**Streaming Data Processing**

**By**

**Karthikeyan K R**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page No.** |
| 1 | Project Statement | 3 |
| 2 | Project Overview | 3 |
| 3 | Project Requirements | 3 |
| 3.1 | Azure Subscription | 3 |
| 3.2 | Data Sources | 3 |
| 3.3 | Azure Event Hub | 3 |
| 3.4 | Azure Stream Analytics Job | 3 |
| 3.5 | Azure Data Lake Storage | 3 |
| 3.6 | Azure Data Factory | 3 |
| 3.7 | Azure Databricks | 3 |
| 4 | Architecture Diagram | 4 |
| 5 | Execution Overview | 4 |
| 6 | Project Implementation | 5 |
| 6.1 | Creating a Python application | 5 |
| 6.2 | Azure Event Hub Namespace | 6 |
| 6.3 | Azure Stream Analytics Job | 9 |
| 6.4 | Azure Data Lake Storage Gen 2 | 15 |
| 6.5 | Azure Data Factory Configuration | 16 |
| 6.6 | Azure Databricks Configuration | 18 |
| 7 | Conclusion | 20 |

**Streaming Data Processing**

1. **Project Statement**:

Set up an end-to-end pipeline with Azure Data Factory for ingesting streaming data and Azure Databricks for real-time processing and analysis of the streaming data.

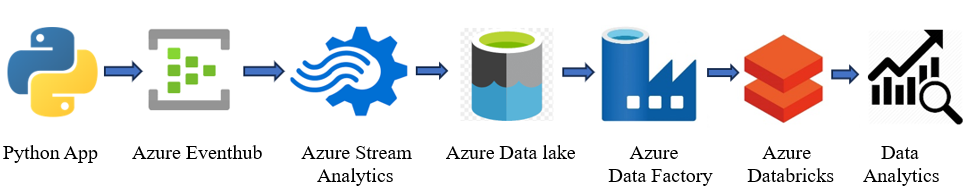
1. **Project Overview**:

The primary goal of this project is to design and implement a robust end-to-end data pipeline using Azure Data Factory (ADF) and Azure Databricks to ingest, process, and analyse streaming data in real-time. The pipeline will enable the extraction of valuable insights from the streaming data source, providing a foundation for timely decision-making and actionable intelligence.

1. **Project Requirements**:
   1. Azure Subscription

* You’re required to have an Azure Subscription to perform this project
  1. Data Source
* Created a python application, which fetches the weather data from the google website
  1. Azure Event Hub
* Azure Event Hubs is a cloud-based messaging service provided by Microsoft Azure which enables large-scale, real-time event streaming for ingesting and processing data from various sources.
  1. Azure Stream Analytics
* Azure Stream Analytics is a cloud-based service provided by Microsoft Azure that helps you analyse and gain insights from real-time data streams.
  1. Azure Data Lake
* Azure Data Lake is like a massive storage space in the cloud where you can store and manage enormous amounts of data. It's designed to handle both large files and lots of small files
  1. Azure Data Factory
* Azure Data Factory is like a digital factory that helps you collect, transform, and move your data from one place to another in a smooth and organized way.
  1. Azure Databricks
* Azure Databricks is like a super-smart workspace in the cloud where you can easily analyse and process large amounts of data using the power of Apache Spark.

1. **Architecture Diagram**:

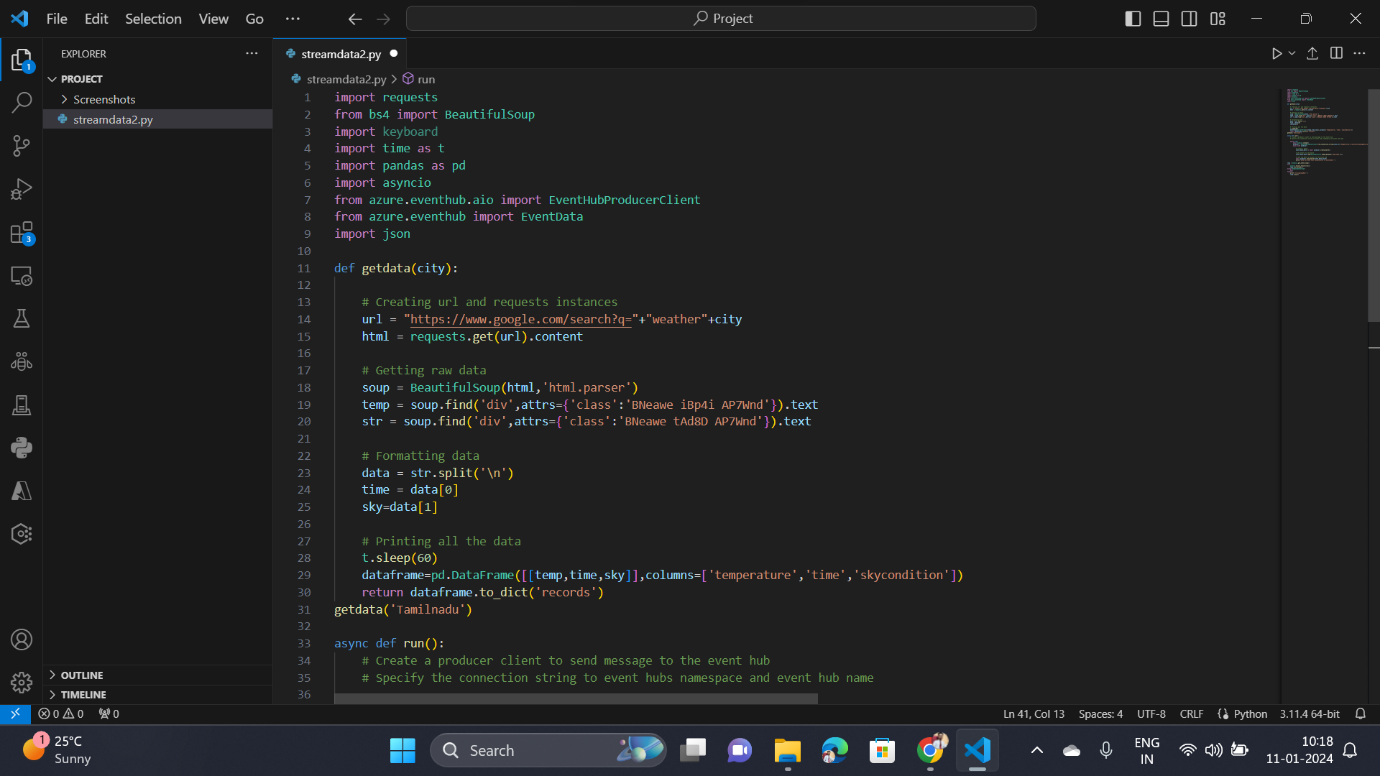


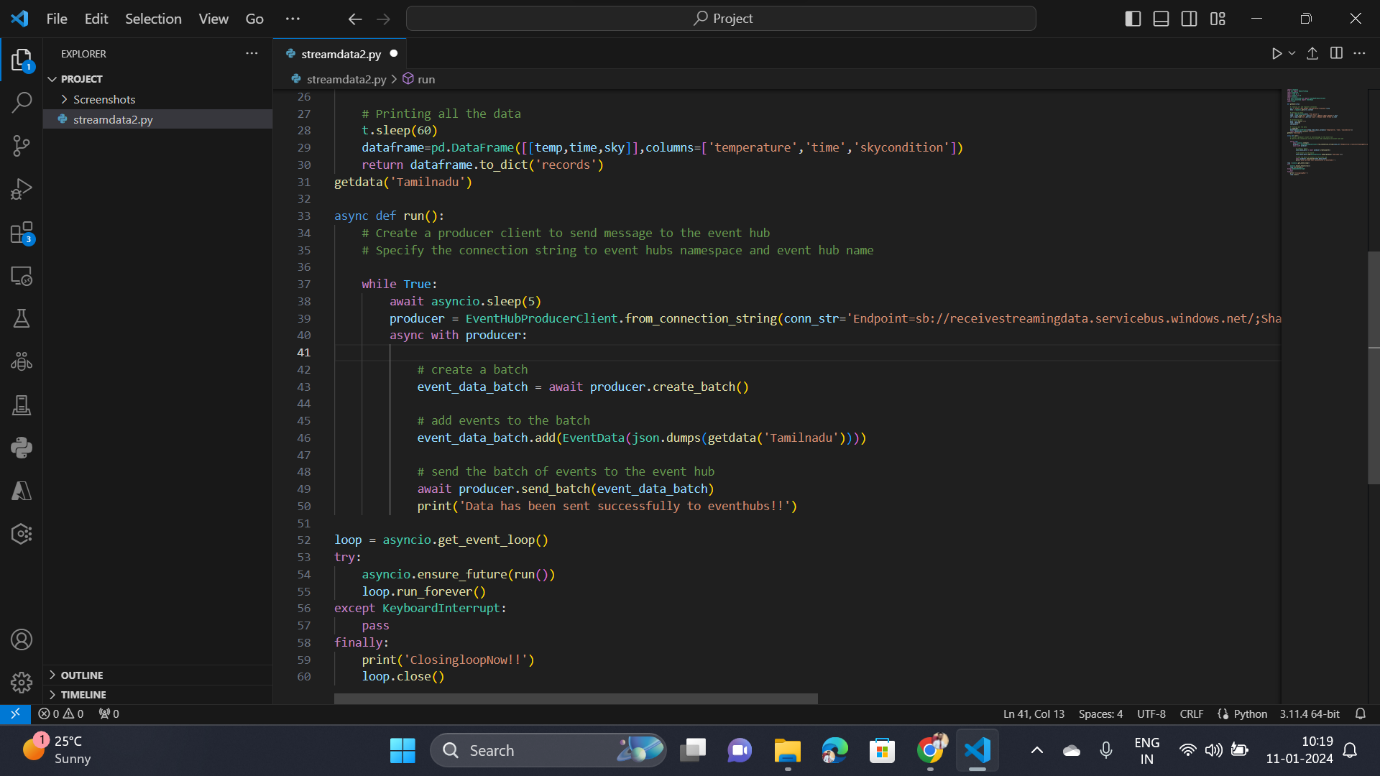
1. **Execution Overview**:

* Create a python application which fetch a stream data from the webpages
* Connect the Python application with the Azure Event Hub and receive the streaming data
* Use Azure Stream Analytics to process the streaming data for storage
* Create an Azure Data Lake Storage Gen2 to store the received date which is in **JSON** format
* Use Azure Data Factory to convert the **JSON** file to **CSV** file
* Connect the Azure Databricks with Azure Data Factory and load the **CSV** file or download and upload the **CSV** file into Azure Databricks
* Use Spark commands to perform data analysis on the **CSV** file

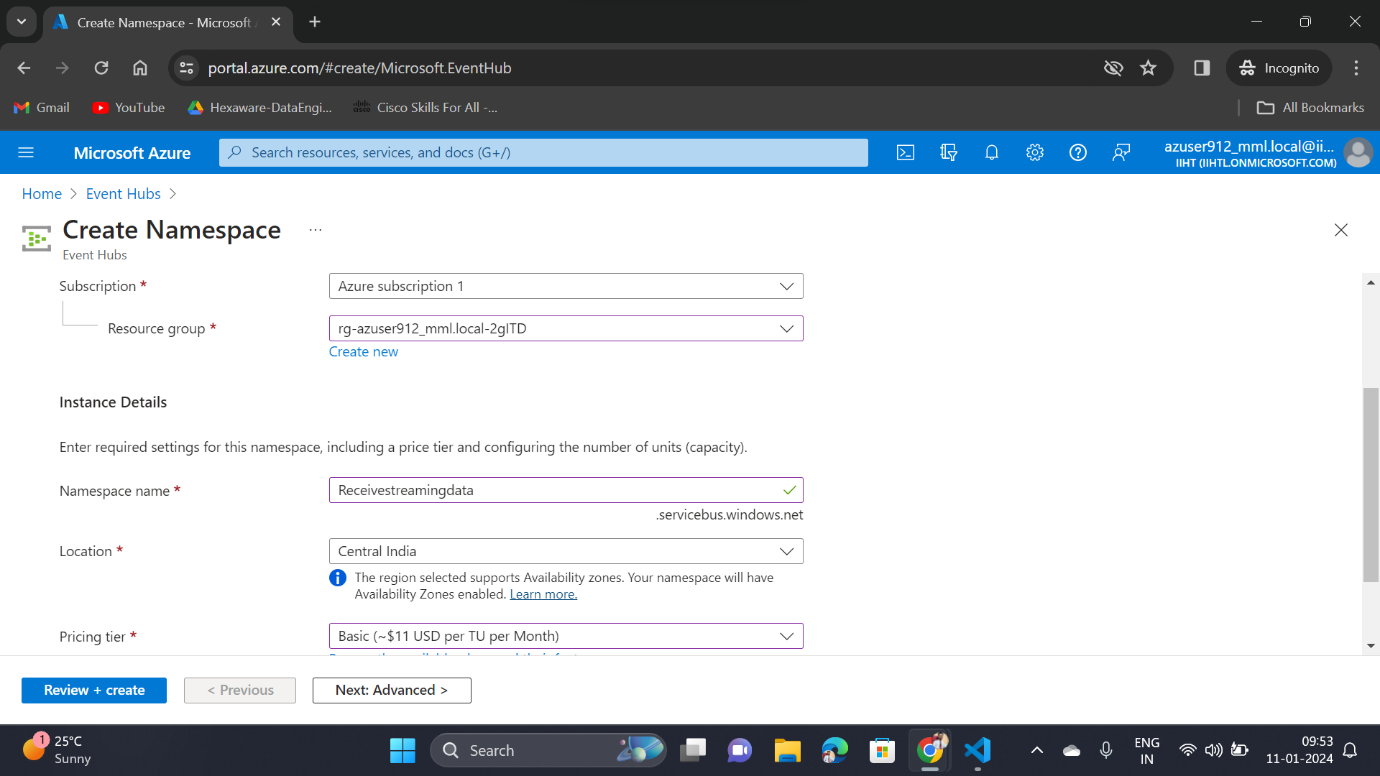
1. **Project Implementation**

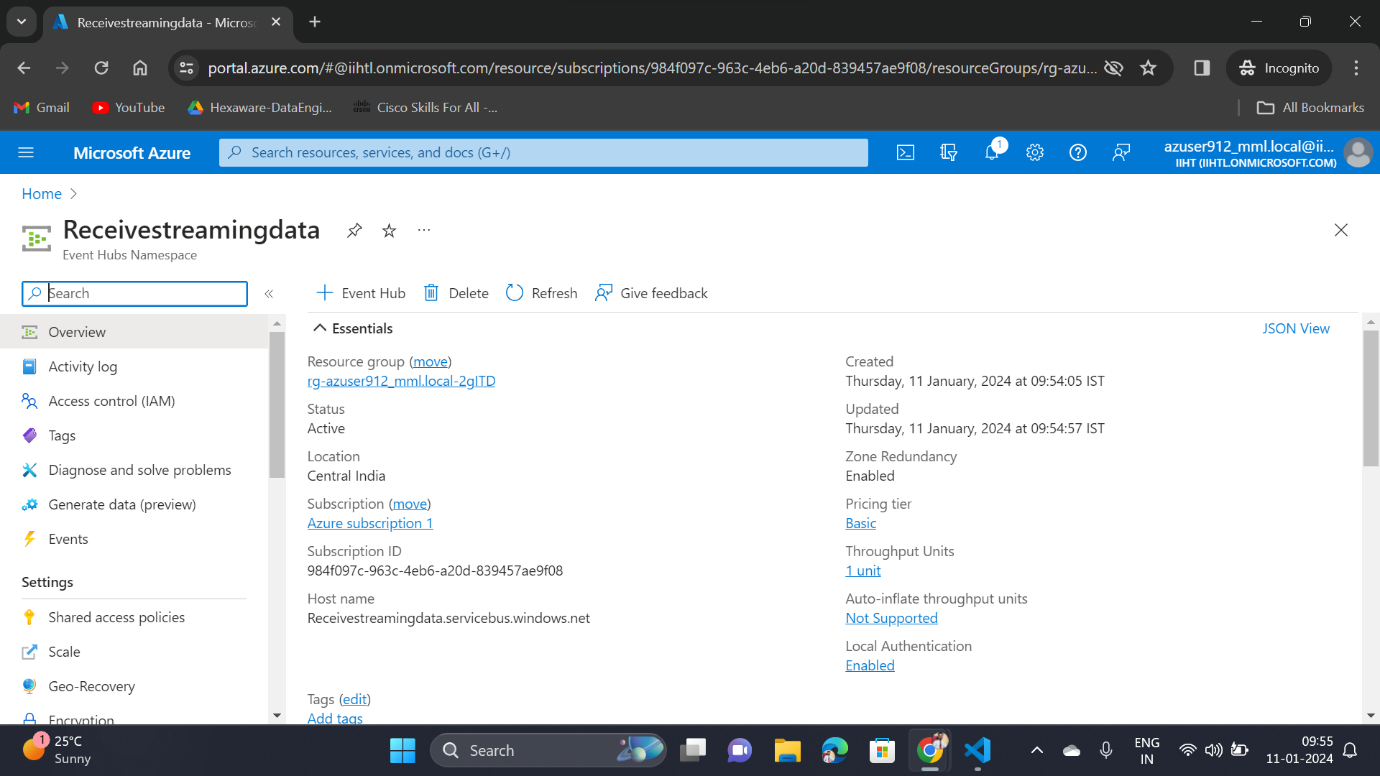
6.1) Creating a Python application

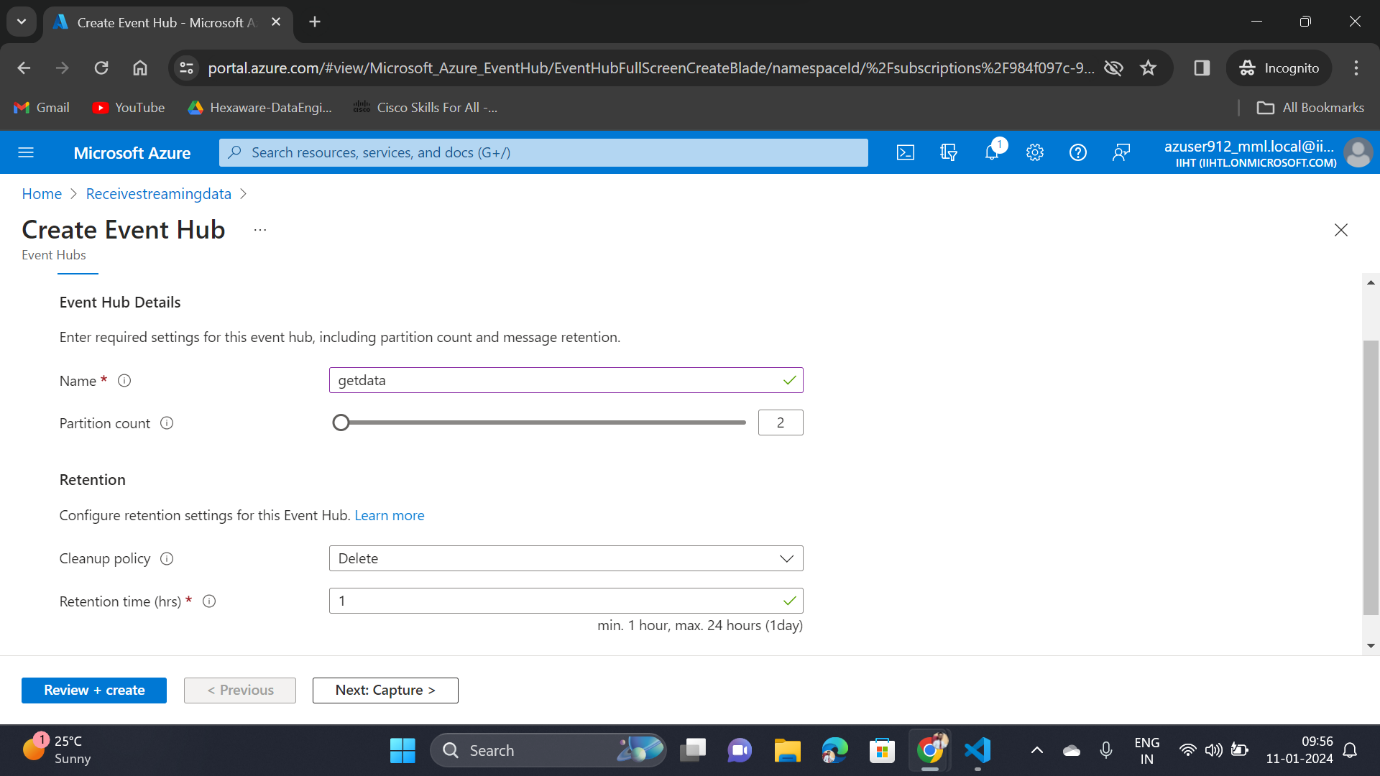




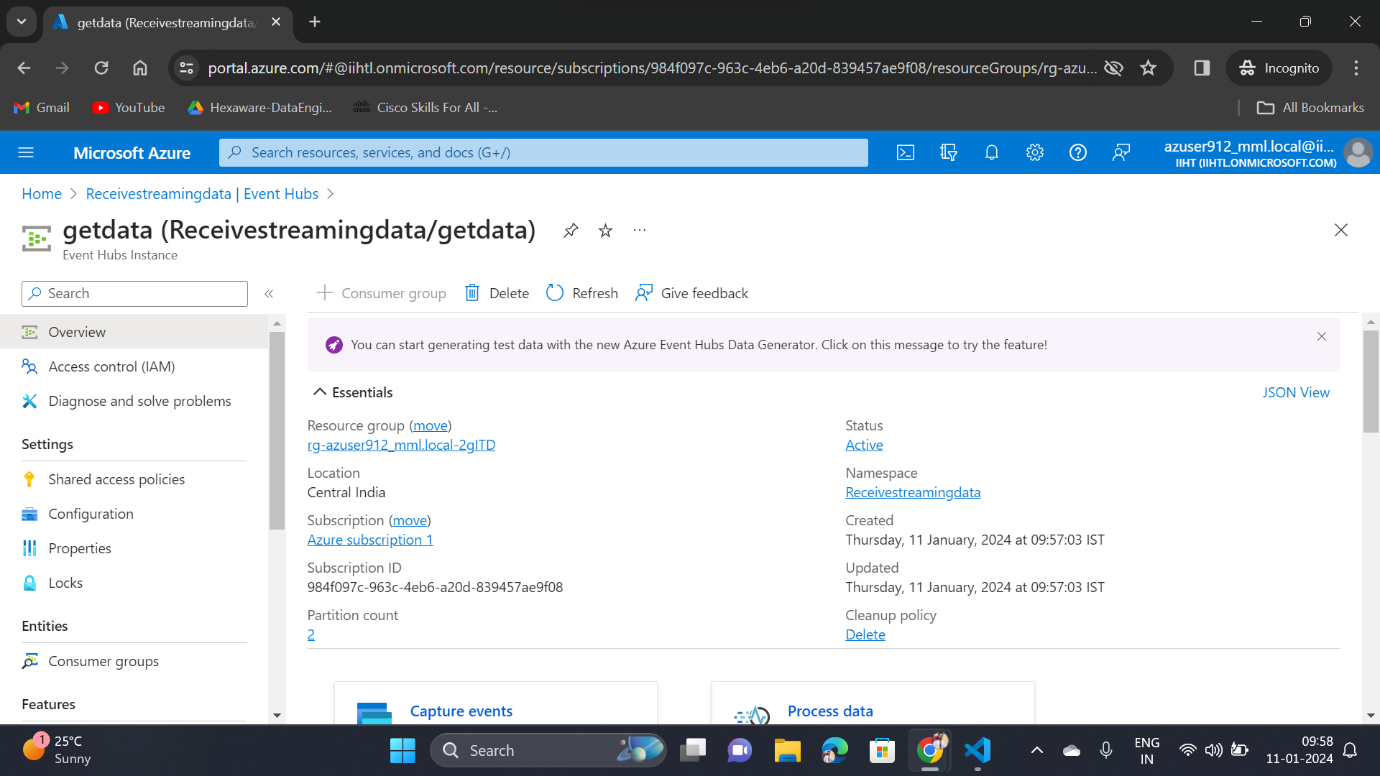
6.2) Creating an Event Hub Namespace



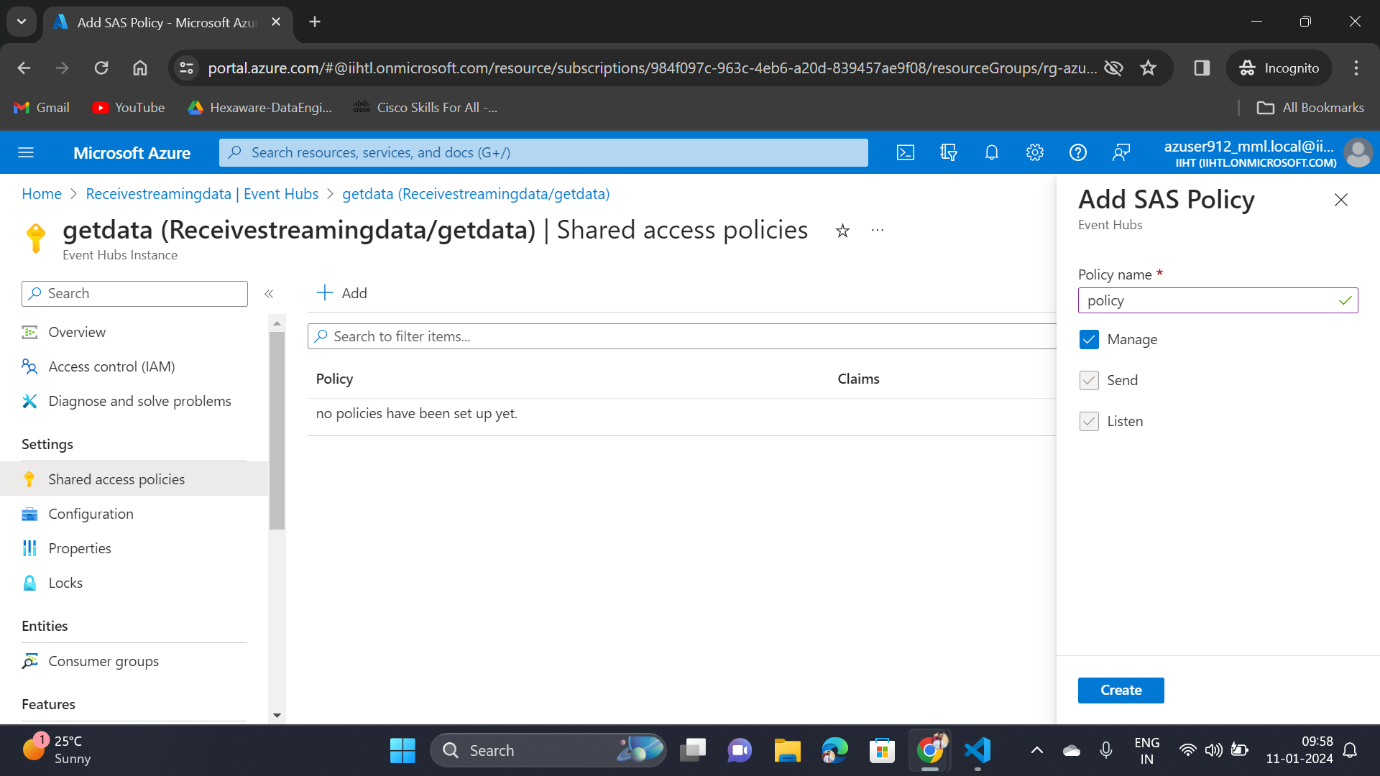
 After the successful creation of namespace,

 Creating an Event Hub to receive the streaming data

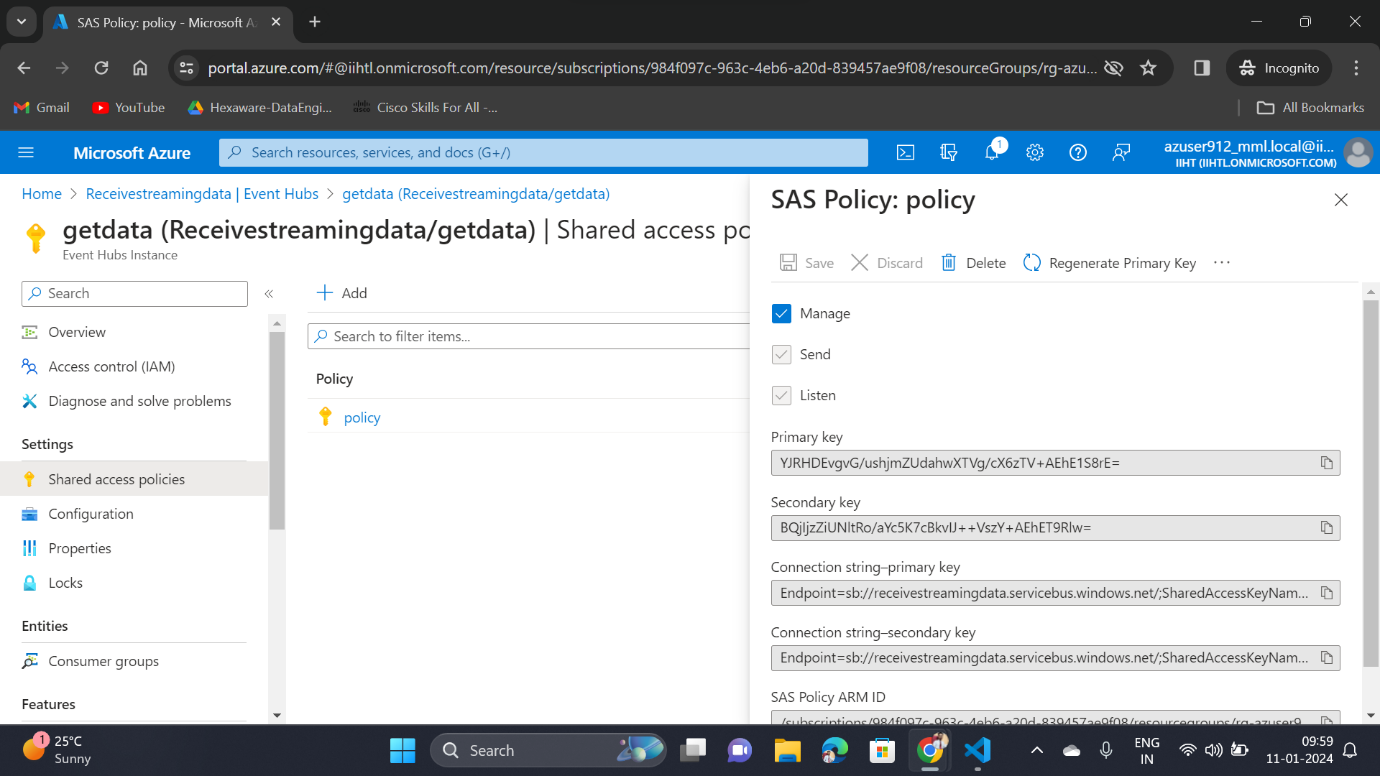
After the successful creation of Event Hub



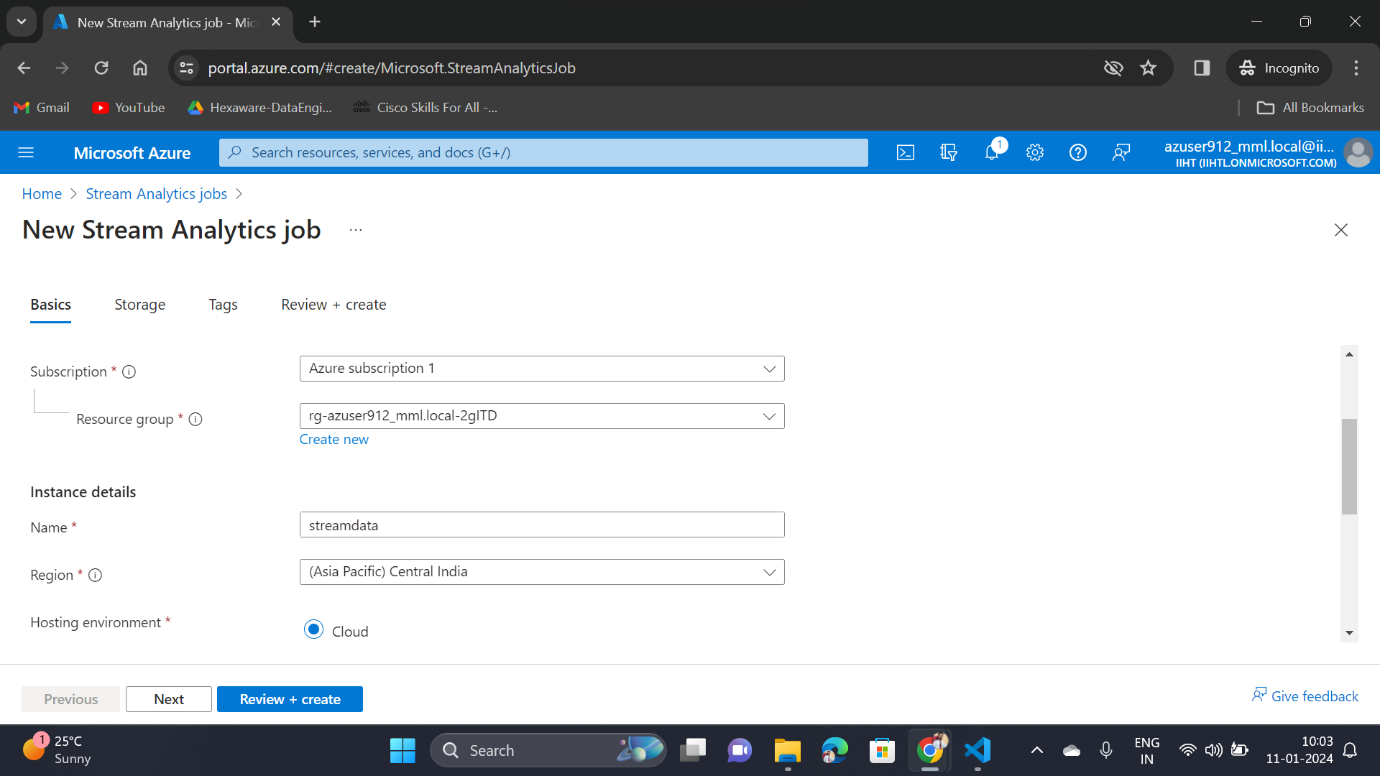
Creating a policy in the created Event Hub



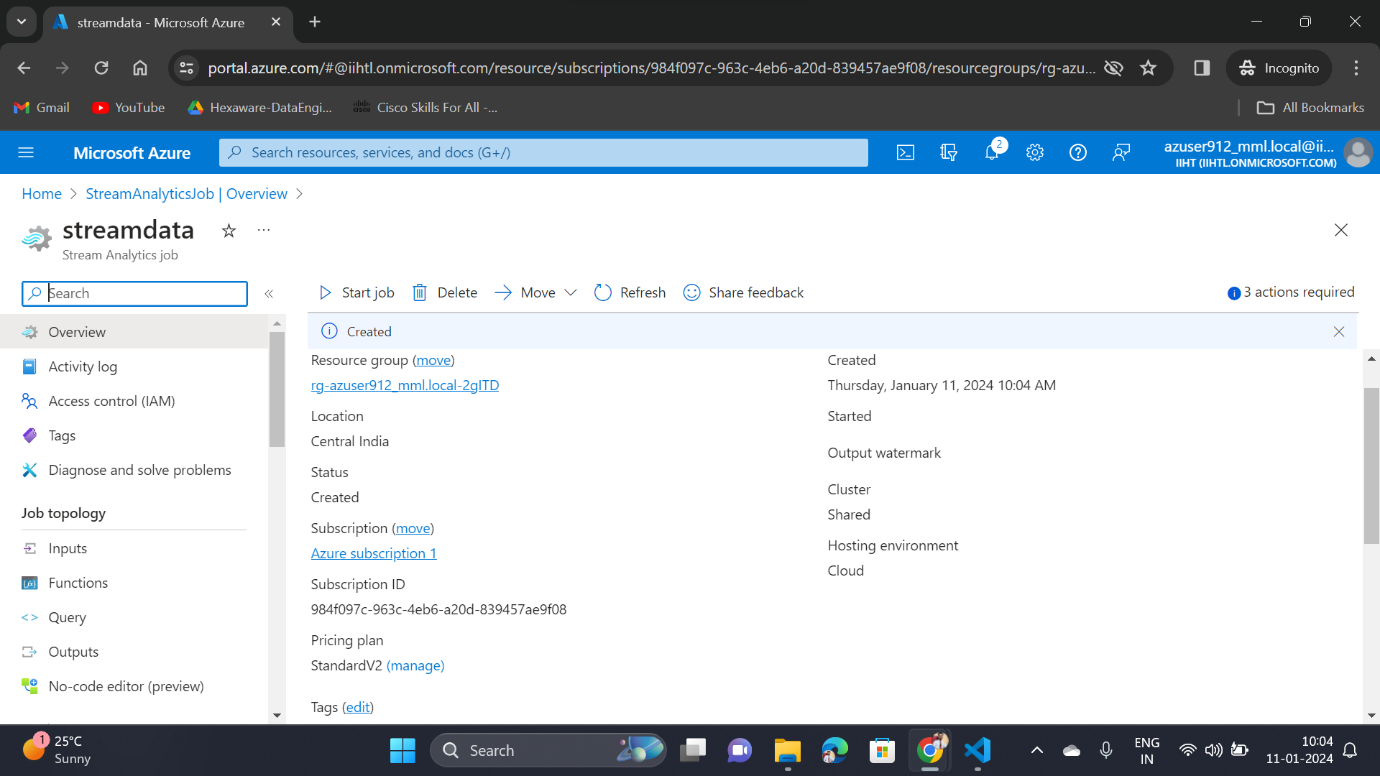
Once created the policy, copy the connection string primary-key and paste that in the python application to establish the connection



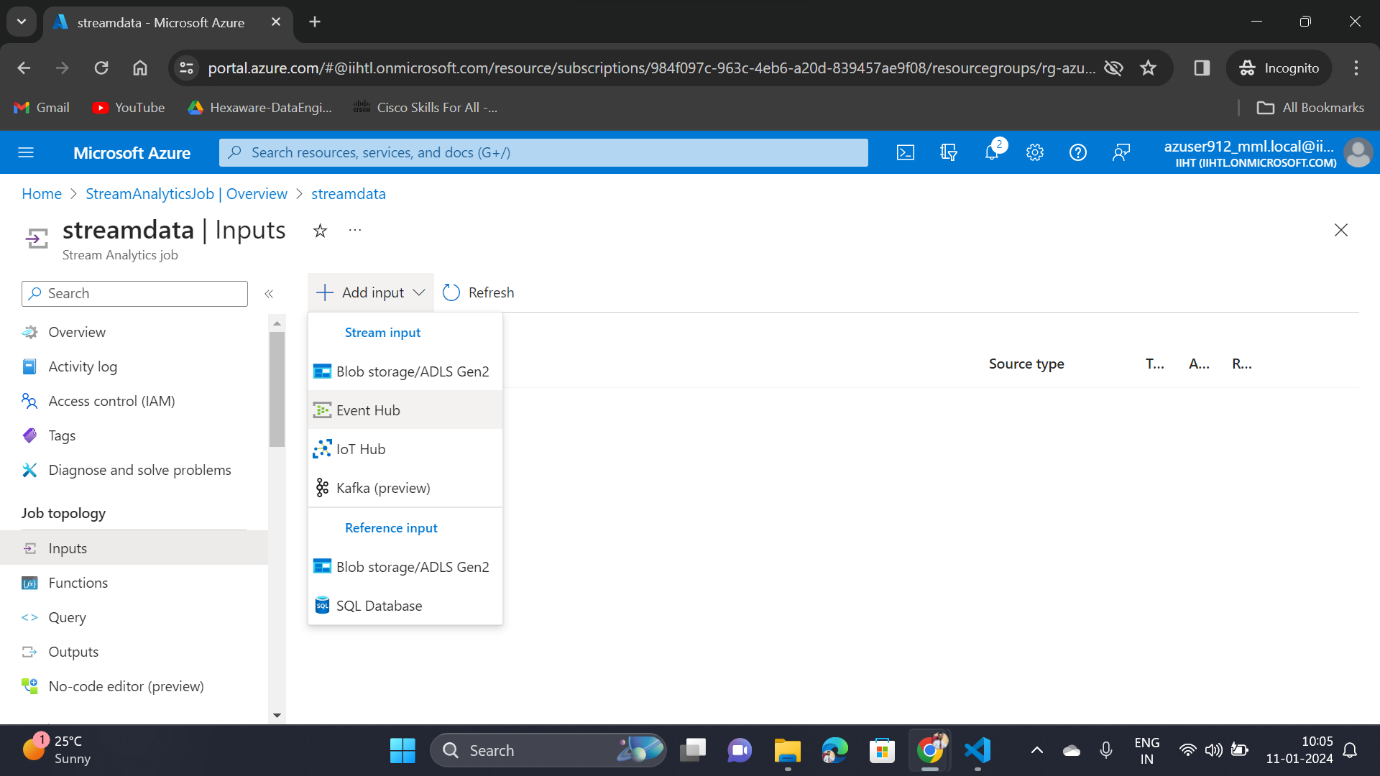
6.3) Creating an Azure Stream Analytics Job



After the successful creation of Azure Stream Analytics Job



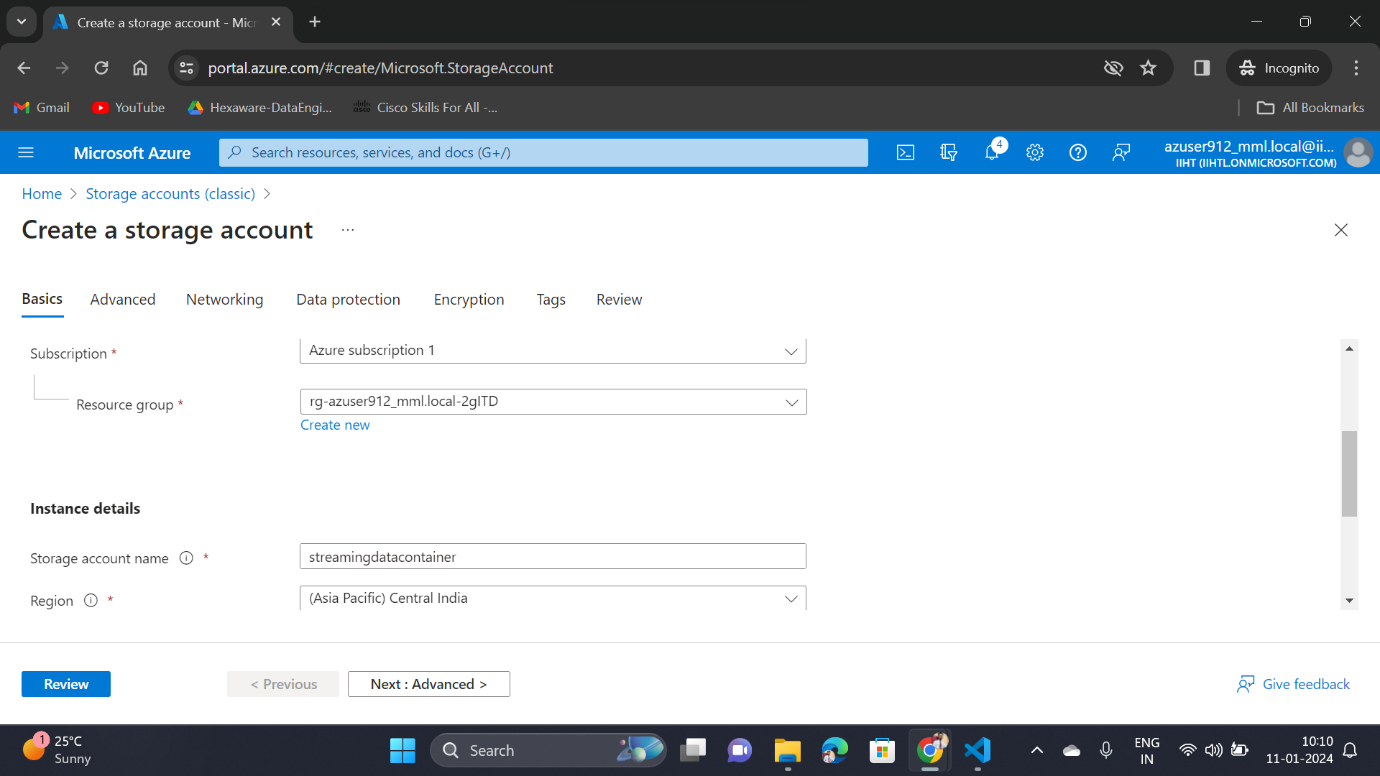
Adding input to the Stream Analytics Job, As the data are receiving through the Event Hub so Choose select the input type as Event Hub



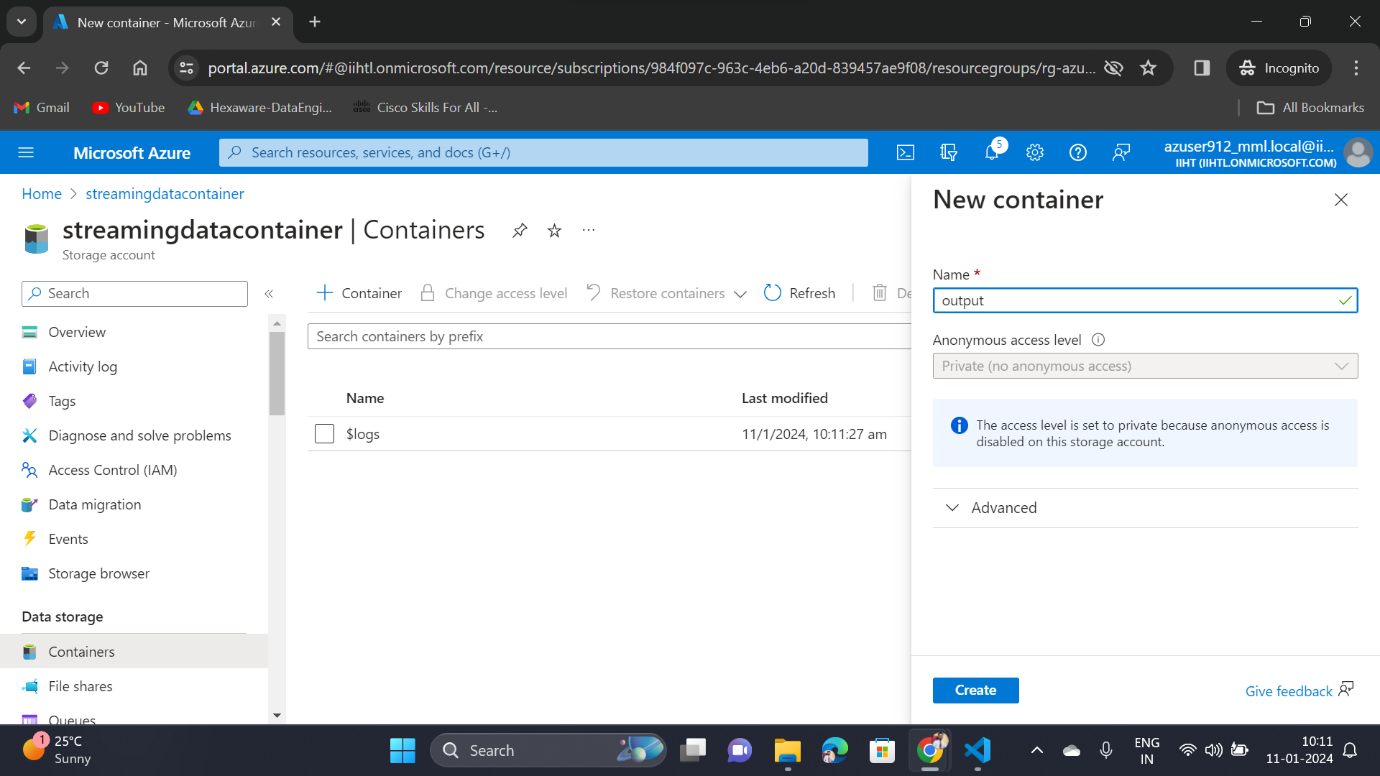
Enter the Event Hub Namespace name and other details in the input field



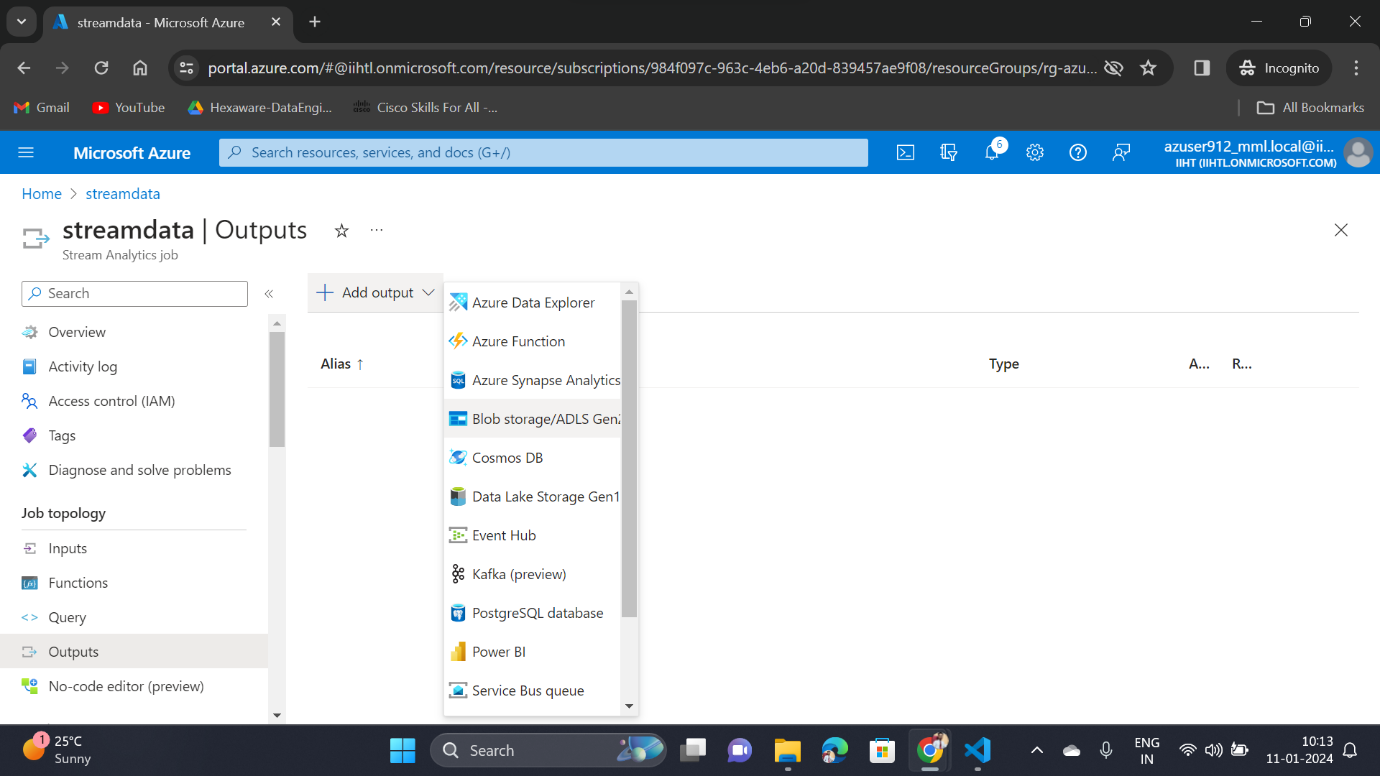
Creating an Azure Data Lake Storage



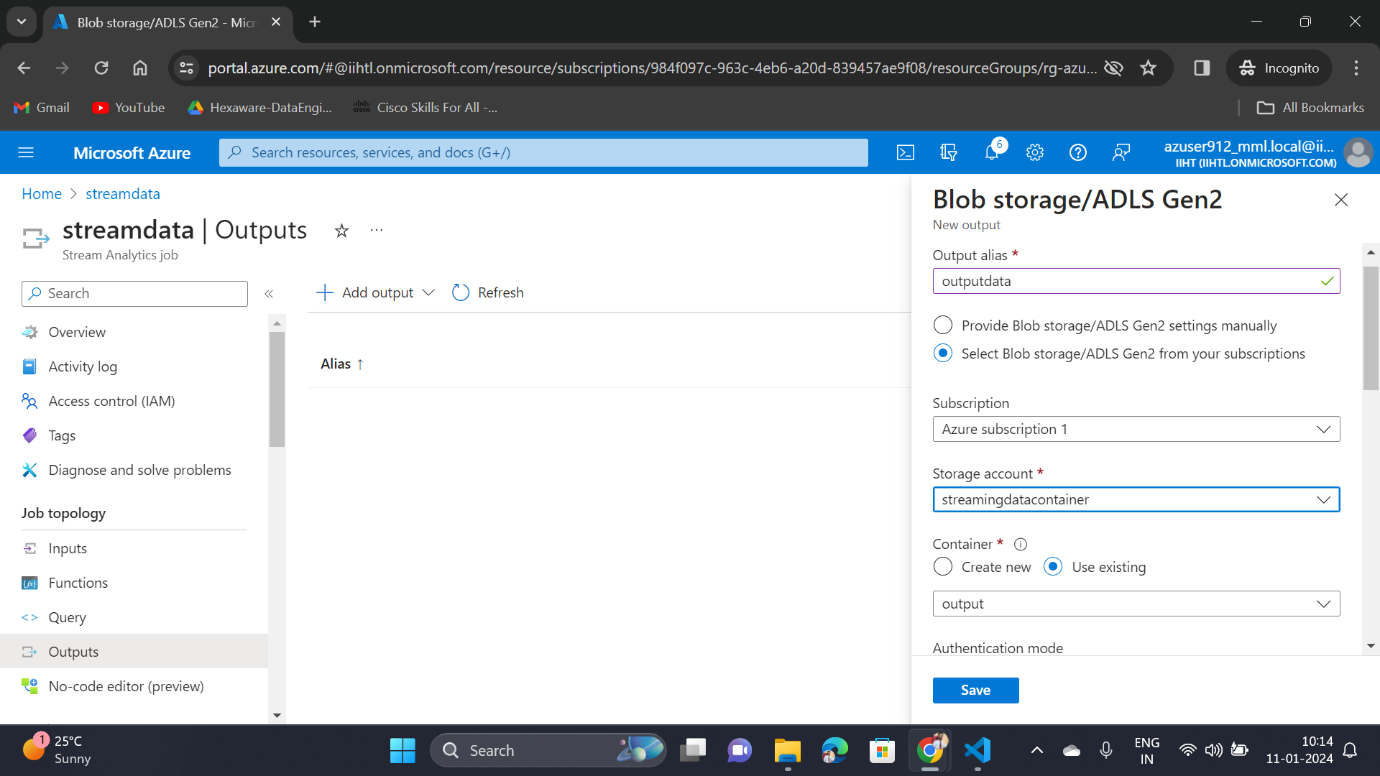
Creating an empty container to ingest the received streaming data



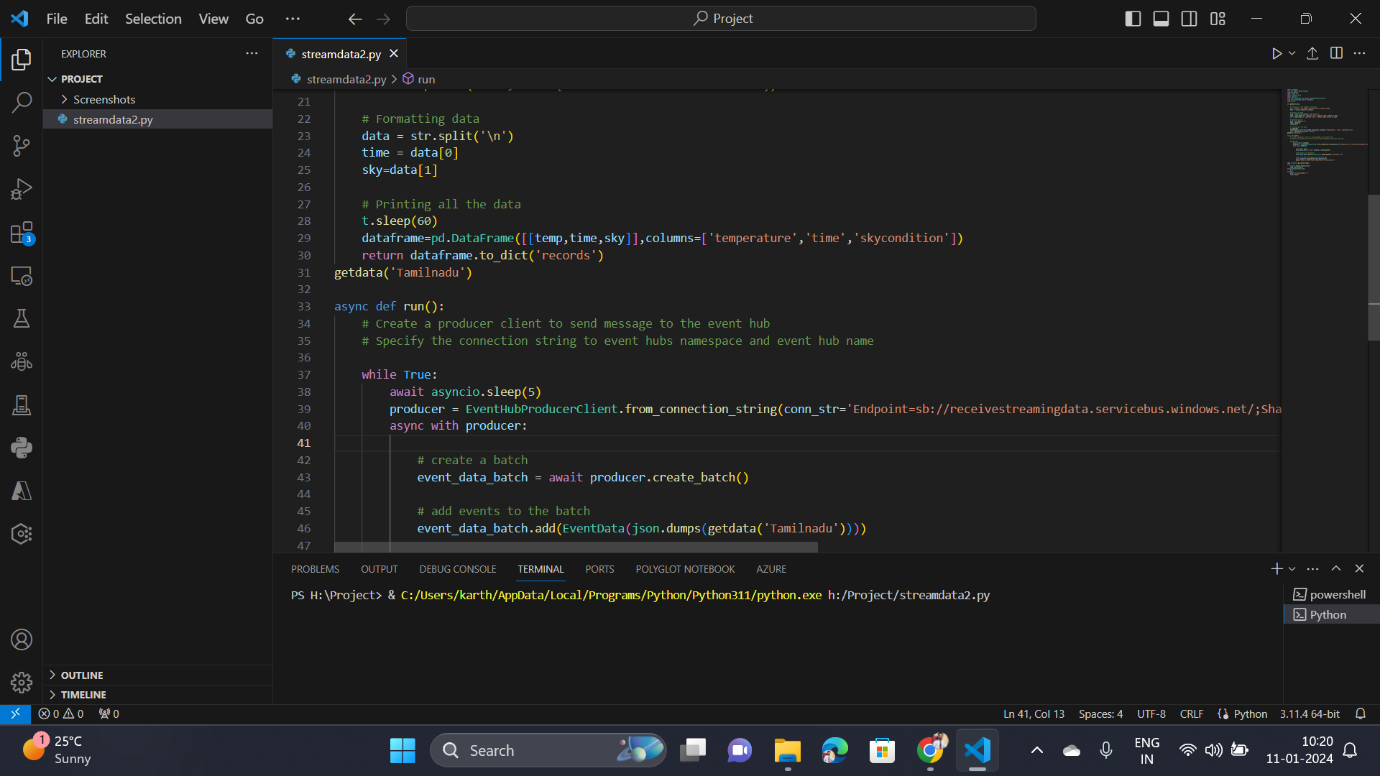
Now go to the output service in the Azure Stream Analytics Job and select output option as Blob Storage / ADLS Gen2



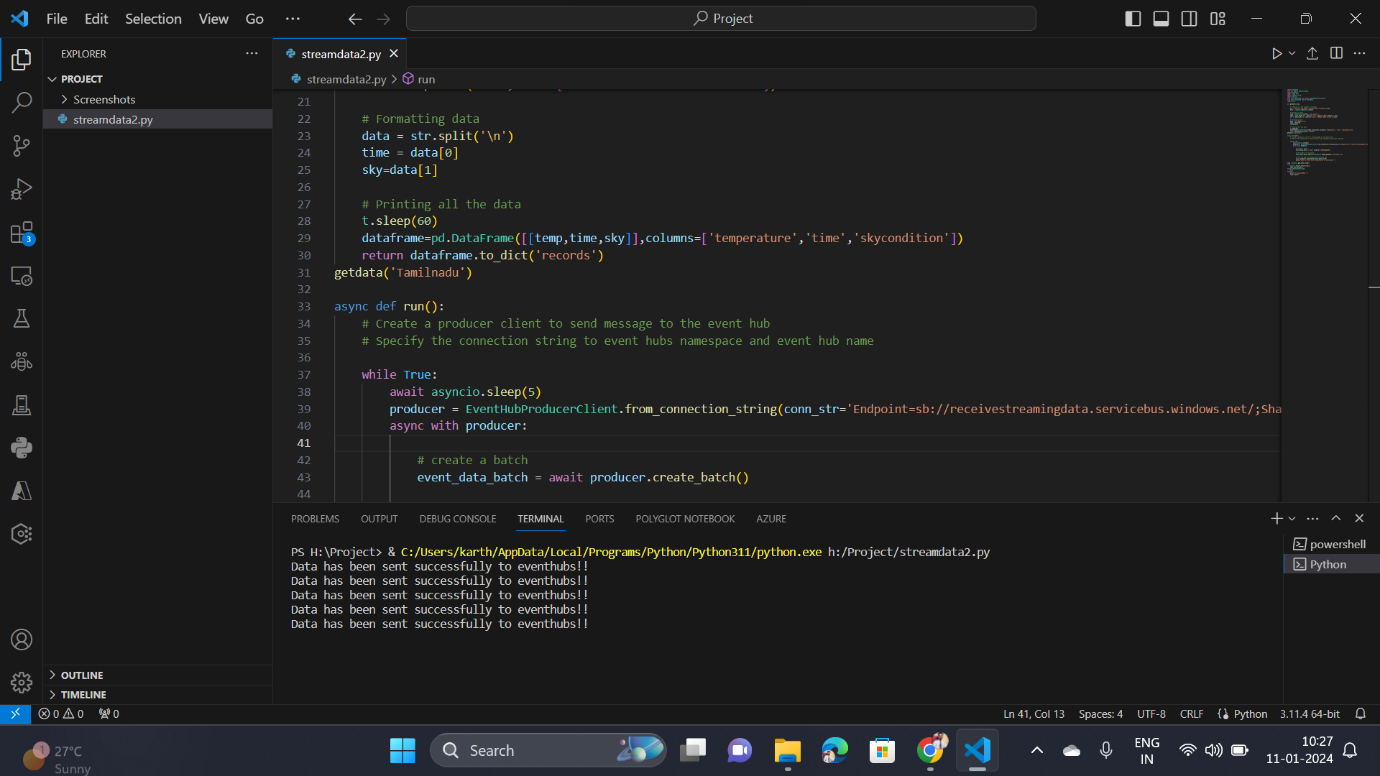
Enter the Azure Data Lake Storage name with container name and other details in the output field



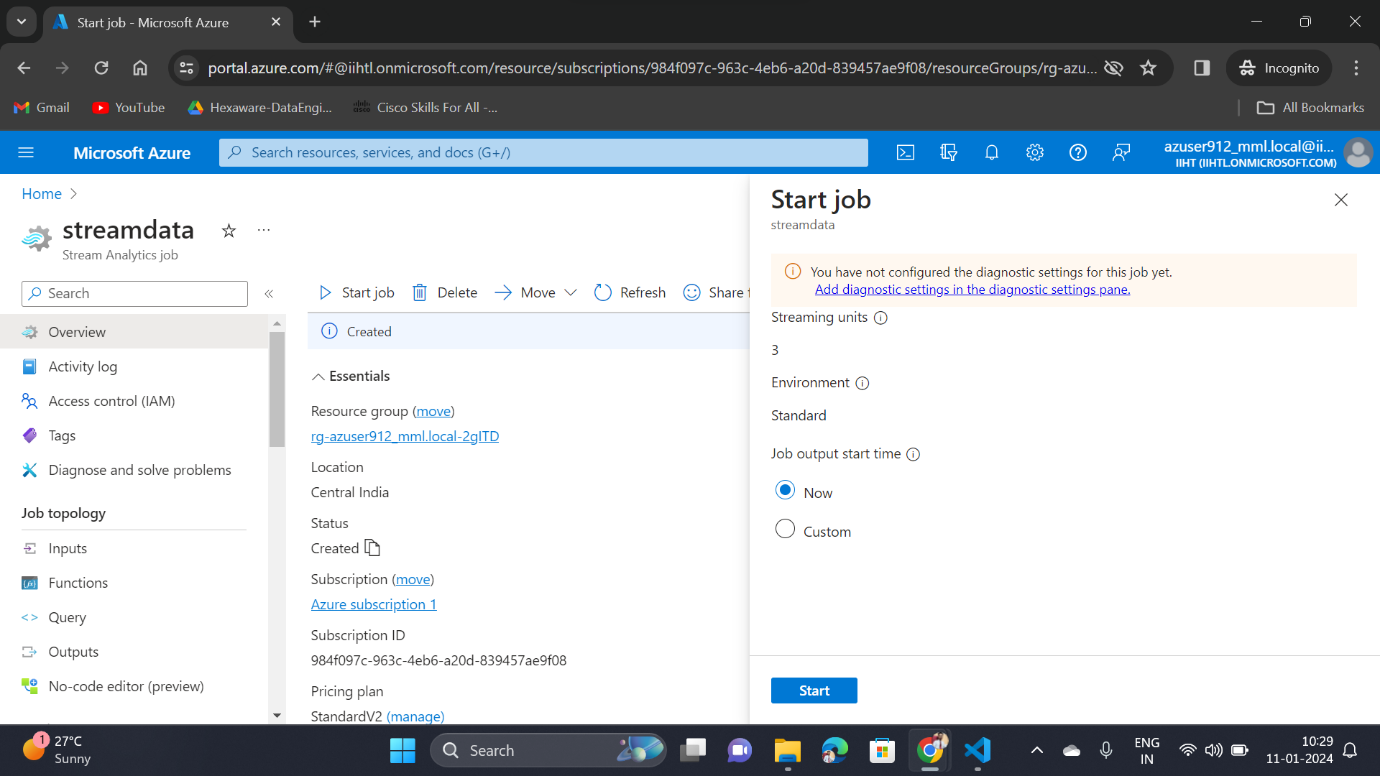
Now, go to the python application and run the application to start the process



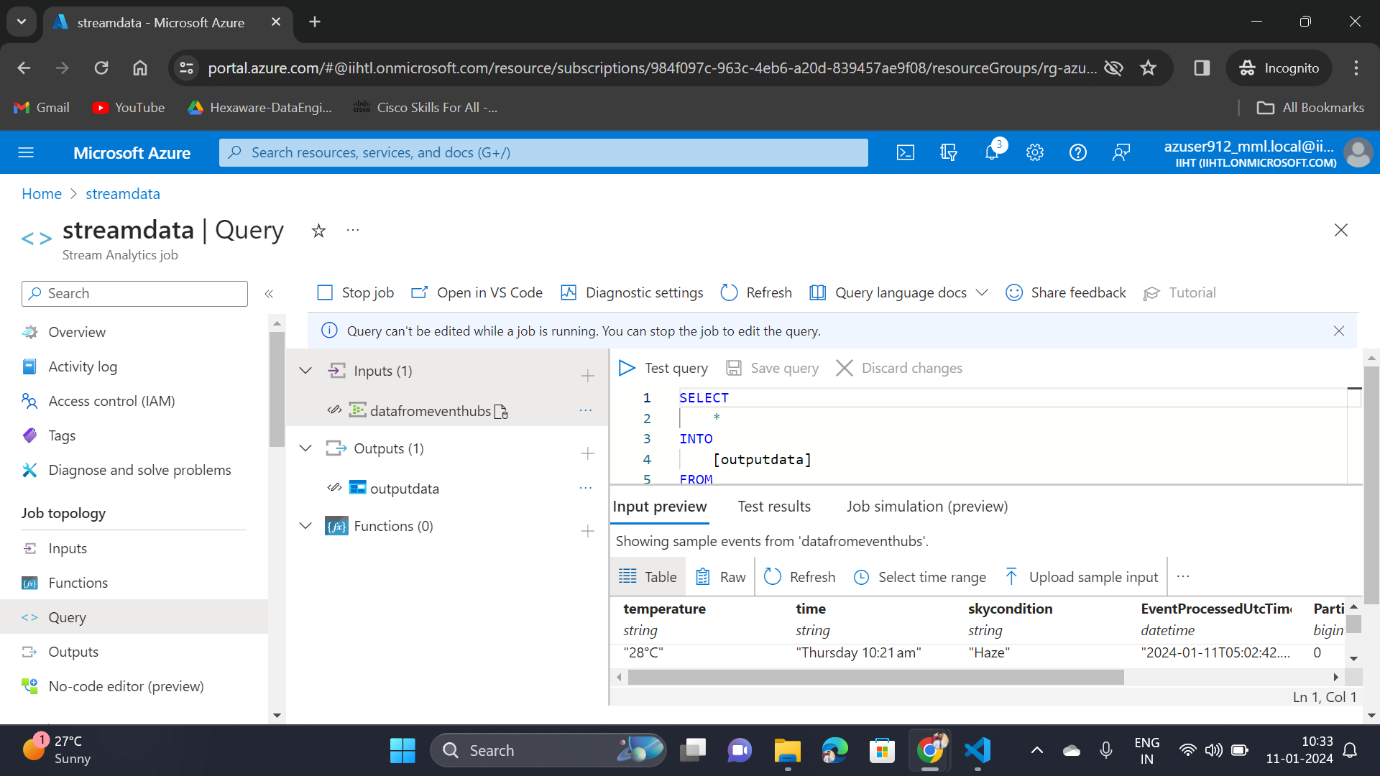
Once the application started running successfully, we can able to see the output as **“Data has been sent successfully to eventhubs !!”**



Now, go to the Azure Stream Analytics Job and start the job

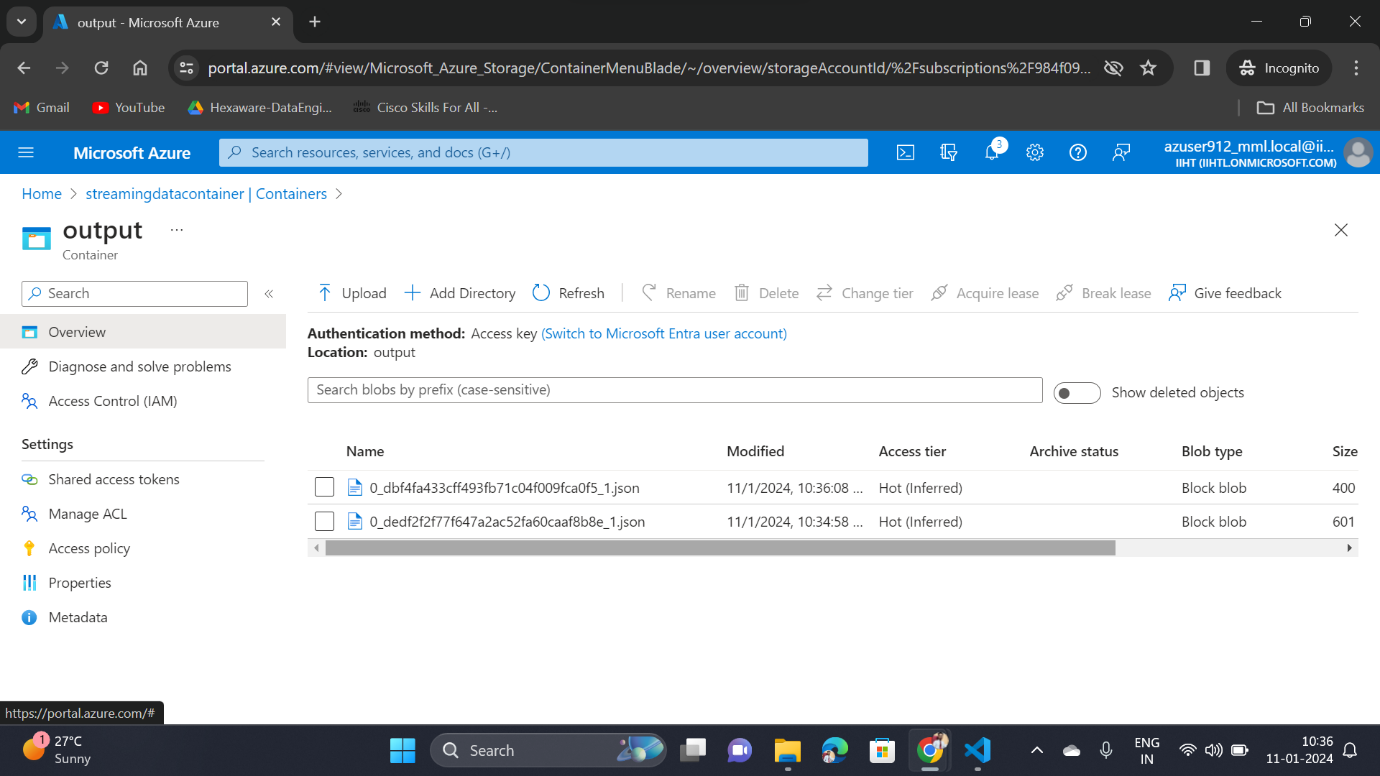


After starting the job go the query service and do refresh, now we can be able to see the streaming data coming from the google webpage

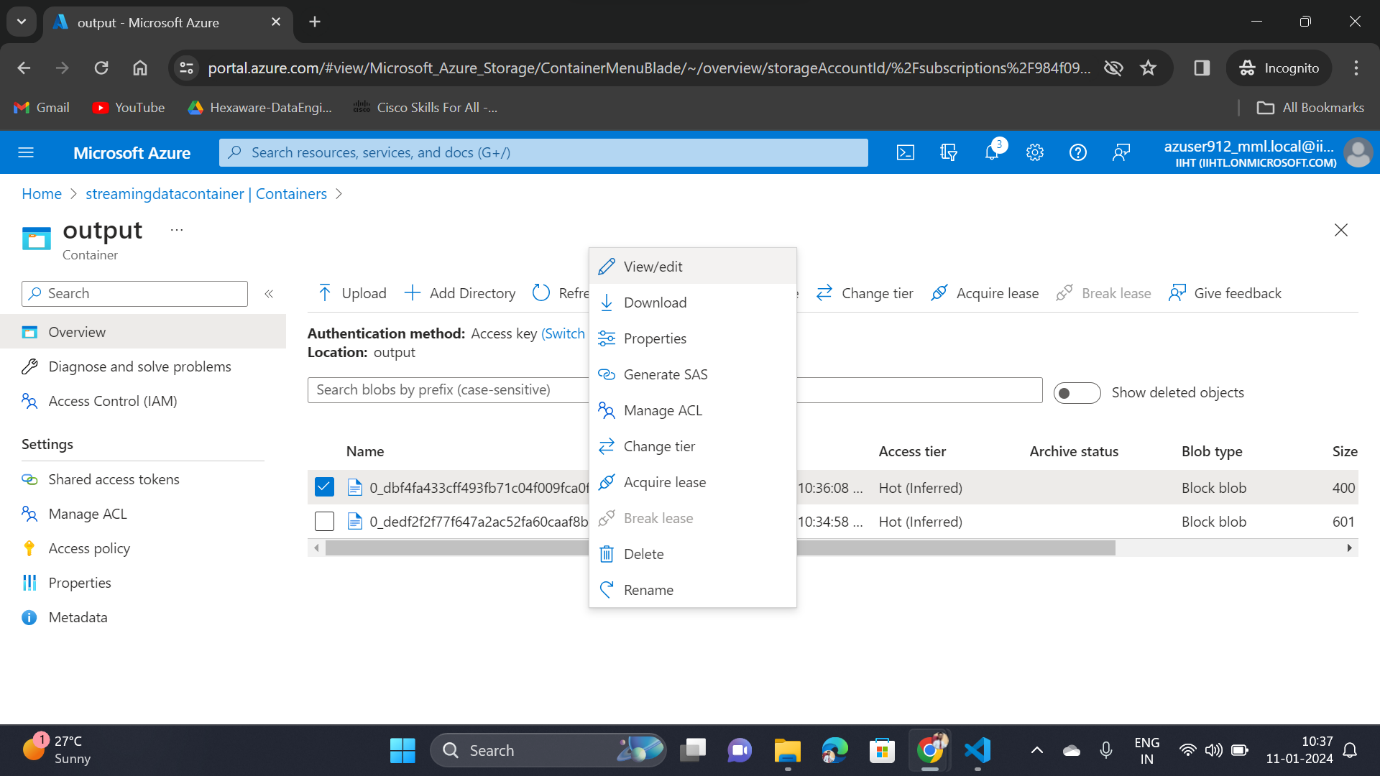


6.4) Azure Data Lake Storage

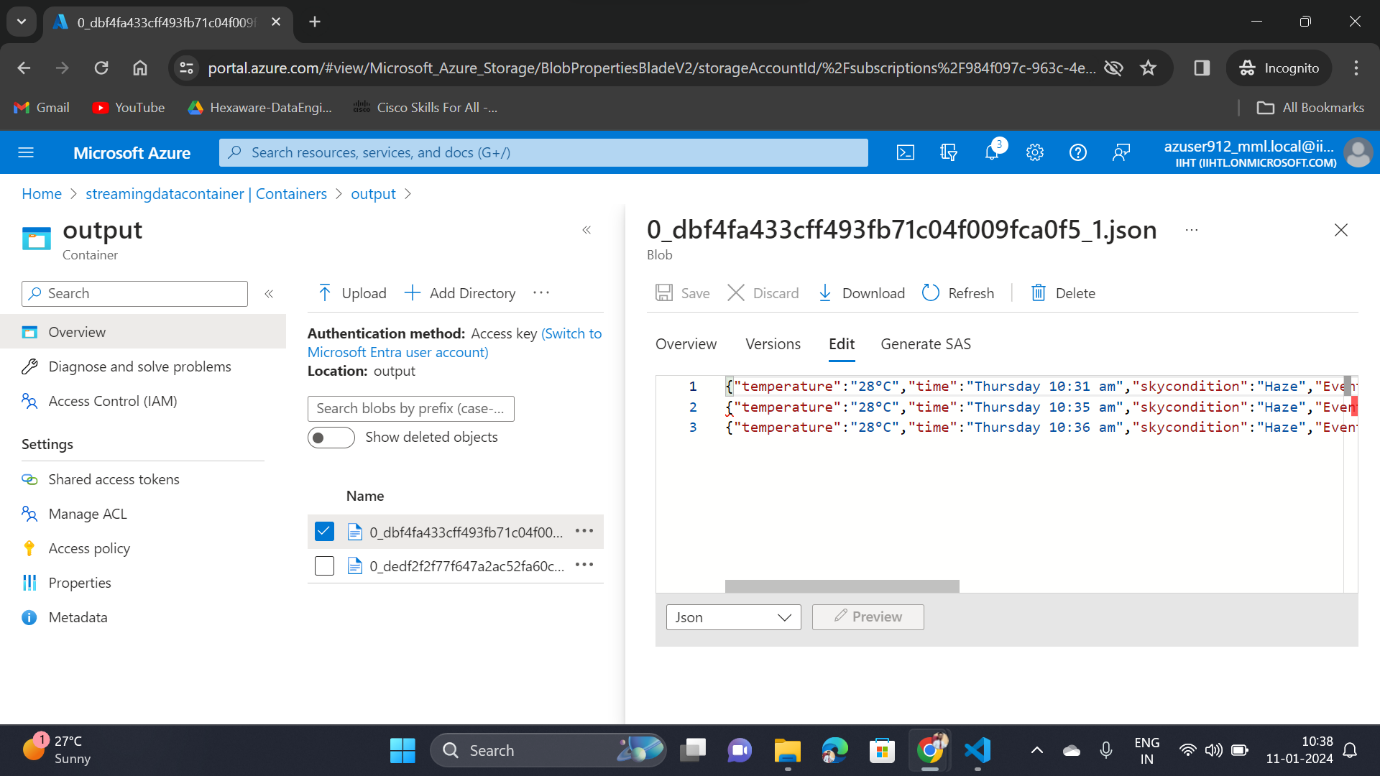
Go to the output container where we will be able to see the **JSON** formatted files present in the container



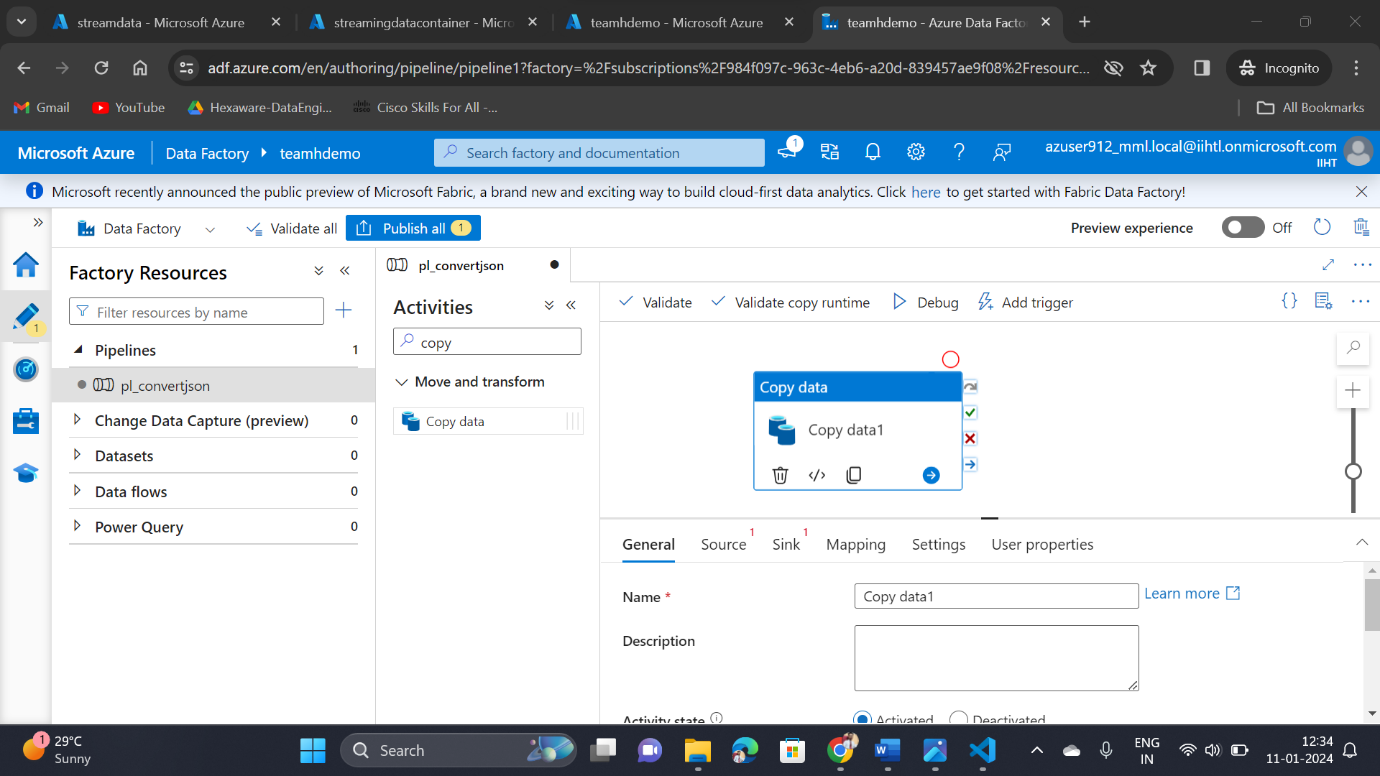
Now choose one file and select View/edit



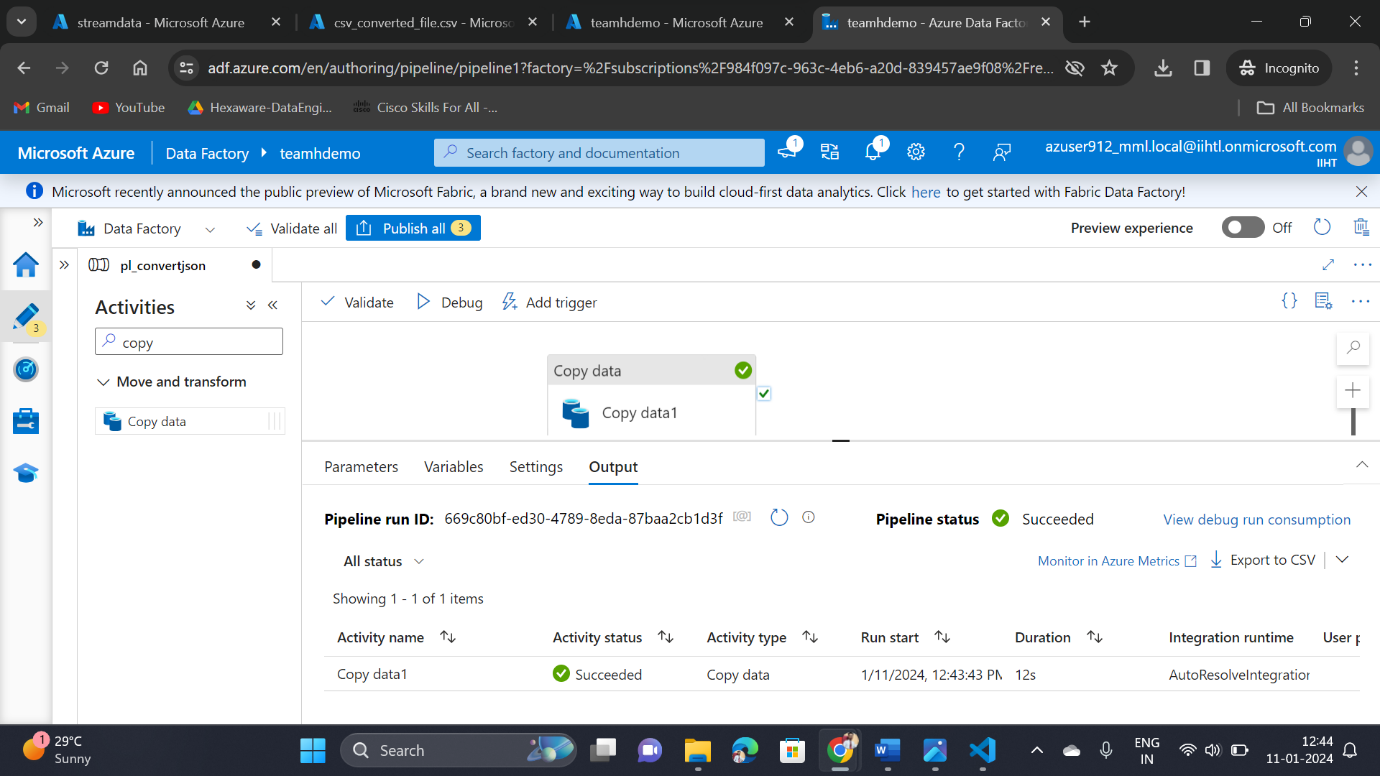
Now we will be able to see the data present in the JSON file



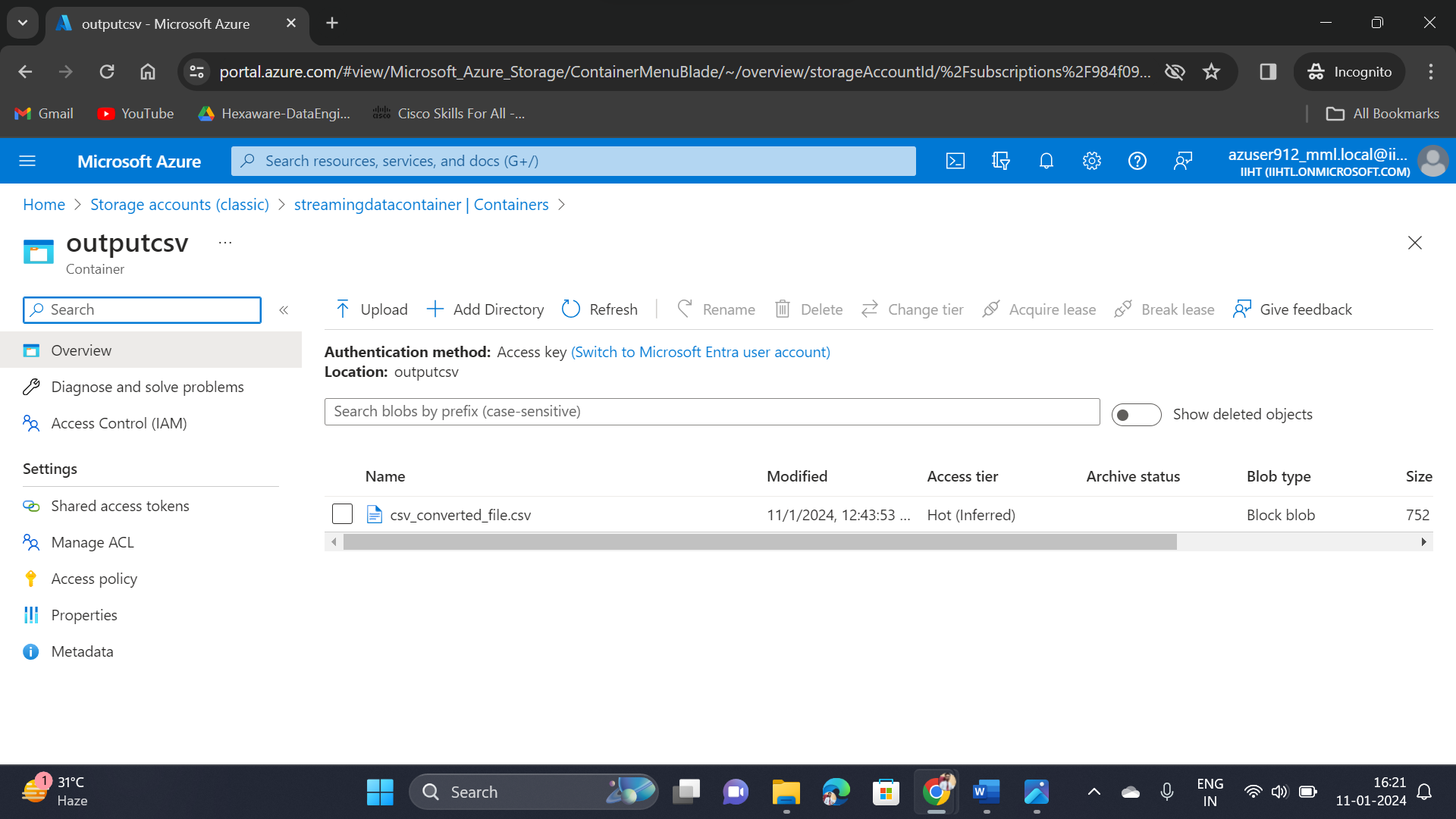
6.5) Create a Data Factory resource and create a pipeline to initiate the copy activity



Fill the parameters which are required and select debug option, after the successfully completion of the activity we can able to see the activity status as succeeded

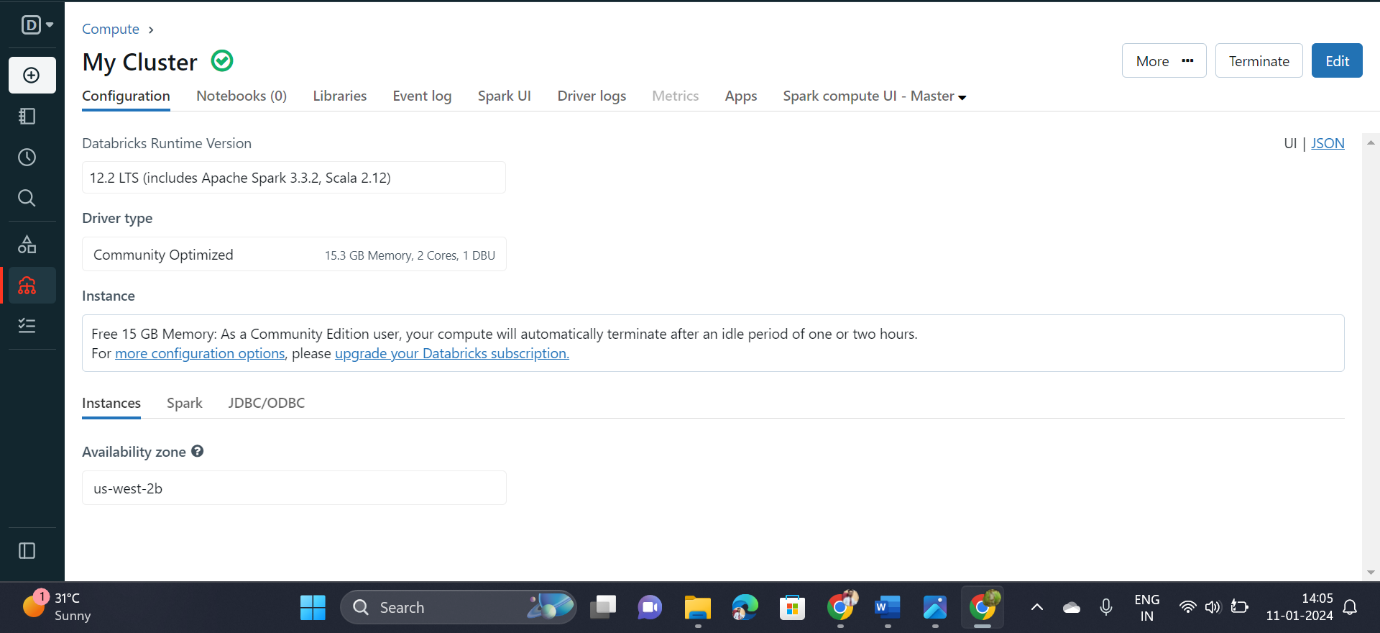


Now, go to the container and we will be able to see the CSV converted file there

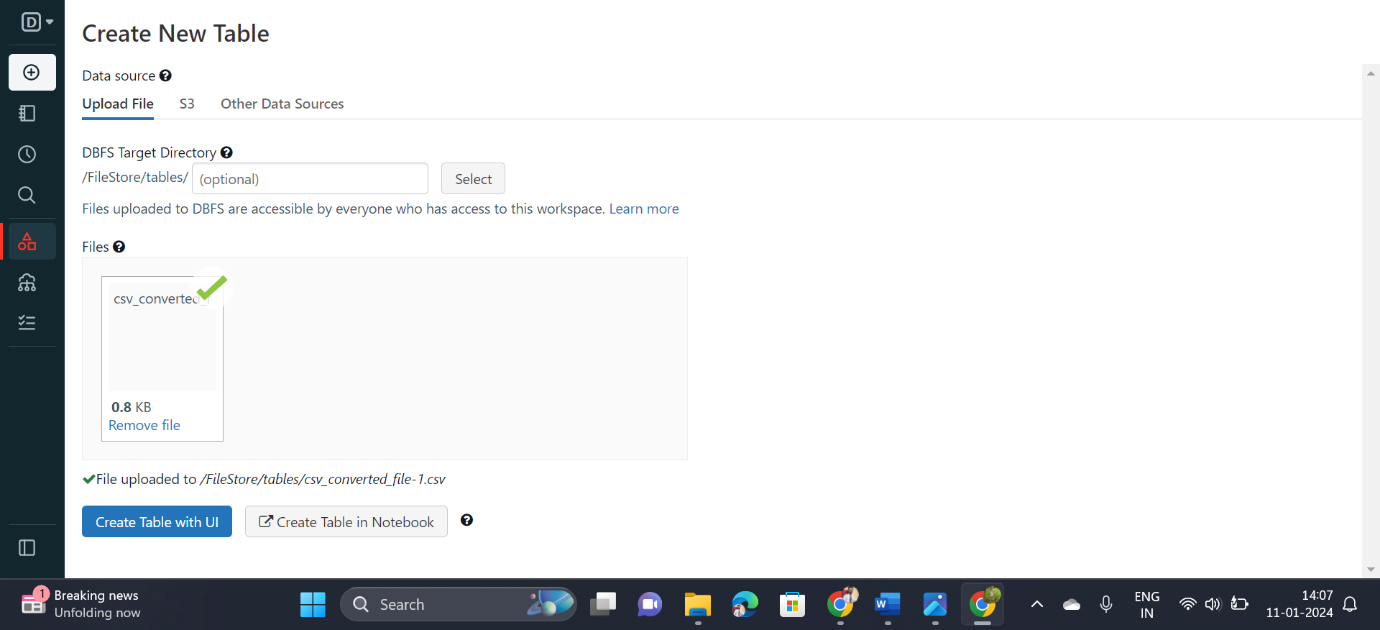


6.6) Azure Databricks

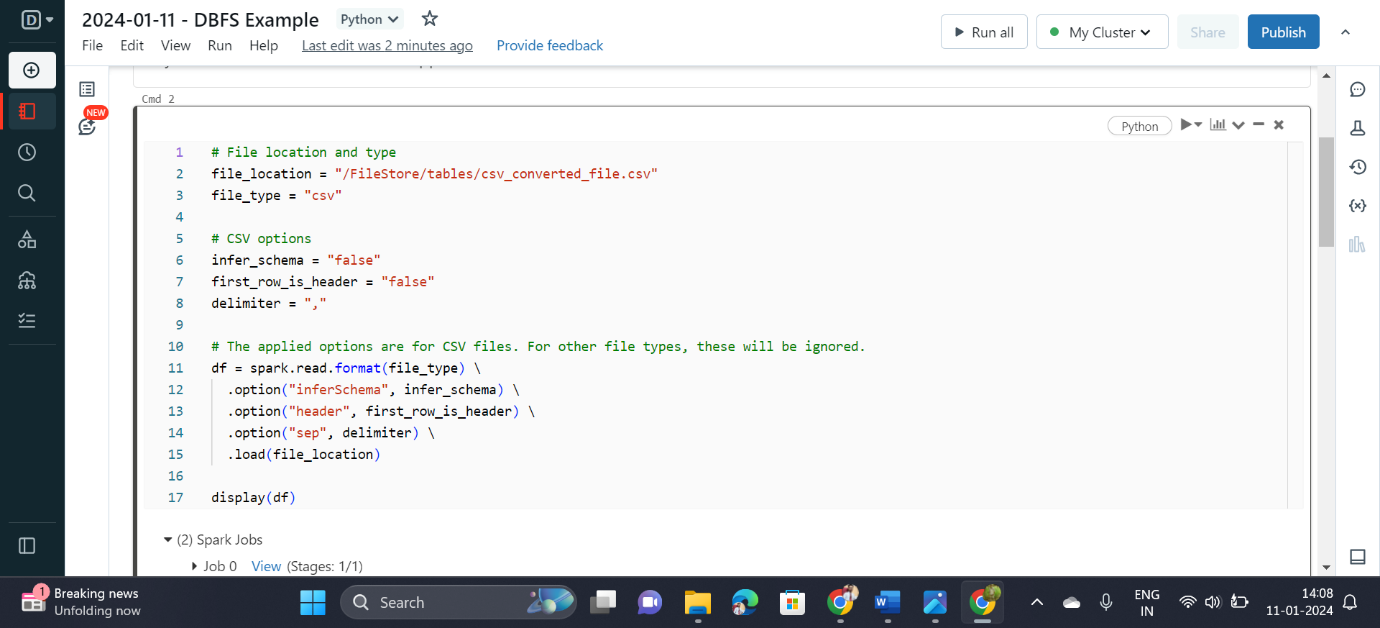
Create one cluster in the Azure Databricks workspace



Then upload the CSV file by using create a new table option and select Create Table in Notebook



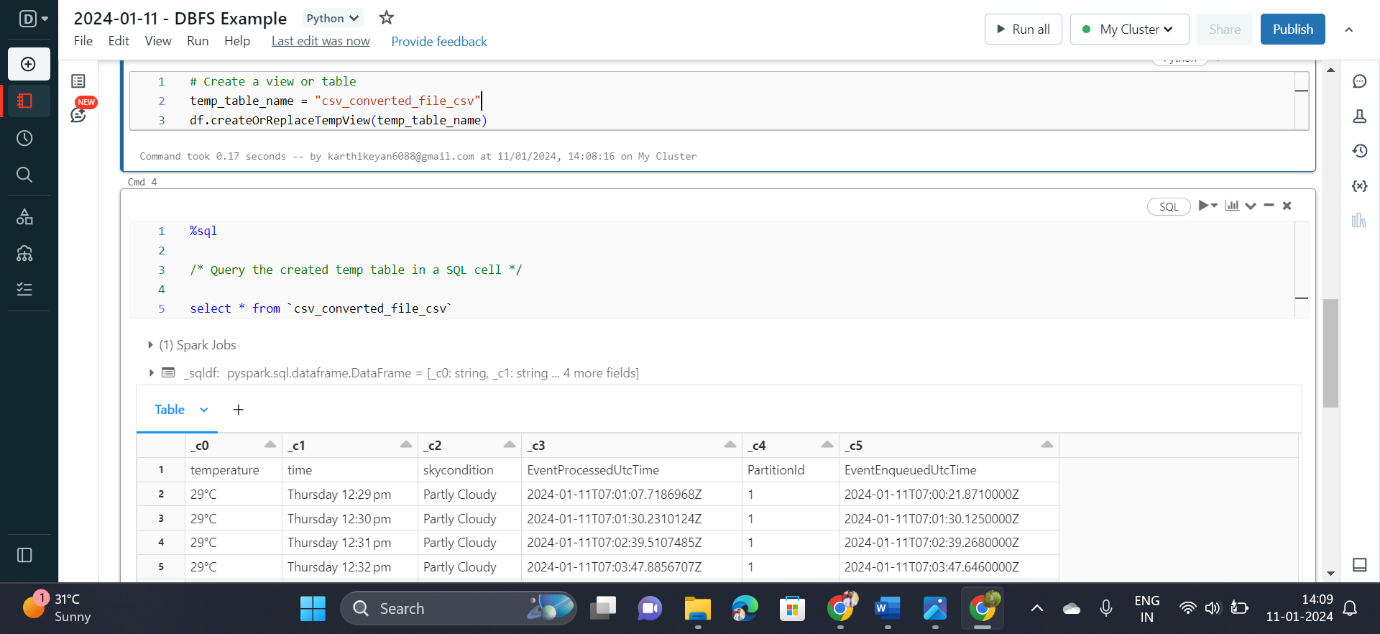
Then execute the python code to create a table for the csv file



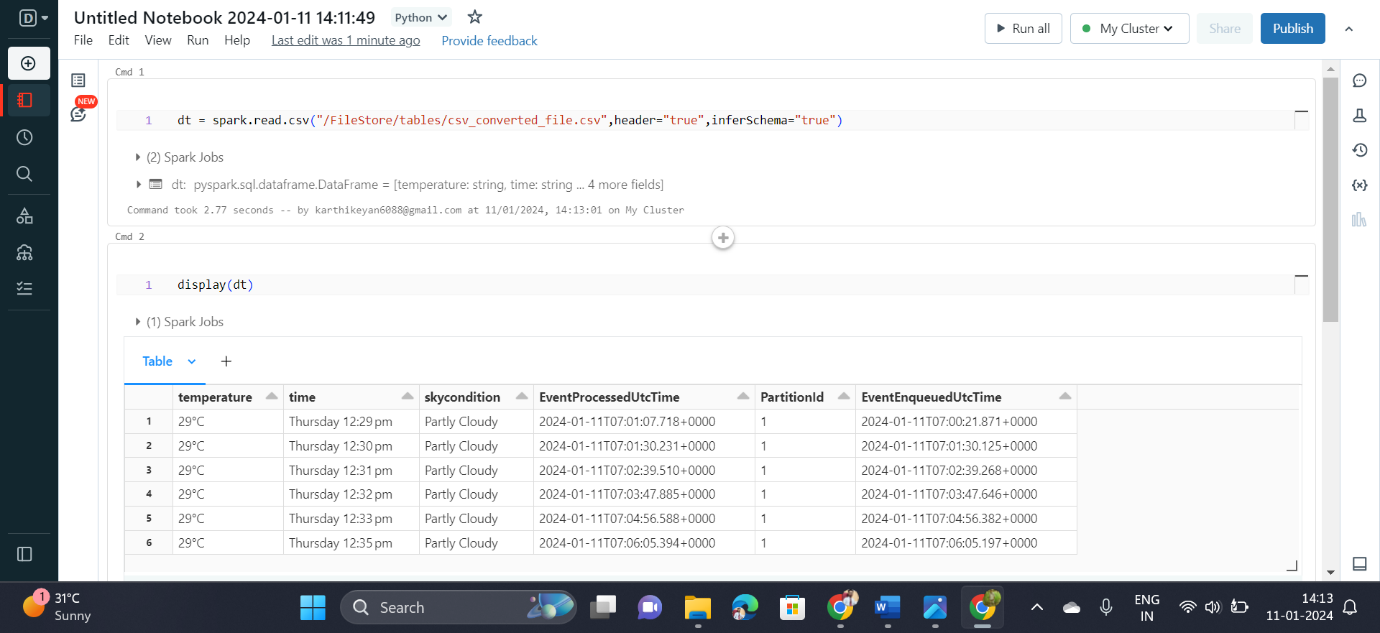
After the successful creation of table, we can be avlr to see the table with weather data



Now create a temporary table and use SQL command to display all the values in the table



Finally, Open one new notebook and execute the spark command to read the csv file data and perform the necessary data analytics process



1. **Conclusion**:

In conclusion, the project to set up an end-to-end pipeline with Azure Data Factory (ADF) for ingesting streaming data and Azure Databricks for real-time processing and analysis of the streaming data is a comprehensive solution for efficiently handling and deriving insights from continuous streams of information. The integration of these Azure services creates a powerful and scalable data processing ecosystem.