Enable Cloud Audit Logs to track access and changes to your resources.
- Review logs for unusual activities and set up alerts for suspicious events.



Conclusion

In conclusion, the integration of Big Data processing, Machine Learning, and robust security practices within cloud platforms like GCP is crucial for organizations aiming to stay competitive in the data-driven era.

Recommendations

Assignments that incorporate diverse programming concepts, such as Machine Learning, cloud computing, and web development, become more effective when approached through teamwork. Collaboration not only saves time but also enhances the flow of knowledge and ideas among team members, leading to better learning outcomes.

References

- https://cloud.google.com/docs/security/encryption-in-transit

- https://sprinto.com/blog/compliance-standards/