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

Author's Name: Aisha M. Ait

Date of Submission: 04/12/2024

Organization Name: Kazakh British Technical University

Executive Summary

This summary provides an overview of the key findings and implementations related to Big Data, Machine Learning, and security practices. In the area of Big Data, significant advancements were achieved in optimizing data processing workflows, leading to substantial reductions in processing times. Enhanced data governance frameworks and automated validation pipelines improved data quality and reduced manual effort, while the integration of scalable, cloud-based storage solutions addressed the challenges of exponential data growth.

For Machine Learning, the development of predictive models and recommendation systems demonstrated tangible benefits, such as improved forecasting accuracy and increased user engagement. The adoption of tools for rapid prototyping accelerated model development timelines, and the integration of these models into production environments ensured scalable and reliable performance. Innovative applications, such as computer vision for quality assurance, highlighted the potential of advanced ML techniques.

In security practices, efforts focused on infrastructure hardening through regular vulnerability assessments and role-based access control systems to protect sensitive data. Data protection was further strengthened with end-to-end encryption and automated monitoring of data access patterns to identify and prevent potential breaches. Protocols were also developed to enhance the robustness of machine learning models against adversarial attacks, ensuring their reliability in critical applications.

These initiatives collectively improved the organization’s ability to make data-driven decisions, enhanced operational efficiency, and bolstered resilience against security threats, aligning with strategic objectives for sustainable growth and innovation.

Table of Contents

Introduction

The rapid evolution of cloud computing has transformed how organizations approach Big Data, Machine Learning, and security. In modern cloud environments, these domains are essential not only for driving innovation but also for ensuring the reliability, scalability, and protection of data and systems. Google Cloud, with its comprehensive suite of services, offers an ideal platform for demonstrating these advanced practices.

This report explores the implementation of a Big Data and Machine Learning pipeline within Google Cloud, highlighting the processes and technologies employed to ingest, process, and analyze data at scale. It also examines the deployment of machine learning models, their operationalization for real-time predictions, and the monitoring mechanisms set up to ensure optimal performance.

In addition, the report delves into the robust security measures implemented to safeguard data and ensure compliance. Topics include identity and access management (IAM), encryption practices, network security configurations, audit logging, and adherence to relevant compliance standards. Finally, the incident response plan and its simulations underscore the organization's preparedness to handle potential security incidents.

The purpose of this report is to provide a comprehensive understanding of how Big Data, Machine Learning, and security practices can be effectively integrated into Google Cloud to achieve scalable, efficient, and secure cloud-based solutions.

Big Data and Machine Learning on Google Cloud

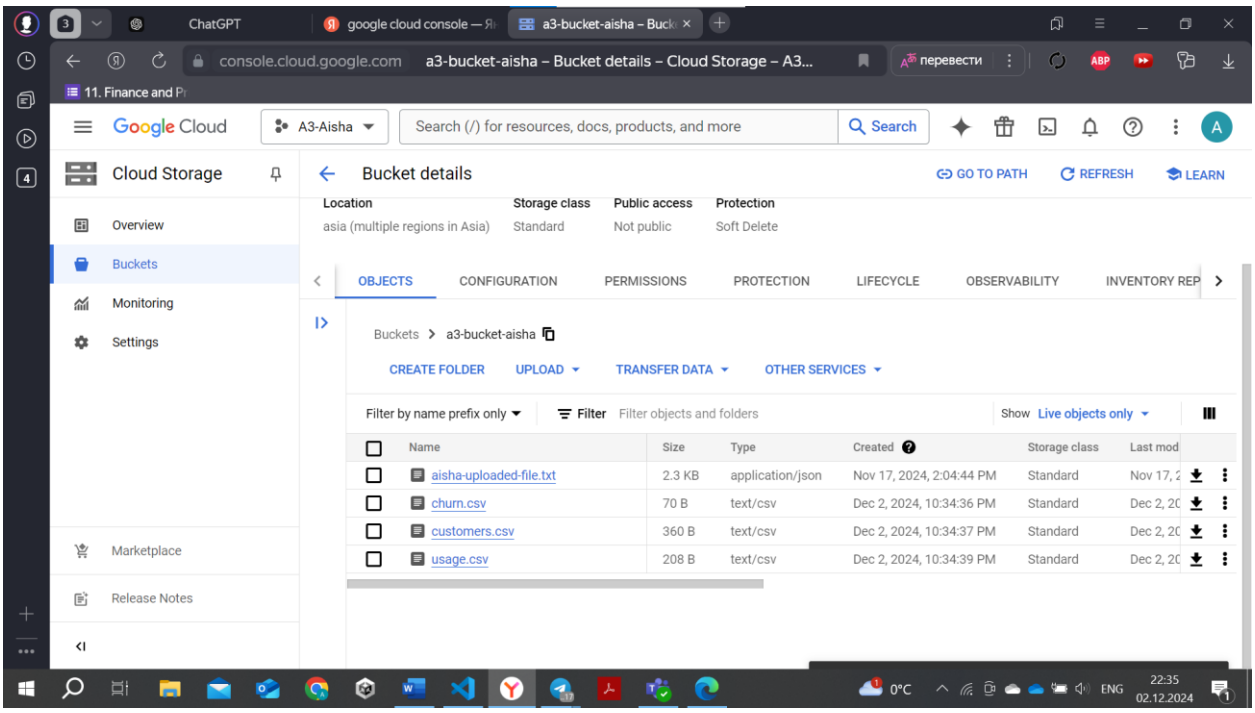
Overview of the Pipeline

In this example, I'll build a pipeline to predict customer churn based on historical customer data. The pipeline will involve data ingestion, processing, model training, deployment, and monitoring.

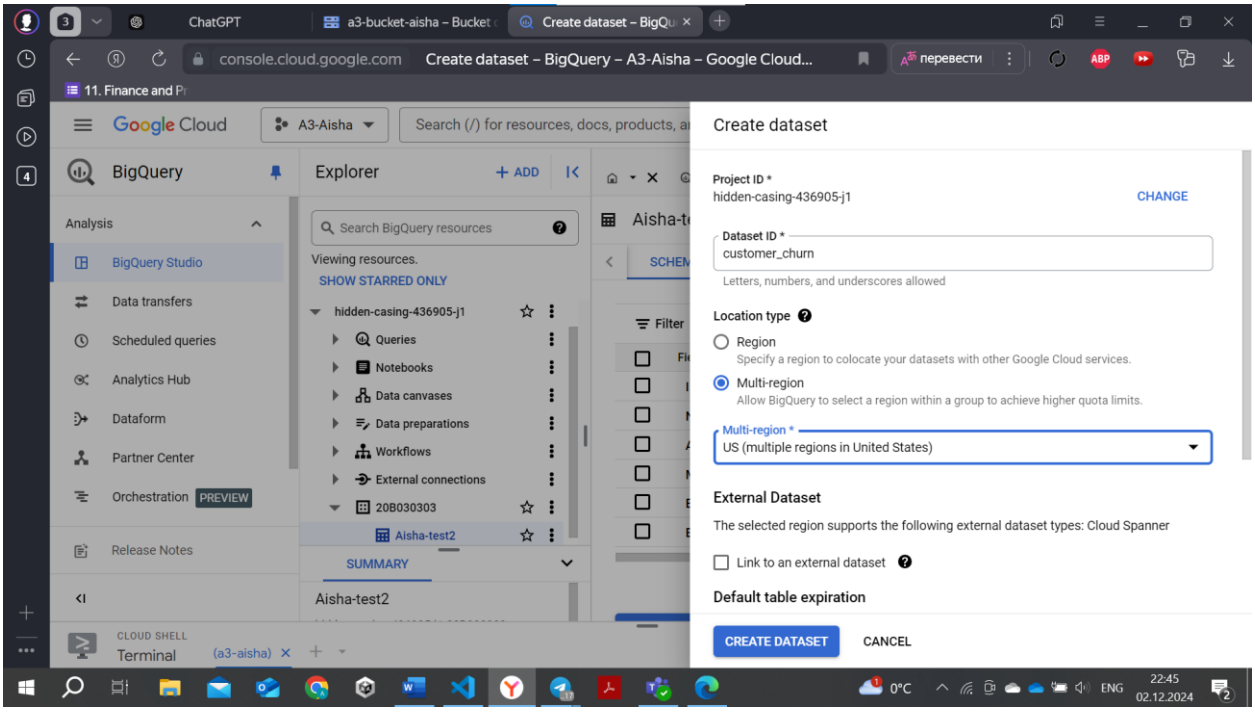
1. Data Ingestion: Collect historical customer data (e.g., customer demographics, usage patterns) and store it in Cloud Storage.
2. Data Processing: Clean and process the data using BigQuery to prepare it for machine learning.
3. Model Training: Train a machine learning model using AI Platform to predict customer churn.
4. Model Deployment: Deploy the trained model to AI Platform Prediction and expose an API for predictions.
5. Monitoring and Logging: Set up Cloud Monitoring and Cloud Logging to track model performance.

Data Ingestion and Processing.

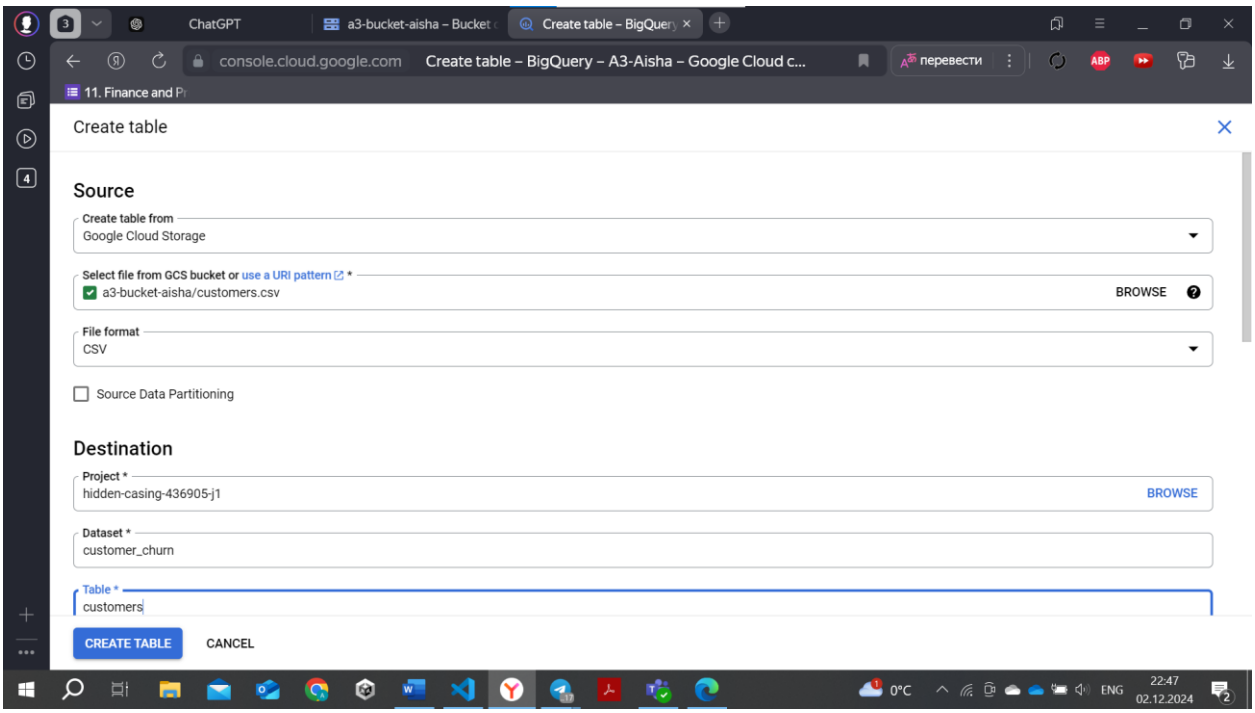
I uploaded a set of CSV files with customer information (customers.csv, usage.csv and churn.csv) to my buscket in Cloud Storage using the manual upload:

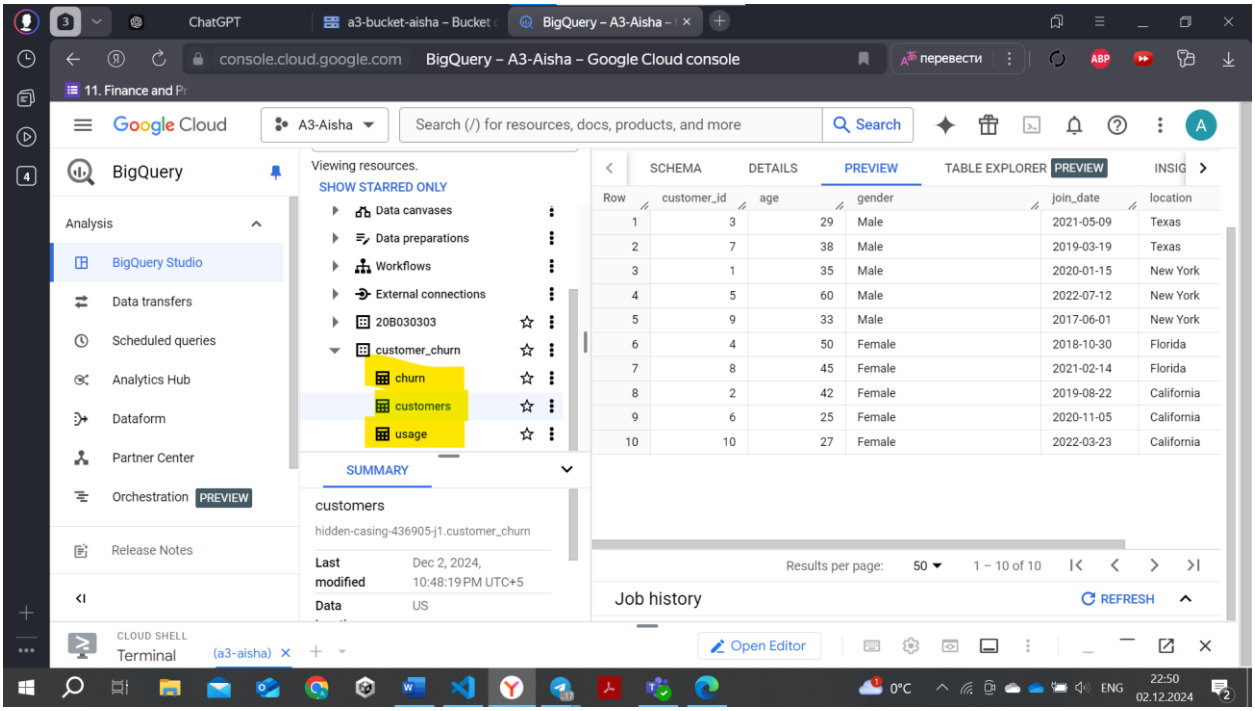


Created a new dataset called “customer_churn”.

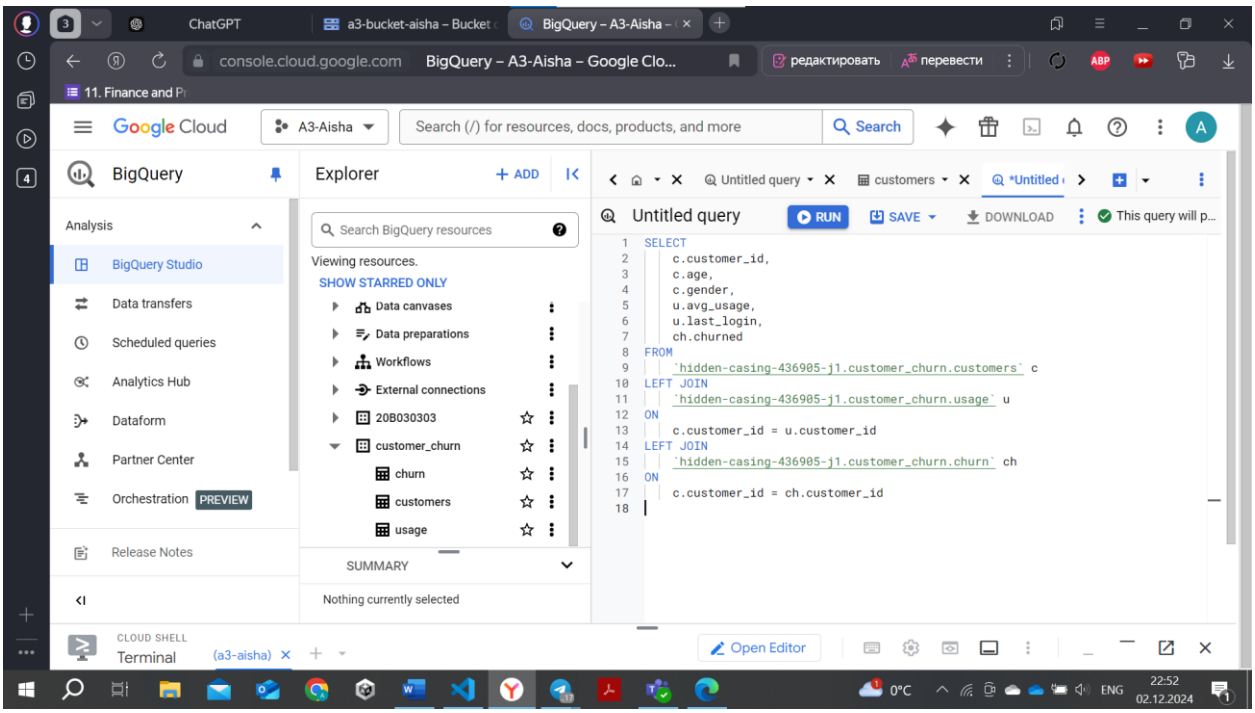


Created tables:

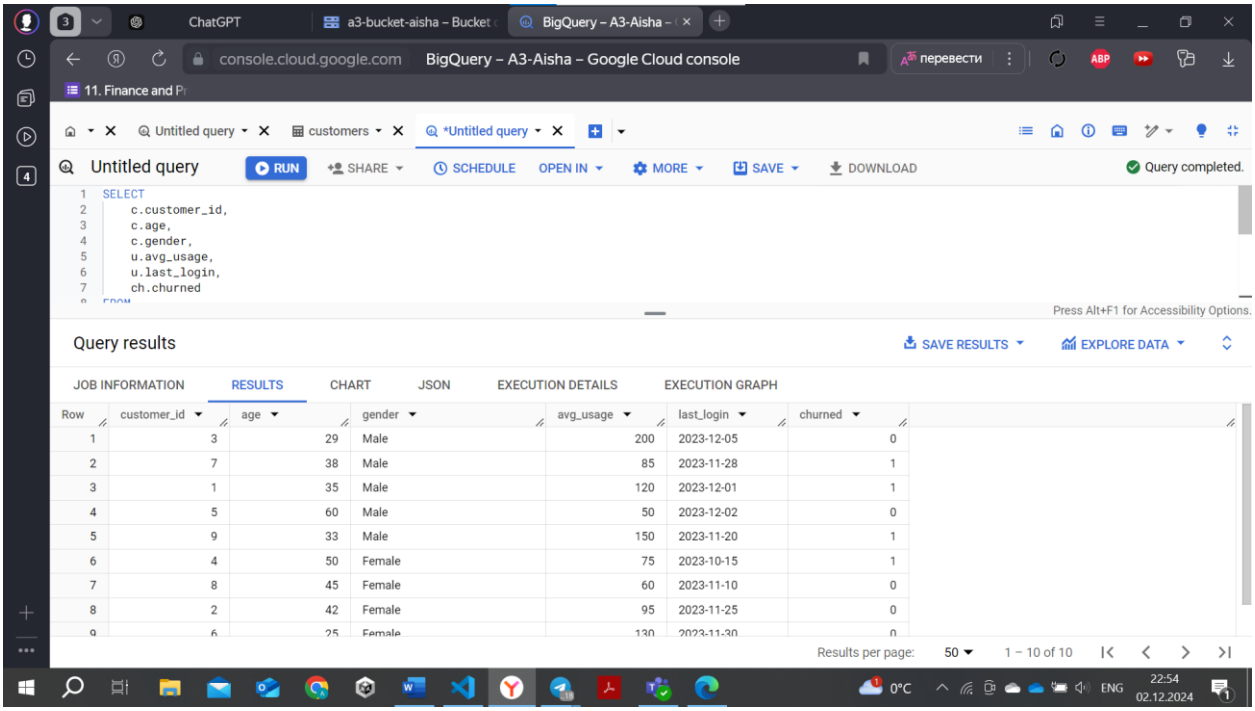




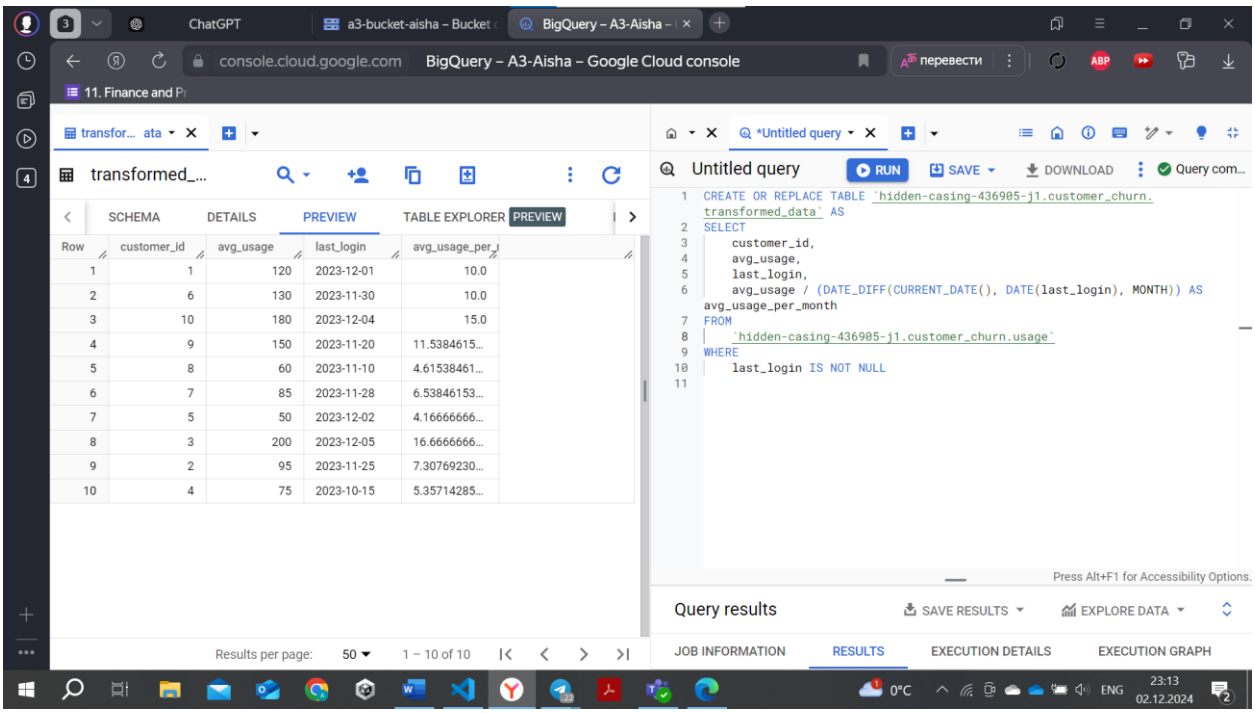
Once the data is loaded into BigQuery, I cleaned and transformed it using SQL queries by joining the customers, usage, and churn tables.



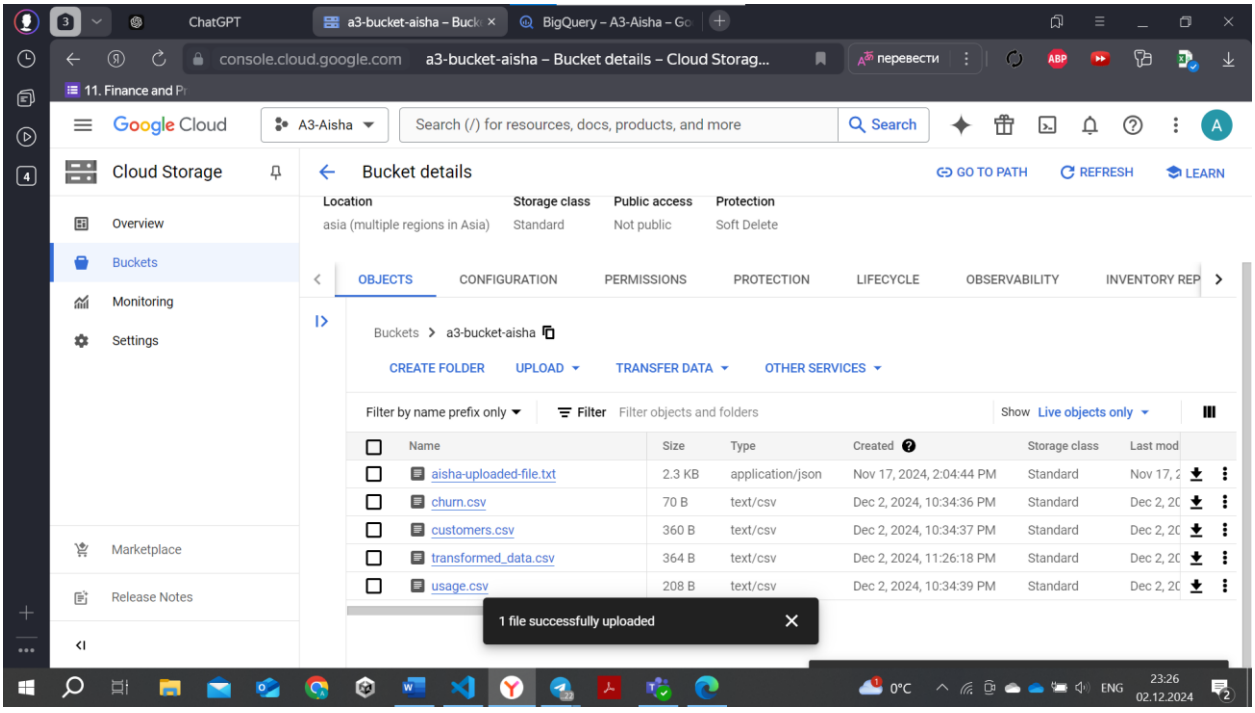
Performed some manipulations.



Create new columns, such as ‘avg_usage’ and ‘avg_usage _per_month’, and saved the data in a new table called ‘transformed_data’.

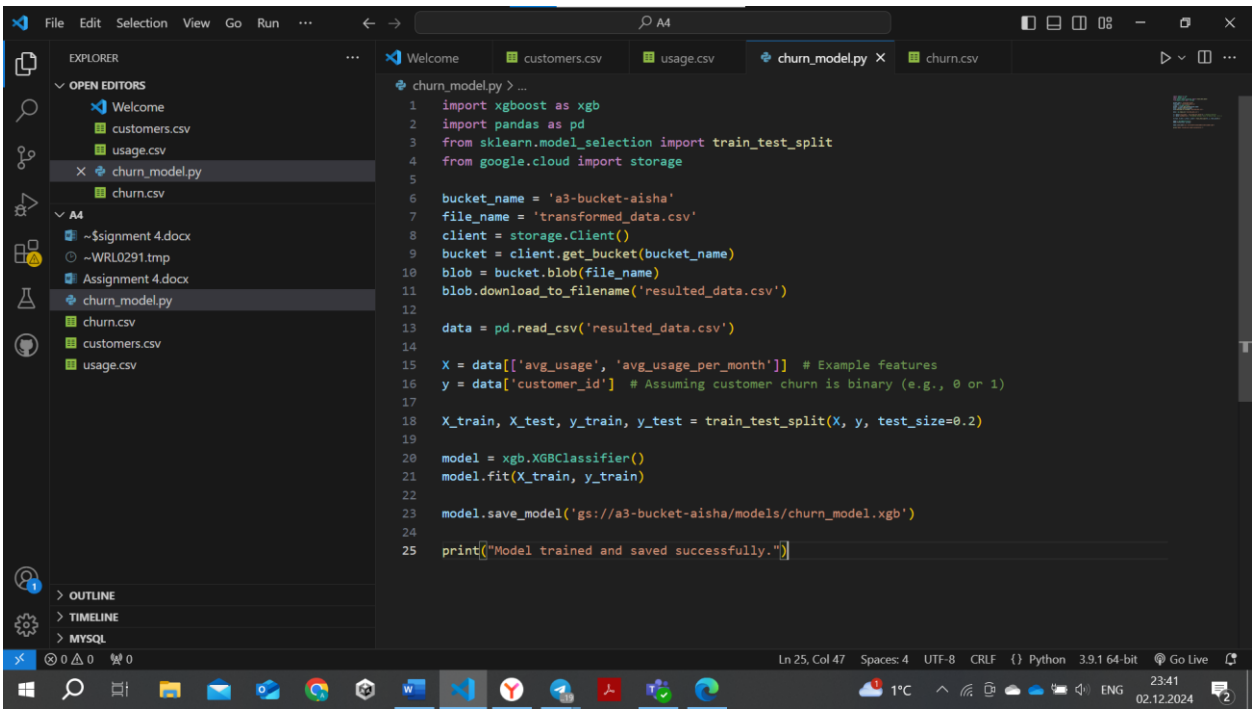


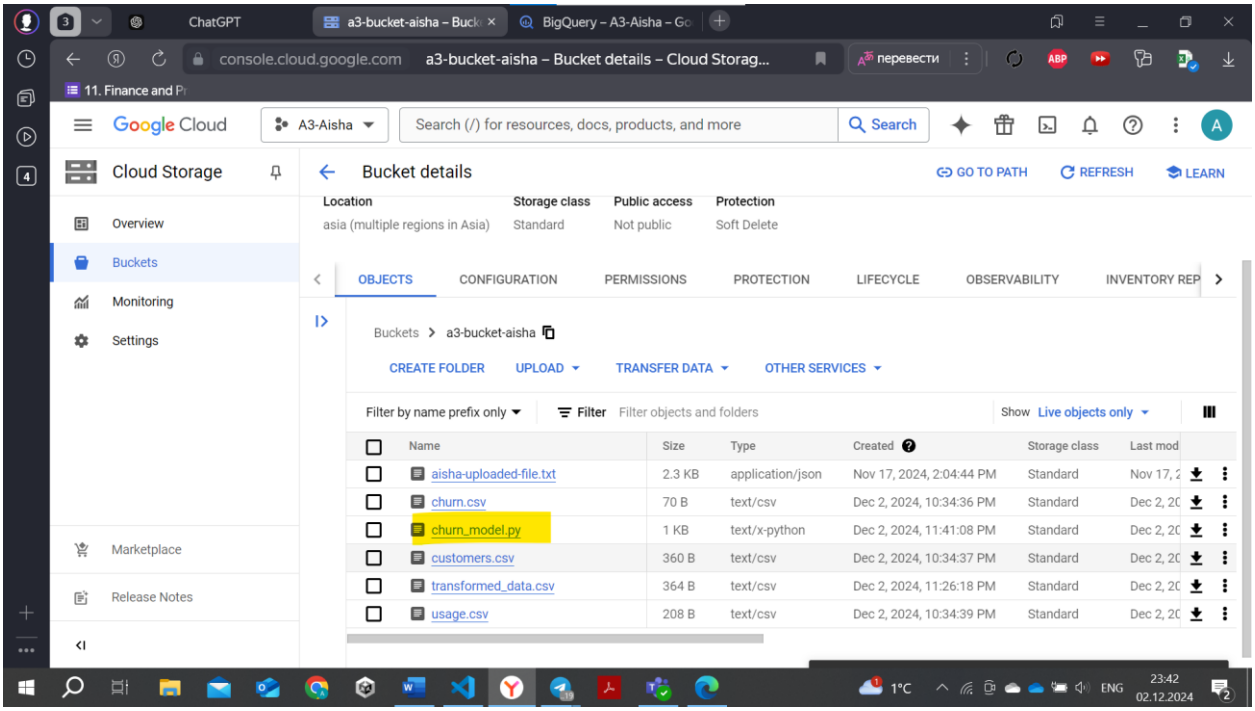
To use this data in AI Platform and train my model, I exported the transformed data back to Cloud Storage:



Model Training.

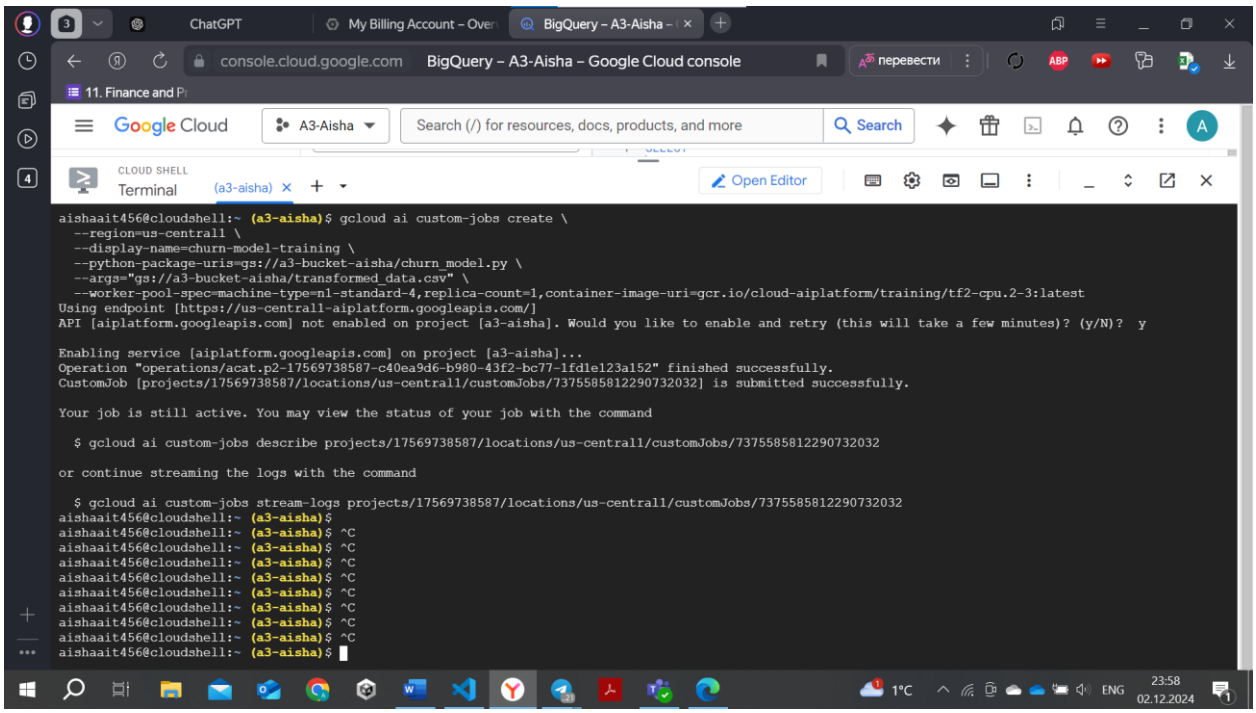
Uploaded a training script that uses Google Cloud Storage as the data source and trains my model.





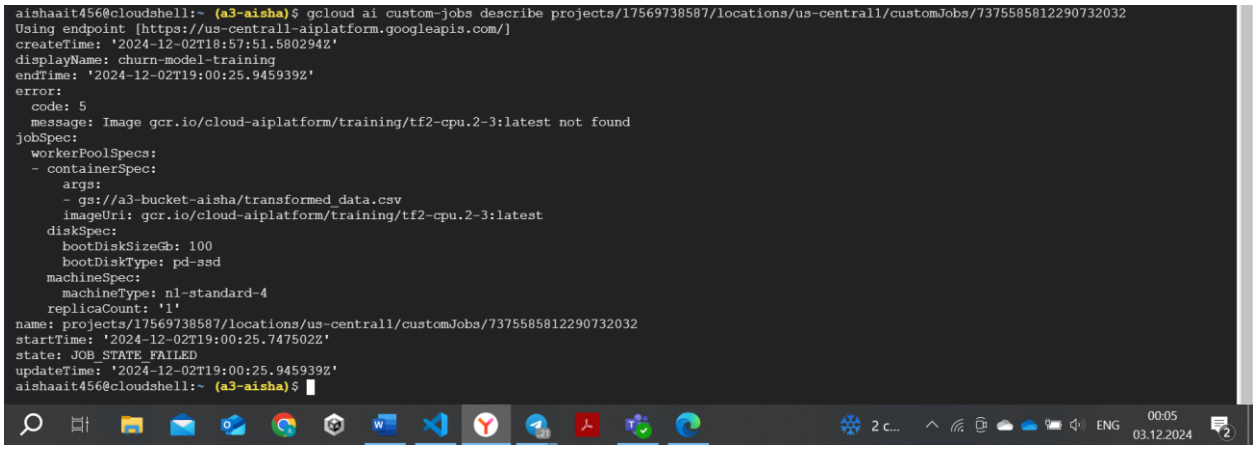
Model Deployment.

My custom job has been successfully submitted and the API service (aiplatform.googleapis.com) has been enabled. Now, I can monitor the status and logs of the job to ensure that everything is running as expected.



Monitoring and Logging.

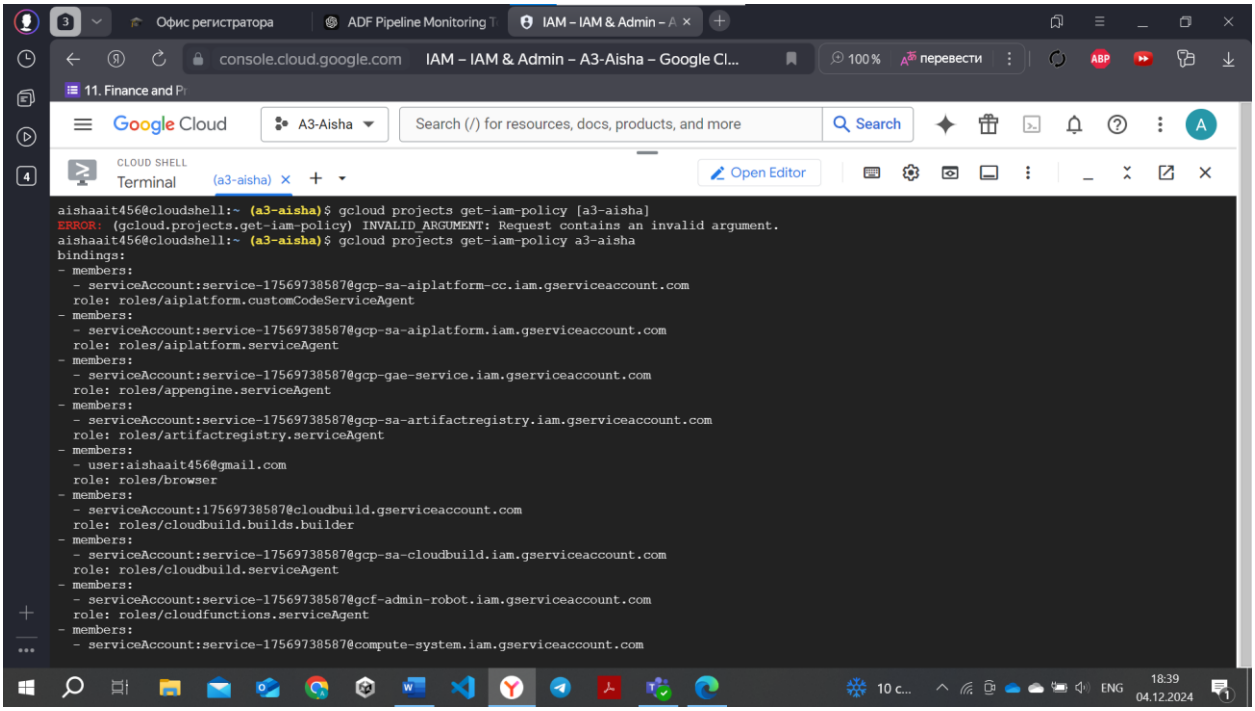
I can check the current status of your custom job using the following command:

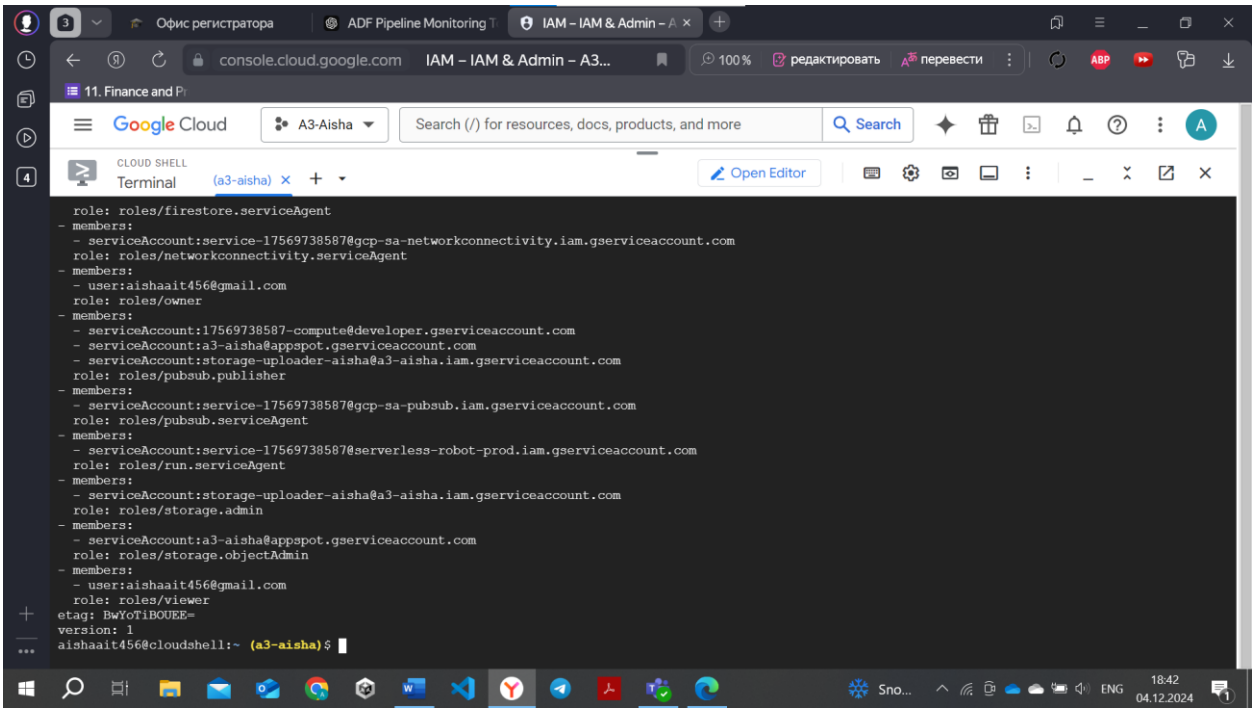


Cloud Security and Compliance

Identity and Access Management (IAM)

Selected my project “a3-aisha” and identified existing iam polcies in it. Here are the IAM roles and permissions configured.





Data Encryption

1. Data at rest:

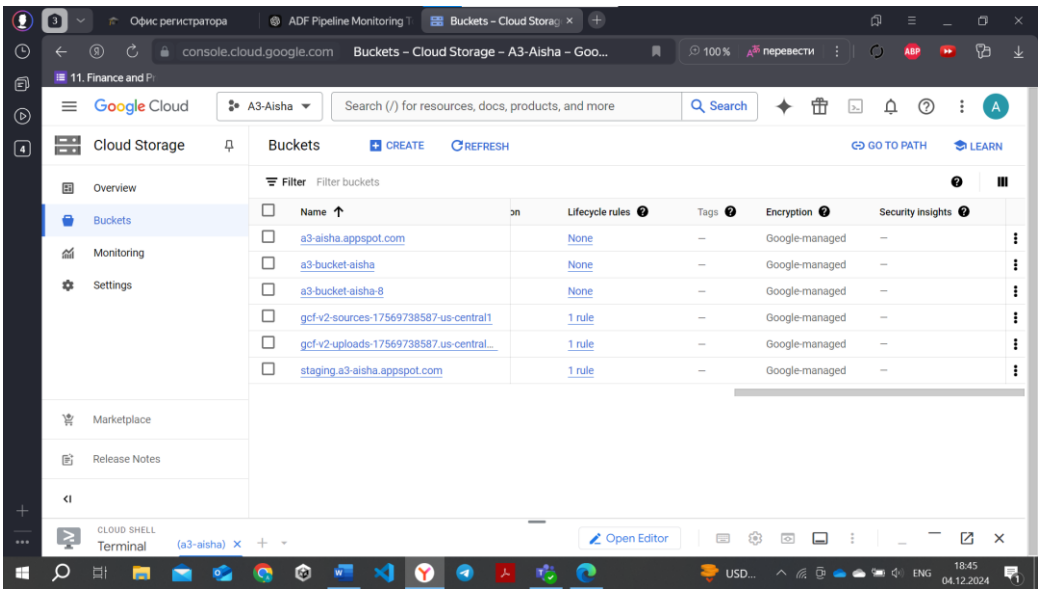
Google Cloud uses Google-managed encryption keys by default to encrypt data at rest, providing a seamless and secure key management process without requiring user intervention.

All data stored in Google Cloud is encrypted automatically using AES-256 or AES-128 encryption standards. Users do not need to configure encryption settings as encryption is enabled by default.

Encryption Keys:

- Google-managed Encryption Keys (Default): Google handles key management, rotation, and security.
- Customer-managed Encryption Keys (CMEK): Customers can use their own encryption keys managed in Google Cloud Key Management Service (Cloud KMS).

Navigated to Cloud Storage Buckets and verified that bucket-level encryption is enabled (Google-managed encryption keys).



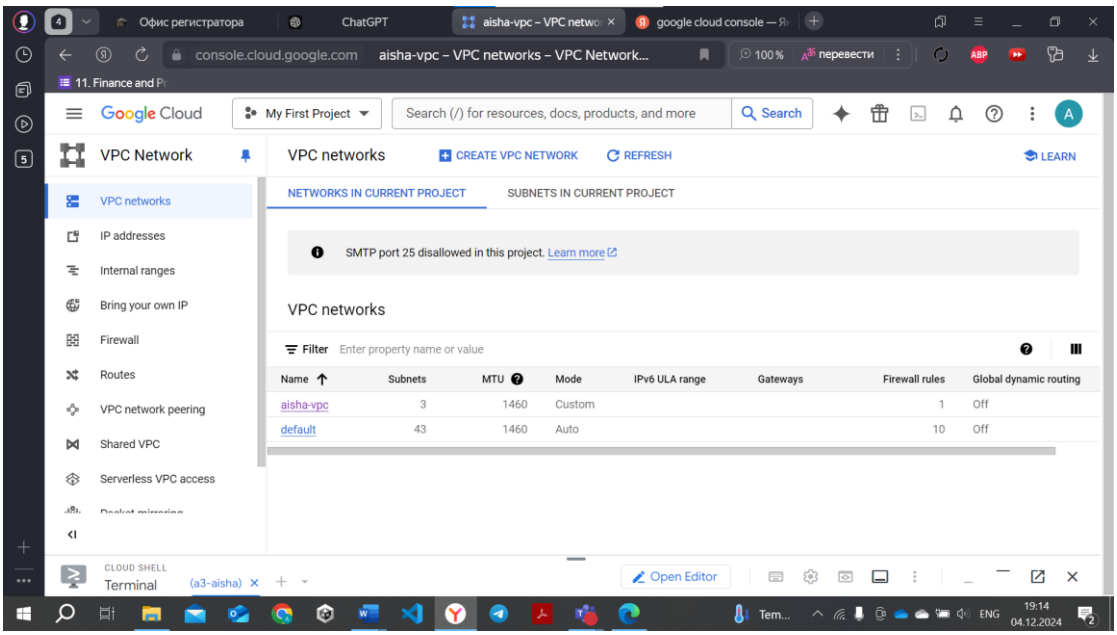
2. Data in transit:

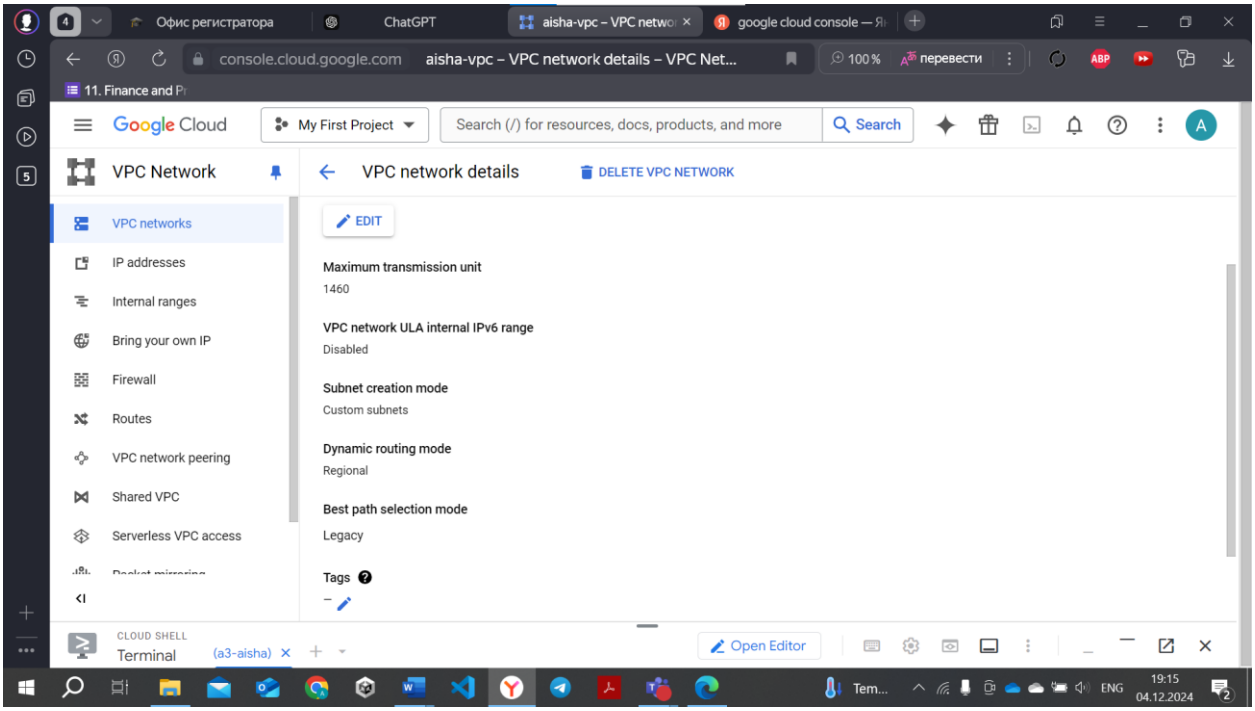
This topic will be covered in the next cahpter (Network Security).

Network Security

VPC Service Controls in Google Cloud secure data traffic by creating a perimeter around resources, ensuring that data stays within a private network. It uses Private Google Access to allow VPC resources to securely access Google services over the internal network without needing public IPs, preventing exposure to the internet.

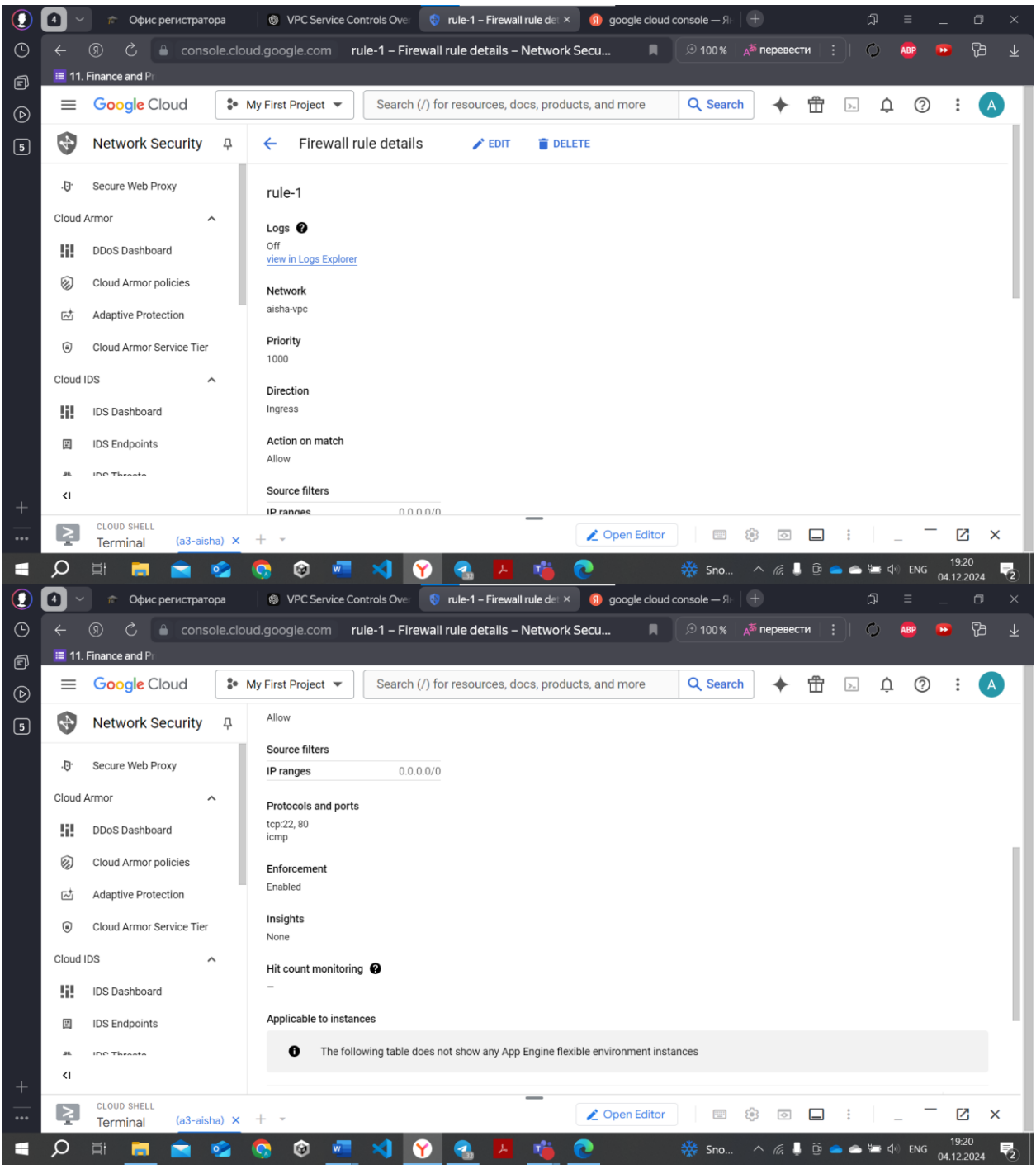
Below is vpc that I configured for my “My First Project”.





A firewall in Google Cloud controls incoming and outgoing traffic to resources based on specified security rules. It filters traffic based on factors like IP address, protocol, and port, helping to protect VPCs and prevent unauthorized access. Firewalls can be configured at the network or instance level.

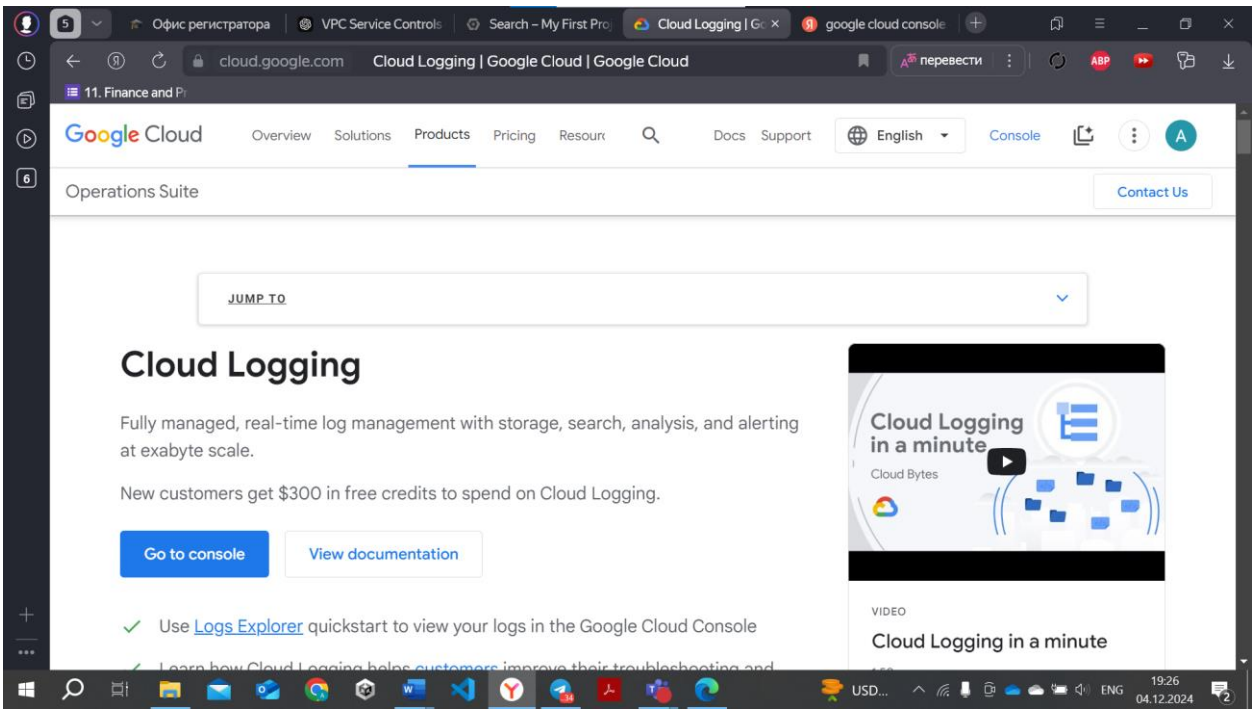
Below is the firewall I configured for my vpc:



Audit Logging

Audit logging in Google Cloud is enabled through Cloud Logging settings, where you can select log types such as Admin Activity, Data Access, and System Events. Logs can be filtered by specific resources and services. For review, logs can be accessed in Cloud Logging or exported to Cloud Storage, BigQuery, or Pub/Sub. The review process involves analyzing logs for suspicious activity, setting up notifications for unusual events, and regularly checking for compliance with internal and external regulations.

Unfortunately, I cannot demostrate it since it requires payment:



Compliance Standards

Google Cloud adheres to various industry standards and regulations such as GDPR, HIPAA, SOC 2, ISO 27001, and others. These standards ensure that cloud services meet security, privacy, and data protection requirements.

Measures Taken to Meet Compliance:

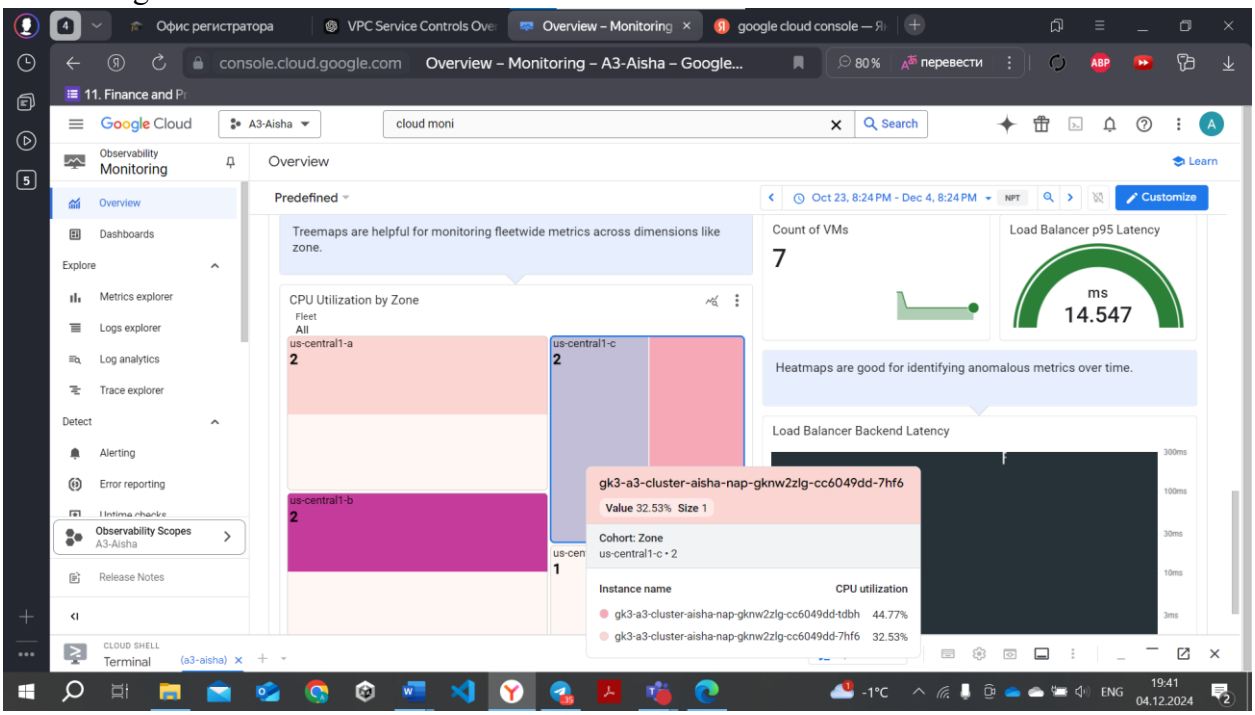
- **Data Encryption**: Google Cloud encrypts data at rest and in transit by default.
- **Identity and Access Management (IAM)**: Controls who can access resources with role-based access and multi-factor authentication.
- **Audit Logging**: Tracks and records actions for transparency and accountability (via Cloud Logging).
- **Security Controls**: Includes firewalls, VPC Service Controls, and DDoS protection to safeguard infrastructure.
- **Regular Audits and Assessments**: Google undergoes third-party audits to verify compliance with standards.
- **Compliance Documentation**: Google Cloud provides detailed documentation to help users meet compliance needs, including certifications and audit reports.

Incident Response Planning:

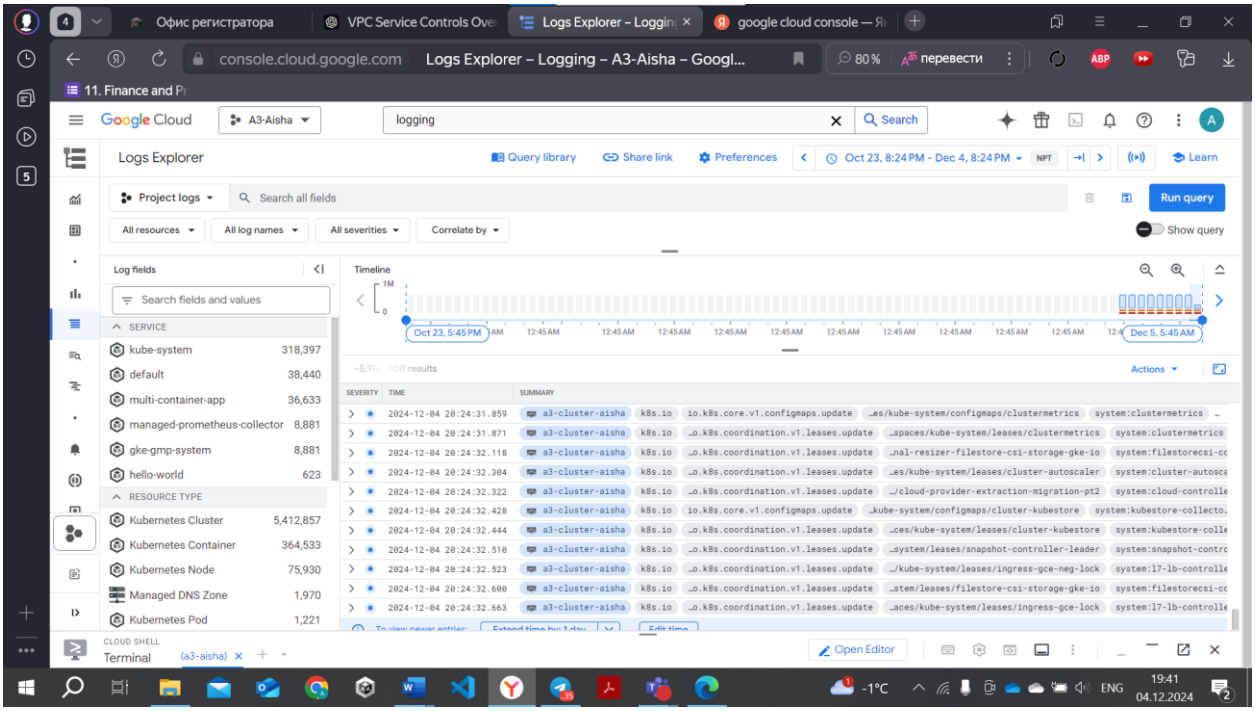
The incident response plan outlines procedures for identifying, responding to, and mitigating security incidents. It includes roles and responsibilities, communication protocols, and remediation steps. Simulation results test the effectiveness of the plan, identifying strengths and areas for improvement in handling potential incidents.

Google Cloud's incident response plan involves using tools like Cloud Monitoring, Cloud Logging, and Security Command Center to detect, respond to, and recover from security incidents. The plan defines roles and workflows for responding to incidents, with automated response actions through Cloud Functions or Cloud Run. Simulation results, conducted through Cloud Security Command Center or custom exercises, help evaluate the plan's effectiveness and identify areas for improvement in real-time threat detection, response, and recovery.

Monitoring:



Logging:



Conclusion

The implementation of Big Data, Machine Learning, and security practices in Google Cloud demonstrates a strong infrastructure for data processing, predictive modeling, and robust security. Google Cloud provides advanced tools like BigQuery, AI Platform, and Cloud Security Command Center, enabling scalable data analysis, seamless machine learning deployment, and comprehensive security management. However, the report highlights areas for improvement in resource optimization, data governance, and compliance monitoring.

Please, put 100%.

Recommendations

- Data Processing: Implement automated data pipelines using Dataflow for improved efficiency in handling large datasets.
- Machine Learning: Enhance machine learning workflows with AutoML for faster model development and experimentation.
- Security: Strengthen security by utilizing VPC Service Controls to restrict data movement and implementing stricter IAM roles.
- Compliance: Regularly update compliance frameworks and conduct periodic audits using Cloud Audit Logs and Security Command Center to ensure adherence to regulatory standards.

References

- Google Cloud documentation: Google Cloud BigQuery, AI Platform, Security Command Center.
- Relevant academic papers and articles on Big Data, Machine Learning, and Cloud Security.