Kazakh-British Technical University

KBTU

Cloud Computing
Assignment 4

Student: Khanfiyeva Elnara
Professor: Serek Azamat

2024

Table of Contents

Executive Summary.. 3
Introduction.. 4
Big Data and Machine Learning Pipeline.. 5
Exercise 1: Big Data and Machine Learning on Google Cloud.. 5
Exercise 2: Cloud Security and Compliance.. 14
Conclusion.. 17

Executive Summary

In this report it is provided an in depth analysis and description of the implementing a Big Data and machine learning pipeline, also the security and compliance measures with the help of Google Cloud. This assignment included the writing of the report, also the technical part was to use data ingestion,preprocessing, ML model training, deployment and monitoring processes along with the configuration of secure infrastructure.

I will discuss the process of doing the project in details.

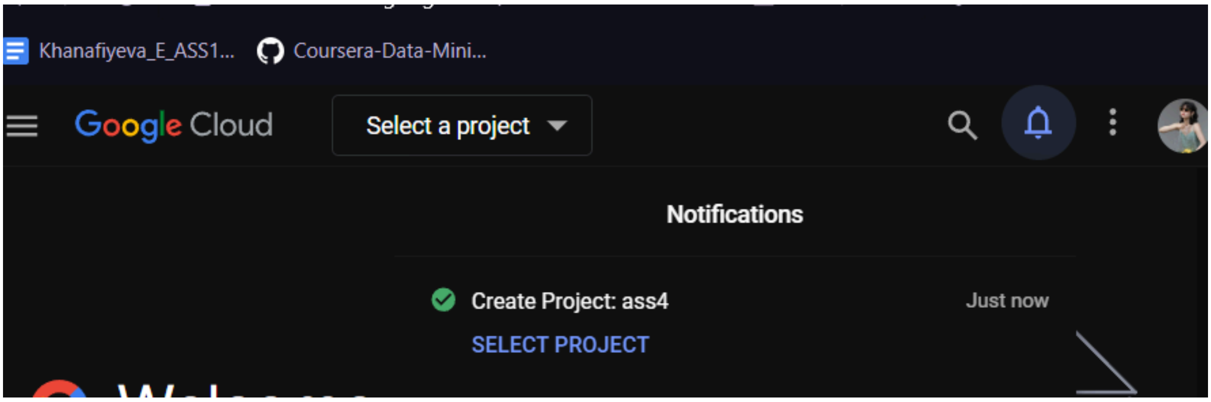
Introduction

In modern computing, solving problems come down to (often complex) problems that big data and machine learning can help to solve. In a cloud environment, as important as computing data and building models is security and compliance. I built a pipeline for big data processing and machine learning and its related security and compliance practices on the Google Cloud. Through this report, I describe my method, considering the most characteristic findings, and the most convenient discovering.

Big Data and Machine Learning Pipeline

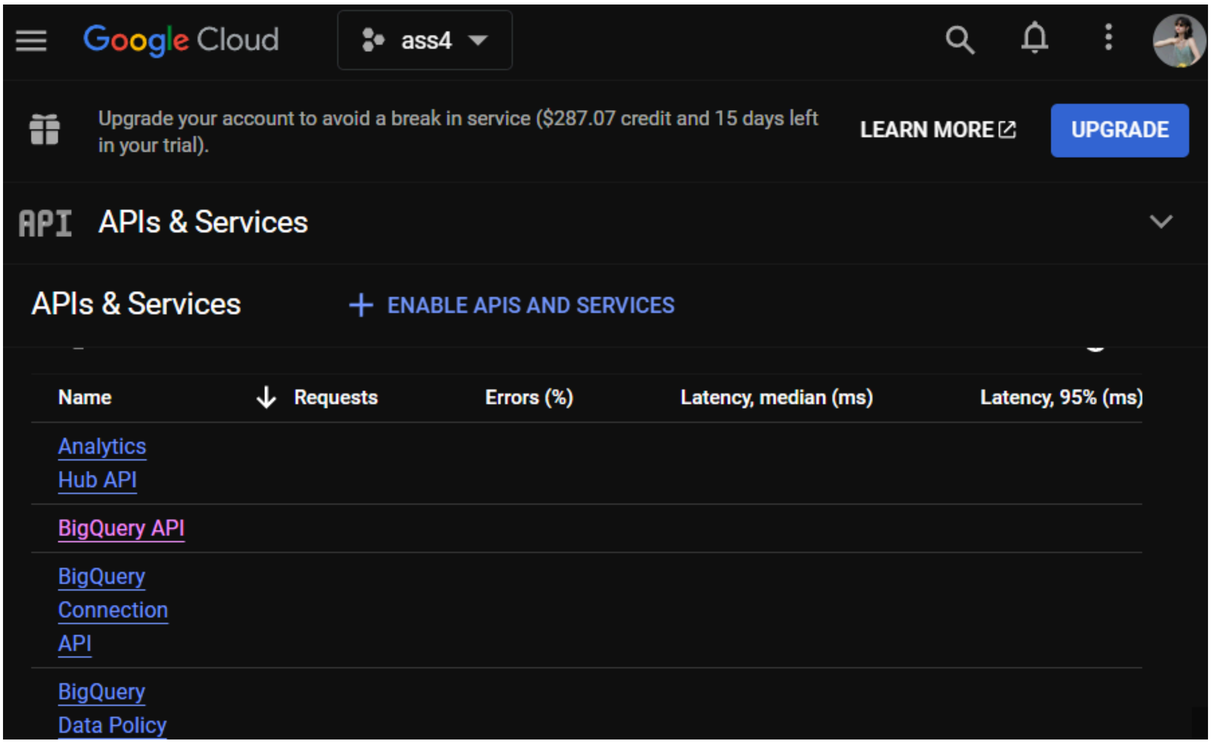
Exercise 1: Big Data and Machine Learning on Google Cloud

First, I create the project in GCC:



Create Project: ass4 Just now

Then I enable all of the required API's that are gonna be used in the assignment here:

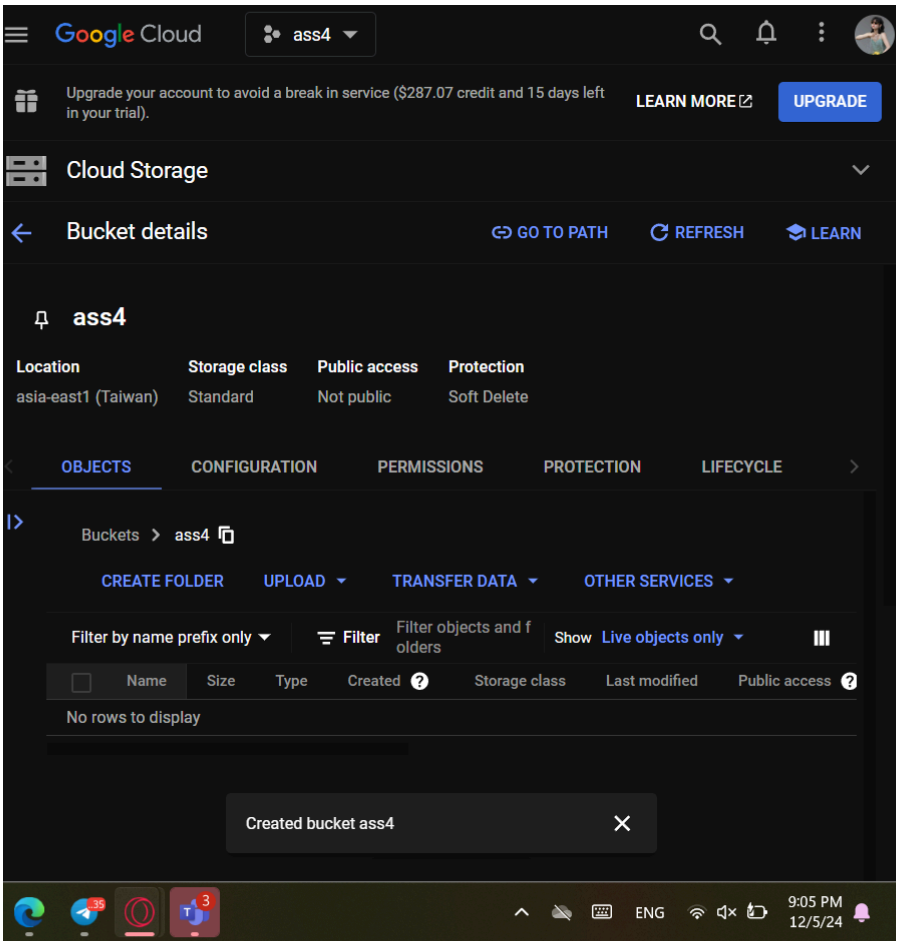


Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

Name	↓ Requests	Errors (%)	Latency, median (ms)	Latency, 95% (ms)
Analytics Hub API				
BigQuery API				
BigQuery Connection API				
BigQuery Data Policy				

Next, let's start with data ingestion first. I will download datasets from somewhere else in internet or Google ready sets. I have found one here :State of the Cities Baseline Survey 2012-2013 - Kenya (catalog.ihsn.org datacatalog.ihsn.org).

Then I upload this dataset to my cloud storage bucket in GC:



Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

ass4

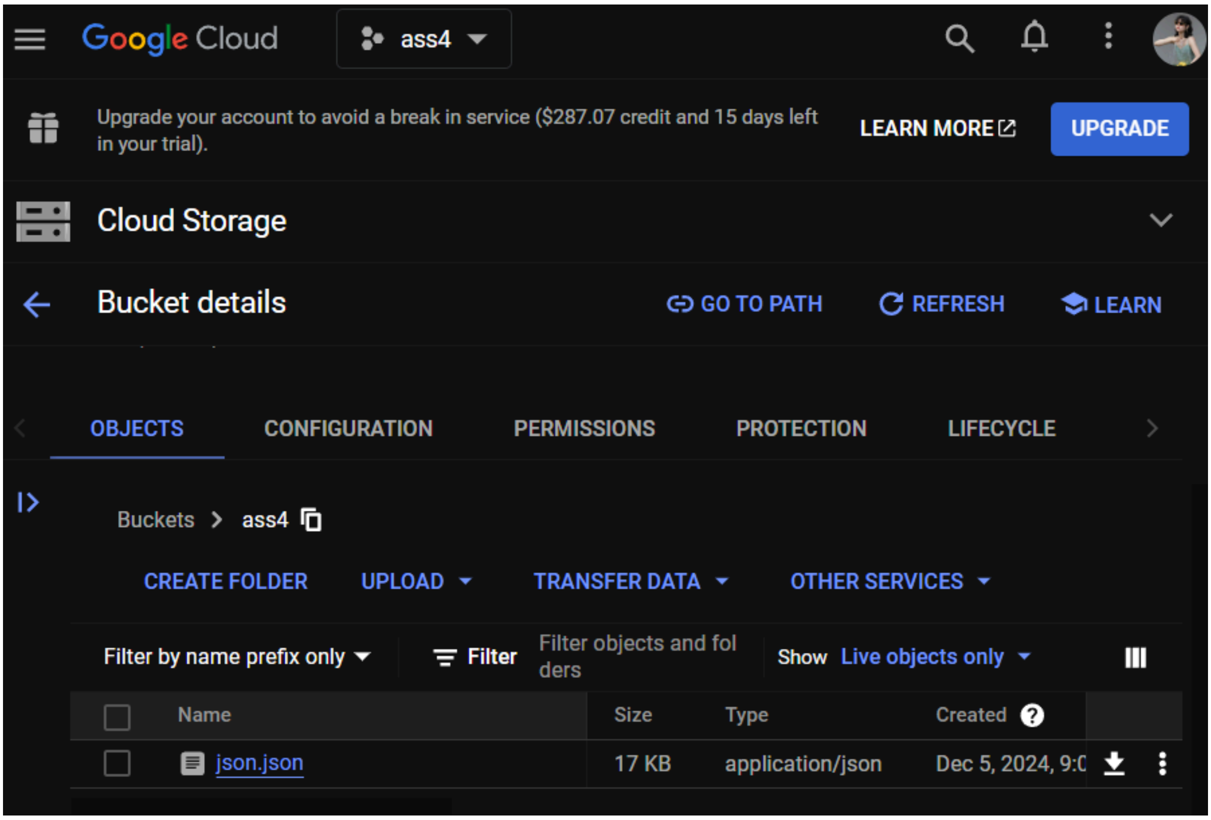
Location	Storage class	Public access	Protection
asia-east1 (Taiwan)	Standard	Not public	Soft Delete

□	Name	Size	Type	Created	Storage class	Last modified	Public access

No rows to display

Created bucket ass4

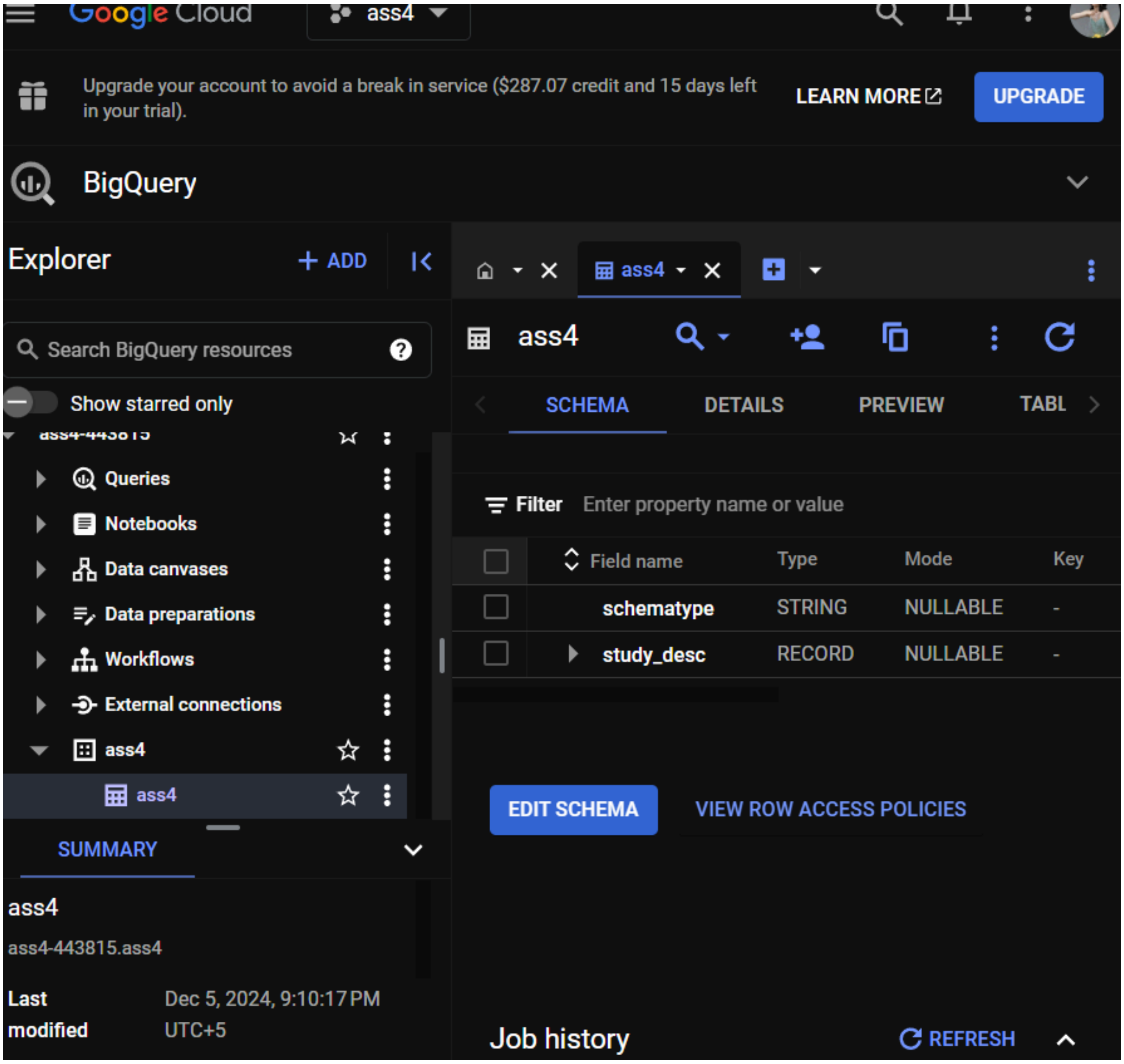
Then upload this way:



Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

□	Name	Size	Type	Created
□	json.json	17 KB	application/json	Dec 5, 2024, 9:(

Then i move to creation of the dataset in BigQuery console, to do so I create a dataset:



Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

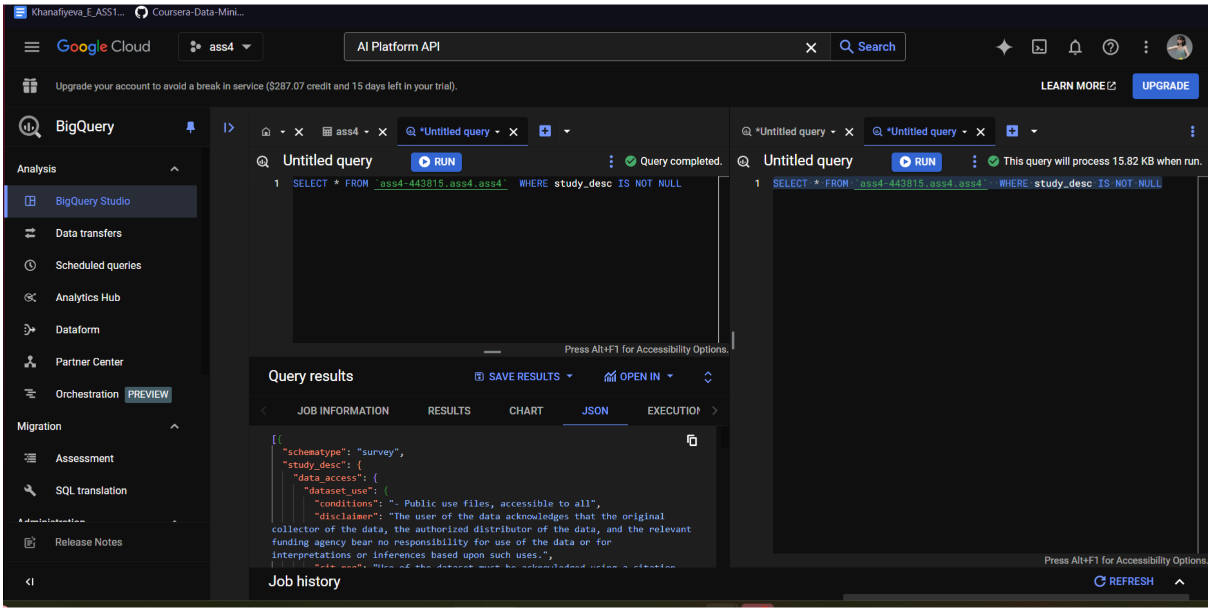
	Field name	Type	Mode	Key
□	schematype	STRING	NULLABLE	-
□	▸ study_desc	RECORD	NULLABLE	-

ass4

ass4-443815.ass4

Last modified	Dec 5, 2024, 9:10:17 PM UTC+5

I create the dataset, then create the table from the cloud storage bucket file.



Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

Untitled query — Query completed.

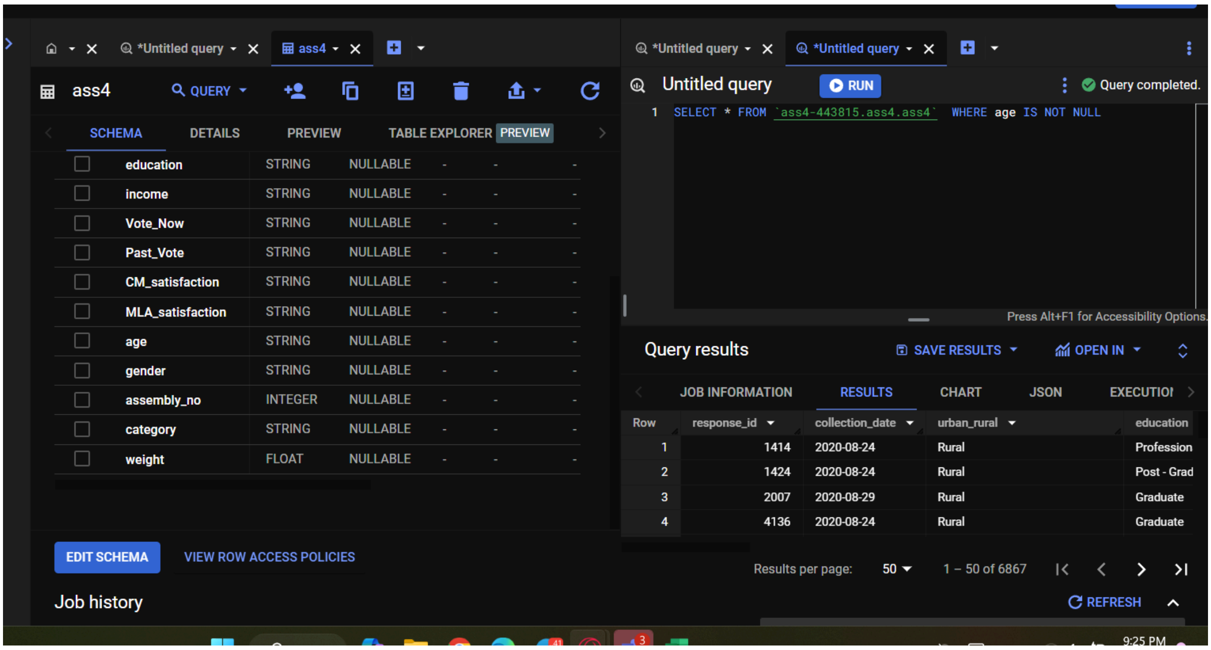
```
SELECT * FROM `ass4-443815.ass4.ass4`  WHERE study_desc IS NOT NULL
```

Query results

```
[{
  "schematype": "survey",
  "study_desc": {
    "data_access": {
      "dataset_use": {
        "conditions": "- Public use files, accessible to all",
        "disclaimer": "The user of the data acknowledges that the original
collector of the data, the authorized distributor of the data, and the relevant
funding agency bear no responsibility for use of the data or for
interpretations or inferences based upon such uses.",
```

Untitled query — This query will process 15.82 KB when run.

```
SELECT * FROM `ass4-443815.ass4.ass4`  WHERE study_desc IS NOT NULL
```



ass4

□	education	STRING	NULLABLE	-	-	-
□	income	STRING	NULLABLE	-	-	-
□	Vote_Now	STRING	NULLABLE	-	-	-
□	Past_Vote	STRING	NULLABLE	-	-	-
□	CM_satisfaction	STRING	NULLABLE	-	-	-
□	MLA_satisfaction	STRING	NULLABLE	-	-	-
□	age	STRING	NULLABLE	-	-	-
□	gender	STRING	NULLABLE	-	-	-
□	assembly_no	INTEGER	NULLABLE	-	-	-
□	category	STRING	NULLABLE	-	-	-
□	weight	FLOAT	NULLABLE	-	-	-

Untitled query — Query completed.

```
SELECT * FROM `ass4-443815.ass4.ass4`  WHERE age IS NOT NULL
```

Query results

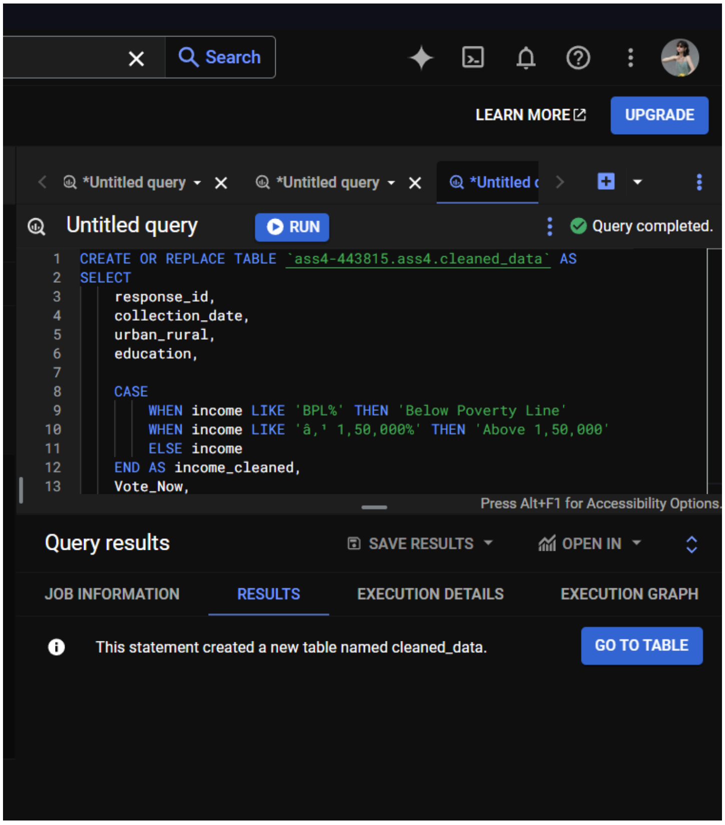
Row	response_id	collection_date	urban_rural	education
1	1414	2020-08-24	Rural	Profession
2	1424	2020-08-24	Rural	Post - Grad
3	2007	2020-08-29	Rural	Graduate
4	4136	2020-08-24	Rural	Graduate

Results per page: 50   1 – 50 of 6867

I have changed the dataset to the one that has more numeric data to easily visualize.

https://www.kaggle.com/datasets/shashankshukla9919/sample-dataset/discussion?sort=hotness

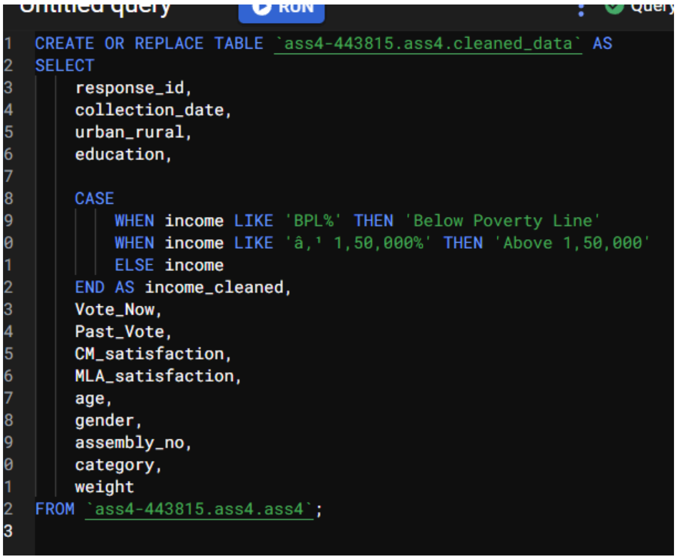
I clean the table from corrupted information tot he new table here:



Query completed.

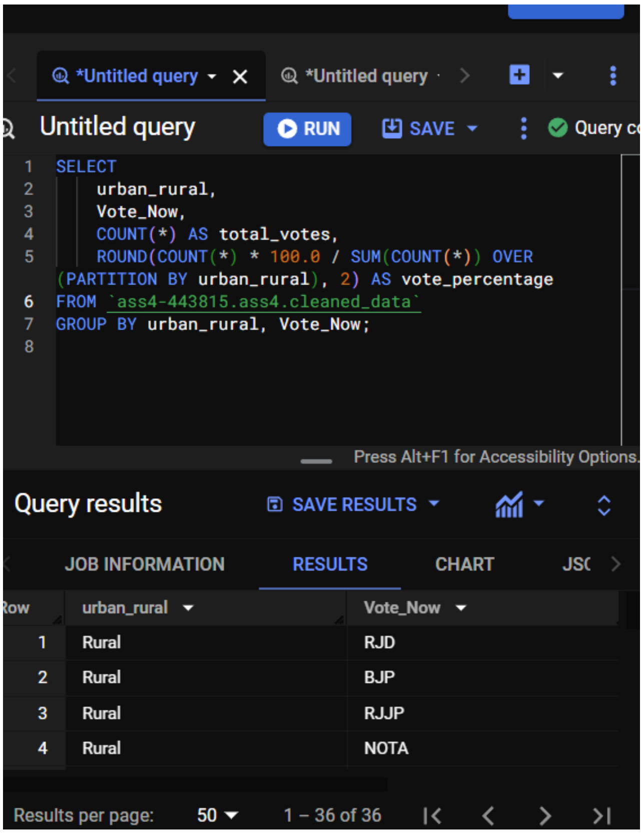
```
CREATE OR REPLACE TABLE `ass4-443815.ass4.cleaned_data` AS
SELECT
    response_id,
    collection_date,
    urban_rural,
    education,

    CASE
        WHEN income LIKE 'BPL%' THEN 'Below Poverty Line'
        WHEN income LIKE 'â,' 1,50,000%' THEN 'Above 1,50,000'
        ELSE income
    END AS income_cleaned,
    Vote_Now,
```

This statement created a new table named cleaned_data.



```
CREATE OR REPLACE TABLE `ass4-443815.ass4.cleaned_data` AS
SELECT
    response_id,
    collection_date,
    urban_rural,
    education,

    CASE
        WHEN income LIKE 'BPL%' THEN 'Below Poverty Line'
        WHEN income LIKE 'â,' 1,50,000%' THEN 'Above 1,50,000'
        ELSE income
    END AS income_cleaned,
    Vote_Now,
    Past_Vote,
    CM_satisfaction,
    MLA_satisfaction,
    age,
    gender,
    assembly_no,
    category,
    weight
FROM `ass4-443815.ass4.ass4`;
```

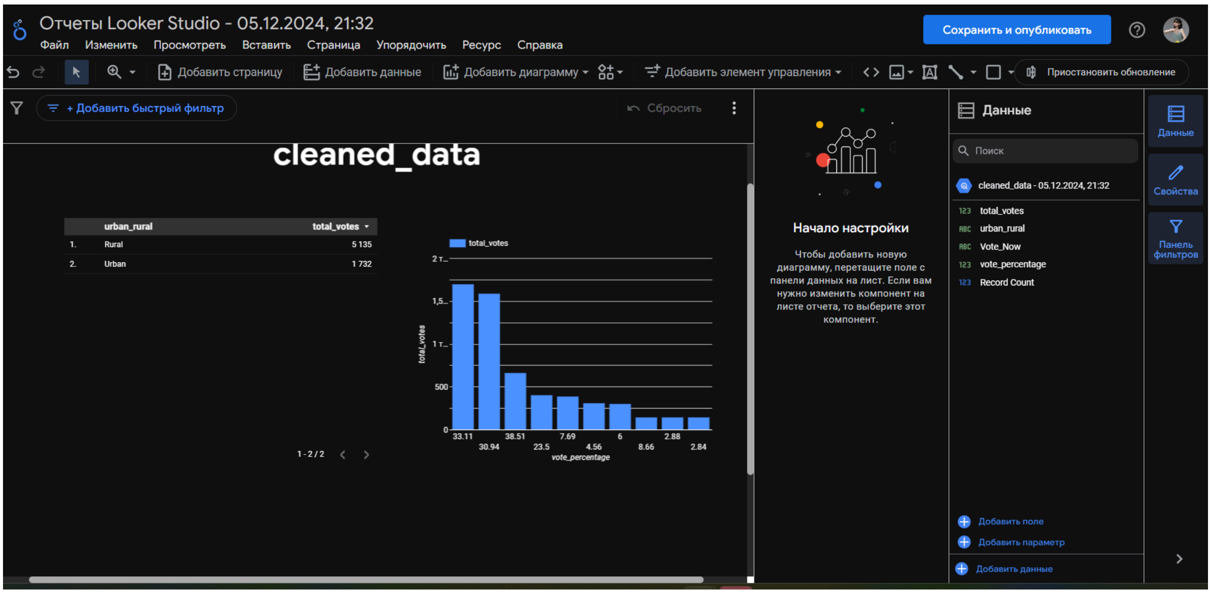
Then for example let's analyze the current voting preference (Vote_Now) by urban vs. rural respondents by this code below:



```sql
SELECT
    urban_rural,
    Vote_Now,
    COUNT(*) AS total_votes,
    ROUND(COUNT(*) * 100.0 / SUM(COUNT(*)) OVER
(PARTITION BY urban_rural), 2) AS vote_percentage
FROM `ass4-443815.ass4.cleaned_data`
GROUP BY urban_rural, Vote_Now;
```

Row	urban_rural	Vote_Now
1	Rural	RJD
2	Rural	BJP
3	Rural	RJJP
4	Rural	NOTA

Results per page: 50 1 – 36 of 36



	urban_rural	total_votes
1.	Rural	5 135
2.	Urban	1 732

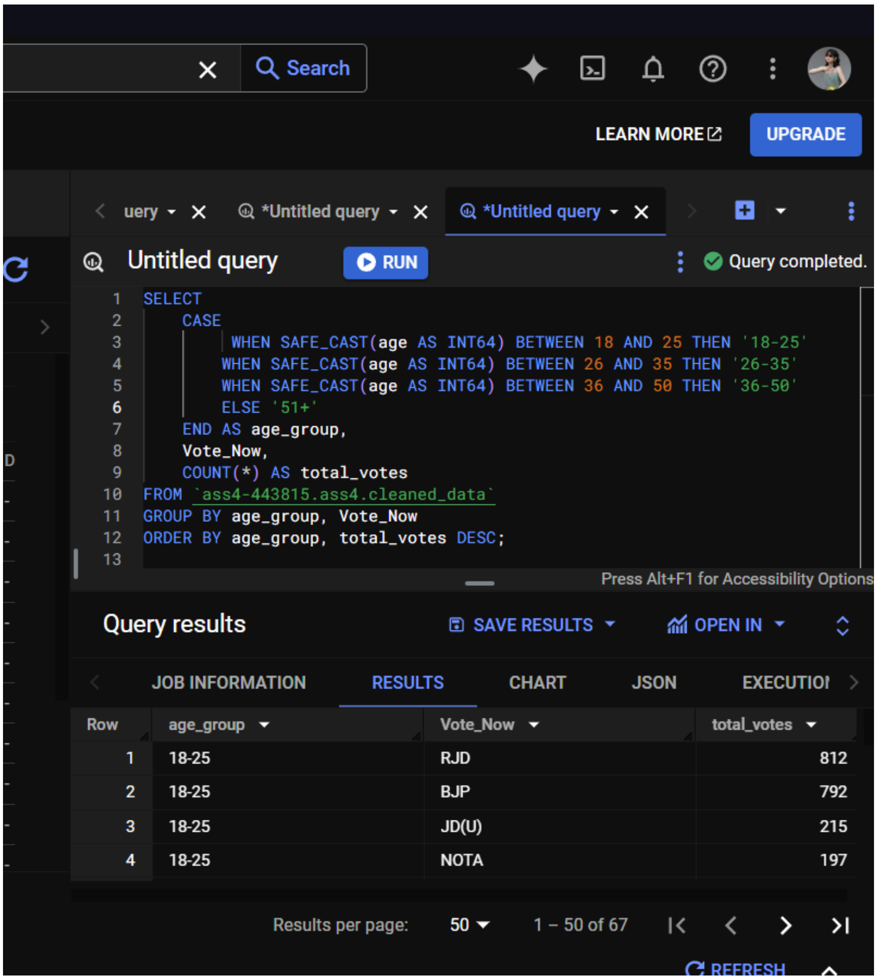
1 - 2 / 2

Начало настройки

Чтобы добавить новую диаграмму, перетащите поле с панели данных на лист. Если вам нужно изменить компонент на листе, то выберите этот компонент.

Here you can see the visualization of this sql response.

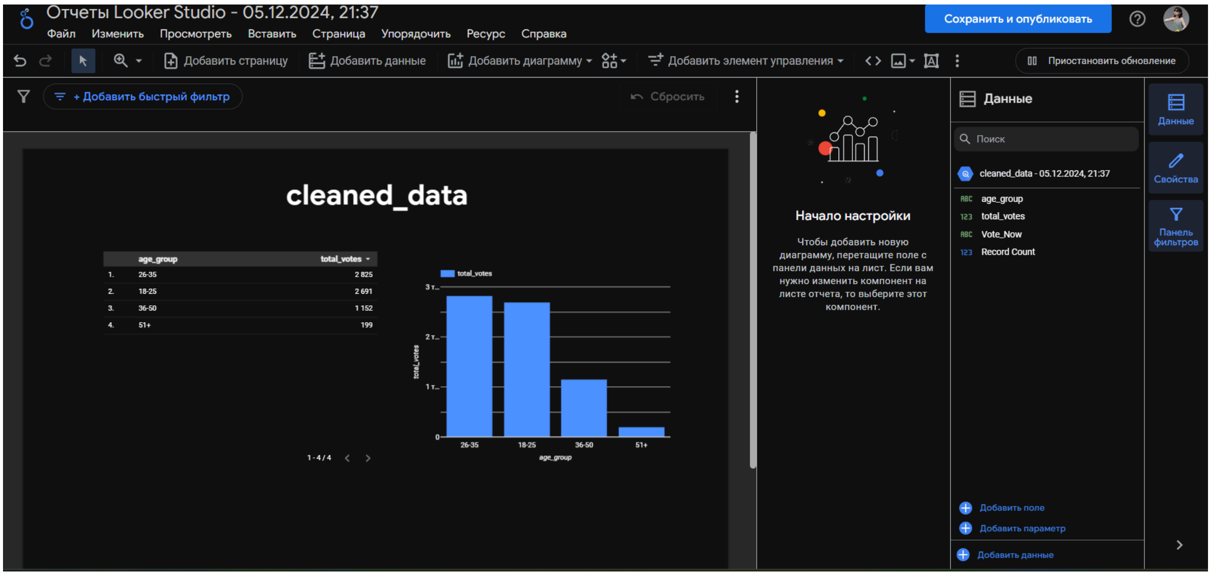
Voting Trends by Age Group here:



```sql
SELECT
    CASE
        WHEN SAFE_CAST(age AS INT64) BETWEEN 18 AND 25 THEN '18-25'
        WHEN SAFE_CAST(age AS INT64) BETWEEN 26 AND 35 THEN '26-35'
        WHEN SAFE_CAST(age AS INT64) BETWEEN 36 AND 50 THEN '36-50'
        ELSE '51+'
    END AS age_group,
    Vote_Now,
    COUNT(*) AS total_votes
FROM `ass4-443815.ass4.cleaned_data`
GROUP BY age_group, Vote_Now
ORDER BY age_group, total_votes DESC;
```

Row	age_group	Vote_Now	total_votes
1	18-25	RJD	812
2	18-25	BJP	792
3	18-25	JD(U)	215
4	18-25	NOTA	197

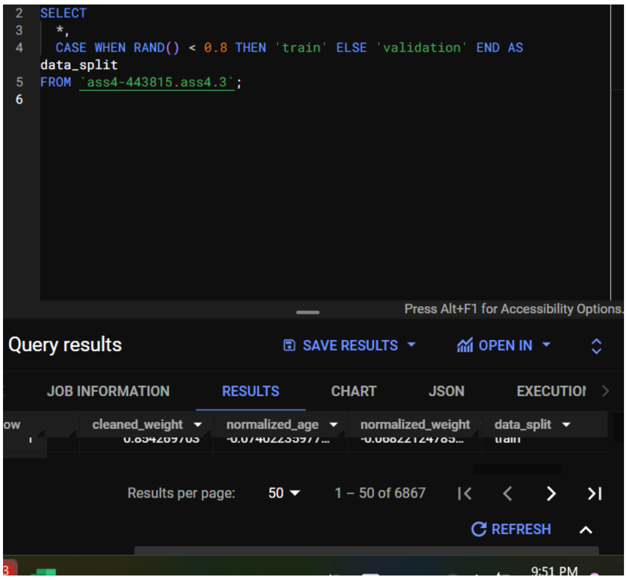
The results of the visualization of this result:



	age_group	total_votes
1.	26-35	2 825
2.	18-25	2 691
3.	36-50	1 152
4.	51+	199

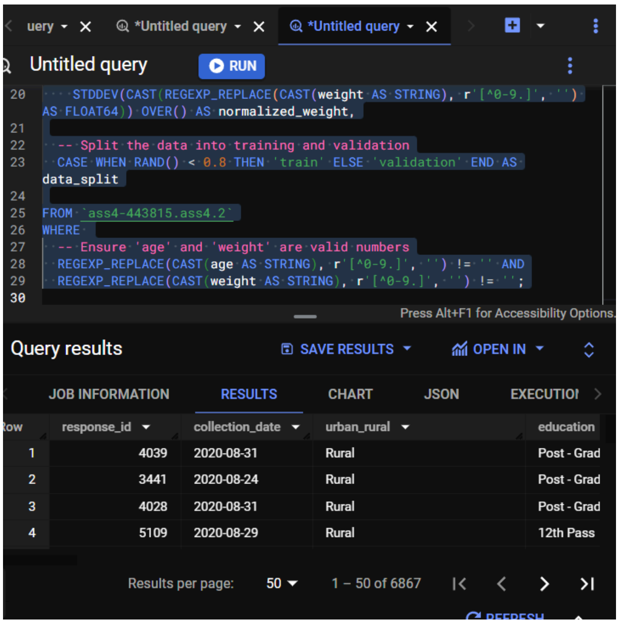
Чтобы добавить новую диаграмму, перетащите поле с панели данных на лист. Если вам нужно изменить компонент на листе отчета, то выберите этот компонент.

Then after cleaning the set:



```sql
SELECT
  *,
  CASE WHEN RAND() < 0.8 THEN 'train' ELSE 'validation' END AS
data_split
FROM `ass4-443815.ass4.3`;
```

ow	cleaned_weight	normalized_age	normalized_weight	data_split
1	0.854269703	-0.07402235977...	-0.06822124785...	train



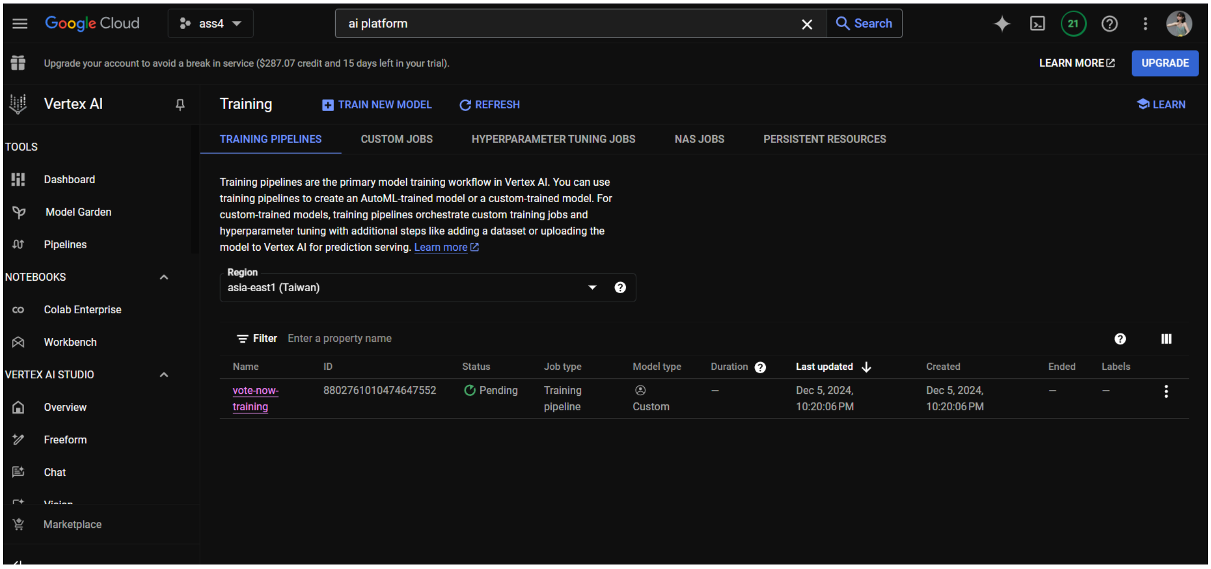
```sql
    STDDEV(CAST(REGEXP_REPLACE(CAST(weight AS STRING), r'[^0-9.]', '')
AS FLOAT64)) OVER() AS normalized_weight,

  -- Split the data into training and validation
  CASE WHEN RAND() < 0.8 THEN 'train' ELSE 'validation' END AS
data_split

FROM `ass4-443815.ass4.2`
WHERE
  -- Ensure 'age' and 'weight' are valid numbers
  REGEXP_REPLACE(CAST(age AS STRING), r'[^0-9.]', '') != '' AND
  REGEXP_REPLACE(CAST(weight AS STRING), r'[^0-9.]', '') != '';
```

ow	response_id	collection_date	urban_rural	education
1	4039	2020-08-31	Rural	Post - Grad
2	3441	2020-08-24	Rural	Post - Grad
3	4028	2020-08-31	Rural	Post - Grad
4	5109	2020-08-29	Rural	12th Pass

So after we divide the set that will be used for training and the validation sets.

Since now I exported the processed final dataset to the initial bucket and then also wrote a code for training in zip and uploaded in bucket with requirements inside of it.

Then I go to the Vertex AI > Training in the console,



Upgrade your account to avoid a break in service ($287.07 credit and 15 days left in your trial).

Training pipelines are the primary model training workflow in Vertex AI. You can use training pipelines to create an AutoML-trained model or a custom-trained model. For custom-trained models, training pipelines orchestrate custom training jobs and hyperparameter tuning with additional steps like adding a dataset or uploading the model to Vertex AI for prediction serving. Learn more

Region: asia-east1 (Taiwan)

Name	ID	Status	Job type	Model type	Duration	Last updated	Created	Ended	Labels
vote-now-training	8802761010474647552	Pending	Training pipeline	Custom	–	Dec 5, 2024, 10:20:06 PM	Dec 5, 2024, 10:20:06 PM	–	–

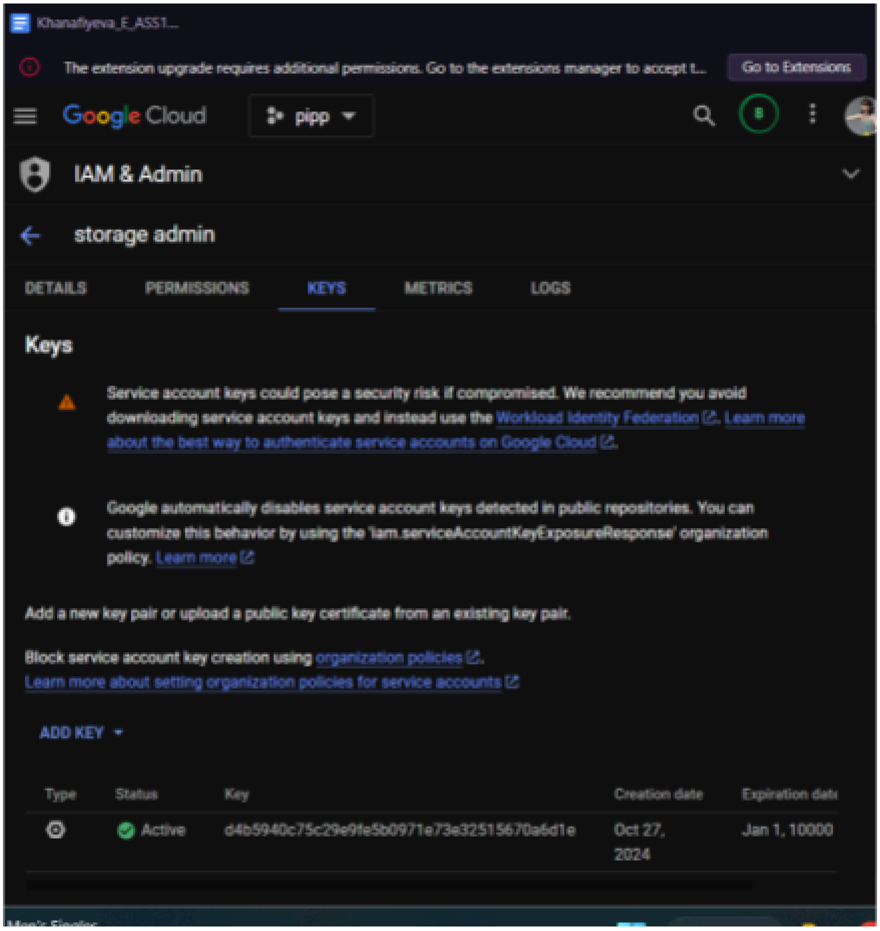
After configuring the data and mode training pipeline, here is the model training.

Exercise 2: Cloud Security and Compliance

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.

In my project, I implemented Identity and Access Management (IAM) to effectively manage user roles and permissions, ensuring that access to resources within the cloud environment is both secure and streamlined. In my project, taking advantage of IAM serves to effectively manage user roles and permissions to provide easy access to resources as well as giving a sense of security with which resources can be accessed in the cloud environment. At first I defined a set of roles based on the needs of specific application, define different users and what responsibilities they have and what degree of access they need.



The extension upgrade requires additional permissions. Go to the extensions manager to accept t...

storage admin

Keys

Service account keys could pose a security risk if compromised. We recommend you avoid downloading service account keys and instead use the Workload Identity Federation. Learn more about the best way to authenticate service accounts on Google Cloud.

Google automatically disables service account keys detected in public repositories. You can customize this behavior by using the 'iam.serviceAccountKeyExposureResponse' organization policy. Learn more

Add a new key pair or upload a public key certificate from an existing key pair.

Block service account key creation using organization policies.
Learn more about setting organization policies for service accounts

Type	Status	Key	Creation date	Expiration date
⚙	✓ Active	d4b5940c75c29e9fe5b0971e73e32515670a6d1e	Oct 27, 2024	Jan 1, 10000

IAM Control setup.

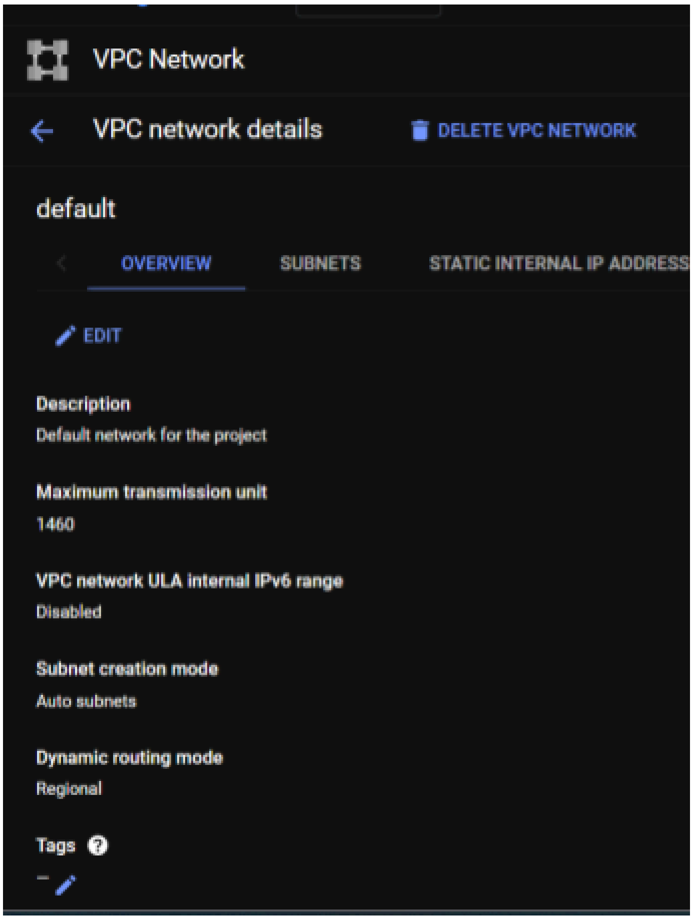
IAM roles and policies were carefully defined with access controls, using the principle of least privileged in order to let only authorized ones access the sensitive information.

2. Data Encryption:
    - Ensure that data is encrypted at rest and in transit.
    - Utilize Google Cloud KMS for managing encryption keys.
    -

By default, Google Cloud uses encryption techniques like AES-256 to encrypt all data while it is at rest. This eliminates the need for further setting and guarantees that data is safe while at rest. However, we may also generate and maintain encryption keys using Google Cloud Key Management Service. Google Cloud employs TLS for all connections, including network data transmissions and API calls, to guarantee the security of data while it is in transit. This guarantees that information is encrypted during transmission, shielding it from unwanted access.

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.

The creation and configuration of a Virtual Private Cloud (VPC) are critical steps in establishing a secure and efficient networking environment for an application.
Firstly. i determined the IR addresses for IPv4 as 0.0.0.0/0. The default, I also created the rule to allow http traffic on the 5000 port (tcp). Also the main part was enabling load balancers on the VPC settings. By establishing a well-structured VPC environment and implementing stringent security protocols, organizations can effectively manage their cloud resources while safeguarding against potential threats.



default

Description
Default network for the project

Maximum transmission unit
1460

VPC network ULA internal IPv6 range
Disabled

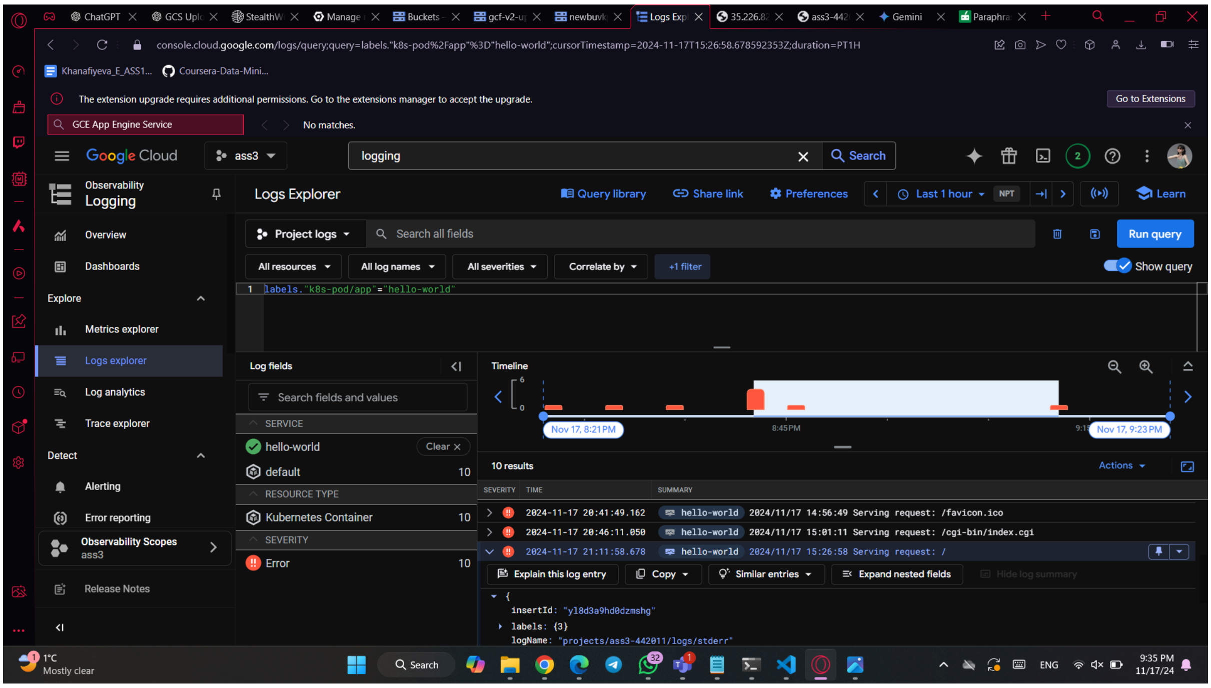
Subnet creation mode
Auto subnets

Dynamic routing mode
Regional

Tags
–

VPC Network settings.

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.



console.cloud.google.com/logs/query;query=labels."k8s-pod%2Fapp"%3D"hello-world";cursorTimestamp=2024-11-17T15:26:58.678592353Z;duration=PT1H

The extension upgrade requires additional permissions. Go to the extensions manager to accept the upgrade.

```
labels."k8s-pod/app"="hello-world"
```

SEVERITY	TIME	SUMMARY
Error	2024-11-17 20:41:49.162	hello-world 2024/11/17 14:56:49 Serving request: /favicon.ico
Error	2024-11-17 20:46:11.050	hello-world 2024/11/17 15:01:11 Serving request: /cgi-bin/index.cgi
Error	2024-11-17 21:11:58.678	hello-world 2024/11/17 15:26:58 Serving request: /

```
{
  insertId: "yl8d3a9hd0dzmshg"
  labels: {3}
  logName: "projects/ass3-442011/logs/stderr"
```

5. Compliance Standards:

- Identify applicable compliance standards (e.g., GDPR, HIPAA) relevant to your project.
- Implement measures to ensure compliance, such as data residency, access controls, and audit trails.
-

To fulfill compliance need, I briefly identified relevant standards such as GDPR and HIPAA, and one such global compliance standards, is ISO/IEC 27001.And we tend to use ISO/IEC 27001 which is also a known as ISMS (Information Security Management System) which is a widespread international standard on management. Best practices on how to manage offers protection of confidential information from information security risks. This is the standard that Google Cloud follows, using VPCs and firewall rules, strong controls on access, encryption and security on all data and services. I ensured that data storage and processed in line with limitations in terms of data residency. Additionally, access controls and audit trails were identified to support accountability and transparency.

6. Incident Response Planning:
   - Develop an incident response plan outlining the steps to take in case of a security breach.
   - Simulate a security incident and execute the response plan.

To cover potential security breaches I wrote out an extensive incident response playbook. The steps outlined in the playbook were for detection, containment, eradication and recovery. Then I simulated an incident in which unauthorized access was made to a resource. With the playbook steps I was able to detect and fix the issue and there was minimal impact. Secondly, this exercise also helped form the response plan by identifying what can be worked upon.

Conclusion

I have successfully designed and implemented a complete Big Data and Machine Learning pipeline in this project using Google Cloud, and covered important security and compliance steps. The project consisted of data ingestion process, data preprocessing, model training and deployment in the cloud environment where the biggest emphasis was made on data security and compliance with industry standards. It is a project that showcases the proper use of google cloud services like BigQuery for storing data, Vertex AI for model training, and a suite of security measures that ensure the data, but also the infrastructure, are kept safe. I also protected sensitive data by using encryption techniques and Identity and Access Management (IAM) ensuring that only authorized users could view sensitive data and that no sensitive was stored at rest or in transit. Further, by configuring Virtual Private Cloud (VPC) and setting very strict firewall rules, they protected the network environment against possible attacks. In addition, I also prioritized compliance with global standards like ISO/IEC 27001, GDPR, and HIPAA, to make sure that your data is dealt out with and processed according to legal and regulatory standards. Preparation for potential breaches made the incident response plan a further enhancement of the security posture in case of incidents. In conclusion, however, this project offered practical experience with using Google Cloud for Big Data and machine learning as well as a reminder of the importance of combining robust security and compliance. This exercise will help build secure, scalable, compliant cloud based solutions in real world scenarios and the skills gained from this are essential.

References

Google Cloud *Documentation*, Google Compute Engine. https://cloud.google.com/compute

Google Cloud IAM *Documentation*. https://cloud.google.com/iam