KAZAKH KBTU BRITISH
TECHNICAL
UNIVERSITY

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

Shayakhmetova Balzhan

Almaty, 2024

Executive Summary

Implementing a Big Data processing and machine learning pipeline requires thorough planning, meticulous data cleaning, preprocessing, and robust model evaluation and deployment strategies. Google Cloud offers a comprehensive suite of tools and services to support these tasks, including BigQuery for analytics, AI Platform for machine learning, Cloud Storage for data handling, and Cloud KMS for encryption.

To ensure the pipeline's security and compliance, best practices were followed. These included configuring Identity and Access Management (IAM) roles and permissions, encrypting data with Cloud KMS, and applying network security measures such as Virtual Private Cloud (VPC) configurations and firewall rules. Cloud Audit Logs were enabled to monitor resource access and modifications, and an incident response plan was established to address potential security breaches.

A secure, scalable, and compliant Big Data processing and machine learning pipeline was successfully built by leveraging these Google Cloud services and adhering to industry best practices.

Assignment contains: 21 pages, 21images

A keywords list: Big Data processing, Machine learning pipeline, Google Cloud services, BigQuery, AI Platform, Cloud Storage, Cloud KMS, Identity and access management (IAM), Data encryption, Network security, VPC, Firewall rules, Cloud Audit Logs, Incident response plan

Introduction

The overall aim of this assignment is to design and implement a secure and compliant big data processing and machine learning pipeline using Google Cloud services. This comprehensive project has a dual focus, requiring the successful completion of two exercises.

Exercise 1 focuses on leveraging Google Cloud tools for data processing and machine learning. This involves setting up a Google Cloud project, collecting and processing a large dataset, training a machine learning model, and deploying it using AI Platform. The exercise also requires the implementation of data cleaning, preprocessing, and visualization using BigQuery and Google Data Studio. In today's data-driven world, integrating Big Data processing with machine learning is crucial for extracting valuable insights and making informed decisions. Big Data processing enables the collection and storage of vast amounts of data, while machine learning algorithms can uncover hidden patterns and relationships within that data. This synergy allows organizations to gain a competitive edge, improve operational efficiency, and drive innovation.

Exercise 2 shifts the focus to ensuring the security and compliance of the pipeline. This involves configuring identity and access management (IAM) roles and permissions, implementing data encryption using Cloud KMS, and setting up network security measures such as VPC and firewall rules. Additionally, Cloud Audit Logs are enabled to track access and changes to resources, and an incident response plan is developed to respond to potential security breaches. However, the importance of security and compliance cannot be overstated. As sensitive data is processed and analyzed, it is essential to safeguard it from unauthorized access, theft, or misuse. This requires implementing robust security measures, such as encryption, access controls, and auditing, to ensure the integrity and confidentiality of the data. Additionally, compliance with industry regulations, such as GDPR, HIPAA, and PCI-DSS, is critical to avoid legal and reputational risks.

The Google Cloud services used in this assignment, including BigQuery, Cloud Storage, AI Platform, and Cloud KMS, play a crucial role in enabling efficient data processing, advanced machine learning capabilities, and robust security measures. BigQuery is a fully-managed enterprise data warehouse service that enables fast and scalable data processing, allowing for the analysis of large datasets and the extraction of valuable insights. Cloud Storage is a highly durable and scalable object storage service that provides a secure and reliable way to store and manage large amounts of data. AI Platform is a managed service that enables the development, deployment, and management of machine learning models, allowing for the creation of advanced predictive models and the automation of decision-making processes. Cloud KMS is a managed service that provides secure key management and encryption capabilities, enabling the secure storage and transmission of sensitive data. These Google Cloud services work together to enable efficient data processing, advanced machine learning capabilities, and robust security measures. By leveraging these services, organizations can:

- Process large datasets quickly and efficiently using BigQuery

- Store and manage large amounts of data securely using Cloud Storage
- Develop and deploy advanced machine learning models using AI Platform
- Securely store and transmit sensitive data using Cloud KMS

Adopting industry best practices in both cloud computing and machine learning is crucial for ensuring scalability, efficiency, and security in today's data-driven world. By leveraging cloud-native services and architectures, organizations can take advantage of scalability, flexibility, and cost-effectiveness. However, this requires implementing robust security measures, such as encryption, access controls, and auditing, to protect sensitive data. In machine learning, following best practices for data preprocessing, feature engineering, and model selection is essential for ensuring accurate and reliable results. This includes implementing robust model validation and testing procedures to ensure model performance and reliability. Additionally, using cloud-based machine learning services, such as Google Cloud AI Platform, can leverage scalability, efficiency, and security. To achieve these goals, organizations should adopt a holistic approach that integrates cloud computing and machine learning best practices. This includes implementing data governance and data quality practices to ensure data accuracy, completeness, and consistency. Using cloud-based data warehousing and analytics services, such as Google BigQuery, can leverage scalability, efficiency, and security.

1 Big Data and Machine Learning on Google Cloud

1.1 Overview of the Pipeline

The implemented pipeline processes a large dataset to derive insights and build a predictive model using Google Cloud services. The pipeline is designed to extract valuable information from the dataset, identify patterns and relationships, and develop a machine learning model that can make accurate predictions. The pipeline consists of several stages, including data ingestion, data processing, and model training. Data ingestion involves collecting and storing the dataset in Google Cloud Storage. Data processing involves using BigQuery to clean, transform, and analyze the data. Model training involves using AI Platform to train a machine learning model on the processed data. The pipeline is designed to be scalable, efficient, and secure, using Google Cloud services such as BigQuery, AI Platform, and Cloud Storage. The pipeline is also designed to be flexible, allowing for easy modification and adaptation to changing requirements. By implementing it, we can gain valuable insights from the dataset, build a predictive model that can make accurate predictions, and improve decision-making processes.

The general idea of the pipeline's structure, including data ingestion, processing, machine learning model training, evaluation, and deployment can be seen in Image 1.

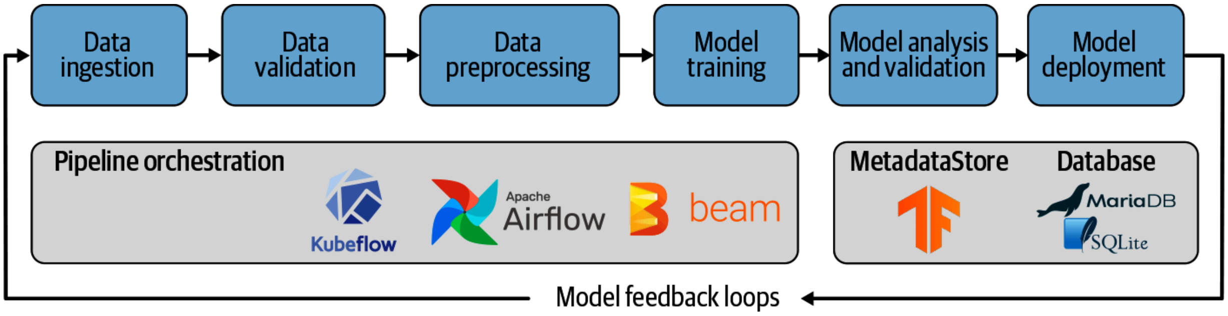
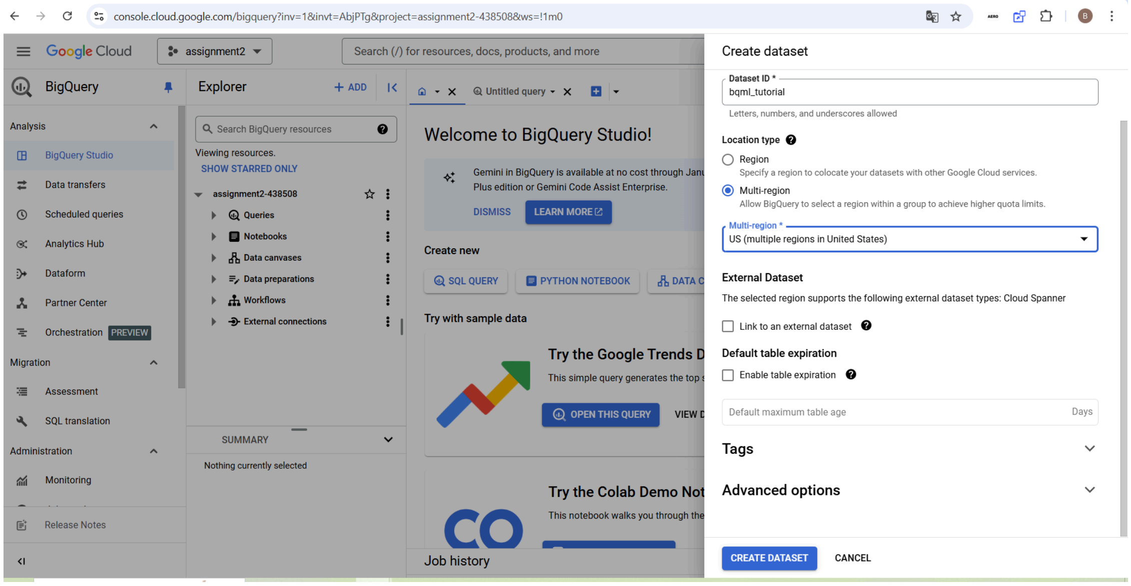


Image 1: Pipeline overview

BigQuery ML lets you create and train machine learning models in BigQuery by using SQL queries. This helps make machine learning more approachable by letting you use familiar tools like the BigQuery SQL editor, and also increases development speed by removing the need to move data into a separate machine learning environment. To create a machine learning model in BigQuery we should:

1. Create a BigQuery dataset to store your ML model in Image 2



Create dataset

Dataset ID *
bqml_tutorial
Letters, numbers, and underscores allowed

Location type

○ Region
Specify a region to colocate your datasets with other Google Cloud services.

● Multi-region
Allow BigQuery to select a region within a group to achieve higher quota limits.

Multi-region *
US (multiple regions in United States)

External Dataset
The selected region supports the following external dataset types: Cloud Spanner

□ Link to an external dataset

Default table expiration
□ Enable table expiration

Default maximum table age Days

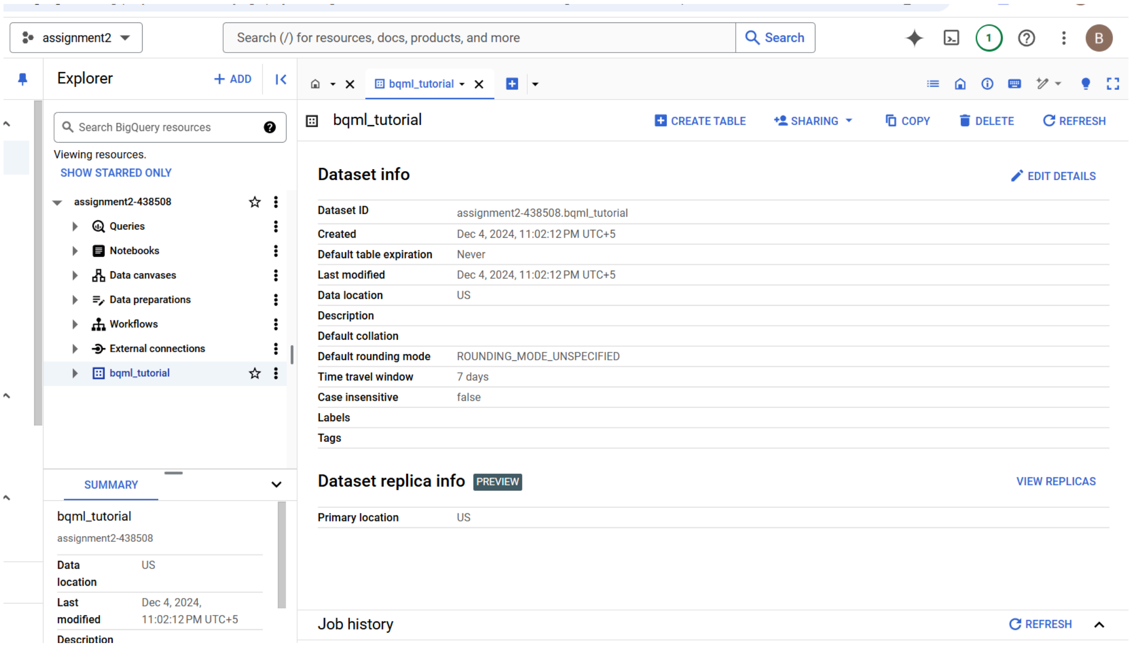
Tags

Advanced options

CREATE DATASET CANCEL

Image 2: Creating dataset

2. The created dataset is shown in the project in Image 3



bqml_tutorial
assignment2-438508

Data location	US
Last modified	Dec 4, 2024, 11:02:12 PM UTC+5
Description	

Dataset info

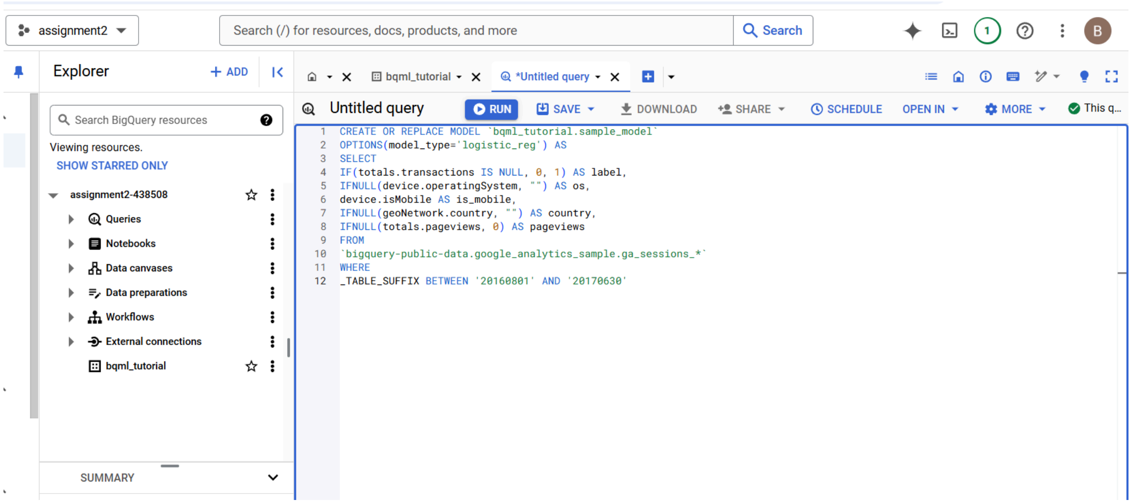
Dataset ID	assignment2-438508.bqml_tutorial
Created	Dec 4, 2024, 11:02:12 PM UTC+5
Default table expiration	Never
Last modified	Dec 4, 2024, 11:02:12 PM UTC+5
Data location	US
Description	
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Time travel window	7 days
Case insensitive	false
Labels	
Tags	

Dataset replica info PREVIEW

Primary location	US

Image 3: Successful creation

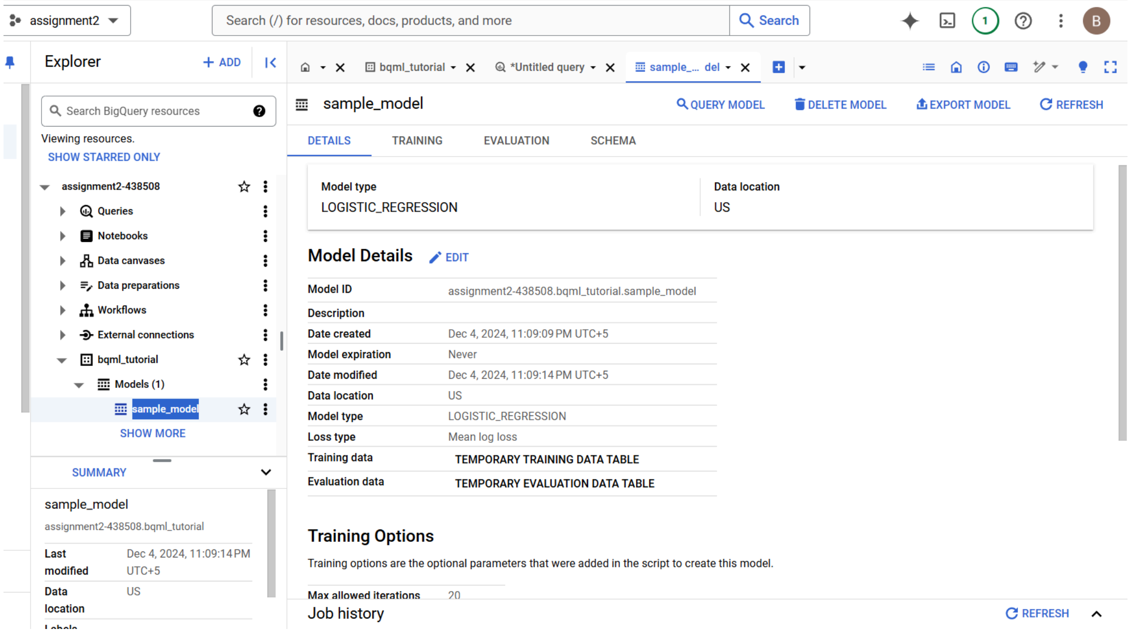
3. Create a logistic regression model using the Analytics sample dataset for BigQuery in Image 4



```sql
CREATE OR REPLACE MODEL `bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
IF(totals.transactions IS NULL, 0, 1) AS label,
IFNULL(device.operatingSystem, "") AS os,
device.isMobile AS is_mobile,
IFNULL(geoNetwork.country, "") AS country,
IFNULL(totals.pageviews, 0) AS pageviews
FROM
`bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
_TABLE_SUFFIX BETWEEN '20160801' AND '20170630'
```

Image 4: Script to create a model

4. The query takes several minutes to complete. After the first iteration is complete, model (sample_model) appears in the navigation panel in Image 5



sample_model

Model type	Data location
LOGISTIC_REGRESSION	US

Model Details

Model ID	assignment2-438508.bqml_tutorial.sample_model
Description	
Date created	Dec 4, 2024, 11:09:09 PM UTC+5
Model expiration	Never
Date modified	Dec 4, 2024, 11:09:14 PM UTC+5
Data location	US
Model type	LOGISTIC_REGRESSION
Loss type	Mean log loss
Training data	TEMPORARY TRAINING DATA TABLE
Evaluation data	TEMPORARY EVALUATION DATA TABLE

Training Options

Training options are the optional parameters that were added in the script to create this model.

Max allowed iterations	20

sample_model

assignment2-438508.bqml_tutorial

Last modified	Dec 4, 2024, 11:09:14 PM UTC+5
Data location	US

Image 5: Successfully created model

1.2 Machine Learning Model Training

Machine learning is about creating a model that can use data to make a prediction. The model is essentially a function that takes inputs and applies calculations to the inputs to produce an output — a prediction.

Machine learning algorithms work by taking several examples where the prediction is already known (such as the historical data of user purchases) and iteratively adjusting various weights in the model so that the model's predictions match the true values. It does this by minimizing how wrong the model is using a metric called loss.

The expectation is that for each iteration, the loss should be decreasing, ideally to zero. A loss of zero means the model is 100% accurate.

When training the model, BigQuery ML automatically splits the input data into training and evaluation sets, in order to avoid overfitting the model. This is necessary so that the training algorithm doesn't fit itself so closely to the training data that it can't generalize to new examples. In Image 6 we see how the model's loss changes over the model's training iterations

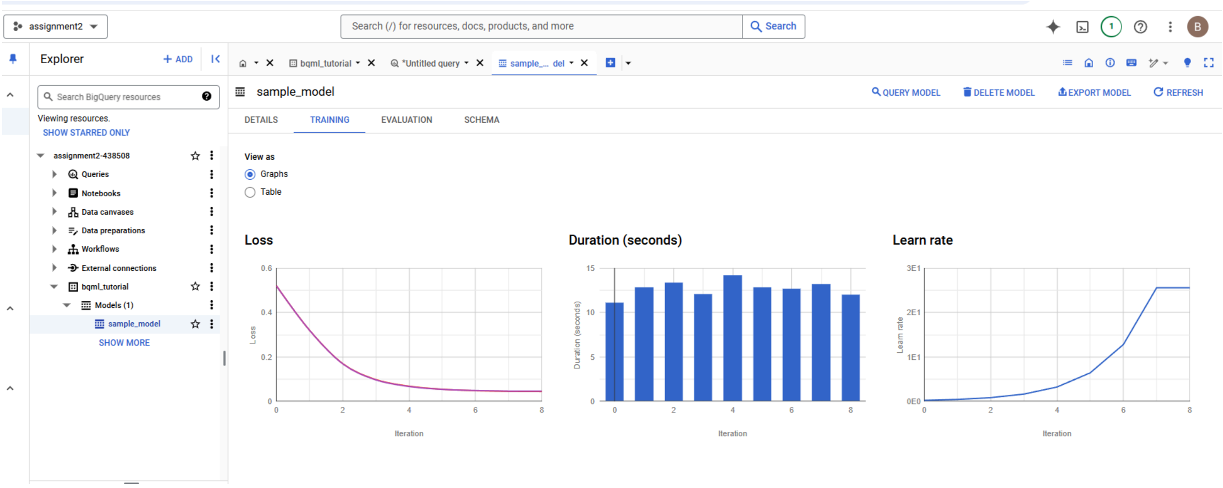
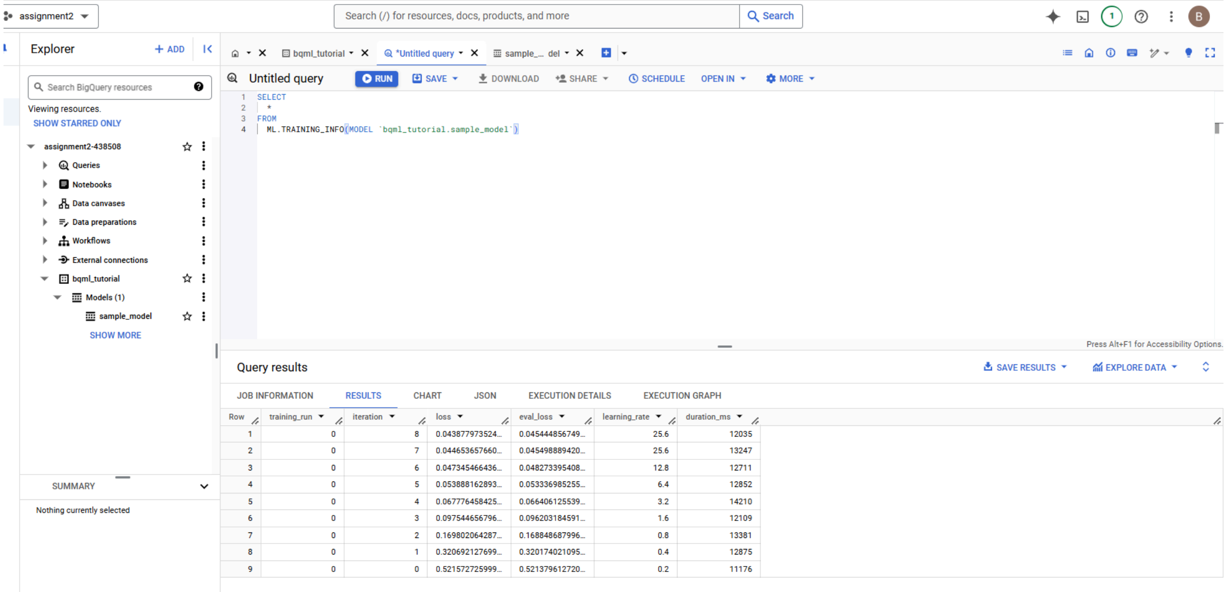


Image 6: Training tab of the model

We can also see the results of the model training by using the ML.TRAINING_INFO function in Image 7



```
SELECT
  *
FROM
  ML.TRAINING_INFO(MODEL `bqml_tutorial.sample_model`)
```

Row	training_run	iteration	loss	eval_loss	learning_rate	duration_ms
1	0	8	0.043877973524...	0.045444856749...	25.6	12035
2	0	7	0.044653657660...	0.045498889420...	25.6	13247
3	0	6	0.047345466436...	0.048273395408...	12.8	12711
4	0	5	0.053888162893...	0.053336985255...	6.4	12852
5	0	4	0.067776458425...	0.066406125539...	3.2	14210
6	0	3	0.097544656796...	0.096203184591...	1.6	12109
7	0	2	0.169802064287...	0.168848687996...	0.8	13381
8	0	1	0.320692127699...	0.320174021095...	0.4	12875
9	0	0	0.521572725999...	0.521379612720...	0.2	11176

Image 7: ML.TRAINING_INFO function

In this assignment I am using binary classification model. Binary Classification works by using a set of training data to learn a model that can then be used to predict outcomes. The data used to train the model contains features (variables) and labels (categories). The model will then use these features to identify patterns and make predictions based on the labeled data. The model is then evaluated based on its ability to accurately predict the correct labels for new data. Binary

Classification is important because it allows businesses to make predictions based on data, which can lead to better decision making. It can help businesses to identify customers who are most likely to buy their products, predict which financial transactions are fraudulent, and prevent equipment failure in manufacturing. It is also useful in natural language processing, sentiment analysis, and image classification.

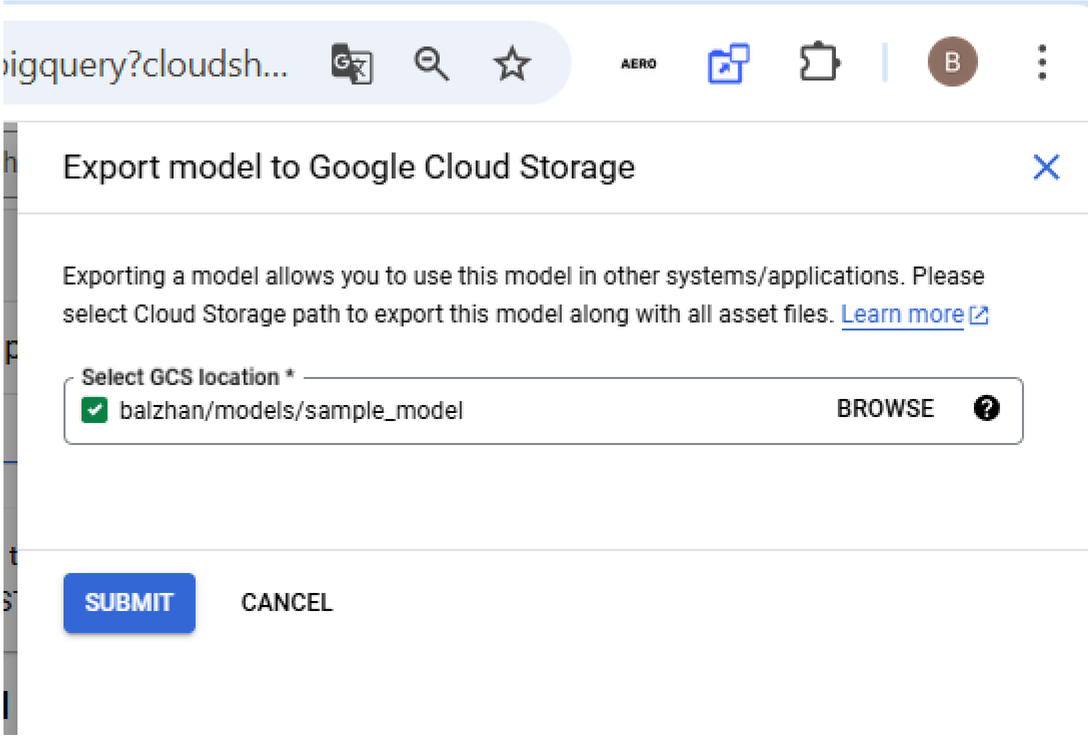
1.3 Data Ingestion and Processing

In this assignment I am using the Google-provided google_analytics_sample.ga_sessions_ table to analyze customer behavior and predict binary outcomes, such as whether a user will make a purchase. The dataset is available directly in BigQuery's public datasets under google_analytics_sample. This dataset contains session-level data from the Google Analytics 360 sample. It includes user activity, traffic source information, and transaction data. Key features:

- **Behavioral Data**: Metrics like session duration, pages viewed, and engagement metrics.
- **Traffic Source Data**: Referral, organic, or paid traffic sources.
- **Transaction Data**: Indicators of purchases (totals.transactions).

Store extracted or processed subsets of the table in Google Cloud Storage for further use.

1. Export model to Google Cloud Storage in Image 8



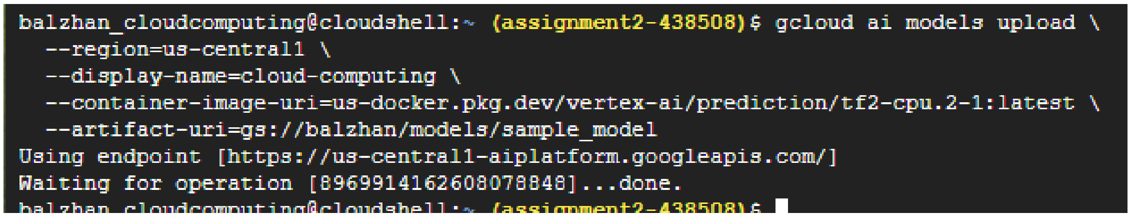
Export model to Google Cloud Storage

Exporting a model allows you to use this model in other systems/applications. Please select Cloud Storage path to export this model along with all asset files. Learn more

Select GCS location *
☑ balzhan/models/sample_model

Image 8: Exporting model

2. Uploading the model now from the bucket to Vertex AI in Image 9



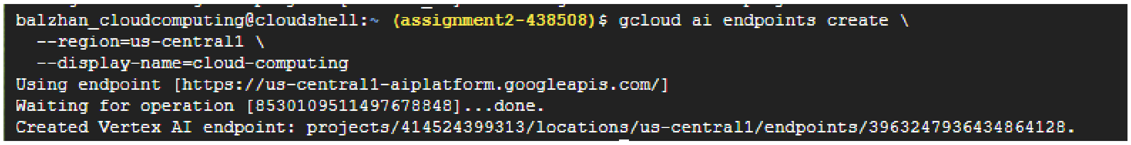
```
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$ gcloud ai models upload \
  --region=us-central1 \
  --display-name=cloud-computing \
  --container-image-uri=us-docker.pkg.dev/vertex-ai/prediction/tf2-cpu.2-1:latest \
  --artifact-uri=gs://balzhan/models/sample_model
Using endpoint [https://us-central1-aiplatform.googleapis.com/]
Waiting for operation [8969914162608078848]...done.
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$
```

Image 9: Uploading model

1.4 Model Deployment

You must deploy a model to an endpoint before that model can be used to serve online predictions. Deploying a model associates physical resources with the model so it can serve online predictions with low latency. To deploy a model:

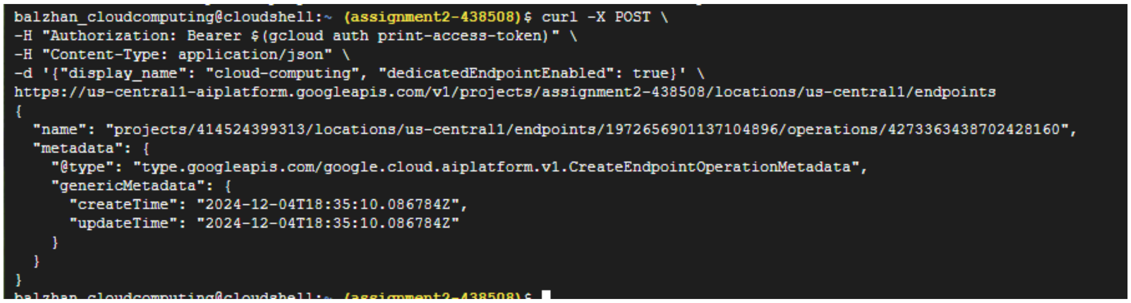
1. Create an endpoint in Image 10. A *dedicated endpoint* is a faster, more stable endpoint with support for larger payload sizes and longer request timeouts.



```
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$ gcloud ai endpoints create \
  --region=us-central1 \
  --display-name=cloud-computing
Using endpoint [https://us-central1-aiplatform.googleapis.com/]
Waiting for operation [8530109511497678848]...done.
Created Vertex AI endpoint: projects/414524399313/locations/us-central1/endpoints/3963247936434864128.
```

Image 10: Creating an endpoint

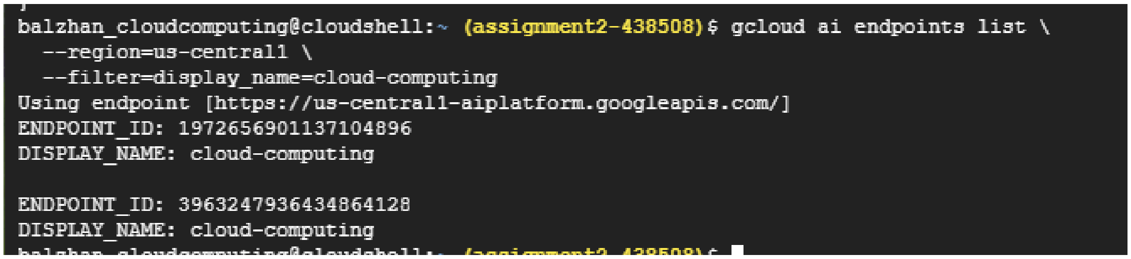
2. To use a dedicated endpoint during Preview, you need to enable it explicitly in Image 11 and you can see the response from the call



```
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$ curl -X POST \
-H "Authorization: Bearer $(gcloud auth print-access-token)" \
-H "Content-Type: application/json" \
-d '{"display_name": "cloud-computing", "dedicatedEndpointEnabled": true}' \
https://us-central1-aiplatform.googleapis.com/v1/projects/assignment2-438508/locations/us-central1/endpoints
{
  "name": "projects/414524399313/locations/us-central1/endpoints/1972656901137104896/operations/4273363438702428160",
  "metadata": {
    "@type": "type.googleapis.com/google.cloud.aiplatform.v1.CreateEndpointOperationMetadata",
    "genericMetadata": {
      "createTime": "2024-12-04T18:35:10.086784Z",
      "updateTime": "2024-12-04T18:35:10.086784Z"
    }
  }
}
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$
```

Image 11: Using dedicated endpoint

3. You need the endpoint ID to deploy the model. Let’s get the list of endpoints in Image 12

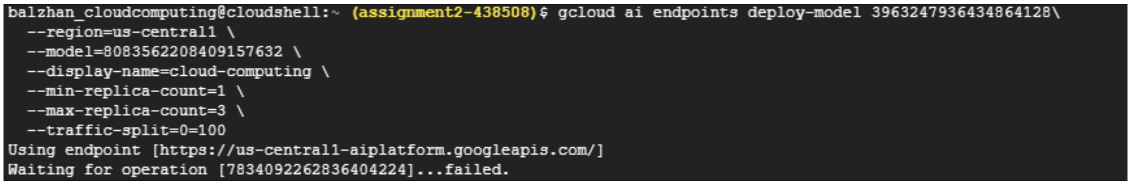


```
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$ gcloud ai endpoints list \
  --region=us-central1 \
  --filter=display_name=cloud-computing
Using endpoint [https://us-central1-aiplatform.googleapis.com/]
ENDPOINT_ID: 1972656901137104896
DISPLAY_NAME: cloud-computing

ENDPOINT_ID: 3963247936434864128
DISPLAY_NAME: cloud-computing
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$
```

Image 12: List of endpoints

4. Deploying the model in Image 13

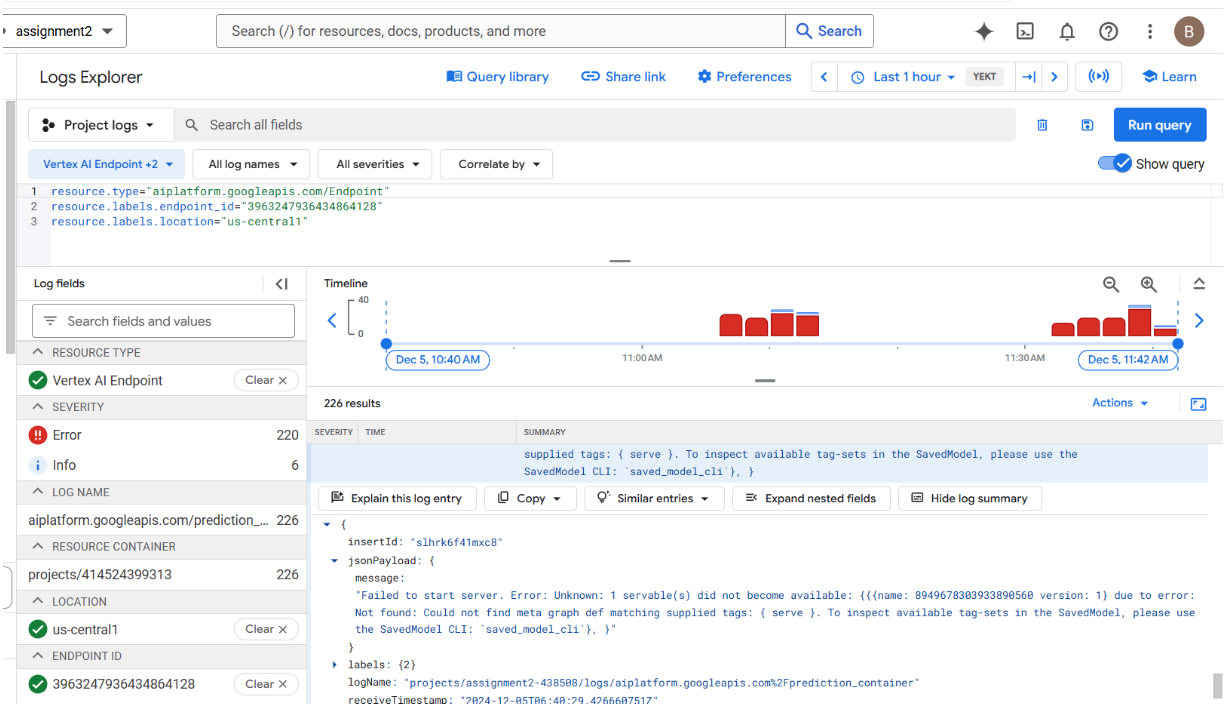


```
balzhan_cloudcomputing@cloudshell:~ (assignment2-438508)$ gcloud ai endpoints deploy-model 3963247936434864128\
  --region=us-central1 \
  --model=8083562208409157632 \
  --display-name=cloud-computing \
  --min-replica-count=1 \
  --max-replica-count=3 \
  --traffic-split=0=100
Using endpoint [https://us-central1-aiplatform.googleapis.com/]
Waiting for operation [7834092262836404224]...failed.
```

Image 13: Deploying the model

1.5 Monitoring and Logging

Monitoring and logging are essential components of any robust data pipeline and machine learning system. For the google_analytics_sample.ga_sessions_ dataset pipeline, monitoring and logging ensure performance tracking, issue detection, and insights into the behavior of the deployed model and data processes in Image 14.



```
resource.type="aiplatform.googleapis.com/Endpoint"
resource.labels.endpoint_id="3963247936434864128"
resource.labels.location="us-central1"
```

Log fields

Search fields and values

RESOURCE TYPE	
Vertex AI Endpoint	Clear
SEVERITY	
Error	220
Info	6
LOG NAME	
aiplatform.googleapis.com/prediction_...	226
RESOURCE CONTAINER	
projects/414524399313	226
LOCATION	
us-central1	Clear
ENDPOINT ID	
3963247936434864128	Clear

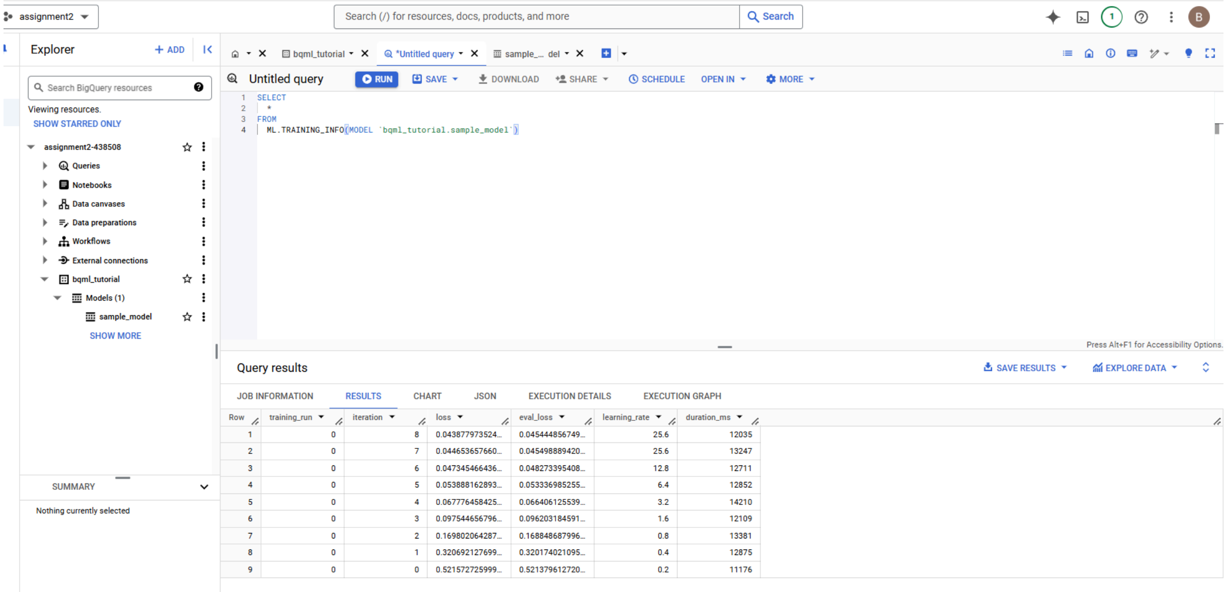
226 results

```
supplied tags: { serve }. To inspect available tag-sets in the SavedModel, please use the
SavedModel CLI: `saved_model_cli`}, }
```

```
{
  insertId: "slhrk6f41mxc8"
  jsonPayload: {
   message:
   "Failed to start server. Error: Unknown: 1 servable(s) did not become available: {{{name: 8949678303933890560 version: 1} due to error:
   Not found: Could not find meta graph def matching supplied tags: { serve }. To inspect available tag-sets in the SavedModel, please use
   the SavedModel CLI: `saved_model_cli`}, }"
  }
  labels: {2}
  logName: "projects/assignment2-438508/logs/aiplatform.googleapis.com%2Fprediction_container"
  receiveTimestamp: "2024-12-05T06:40:29.4266607517"
```

Image 14: Logging

I can see the results of the model training by using the ML.TRAINING_INFO function in Image 15



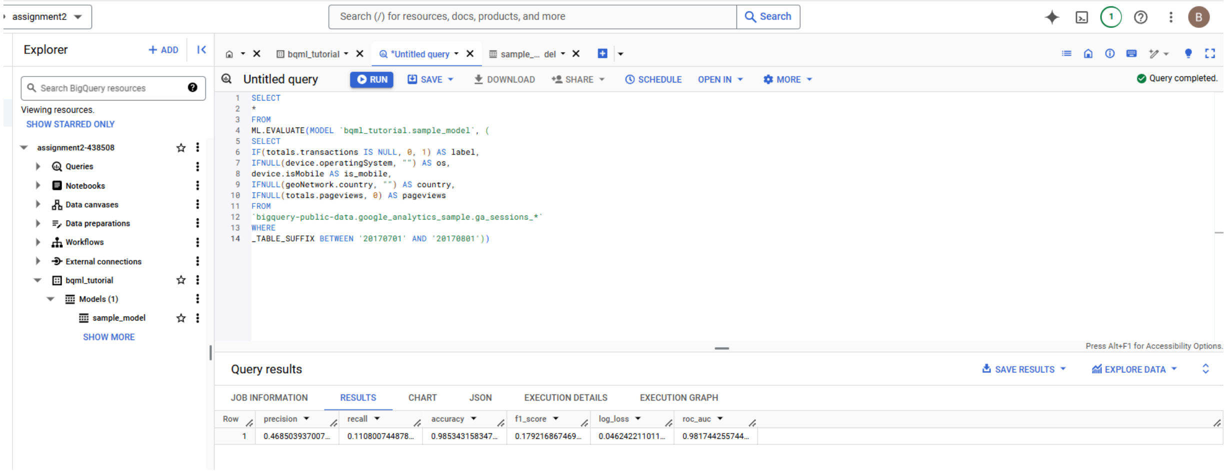
```
SELECT
  *
FROM
  ML.TRAINING_INFO(MODEL `bqml_tutorial.sample_model`)
```

Query results

Row	training_run	iteration	loss	eval_loss	learning_rate	duration_ms
1	0	8	0.043877973524...	0.045444856749...	25.6	12035
2	0	7	0.044653657660...	0.045498889420...	25.6	13247
3	0	6	0.047345466436...	0.048273395408...	12.8	12711
4	0	5	0.053888162893...	0.053336985255...	6.4	12852
5	0	4	0.067776458425...	0.066406125539...	3.2	14210
6	0	3	0.097544656796...	0.096203184591...	1.6	12109
7	0	2	0.169802064287...	0.168848687996...	0.8	13381
8	0	1	0.320692127699...	0.320174021095...	0.4	12875
9	0	0	0.521572725999...	0.521379612720...	0.2	11176

Image 15: ML training

In this assignment, you are using a binary classification model that detects transactions. The values in the label column are the two classes generated by the model: 0 (no transactions) and 1 (transaction made) in Image 16



```sql
SELECT
*
FROM
ML.EVALUATE(MODEL `bqml_tutorial.sample_model`, (
SELECT
IF(totals.transactions IS NULL, 0, 1) AS label,
IFNULL(device.operatingSystem, "") AS os,
device.isMobile AS is_mobile,
IFNULL(geoNetwork.country, "") AS country,
IFNULL(totals.pageviews, 0) AS pageviews
FROM
`bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
_TABLE_SUFFIX BETWEEN '20170701' AND '20170801'))
```

Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.468503937007...	0.110800744878...	0.985343158347...	0.179216867469...	0.046242211011...	0.981744255744...

Image 16: The result prediction

2 Cloud Security and Compliance

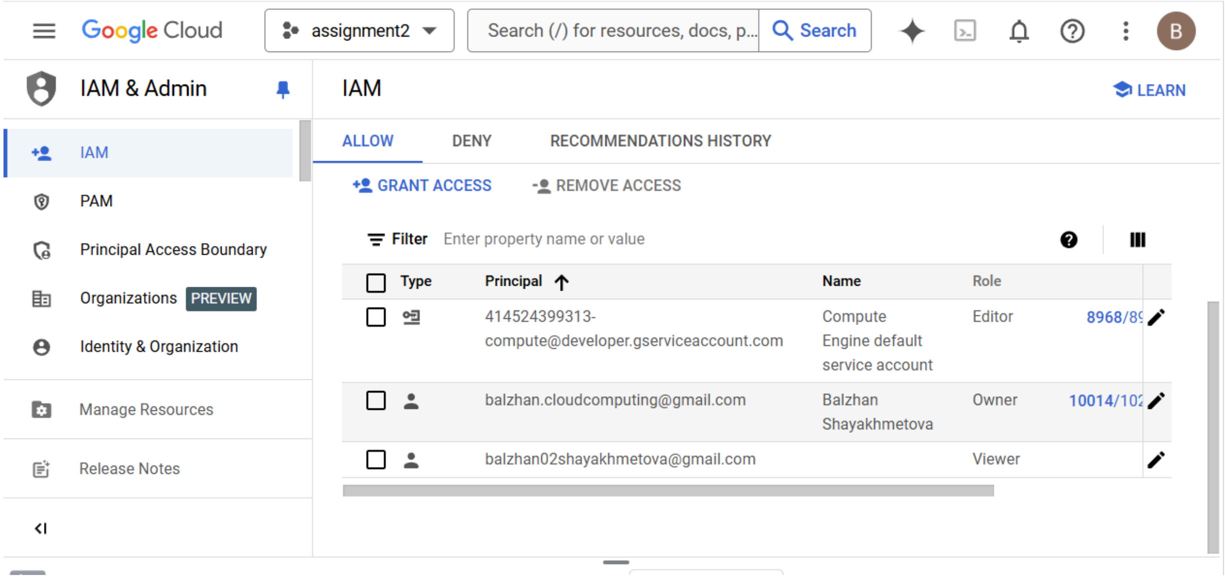
2.1 Identity and Access Management (IAM)

Identity and Access Management (IAM) is critical for securing the Google Cloud environment by ensuring that resources are accessible only to authorized users and applications. Implementing proper IAM policies in the project ensures adherence to the principle of least privilege and compliance with security standards.

Role	Description	Permissions
Viewer	View-only access to project resources	resourcemanager.projects.get, storage.buckets.list, compute.instances.list
Editor	Full control over resources, except IAM	Viewer permissions + resourcemanager.projects.update, compute.instances.create, storage.buckets.create

Owner	Full control, including IAM and billing	Editor permissions + resourcemanager.projects.setIamPolicy, billing.accounts.get, billing.accounts.update
Custom	Fine-grained permissions for specific needs	Define custom roles by selecting specific permissions from IAM Permissions List.

The configured roles of the project in my current project in Image 17



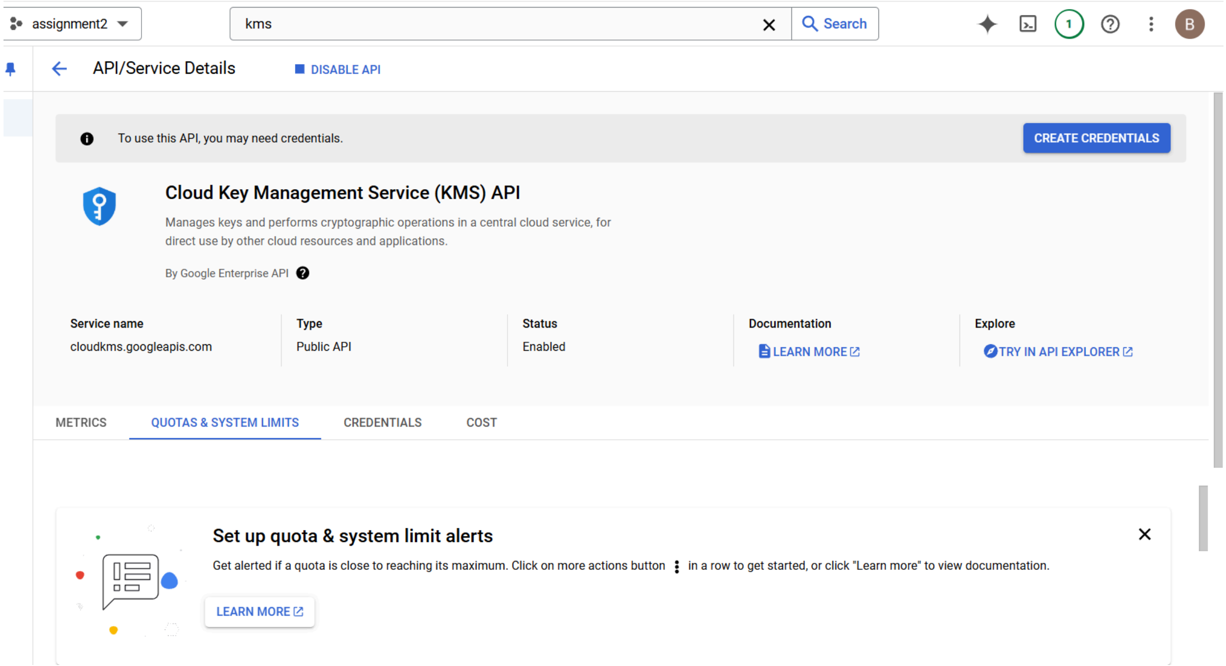
□ Type	Principal ↑	Name	Role	
□	414524399313-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	8968/8
□	balzhan.cloudcomputing@gmail.com	Balzhan Shayakhmetova	Owner	10014/102
□	balzhan02shayakhmetova@gmail.com		Viewer	

Image 17: IAM roles

2.2 Data Encription

Data encryption is a fundamental practice for protecting sensitive information, ensuring compliance with security standards, and safeguarding against unauthorized access. On Google Cloud, encryption protects data at rest, in transit, and during processing.

Data Encryption at Rest. Google Cloud automatically encrypts all data at rest using AES-256 encryption. This includes data stored in Cloud Storage and BigQuery. For enhanced security and compliance, Customer-Managed Encryption Keys (CMEK) were implemented using Google Cloud Key Management Service (Cloud KMS). CMEK provides greater control over encryption keys, allowing for key rotation and usage auditing in Image 18



To use this API, you may need credentials.

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

Set up quota & system limit alerts

Get alerted if a quota is close to reaching its maximum. Click on more actions button ⋮ in a row to get started, or click "Learn more" to view documentation.

Image 18: KMS API

Encryption keys were created and managed in a key ring within Cloud KMS. These keys were then linked to Cloud Storage buckets and BigQuery datasets. This ensured that even though Google automatically encrypts data, additional security measures were enforced for sensitive information.

Data Encryption in Transit. To secure data during transmission, Transport Layer Security (TLS) was used for all communications between users, applications, and Google Cloud services. Additionally, public access to storage buckets and BigQuery datasets was disabled, and data transfer was restricted to private network connections using Virtual Private Cloud (VPC). These measures ensured secure communication and minimized exposure to potential threats.

Key management was handled carefully to avoid unauthorized access. Access to encryption keys was restricted using IAM roles, granting only the necessary permissions to users and services. Key rotation policies were implemented to renew encryption keys periodically, further reducing risks of key compromise. Audit logs were enabled to monitor key usage and to track encryption and decryption events for compliance purposes.

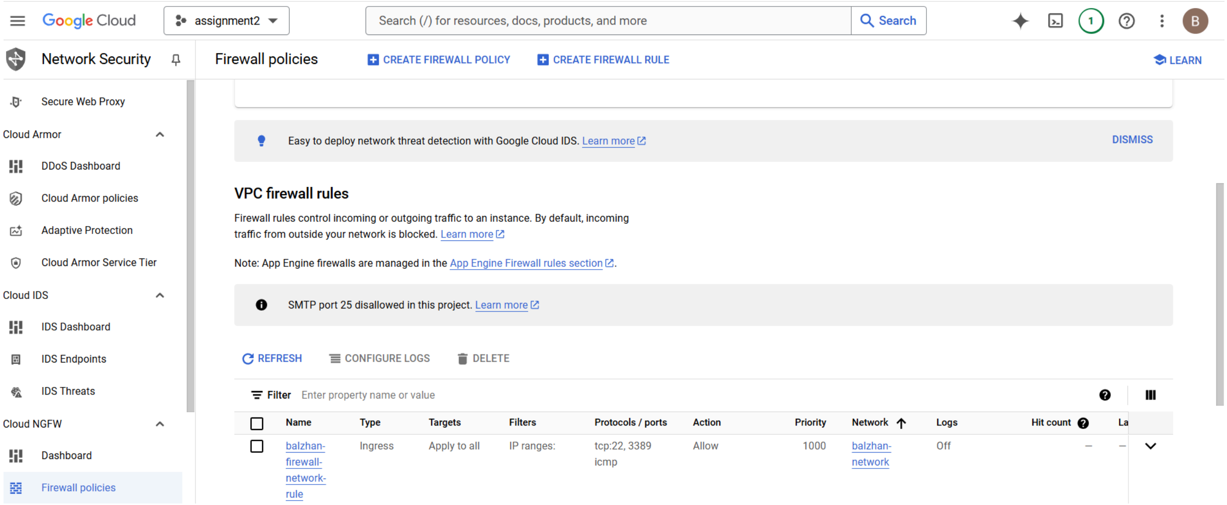
Compliance and Security. These encryption measures align with industry standards and regulations such as GDPR and HIPAA. By encrypting data both at rest and in transit, the pipeline ensures compliance while maintaining robust security practices.

2.3 Network Security

Network security is a crucial aspect of protecting cloud resources from unauthorized access and mitigating potential threats. In the pipeline, network security was implemented through a Virtual Private Cloud (VPC) configuration and firewall rules to control access to resources and ensure secure communication between components.

1. The VPC ensured that all sensitive resources, including storage and compute instances, were hosted in a private network, isolated from public access.
2. The VPC allowed internal communication between services without exposing resources to the outside world.
3. Subnets were defined within the VPC to segregate resources based on their function and security requirements. For example, separate subnets were used for storage services (Cloud Storage), data processing (BigQuery), and AI services (AI Platform).
4. Private Google access was enabled to allow services within the VPC to access Google Cloud APIs securely without requiring public IPs.
5. Custom routing was configured to ensure that traffic between different subnets within the VPC is correctly managed, with appropriate access controls based on the subnet's function.

Firewall rules were configured to control inbound and outbound traffic to and from resources in the VPC. These rules were essential to restrict access to sensitive data and services, ensuring that only authorized users or services could access them in Image 19



Easy to deploy network threat detection with Google Cloud IDS. Learn more

VPC firewall rules

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

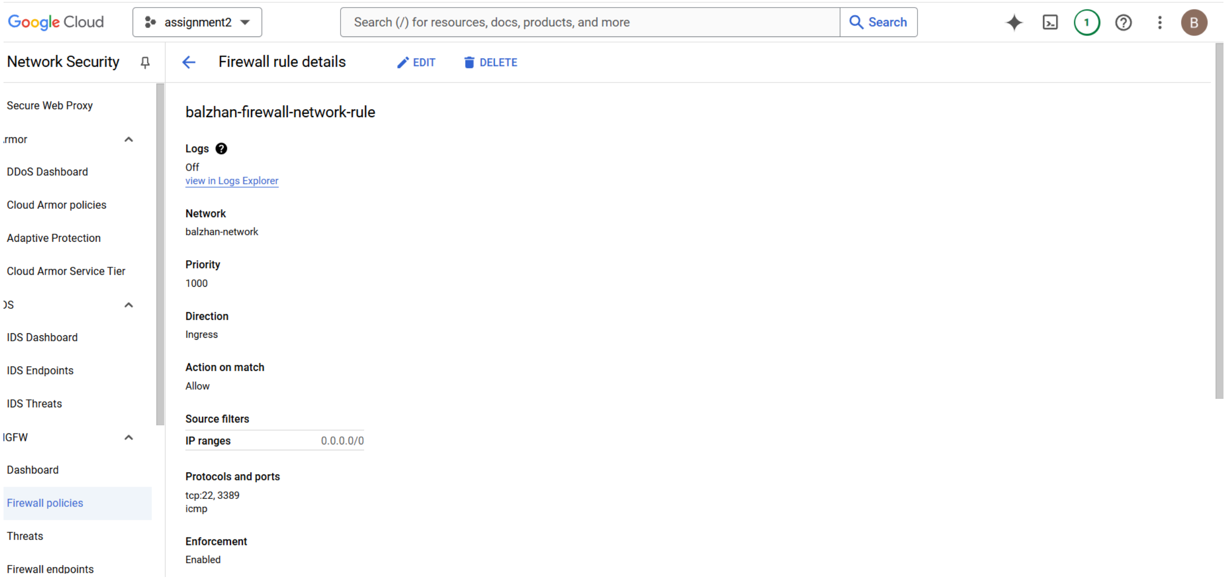
Note: App Engine firewalls are managed in the App Engine Firewall rules section.

SMTP port 25 disallowed in this project. Learn more

Name	Type	Targets	Filters	Protocols / ports	Action	Priority	Network	Logs	Hit count	La
balzhan-firewall-network-rule	Ingress	Apply to all	IP ranges:	tcp:22, 3389 icmp	Allow	1000	balzhan-network	Off	–	–

Image 19: Firewall rules on network

Firewall rule details in Image 20



Firewall rule details

balzhan-firewall-network-rule

Logs
Off
view in Logs Explorer

Network
balzhan-network

Priority
1000

Direction
Ingress

Action on match
Allow

Source filters

IP ranges	0.0.0.0/0

Protocols and ports
tcp:22, 3389
icmp

Enforcement
Enabled

Image 20: Firewall rule details

2.4 Audit Logging

Audit logging is a vital component of cloud security and compliance, allowing organizations to monitor access to sensitive resources and track changes to infrastructure and data. In the pipeline, audit logging was implemented to ensure transparency, detect potential security breaches, and maintain a record of who accessed or modified resources. Google Cloud's Cloud Audit Logs were utilized for this purpose.

Objective of Audit Logging

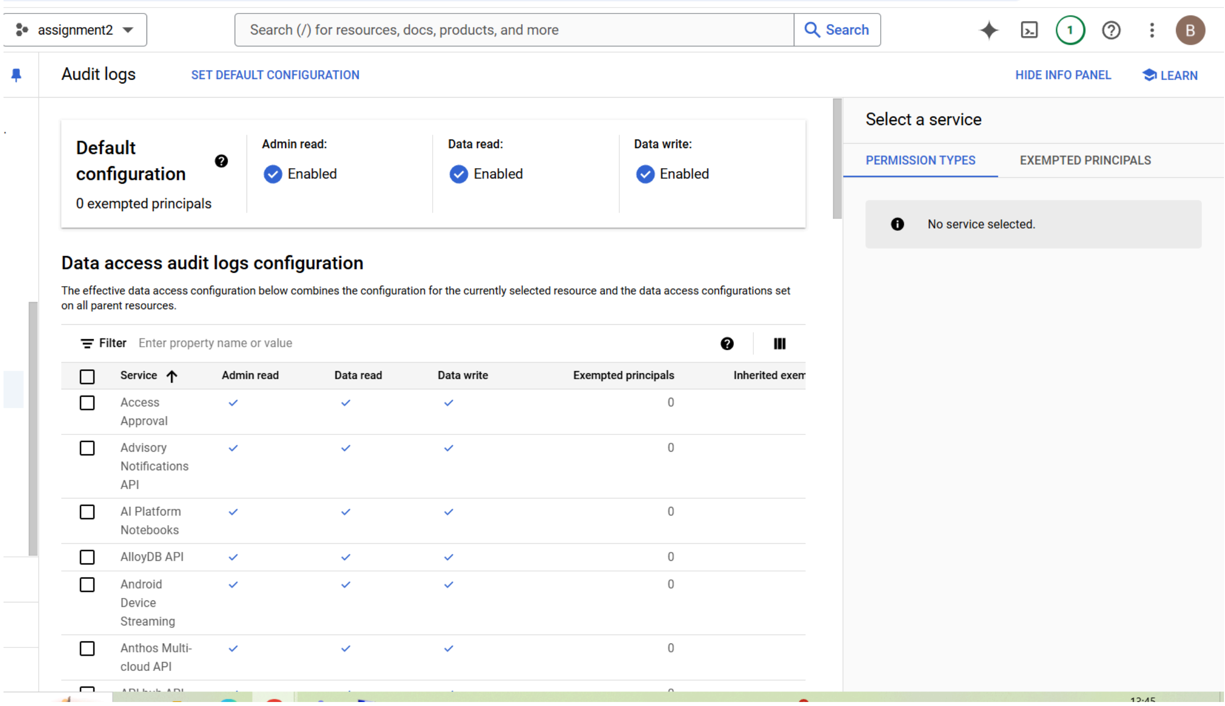
- **Security and Compliance**: Track who accessed resources, what actions they performed, and when these actions occurred, ensuring compliance with security standards (e.g., GDPR, HIPAA).
- **Transparency and Accountability**: Provide detailed records for investigation in case of suspicious activity or breaches.
- **Operational Monitoring**: Monitor infrastructure changes, data access, and API calls to ensure that the pipeline is operating as expected.

Audit Logging Setup

1. **Enable Cloud Audit Logs**:
    - **Cloud Audit Logs** were enabled for all services used in the pipeline, including BigQuery, Cloud Storage, and AI Platform. This provided visibility into all administrative actions (e.g., who created datasets, accessed storage, or deployed models) as well as data access events (e.g., querying BigQuery, retrieving files from Cloud Storage).
    - **Audit Log Types**:

        - **Admin Activity Logs**: Track actions that modify cloud resources (e.g., creating, updating, or deleting datasets or models).
        - **Data Access Logs**: Capture access to sensitive data, such as queries run on BigQuery or file retrieval from Cloud Storage.
        - **System Event Logs**: Monitor system-generated events, such as automatic scaling or updates to resources.

The configuration is set in Image 21


assignment2
Search (/) for resources, docs, products, and more
Search
B
SET DEFAULT CONFIGURATION
HIDE INFO PANEL
LEARN
PERMISSION TYPES
EXEMPTED PRINCIPALS
Filter
Enter property name or value

Audit logs

Default configuration

0 exempted principals

Admin read: Enabled

Data read: Enabled

Data write: Enabled

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 exem
□	Access Approval	✓	✓	✓	0	
□	Advisory Notifications API	✓	✓	✓	0	
□	AI Platform Notebooks	✓	✓	✓	0	
□	AlloyDB API	✓	✓	✓	0	
□	Android Device Streaming	✓	✓	✓	0	
□	Anthos Multi-cloud API	✓	✓	✓	0	

Select a service

No service selected.

Image 21: Configuration of logs

2. **Configure Log Retention**:
    - Set up **log retention policies** to retain logs for the required duration (e.g., 90 days, 1 year) in compliance with data governance policies.
    - Logs were stored in **Cloud Storage** or **BigQuery** for analysis and long-term storage.
3. **Access Control**:
    - Configured IAM roles to control who can access the audit logs. Only authorized users, such as project admins and security officers, were granted permissions to view or modify audit logs.
    - Created **viewer roles** for non-administrative users who needed read-only access to logs for monitoring purposes.

2.5 Compliance Standards

Ensuring compliance with industry standards is essential for protecting sensitive data, maintaining trust, and avoiding legal or regulatory penalties. In the context of the implemented Big Data processing and machine learning pipeline, several key compliance standards were identified, and specific measures were taken to meet them. These standards primarily relate to data protection, privacy, and security.

Measures Taken to Meet Compliance Standards

1. **Data Encryption**
    - **GDPR and HIPAA**: Both regulations require the encryption of personal and sensitive data at rest and in transit. Google Cloud automatically encrypts data using AES-256 encryption, and Customer-Managed Encryption Keys (CMEK) were implemented for additional control over encryption.
    - **ISO/IEC 27001**: Encryption was part of the broader security controls to protect sensitive data. This measure was aligned with the ISO/IEC 27001 standard's data confidentiality requirements.
2. **Access Control and IAM (Identity and Access Management)**
    - **GDPR**: Ensured that only authorized users had access to personal data by configuring IAM roles based on the principle of least privilege. Sensitive data, including personal information, was only accessible by users who required it for their job functions.
    - **HIPAA**: Limited access to healthcare-related data to authorized personnel using strict IAM policies and roles. Service accounts were granted minimal access to reduce the risk of unauthorized exposure.
    - **ISO/IEC 27001**: Used detailed access controls, including multi-factor authentication (MFA) for critical access points, to prevent unauthorized access and meet security best practices as outlined in the ISO standard.
3. **Data Residency and Sovereignty**
    - **GDPR**: Ensured that personal data was stored and processed within the EU region or other approved locations. Google Cloud's ability to select the region for data storage, such as us-central1, ensured compliance with data residency requirements.
    - **ISO/IEC 27001**: Defined the data residency policies as part of the ISMS, ensuring that data was stored in secure locations and in compliance with regulatory requirements.
4. **Logging and Monitoring**
    - **GDPR**: Logging all access to personal data ensured that it could be traced back to a specific user, maintaining accountability and transparency in line with GDPR's requirements for data access records.
    - **HIPAA**: Set up detailed logging and monitoring of access to healthcare-related data, enabling compliance with HIPAA's audit trail requirements.

   - **ISO/IEC 27001**: Implemented audit logging and continuous monitoring as part of the ISMS to ensure that security controls were operating as expected, and any unauthorized actions were detected promptly.
5. **Incident Response Planning**
   - **GDPR**: A well-defined incident response plan was in place to notify affected individuals in the event of a data breach, as required by GDPR. This plan includes data breach detection, reporting, and resolution procedures.
   - **HIPAA**: Ensured that a comprehensive incident response plan was in place to address any security breaches involving healthcare data, which is required under HIPAA regulations.
   - **ISO/IEC 27001**: The incident response plan was regularly tested and updated to align with ISO/IEC 27001 standards for information security management.
6. **Data Retention and Deletion**
   - **GDPR**: The pipeline was designed to ensure that personal data was retained only for the necessary duration, with policies in place for data deletion or anonymization when no longer needed.
   - **HIPAA**: Ensured that healthcare-related data was retained according to HIPAA's retention schedules, and securely deleted when no longer required for legal or operational purposes.
   - **ISO/IEC 27001**: Data retention and deletion policies were aligned with the ISO/IEC 27001 framework to ensure sensitive data was managed and disposed of securely.

2.6 **Incident Response Planning**

Incident response planning is a vital component of ensuring that an organization is prepared to effectively respond to security breaches or data-related incidents. In this pipeline, an incident response plan was developed to detect, respond to, and recover from potential security threats or data breaches. The plan was structured to minimize the impact of incidents and ensure rapid recovery while maintaining compliance with regulatory standards.

Incident Response Plan Overview

The incident response plan for the Big Data and machine learning pipeline followed a structured, step-by-step process based on industry best practices, including those outlined by NIST (National Institute of Standards and Technology) and ISO/IEC 27035. The key components of the plan included:

7. **Preparation**:
   - **Establish an Incident Response Team (IRT)**: The IRT consisted of security professionals, cloud architects, and compliance officers. They were responsible for handling security incidents, investigating threats, and implementing necessary corrective actions.

- **Develop Incident Response Tools**: Automated tools were set up for real-time detection of potential security incidents, including anomaly detection through audit logs and monitoring of network traffic.

8. **Identification**:
   - **Alerting**: Alerts were configured to notify the IRT in the event of suspicious activities such as unauthorized data access, unusual changes in resource configurations, or any breach of security controls.
   - **Monitoring**: Real-time monitoring through Cloud Logging and Cloud Monitoring provided visibility into cloud resources, allowing the team to identify incidents quickly.
   - **Initial Investigation**: Once an alert was triggered, an initial investigation was conducted to verify the incident, assess its severity, and determine the scope of the breach.
9. **Containment**:
   - **Short-Term Containment**: The immediate goal was to contain the incident to prevent further damage. This involved actions such as isolating affected systems (e.g., shutting down a compromised VM or revoking unauthorized IAM roles).
   - **Long-Term Containment**: Steps were taken to ensure that similar incidents would not occur again, such as implementing more restrictive firewall rules or enhancing encryption methods.
10. **Eradication**:
    - **Root Cause Analysis**: After containment, the root cause of the incident was identified. For example, if an IAM role was misconfigured, the misconfiguration was corrected, and the permissions were properly adjusted.
    - **Patching**: Vulnerabilities were patched, and any malicious activity was removed from the environment (e.g., stopping malicious scripts or reversing unauthorized changes to resources).
11. **Recovery**:
    - **System Restoration**: Systems and services were restored to a secure state. This involved restoring data from backups, reconfiguring resources, and applying necessary security patches.
    - **Testing**: Once systems were restored, testing was conducted to ensure the environment was secure and that no further breaches occurred. The restored services were closely monitored during the recovery phase.
    - **Communication**: Throughout the process, internal stakeholders and affected users were notified, ensuring transparency and keeping them informed about the actions being taken.
12. **Lessons Learned**:
    - After the incident was resolved, a post-incident review was conducted. The IRT evaluated the response process, identified any gaps, and updated the incident response plan to improve future responses.
    - Root cause analysis was shared with the team to prevent similar incidents and improve overall security posture.

- **Reporting**: Detailed reports of the incident and the response actions taken were generated, and relevant compliance authorities were notified, if required (e.g., GDPR breach notification).

Conclusion

This assignment successfully demonstrated the integration of Google Cloud services to build a secure, scalable, and efficient Big Data processing and machine learning pipeline. By leveraging tools like BigQuery for data processing, Cloud Storage for secure data storage, AI Platform for machine learning, and Cloud KMS for key management, the project showcased how cloud-native services can simplify complex workflows, enhance machine learning capabilities, and ensure data security.

The focus on security, through measures such as IAM roles, data encryption, network security, and audit logging, was essential to protecting sensitive data and ensuring compliance with regulations like GDPR and HIPAA. The incident response plan ensured that the system was prepared to respond effectively to any potential security threats.

By adopting cloud computing and machine learning best practices, organizations can achieve scalability, flexibility, and cost-effectiveness while ensuring the security and integrity of their data. This holistic approach, combining secure data processing, machine learning, and robust security practices, provides a solid foundation for organizations to leverage big data and machine learning to gain valuable insights and drive innovation, all while maintaining compliance and minimizing risks.

References

1. https://cloud.google.com/vertex-ai/docs/general/deployment#api
2. https://cloud.google.com/vertex-ai/docs/model-registry/import-model
3. https://cloud.google.com/vertex-ai/docs/general/locations
4. https://supertype.ai/notes/deploying-machine-learning-models-with-vertex-ai-on-google-cloud-platform/
5. https://cloud.google.com/bigquery/docs/create-machine-learning-model
6. https://h2o.ai/wiki/binary-classification/#:~:text=Binary%20Classification%20is%20a%20type,either%20be%20positive%20or%20negative.