

**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

Introduction	4
Set Up a Google Cloud Project	5
Data Ingestion	5
Data processing with BigQuery	6
Machine Learning Model Training	9
Cloud Security and Compliance	
Identity and Access Management (IAM)	12
Data Encryption	13
Network Security	14
Audit Logging	15
Conclusion	16
Recommendations	16
References	16

## 1. Introduction

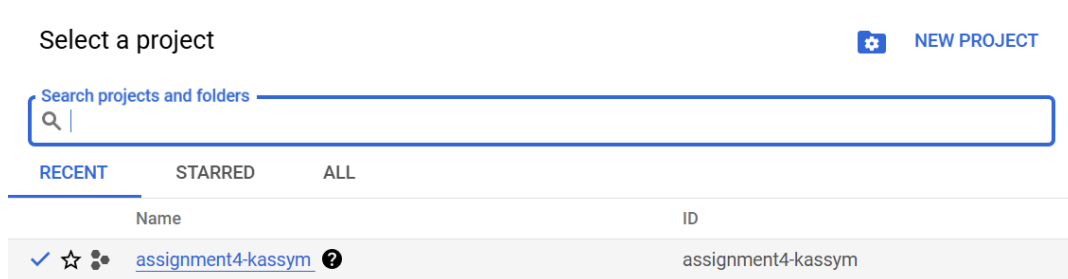
The importance of processing Big Data continues to grow as humanity generates vast amounts of data every minute. Simultaneously, cloud computing and storage technologies are becoming increasingly popular worldwide. This makes it essential to integrate modern concepts to process and analyze data effectively, 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.

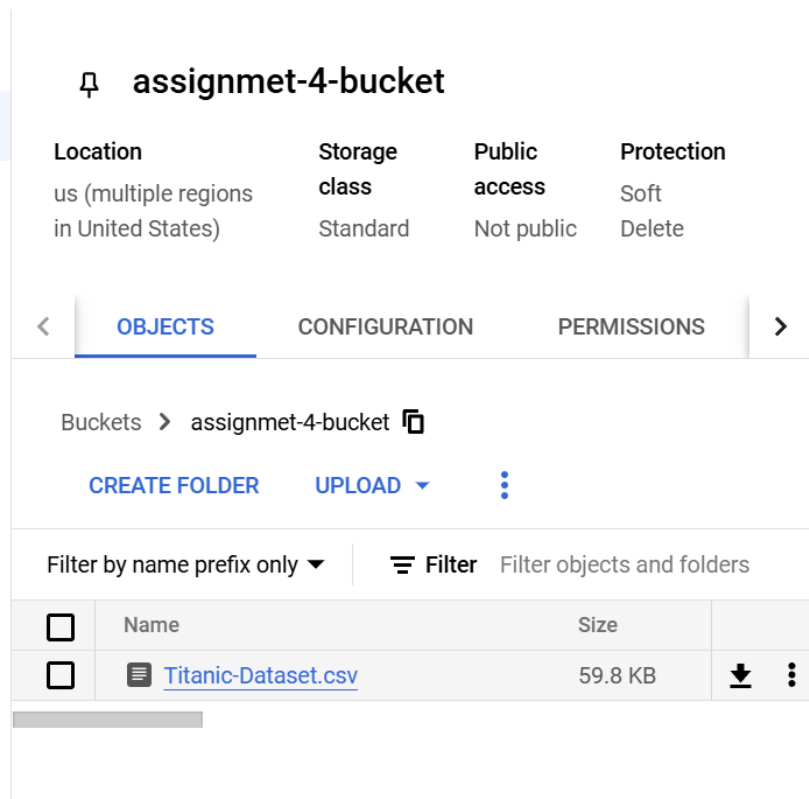
## 1. Set Up a Google Cloud Project:

- 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.

The screenshot shows the BigQuery Schema Explorer interface for a dataset named 'titanic'. The interface includes a top navigation bar with options like QUERY, SHARE, COPY, SNAPSHOT, DELETE, and EXPORT. Below this is a tabbed interface with tabs for SCHEMA, DETAILS, PREVIEW, TABLE EXPLORER, PREVIEW (highlighted), INSIGHTS, LINEAGE, DATA PROFILE, and DATA QUALITY. The SCHEMA tab is active, displaying a table of fields with columns: Field name, Type, Mode, Key, Collation, Default Value, Policy Tags, and Description. The fields listed are PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. Each field has a checkbox to its left. Below the table are two buttons: EDIT SCHEMA and VIEW ROW ACCESS POLICIES.

<input type="checkbox"/>	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
<input type="checkbox"/>	PassengerId	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Survived	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Pclass	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Name	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Sex	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Age	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	SibSp	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Parch	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Ticket	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Fare	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Cabin	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Embarked	STRING	NULLABLE	-	-	-	-	-

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

The screenshot shows the BigQuery Query Editor interface. At the top, there's a header with 'Untitled query' and buttons for RUN, SAVE, DOWNLOAD, SHARE, SCHEDULE, OPEN IN, and MORE. Below the header is a text area containing a SQL query: `SELECT sex, avg(Age) as avg_age FROM `assignment4-kassym.TitanicDataset.titanic` group by sex`. Below the query is a section titled 'Query results' with tabs for JOB INFORMATION, RESULTS (highlighted), CHART, JSON, EXECUTION DETAILS, and EXECUTION GRAPH. The RESULTS tab shows a table with two columns: 'sex' and 'avg\_age'. The table has two rows: one for 'male' with an average age of 30.72664459161... and one for 'female' with an average age of 27.91570881226....

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

[titanic](#)
[\\*Untitled query](#)
[\\*Untitled query](#)

Untitled query
 [RUN](#)
[SAVE](#)
[DOWNLOAD](#)
[SHARE](#)
[SCHEDULE](#)

```
1 delete from `assignment4-kassym.TitanicDataset.titanic` where age is null
```

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	f0_					
1	177					

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

Untitled query
 [RUN](#)
[SAVE](#)
[DOWNLOAD](#)
[SHARE](#)
[SCHEDULE](#)
[OPEN IN](#)
[MORE](#)

```
1 select survived, count(sex), sex from `assignment4-kassym.TitanicDataset.titanic` group by survived, sex
```

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	survived	f0_	sex			
1	0	360	male			
2	0	64	female			
3	1	93	male			
4	1	107	female			

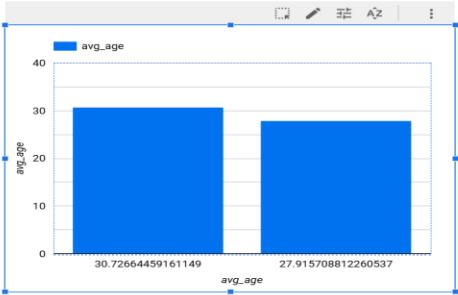
f0_	survived
93	1
197	1
360	0
64	0

1 - 4 / 4 < >

# titanic

	avg_age	Record Count
1.	30.72664459161149	1
2.	27.915708812260537	1





## 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).

Jobs									
<div><div>+</div> NEW TRAINING JOB</div> <div><div>↺</div> REFRESH</div> <div><div>■</div> CANCEL</div>									
<div><div>⚠</div> Vertex AI is our next generation AI Platform, with many new features that are unavailable in the current platform. Migrate your resources to Vertex AI to get the latest machine learning to-end journeys, and productionize models with MLOps. <a href="#">Learn More</a></div> <div><div>MIGRATE TO VERTEX AI</div><div>GO TO VERTEX AI</div></div>									
<div><div>≡</div> Filter Filter by prefix...</div>									
<input type="checkbox"/>	Job ID	Type	HyperTune	HyperTune parameters	Target metric	Create time	Elapsed time	Logs	Labels
<input type="checkbox"/>	<div><div>🔍</div> <a href="#">tutanickassym</a></div>	Custom code training	No			Dec 3, 2024, 5:15:15 PM	8 min 56 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>❗</div> <a href="#">pytanicititanic</a></div>	Custom code training	No			Dec 3, 2024, 4:31:49 PM	9 min 37 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>❗</div> <a href="#">TitanicLinearKassym</a></div>	Built-in algorithm training	No			Dec 3, 2024, 3:17:28 PM	7 min 25 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>🔍</div> <a href="#">Kassym</a></div>	Built-in algorithm training	No			Dec 3, 2024, 3:14:49 PM	7 min 26 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>🔍</div> <a href="#">kassymtitanic</a></div>	Custom code training	No			Dec 3, 2024, 2:49:48 PM	7 min 46 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>🔍</div> <a href="#">titanickassym</a></div>	Built-in algorithm training	No			Dec 3, 2024, 2:45:08 PM	7 min 45 sec	<a href="#">View</a> <a href="#">Logs</a>	
<input type="checkbox"/>	<div><div>❗</div> <a href="#">titanickassym</a></div>	Built-in algorithm training	No			Dec 3, 2024, 12:14:11 PM	7 min 28 sec	<a href="#">View</a> <a href="#">Logs</a>	

Region					
us-central1					
<div><div>≡</div> Filter Filter by prefix...</div>					
<input type="checkbox"/>	Name	Default version	Description	Endpoint	Labels
<input type="checkbox"/>	<a href="#">titanickassym</a>	-		us-central1-ml.googleapis.com	⋮


## 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 

 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",
    "--undata=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.

VIEW BY PRINCIPALS					
VIEW BY ROLES					
+ GRANT ACCESS - REMOVE ACCESS					
Filter Enter property name or value					
<input type="checkbox"/> Type	Principal ↑	Name	Role	Security insights ?	
<input type="checkbox"/> 	703368678924-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor		
<input type="checkbox"/> 	alihan230801@gmail.com		Viewer		
<input type="checkbox"/> 	baknur.samgat@gmail.com		Editor		
<input type="checkbox"/> 	Kassym914@gmail.com	Kassym Orakbay	Owner		

#### Roles for "assignment4-kassym" project

A role is a group of permissions that you can assign to principals. You can create a role and add permissions to it, or copy an existing role and adjust its permissions. [Learn more](#)

Filter Enter property name or value				
<input type="checkbox"/> Type	Title	Used in	Status	
<input type="checkbox"/> 	<a href="#">LeastPrivilege</a>	Custom	Enabled	

<input type="checkbox"/> 	baknur.samgat@gmail.com	LeastPrivilege	
--	-------------------------	----------------	---

## 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.









[←](#) Key ring details [+ CREATE KEY](#) [+ CREATE IMPORT JOB](#)

[KEYS](#) [IMPORT JOBS](#)

### Keys for "assignment4" key ring

A cryptographic key is a resource that is used for encrypting and decrypting data or for producing and verifying digital signatures. To perform operations on data with a key, use the Cloud KMS API. [Learn more](#)

 **Filter** Enter property name or value

<input type="checkbox"/>	Name 	Status  	Protection level 	Purpose 	Next rotation 	Actions
<input type="checkbox"/>	<a href="#">assignment4</a>	 Available	Software	Symmetric encrypt/decrypt	Mar 5, 2025	

No keys selected

### 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.

^ New subnet

Name \*

assignment4

Lowercase letters, numbers, hyphens allowed

Description

Region \*

IP stack type

☒ IPv4 (single-stack)  
☐ IPv4 and IPv6 (dual-stack)  
☐ IPv6 (single-stack)

IPv4 range \*

123.123.123.0/25

Ex: 10.0.0.0/24

REFRESH

CONFIGURE LOGS

DELETE

Filter Enter property name or value

	Name	Type	Targets	Filters	Protocols / ports	Action	Priority	Network	Logs	
	assignment4	Ingress	assignment4	IP ranges:	All	Allow	1000	default	Off	

## 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.

Logs Explorer

[Query library](#) [Share link](#)

Project logs ▾

Search all fields

Audited Resource ▾

All log names ▾

All severities ▾

Correlate by ▾

1 resource.type="audited\_resource"

Log fields

<|

Search fields and values

^ RESOURCE TYPE

✓ Audited Resource Clear ×

^ SEVERITY

i Notice 8

Timeline

< [ 4 0

Dec 5, 3:20 PM 3:4

8 results

SEVERITY	TIME	SUMMARY
----------	------	---------

## **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/>