



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

Author's Name: Kerimbay Kairuddin

Date of Submission: December 04, 2024

Organization Name: Kazakh British Technical University (KBTU)

Almaty 2024

Exercise 1: Big Data and Machine Learning on Google Cloud

Tasks:

1. Set Up a Google Cloud Project:

I utilized a previously established project from Assignment 1.

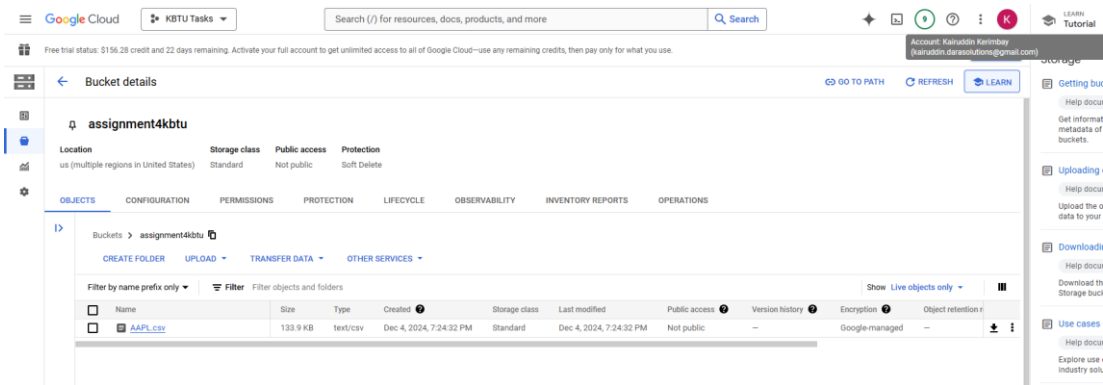
Within the 'APIs & Services' section, under 'Library,' I enabled the following APIs:

The image contains two screenshots from the Google Cloud console. The top screenshot shows the 'API/Service Details' page for the 'AI Platform Training & Prediction API'. It indicates the API is enabled and provides links to documentation and the API Explorer. The bottom screenshot shows the 'Product details' page for 'Cloud Storage' and 'BigQuery API', both of which are also enabled. The interface includes navigation menus on the left and top, and a user account dropdown in the top right corner.

2. Data Ingestion:

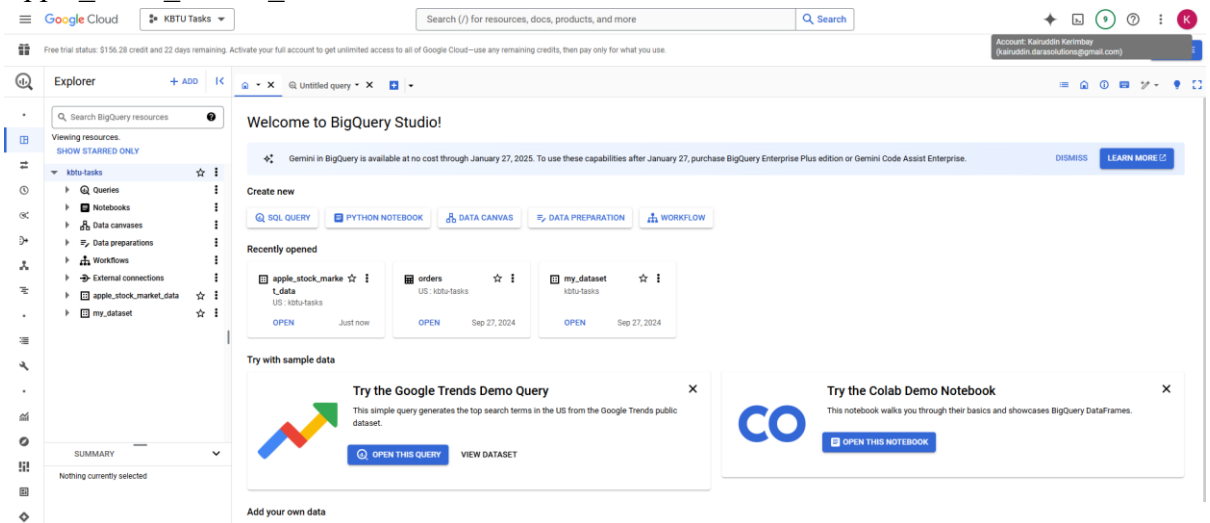
I sourced a dataset from Kaggle.com and subsequently uploaded it to a Google Cloud Storage bucket.

The image is a screenshot of a Kaggle dataset page titled 'Apple Stock Market Data (2020-2024)'. The page shows the dataset's description, a download button, and a 'New Notebook' option. The dataset is categorized under 'Data Card', 'Code (1)', 'Discussion (0)', and 'Suggestions (0)'. The Kaggle logo and navigation menu are visible on the left side of the page.

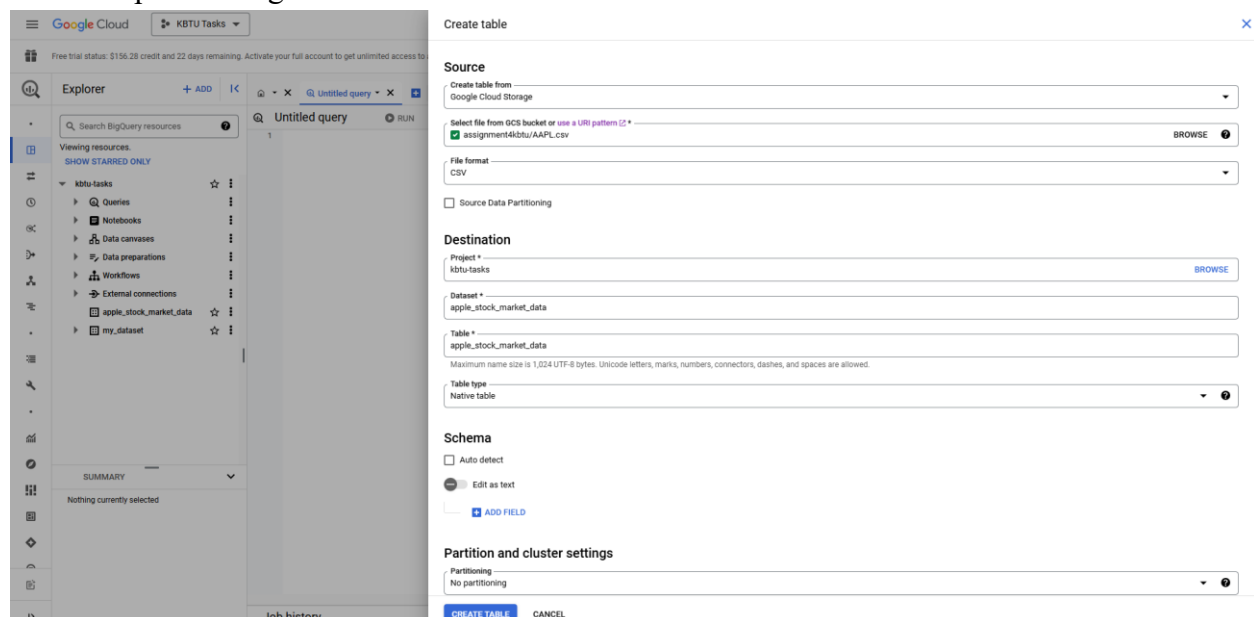


3. Data Processing with BigQuery:

In the BigQuery panel of the 'kbtu-tasks' project, I created a new dataset named 'apple_stock_market_data'.



In the BigQuery interface within the 'kbtu-tasks' project, I configured a new table named 'apple_stock_market_data.' This table was set up to import data from a Google Cloud Storage CSV file, leveraging the comprehensive settings available for schema definition and table partitioning.



Google Cloud

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

Free trial status: \$156.28 credit and 22 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Account: Karutdin Karimay (karutdin.darazsolutions@gmail.com)

Explorer

Search BigQuery resources

Showing resources

SHOW STARRED ONLY

- ↳ Kbtu-tasks
- ↳ Queries
- ↳ Notebooks
- ↳ Data canvases
- ↳ Data preparations
- ↳ Workflows
- ↳ External connections
- ↳ **apple_stock_market_data**
- ↳ **apple_stock_market_data**
- ↳ **my_dataset**

Summary

Nothing currently selected

Query results

SAVE RESULTS EXPLORE DATA

Press Alt+H1 for Accessibility O

Row	Price	Adj. Close	Close	High	Low	Open	Volume
1	2020-07-21	94.51815795898438	97.0	99.25	96.74250030517578	99.17250061035156	103433200
2	2020-09-11	109.33113098144531	112.0	115.2300033569336	110.0	114.5699996482422	180860300
3	2020-10-05	113.72389221191406	116.5	116.6500015258789	113.55000305175781	113.91000366210938	106243800
4	2021-04-01	120.45589447021484	123.0	124.18000030517578	122.48999786376953	123.66000366210938	75089100
5	2021-02-02	123.39389223388672	126.0	129.72000122070312	125.5999984741211	128.00999450683594	103916400
6	2021-04-09	130.24903395507812	133.0	133.0999932861328	129.47000122070312	129.8000030517578	106686700
7	2021-04-21	130.7387237548828	133.5	133.75	131.3000030517578	132.36000061035156	68847100
8	2021-04-15	131.71798706054688	134.5	135.0	133.63999938964844	133.82000732421875	89347100
9	2022-12-15	135.0795135498047	136.5	141.8000030517578	136.02999877929688	141.11000061035156	98931900
10	2022-11-08	138.0482940673828	139.5	141.42999257578125	137.490000549316406	140.01000366210938	89908500
11	2021-09-30	139.0166015625	141.5	144.3800048828125	141.27999877929688	143.66000366210938	89056700
12	2021-10-06	139.5078125	142.0	142.14999389648438	138.3699951171875	139.47000122070312	83221100
13	2023-01-30	141.51187133789062	143.0	145.5500030517578	142.85000610351562	144.9600067138672	64015300
14	2021-07-12	141.75155639648438	144.5	146.32000732421875	144.0	146.2100067138672	76299700
15	2022-05-11	144.53457641601562	146.5	155.4499969482422	145.80999755859375	153.5	142689800

Results per page: 50 1 - 50 of 1000

REFRESH

.inh histrv

I generated a new report in Google Data Studio, selecting BigQuery as the data source." This version enhances the professionalism and clarity of the statement, making it suitable for a formal document.

lookerstudio.google.com/u/0/reporting/5e14e9cf-3992-4ca8-8803-15d37a8844ab/page/f2XE/edit

Отчет без названия

Файл Просмотреть Страница Справка

Сбросить Поделиться Открыть

Добавить страницу Добавить данные Добавить диаграмму Добавить элемент управления Тема и шаблон Приостановить обновление

Добавьте данные в отчет

Учетные данные: X

Загружайте отчеты BigQuery быстрее с BigQuery VE Engine. Подробнее

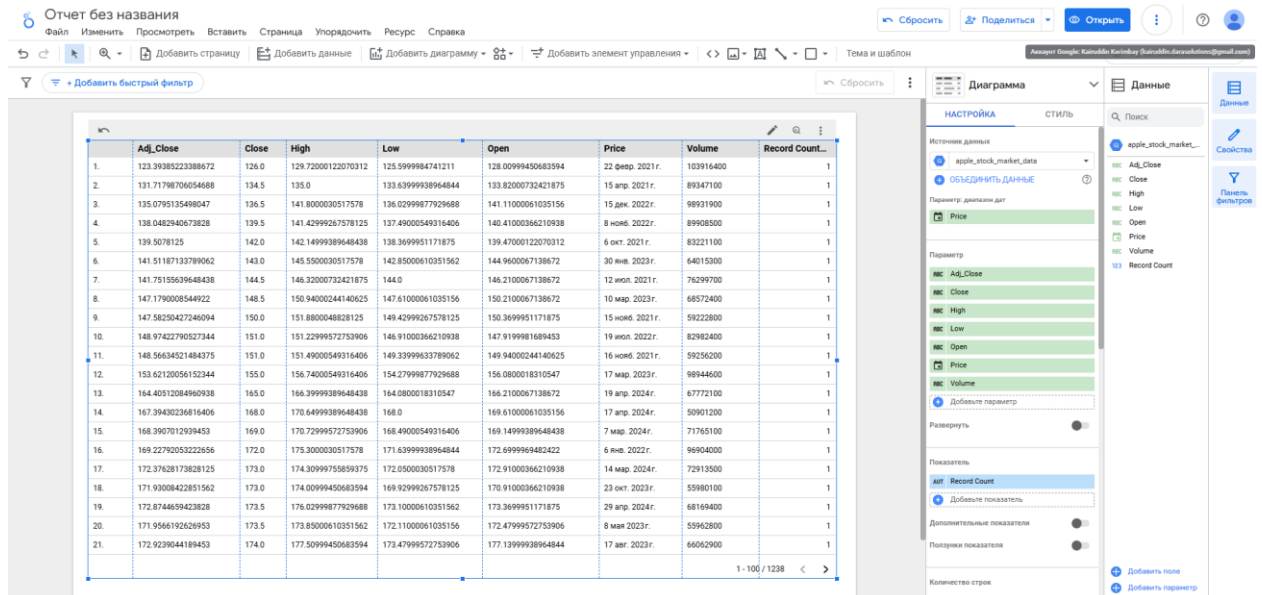
BigQuery

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.

ПОДРОБНЕЕ... СООБЩИТЬ О ПРОБЛЕМЕ

НЕДАВНИЕ ПРОЕКТЫ	Project	Набор данных	Table
МОИ ПРОЕКТЫ	Search Projects	Search Datasets	Search Tables
ДОСТУПНЫЕ МНЕ ПРОЕКТЫ	Enter Project Id manually	apple_stock_market_data	apple_stock_market_data
ПОЛЬЗОВАТЕЛЬСКИЙ ЗАПРОС		my_dataset	
НАБОРЫ ОБЩЕДОСТУПНЫХ ДАННЫХ	KBTU Tasks		
	Second KBTU		



4. Machine Learning Model Training:

Train new model

- Training method
- Model details
- Training container
- Hyperparameters (optional)
- Compute and pricing
- Prediction container (optional)

Custom container

Build a custom Docker container. Must be stored in [Container Registry](#)

Model framework *
TensorFlow

Model framework version *
2.1

Pre-built container settings

Before you begin, you need to package and upload your application code and dependencies to a Cloud Storage bucket. [Learn more](#)

Package location (Cloud Storage path) *
☒ gs://assignment4kbtu/training_model.zip [BROWSE](#)

[Learn how to package and upload](#) your application code and dependencies

[+ ADD PACKAGE](#)

Python module *

Model output directory
☒ gs://assignment4kbtu [BROWSE](#)

Your model artifacts and other data needed for training will be stored on Cloud Storage. You should specify a path here if you do not set an output directory in your application code or arguments.

Arguments

Optional. Add arguments for the command that runs when the container starts. Overrides the container's CMD instruction. Enter one parameter and its argument per line.

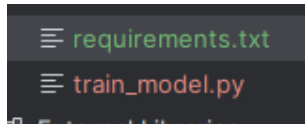
For parameters you want to tune with HyperTune, enter arguments of the hyperparameters you defined in the training code in the HyperTune setting below. If none, click Next to skip this step.

[CONTINUE](#)

Here the python code:

```
import tensorflow as tf
from sklearn.model_selection import train_test_split
import pandas as pd
df = pd.read_csv('gs://assignment4kbtu/AAPL.csv')
df.fillna(0, inplace=True)
X_train, X_test, y_train, y_test = train_test_split(df.drop('target', axis=1), df['target'], test_size=0.2)
model = tf.keras.models.Sequential([
    tf.keras.layers.Dense(128, activation='relu', input_shape=(X_train.shape[1],)),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test))
model.save('gs://assignment4kbtu/model_output_directory')
```

In than zip folder there these files



In requirements.txt there is this data, It is necessary to import dependencies:

tensorflow==2.6.0

pandas

scikit-learn

Here I chose basic machine, accelerator and disk types

A screenshot of the Google Cloud AI Platform 'Train new model' configuration page. The left sidebar shows a progress bar with steps: Training method, Model details, Training container, Hyperparameters (optional), Compute and pricing (selected), and Prediction container (optional). The main content area is titled 'Worker pool 0 (chief)' and contains several configuration sections. The 'Machine type' is set to 'n1-standard-4, 4 vCPUs, 15 GiB memory'. The 'Accelerator type' is set to 'NVIDIA_TESLA_T4'. The 'Accelerator count' is set to 1. The 'Worker count' is set to 1. The 'Disk type' is set to 'Standard'. The 'Disk size' is set to 100 GB. There is a 'Reservations' section with a warning that only A2 and A3 machine types are supported. The 'Reservation type' is set to 'On-demand'. The 'Availability policies' section shows the 'VM provisioning model' set to 'Standard'. The bottom of the page shows a Windows taskbar with various application icons.

Train new model

- ✓ Training method
- ✓ Model details
- ✓ Training container
- ✓ Hyperparameters (optional)
- ✓ **Compute and pricing**
- 6 Prediction container (optional)

START TRAINING **CANCEL**

will eliminate the job startup time that's otherwise needed for compute resource creation. It does not incur additional cost to run jobs on an existing persistent resource. [Learn more](#)

☒ **Deploy to new worker pool:** The training service will find resources from the Compute Engine resource pool based on the specifications you provided. You will pay for the compute resources throughout job training time. [Learn more](#)

Worker pool 0 (chief)

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

Accelerator type
NVIDIA_TESLA_T4

Accelerators can speed up model training that involves intensive compute tasks. [Learn more](#)

Accelerator count
1

Worker count
1

Disk type
Standard

Disk size
100 GB

Reservations

i Only A2 and A3 machine types are supported to use with reservations. [LEARN MORE](#)

Reservation type
On-demand

Restrictions may apply when using reservations. [Learn more](#)

Availability policies

VM provisioning model
Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

The screenshot shows the Google Cloud Vertex AI console. The 'Training' tab is selected, displaying a table of training pipelines. A single pipeline named 'assignment4' is listed with a status of 'Pending'.

Name	ID	Status	Job type	Model type	Duration	Last updated	Created	Ended	Labels
assignment4	9166893972826619904	Pending	Training pipeline	Custom	—	Dec 5, 2024, 10:36:31 AM	Dec 5, 2024, 10:36:31 AM	—	—

The screenshot shows the Google Cloud Vertex AI console. The 'Training' tab is selected, displaying a table of training pipelines. A single pipeline named 'assignment4' is listed with a status of 'Failed'.

Name	ID	Status	Job type	Model type	Duration	Last updated	Created	Ended	Labels
assignment4	9166893972826619904	Failed	Training pipeline	Custom	58 sec	Dec 5, 2024, 10:39:11 AM	Dec 5, 2024, 10:36:31 AM	—	—

This is the result. And I tried several times, but only got a glitch.

The screenshot shows the Google Cloud Vertex AI console. The 'Training' tab is selected, displaying a table of training pipelines. A single pipeline named 'assignment4' is listed with a status of 'Failed'.

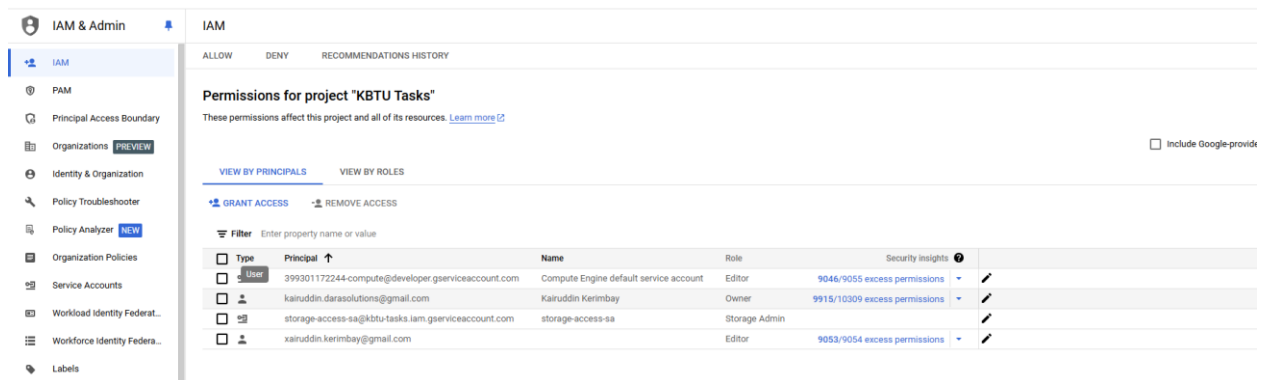
Name	ID	Status	Job type	Model type	Duration	Last updated	Created	Ended	Labels
assignment4	9166893972826619904	Failed	Training pipeline	Custom	15 min 35 sec	Dec 5, 2024, 10:54:46 AM	Dec 5, 2024, 10:36:31 AM	Dec 5, 2024, 10:54:46 AM	—

Exercise 2: Cloud Security and Compliance

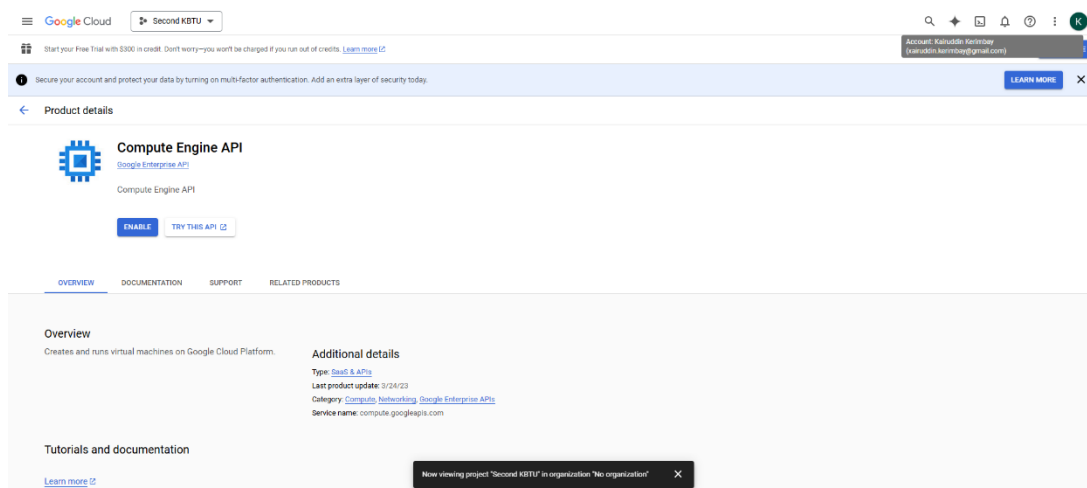
1. Identity and Access Management (IAM):

In the IAM & Admin section of the Google Cloud Console for the 'KBTU Tasks' project, I configured permissions for specific users. This involved assigning tailored roles to each user to control their access levels within the project. For example, one user was given the 'Editor' role, allowing them read-write access across most services, while another was assigned the 'Storage

Admin' role, focusing their permissions on cloud storage management.



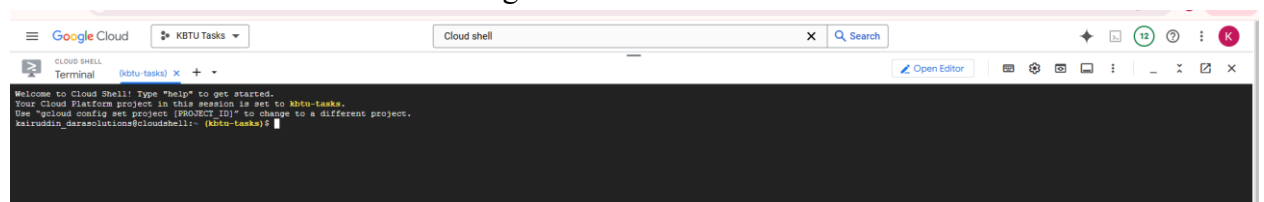
This screenshot depicts the user interface of the Google Cloud Platform as viewed from the account 'xairuddin.kerimbay'. The display shows the Compute Engine API overview page, indicating that this account has access to manage and view details related to Compute Engine services.



For more focused access management, there is a service account specifically dedicated to handling cloud storage operations: 'storage-access-sa@kbtu-tasks.iam.gserviceaccount.com'. This account is configured to limit its permissions strictly to cloud storage services, ensuring a higher level of security and compliance with the principle of least privilege.

2. Data Encryption:

To do these tasks I decided to use Google Cloud Shell



In the process of setting up data encryption, I utilized the Google Cloud Shell from within the 'kbtu-tasks' project environment to activate necessary services. As illustrated in the screenshot, I executed the command `gcloud services enable cloudkms.googleapis.com` to

enable the Google Cloud Key Management Service (KMS). This action was successfully completed as confirmed by the operation status message displayed in the terminal.

```
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ gcloud services enable cloudkms.googleapis.com
Operation "operations/acat.p2-399301172244-3b896cd4-1135-4dcc-ad3f-bc6e198d3baa" finished successfully.
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ ^C
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $
```

After enabling the necessary services, I proceeded to configure encryption keys using Google Cloud Key Management Service (KMS). Initially, I created a key ring named 'my-keyring' in the global location, as demonstrated by the command:

```
gcloud kms keyrings create "my-keyring" \ --location "global"
```

Following this, I created a symmetric encryption key named 'my-symmetric-key' within this key ring, designated specifically for encryption purposes. The command executed was:

```
gcloud kms keys create "my-symmetric-key" \  
  --location "global" \  
  --keyring "my-keyring" \  
  --purpose "encryption"
```

```
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ gcloud kms keyrings create "my-keyring" \
  --location "global"
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ gcloud kms keys create "my-symmetric-key" \
  --location "global" \
  --keyring "my-keyring" \
  --purpose "encryption"
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $
```

These commands were successfully run in the Google Cloud Shell, setting up the foundational security infrastructure for data encryption.

Created a plaintext file named 'data.txt' containing the string 'my-contents' using the echo command in the Google Cloud Shell. This file was then encrypted using the Google Cloud Key Management Service. The encryption process was executed with the following command sequence:

Create a plaintext file:

```
$ echo "my-contents" > ./data.txt
```

Encrypt the file:

```
$ gcloud kms encrypt \  
  --location "global" \  
  --keyring "my-keyring" \  
  --key "my-symmetric-key" \  
  --plaintext-file ./data.txt \  
  --ciphertext-file ./data.txt.enc
```

The output shows that the file 'data.txt.enc' was successfully created and lists its properties, confirming the encryption process was executed correctly. Subsequent verification commands display the

contents of both the plaintext and the encrypted file to demonstrate the successful encryption.

```
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ echo "my-contents" > ./data.txt
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ gcloud kms encrypt \
  --location "global" \
  --keyring "my-keyring" \
  --key "my-symmetric-key" \
  --plaintext-file ./data.txt \
  --ciphertext-file ./data.txt.enc
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ ls -l ./data.txt.enc
-rw-rw-r-- 1 kairuddin_darasolutions kairuddin_darasolutions 93 Dec  4 18:35 ./data.txt.enc
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ cat ./data.txt.enc

[?2004hkairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ sudo cat ./data.txt
my-contents
```

Following the encryption of the data, I proceeded to decrypt the file to verify the integrity and success of the encryption process. This was achieved using the Google Cloud Key Management Service with the command shown below

```
gcloud kms decrypt \
  --location "global" \
  --keyring "my-keyring" \
  --key "my-symmetric-key" \
  --plaintext-file - \
  --ciphertext-file ./data.txt.enc
```

```
kairuddin_darasolutions@cloudshell:~ (kbtu-tasks) $ gcloud kms decrypt \
  --location "global" \
  --keyring "my-keyring" \
  --key "my-symmetric-key" \
  --plaintext-file - \
  --ciphertext-file ./data.txt.enc
my-contents
```

The command specifies the encrypted file (data.txt.enc) and outputs the decrypted contents directly to the console, confirming that the original data ('my-contents') was successfully restored. This demonstrates the effective use of encryption and decryption processes managed through Google Cloud KMS.

3. Network Security:

In the setup of network security for the project, I configured a new Virtual Private Cloud (VPC) network named 'assignment4kairuddin' in Google Cloud. This configuration included setting up various subnets to ensure a structured and secure network environment across multiple regions, as shown in the screenshot.

The VPC was set to custom mode, allowing for the explicit definition of subnets rather than using automatic subnet creation. This approach gives precise control over the IP ranges and locations, enhancing the security and efficiency of network resource allocation.

Key details of the setup included:

- Name of the VPC: assignment4kairuddin
- Subnet creation mode: Custom, allowing for tailored subnet management.
- IP ranges: Configured to align with regional requirements, ensuring optimal performance and compliance with data residency regulations.
- MTU settings: Set at 1460, the default for Google Cloud, ensuring effective data transmission.

This configuration underscores the project's commitment to robust network security practices by segmenting network traffic and applying region-specific settings to meet operational and compliance needs.

Free trial status: \$156.23 credit and 22 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Account: kairuddin.kerimbay (kairuddin.darasolutions@gmail.com)

VPC Network

Create a VPC network

VPC networks

IP addresses

Internal ranges

Bring your own IP

Firewall

Routes

VPC network peering

Shared VPC

Serverless VPC access

Packet mirroring

VPC Flow Logs

Name *

assignment4kairuddin

Description

Maximum transmission unit (MTU)

1460

Subnet creation mode

Custom

Automatic

Subnets

Subnets let you create your own private cloud topology within Google Cloud. Click Automatic to create a subnet in each region, or click Custom to manually define the subnets. [Learn more](#)

IP stack type

IPv4 (single-stack)

These IP address ranges will be assigned to each region in your VPC network. When an instance is created for your VPC network, it will be assigned an IP from the appropriate region's address range.

Region	IP address range
africa-south1	10.218.0.0/20
asia-east1	10.140.0.0/20
asia-east2	10.170.0.0/20
asia-northeast1	10.146.0.0/20
asia-northeast2	10.174.0.0/20

VPC networks

Filter

Enter property name or value

Name	Subnets	MTU	Mode	IPv6 ULA range	Gateways	Firewall rules	Global dynamic routing
assignment4kairuddin	40	1460	Auto			0	Off
default	43	1460	Auto			14	Off

4. Audit Logging:

In this segment of the project, I focused on setting up and managing Audit Logging to enhance the security and compliance measures of our Google Cloud environment. The screenshot illustrates two main aspects:

Audit Logs Configuration:

Log Management and Review:

IAM & Admin | Audit logs | SET DEFAULT CONFIGURATION

Default configuration
0 exempted principals

Admin read: Disabled | Data read: Disabled | Data write: Disabled

Data access audit logs configuration
The effective data access configuration below combines the configuration for the currently selected resource and the data access configurations set on all parent resources.

Service	Admin read	Data read	Data write	Exempted principals	Inherited exempted principals
Access Approval	✓			0	0
Advisory Notifications API	✓			0	0
AI Platform Notebooks	✓			0	0
AlloyDB API	✓			0	0
Android Device Streaming	✓			0	0
Anthos Multi-cloud API	✓			0	0
API hub API	✓			0	0

50 services selected

PERMISSION TYPES | EXEMPTED PRINCIPALS

You can configure what types of operations are recorded in your data access audit logs for the selected services. There are several subtypes of data access audit logs:

- ☒ **Admin read**
Records operations that read metadata or configuration information.
- ☐ **Data read**
Records operations that read user-provided data.
- ☐ **Data write**
Records operations that write user-provided data.

SAVE

The IAM & Admin > Audit Logs panel displays the default configuration settings for various services. I tailored the logging settings to ensure comprehensive coverage across administrative read, data read, and data write activities, which are essential for maintaining rigorous oversight.

Log Management and Review:

Google Cloud | KBTU Tasks | logging | Search

Free trial status: \$156.23 credit and 22 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Logs Explorer | Query library | Share link | Preferences | Last 1 hour | Run query | Show query

Project logs | Search all fields | All log names | All severities | Correlate by

1 resource.type="k8s.cluster"

Log fields | Timeline

Search fields and values | Collapse log fields

RESOURCE TYPE
Kubernetes Cluster

SEVERITY
Default

LOG NAME
cloudaudit.googleapis.com/activity

PROJECT ID
kbtu-tasks

LOCATION
us-central1

CLUSTER NAME
hello-world-cluster

52,624 results

85% of results are similar and can be hidden. Hide similar entries | Preview

Summary of log entries showing timestamps, cluster names, and log messages.

In the Logs Explorer section, you can see the activity logs filtered for a particular project. This includes logs related to Kubernetes Engine and other critical components, providing a granular view of operations within the platform.

These configurations and the proactive review of audit logs are crucial for detecting potential security incidents early and providing audit trails necessary for compliance with industry standards.

Executive Summary

In this assignment, I embarked on a comprehensive journey to understand and implement key technologies within Google Cloud, focusing on Big Data, Machine Learning, and robust security measures. Here's a concise overview of what I have learned and achieved:

Big Data and Machine Learning: I developed a hands-on understanding of how to construct and manage a data processing pipeline using Google Cloud's BigQuery and Cloud Storage. This experience allowed me to appreciate the complexities of managing large datasets and the power of cloud computing in processing and analyzing data at scale. I also engaged in building and training machine learning models, which enhanced my skills in applying theoretical knowledge to practical, real-world data problems.

Security Practices: I learned the critical importance of securing data and applications in the cloud environment. Through the implementation of IAM roles, data encryption, and network security measures such as VPC configurations and firewall rules, I gained practical insights into the layered security approach necessary for protecting sensitive information and ensuring compliance with data protection standards.

Audit and Compliance: Setting up and managing audit logs helped me understand the mechanisms for monitoring and recording activities within cloud environments, which is crucial for compliance and security governance.

This assignment not only enhanced my technical skills but also deepened my understanding of the operational and security challenges in cloud computing, preparing me for future roles in technology and data management.

References

- Google Cloud Documentation. (n.d.). *Creating a training pipeline with Vertex AI*. Retrieved from <https://cloud.google.com/vertex-ai/docs/training/create-training-pipeline>
- Kaggle. (n.d.). *Models*. Retrieved from <https://www.kaggle.com/models>
- [HsjTlrCiAmQc](https://www.youtube.com/watch?v=HsjTlrCiAmQc). (n.d.). *YouTube*. Retrieved from <https://www.youtube.com/watch?v=HsjTlrCiAmQc>
- Looker Studio. (n.d.). *Document editing page*. Retrieved from <https://lookerstudio.google.com/u/0/reporting/5e14e9cf-3992-4ca8-8803-15d37a38844b/page/Ej2XE/edit>
- Google Cloud Documentation. (n.d.). *Audit logs*. Retrieved from <https://cloud.google.com/logging/docs/audit>