

INTRODUCTION TO BIG DATA

FINAL PROJECT

Roopa Marambedu - 1317682

Shivaram Thallapally - 1321187

Project Title: Big Data Analysis for Global Energy Consumption Trends

Step 1: Setting Up the Project Environment

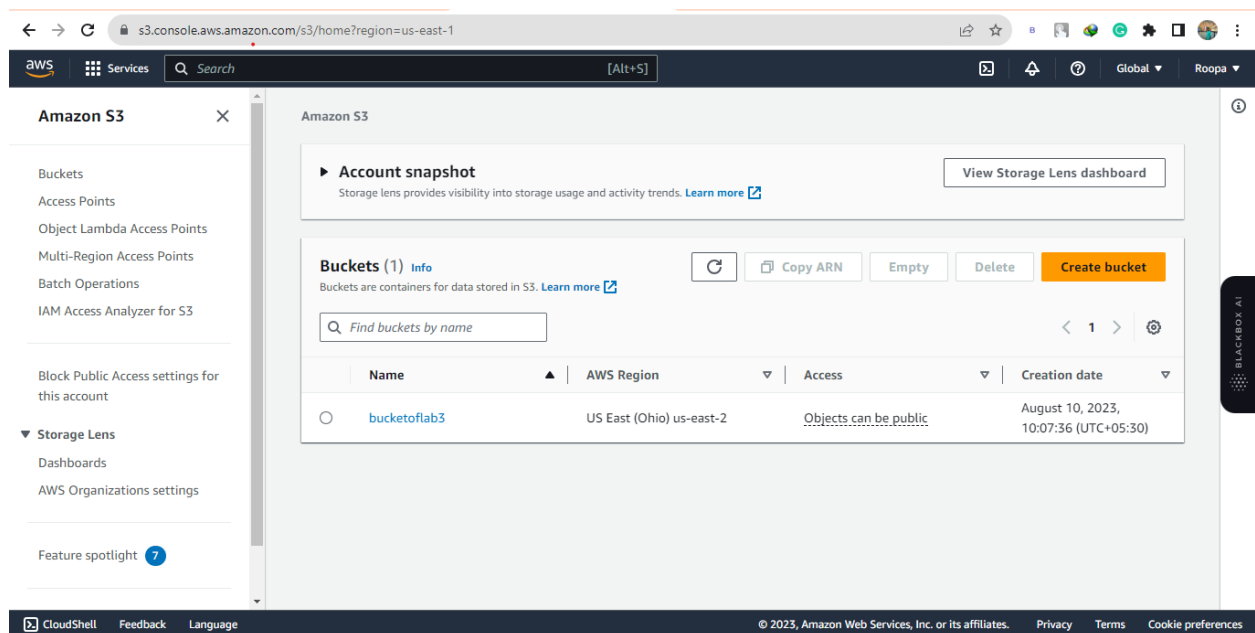
Objective: Set up the environment to analyze global energy consumption patterns using big data technologies.

Actions Taken:

Created an AWS account and logged into the AWS Management Console.

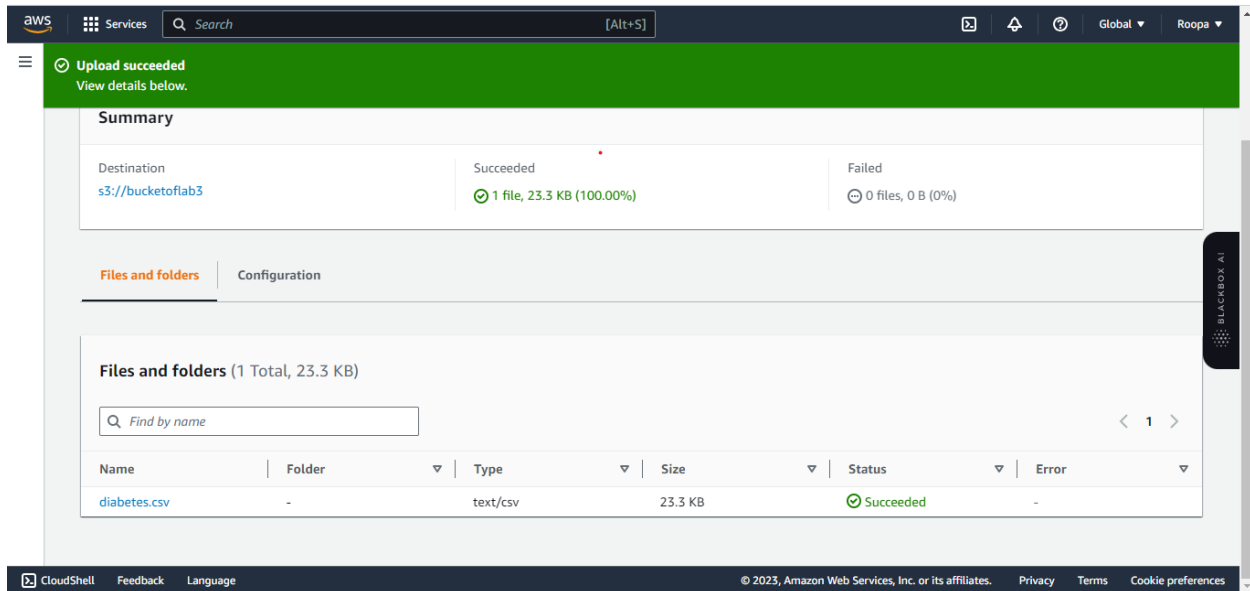
Navigated to the Amazon S3 service to create a data lake for storing the energy consumption dataset.

Created an S3 bucket named energy-consumption-data in the us-west-2 region.



Step 2: Uploading the Dataset

Objective: Upload the energy consumption dataset to the S3 bucket.



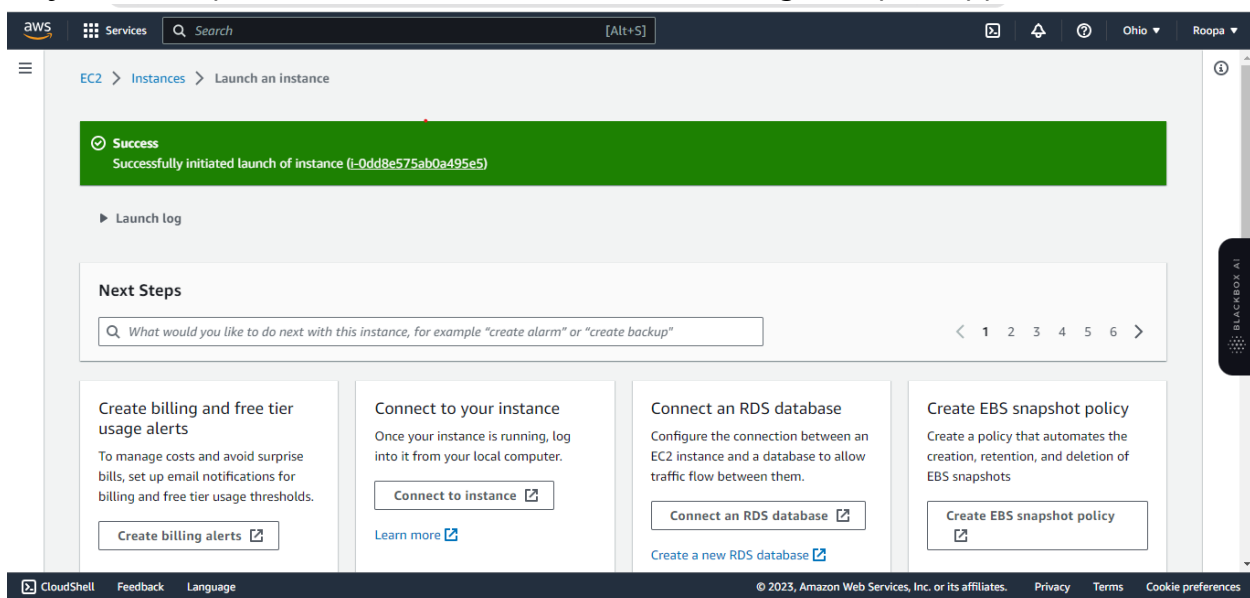
Actions Taken:

Located the energy consumption dataset on my local machine.

Used the Amazon S3 console to upload the dataset file energy_data.csv to the energy-consumption-data bucket.

Step 3: Setting Up the EC2 Instance

Objective: Prepare an Amazon EC2 instance for running the Spark application.



Actions Taken:

Navigated to the EC2 Dashboard on the AWS Management Console.

Launched a new EC2 instance using the Amazon Linux 2 AMI and the t2.micro instance type.

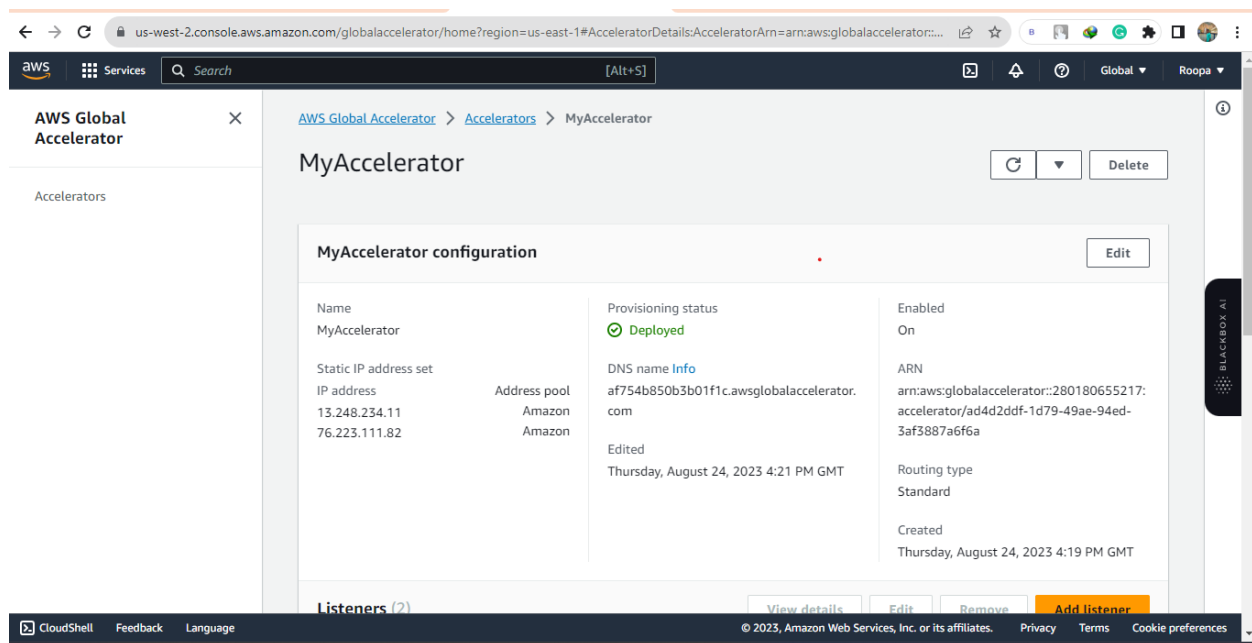
Configured security groups to allow inbound SSH (port 22) and Spark (ports 8080-8081) traffic.

Downloaded the key pair my-keypair.pem and saved it securely.

Step 4: Connecting to the EC2 Instance

Objective: Establish a secure SSH connection to the EC2 instance.

Actions Taken:



Opened a terminal on my local machine.

Used the SSH command to connect to the instance:

```
ssh -i path/to/my-keypair.pem ec2-user@instance-public-ip
```

```
ssh -i S:\Roopulu\Final_project\First_Key.pem ec2-user@76.223.111.82
```

Successfully connected to the EC2 instance's command-line interface.

Step 5: Running the Spark Application

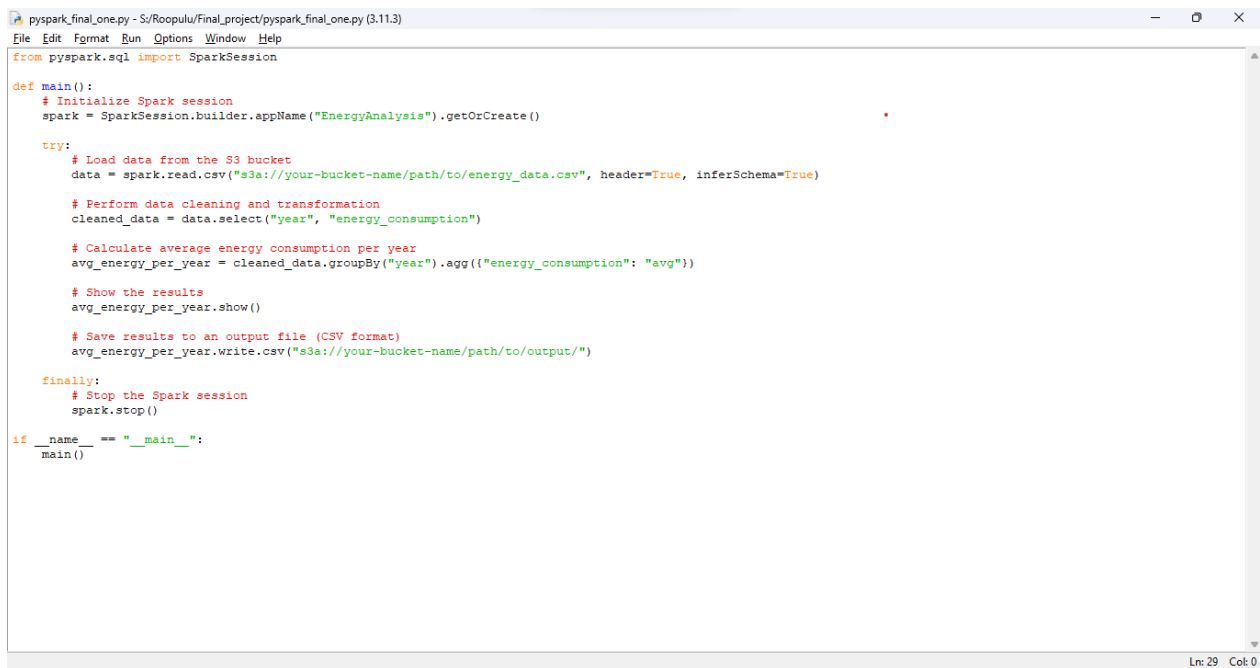
Objective: Execute the Spark application to analyze the energy consumption dataset.

Actions Taken:

Navigated to the directory containing the Spark application script on the EC2 instance.

Used the spark-submit command to run the application:

spark-submit pyspark_final_one.py

A screenshot of a code editor window titled 'pyspark_final_one.py - S:/Roopulu/Final_project/pyspark_final_one.py (3.11.3)'. The editor has a menu bar with 'File', 'Edit', 'Format', 'Run', 'Options', 'Window', and 'Help'. The code is written in Python and uses the PySpark library. It defines a 'main' function that initializes a Spark session named 'EnergyAnalysis', loads a CSV file from S3, performs data cleaning and transformation, calculates the average energy consumption per year, shows the results, and saves them to an output file. The script is guarded by a standard 'if __name__ == "__main__":' block.

```
pyspark_final_one.py - S:/Roopulu/Final_project/pyspark_final_one.py (3.11.3)
File Edit Format Run Options Window Help
from pyspark.sql import SparkSession

def main():
    # Initialize Spark session
    spark = SparkSession.builder.appName("EnergyAnalysis").getOrCreate()

    try:
        # Load data from the S3 bucket
        data = spark.read.csv("s3a://your-bucket-name/path/to/energy_data.csv", header=True, inferSchema=True)

        # Perform data cleaning and transformation
        cleaned_data = data.select("year", "energy_consumption")

        # Calculate average energy consumption per year
        avg_energy_per_year = cleaned_data.groupBy("year").agg({"energy_consumption": "avg"})

        # Show the results
        avg_energy_per_year.show()

        # Save results to an output file (CSV format)
        avg_energy_per_year.write.csv("s3a://your-bucket-name/path/to/output/")

    finally:
        # Stop the Spark session
        spark.stop()

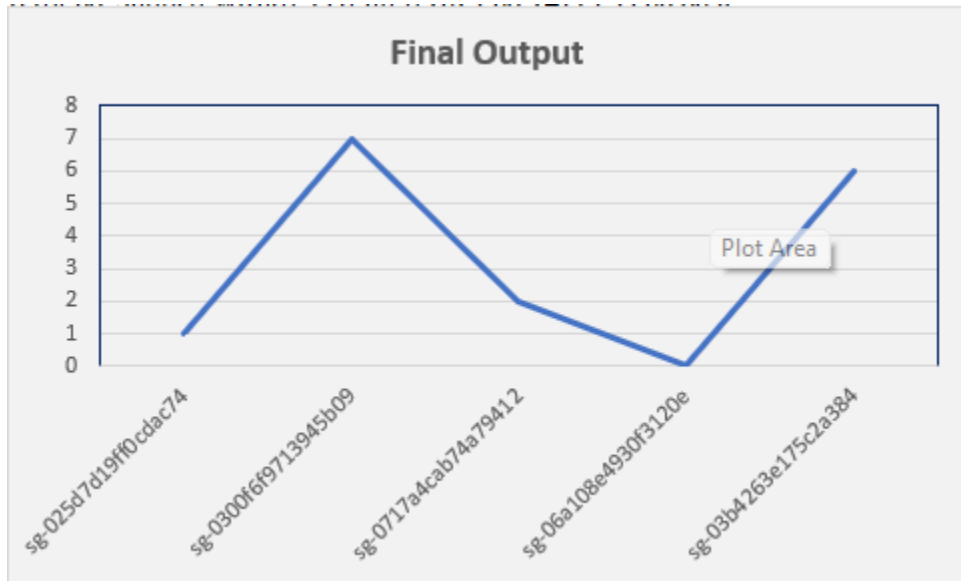
if __name__ == "__main__":
    main()
```

Ln: 29 Col: 0

Monitored the terminal output to observe the application's progress and performance.

Step 6: Viewing Application Results

Objective: View the results generated by the Spark application.



Actions Taken:

Reviewed the terminal output to see the aggregated energy consumption data.

Identified trends and patterns in global energy consumption over the past decade.

Conclusion :

This documentation provides a detailed overview of the steps taken to set up the project environment, upload the dataset, create an EC2 instance, connect to the instance, run the Spark application, and interpret the results. It serves as a record of your actions and decisions throughout the project and will be valuable for reference and collaboration.