# **PROJECT 7: Designing Data Solutions for Different Data-Related Challenges**

**PROJECT OBJECTIVE:** To build Data Solutions using Azure Technologies, including Cosmos DB, Dedicated SQL Pool, Spark Pool, Event Hub, and Stream Analytics.

**Technologies Used:**

● Azure Dedicated SQL Pool – Data warehousing for structured data.

● Azure Synapse Spark Pool – Large-scale data processing with Spark.

● Azure Event Hub – Real-time data ingestion.

● Azure Stream Analytics – Stream processing for real-time analytics.

**Components and Tasks:**

1. **Azure Dedicated SQL Pool**:

**Purpose**: The Azure Dedicated SQL Pool (formerly known as SQL Data Warehouse) is designed to perform batch processing and analytics on structured data. It allows organizations to analyze large volumes of data quickly and efficiently.

**Key Features:**

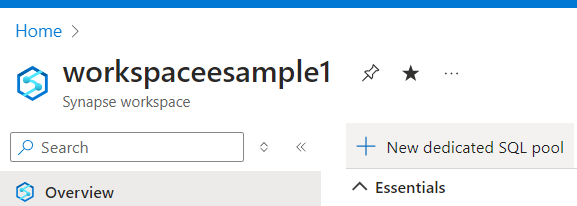
* **Massively Parallel Processing (MPP)**: Distributes data and processing across multiple nodes(60) to enhance performance for large queries.
* **Scalability**: Easily scale up or down resources based on workload demands, allowing for cost-effective performance tuning.
* **Structured Data Analytics**: Supports T-SQL for querying data, making it accessible for SQL-based analytics.
* **Integration with Azure Services**: Seamlessly integrates with other Azure services like Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, and Power BI.

**Use Cases:**

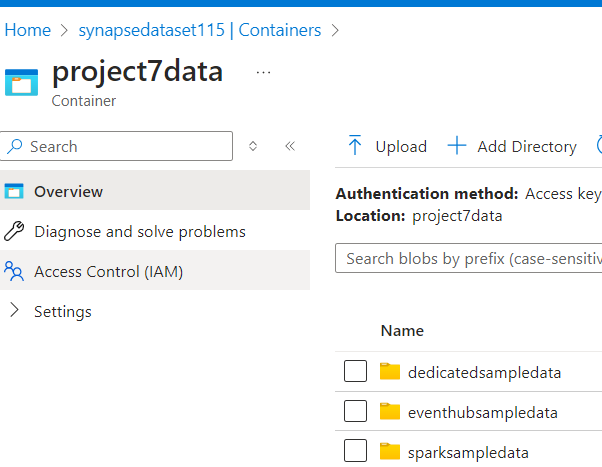
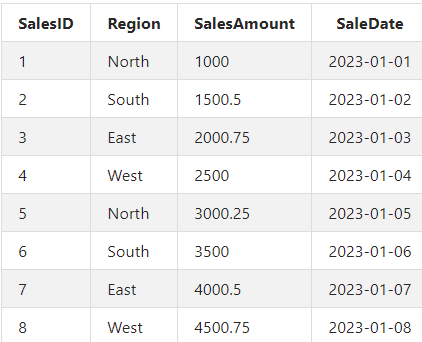
* + **Data Warehousing**: Centralize and analyze large datasets from various sources.
  + **Business Intelligence**: Enable reporting and visualization tools like Power BI to access and analyze data.
  + **Machine Learning**: Prepare data for machine learning models by performing necessary transformations and aggregations.

**Tasks**: **To perform** **Sales Analysis with Dedicated SQL Pool**

* + - Set up the Dedicated SQL Pool ‘samplededicatedsqlpool’ in a synapse workspace

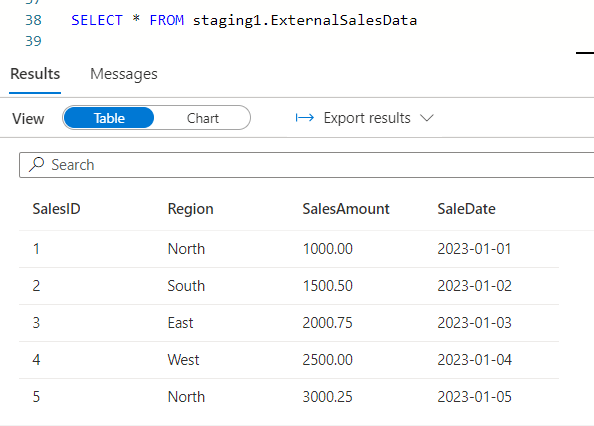
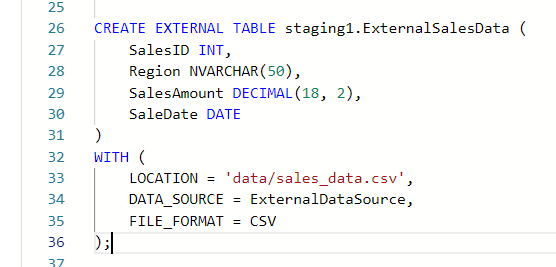


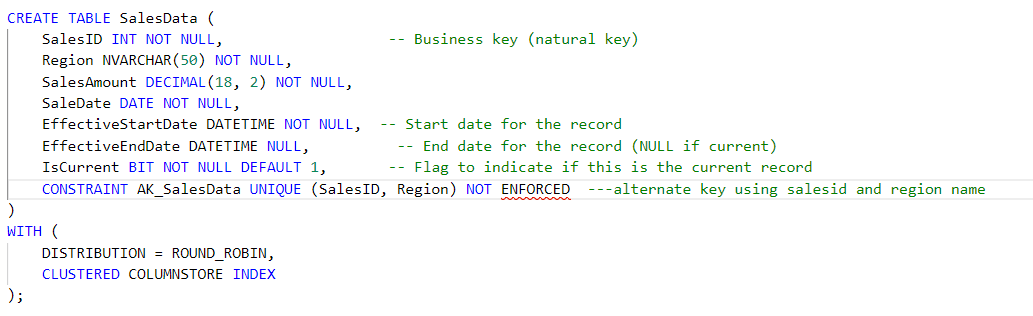
* + - Upload a sample file in ADLS: dedicatedsampledata/data/sales\_data.csv

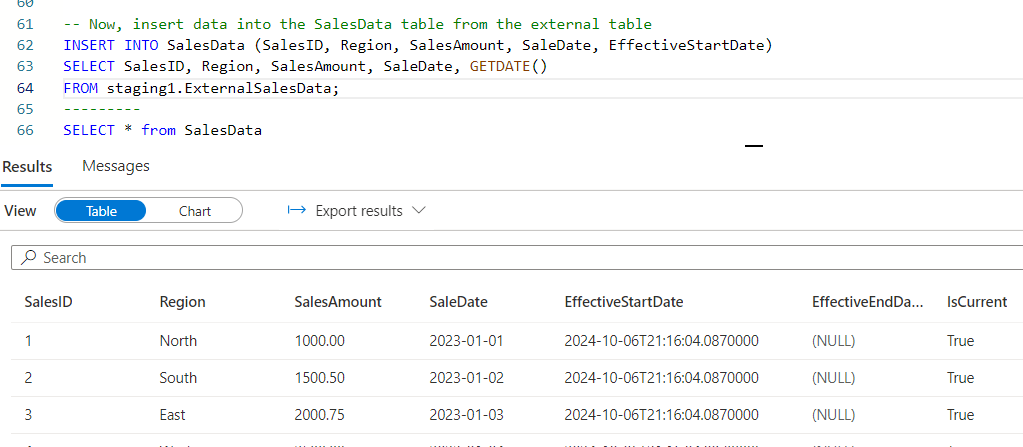
 

Develop SQL queries for data extraction and transformation

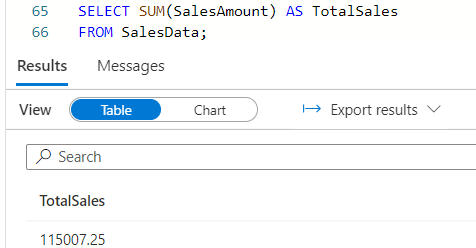
* + - Create an external table ‘ExternalSalesData’ to read sales data from ADLS:

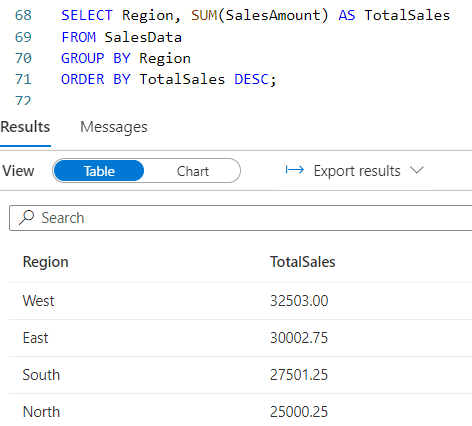


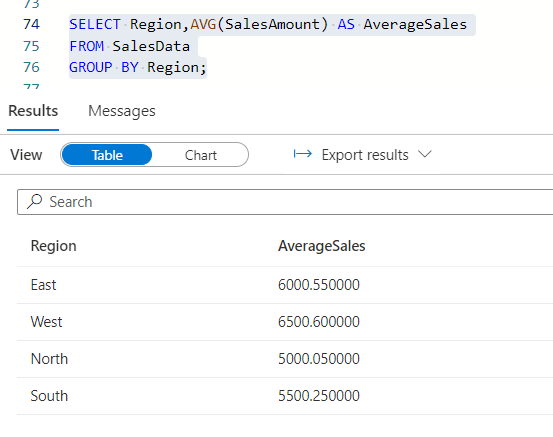
* + - create ‘SalesData’ table with following features: Type 2 SCD, alternate key
    - Insert data to SalesData table from the external table:

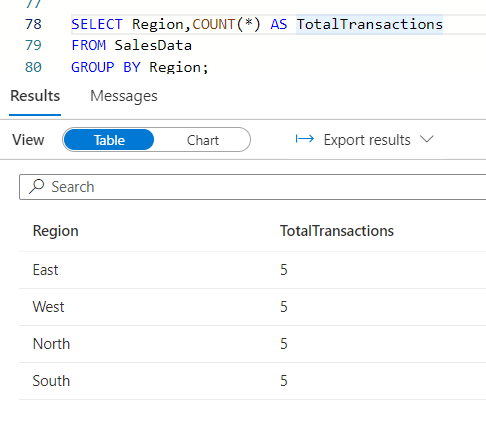


* + Perform some transformations on sales data:

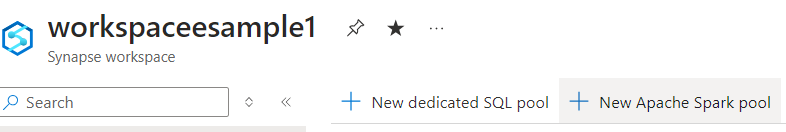
 TOTAL SALES

 REGION WISE TOTAL SALES

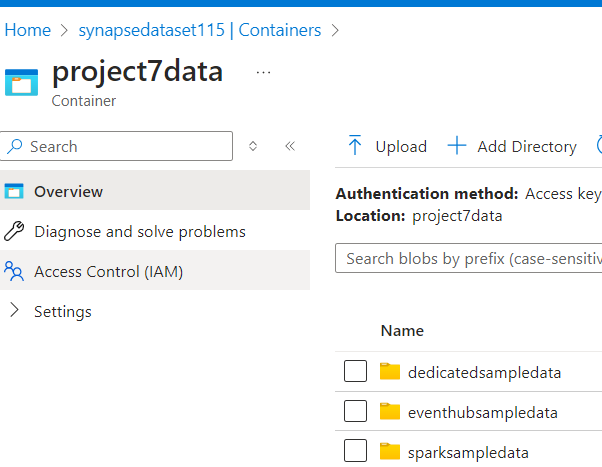
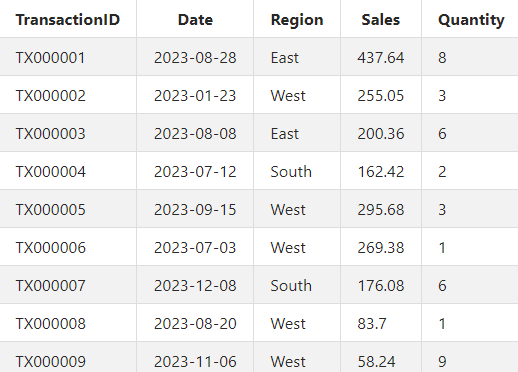
 REGION WISE AVG SALES

 TOTAL TRANSACTIONS IN EACH REGION

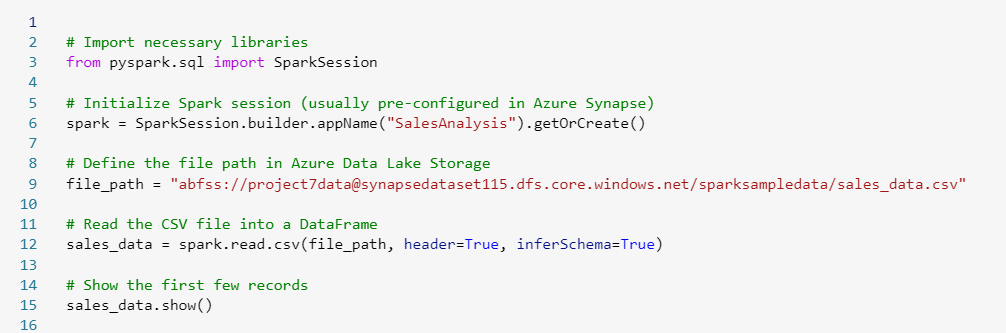
1. **Azure Synapse Spark Pool**:
   * **Purpose**: The Azure Synapse Spark Pool is designed to conduct large-scale data transformations. It leverages Apache Spark, enabling users to process vast amounts of data efficiently, whether in near real-time or in batch mode.
   * **Key Features:** scalability, multiple language support, seamless Integration with Azure Services, high performance
   * **Tasks**:
     + Set up the Spark Pool ‘samplesparkpool’ in synapse workspace:

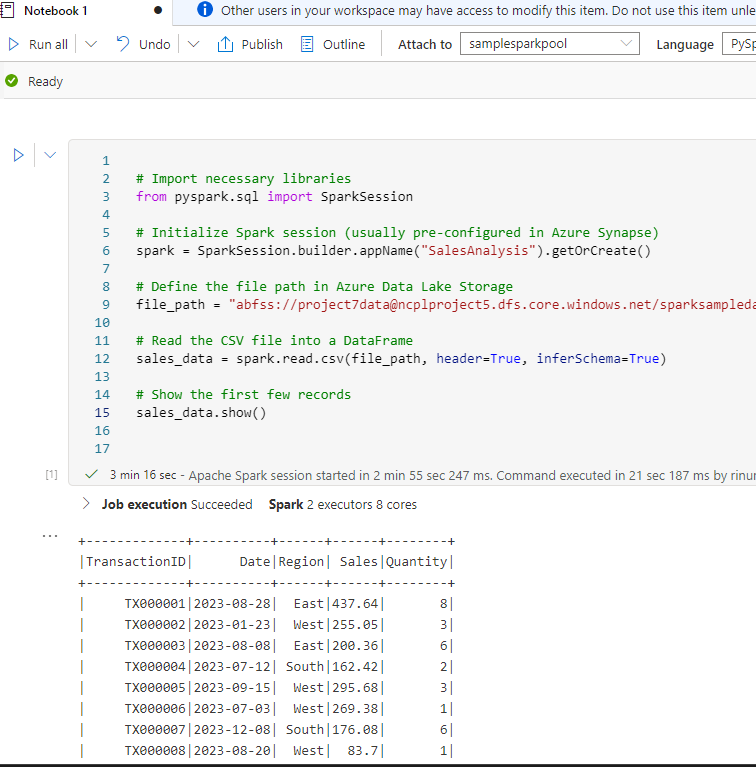


* + - Upload a sample file in ADLS: sparksampledata/sales\_data.csv

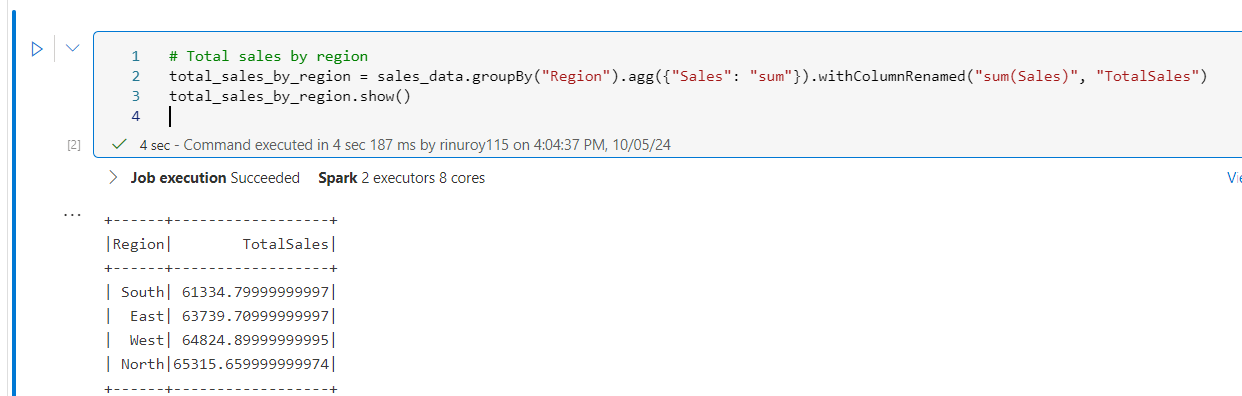
 

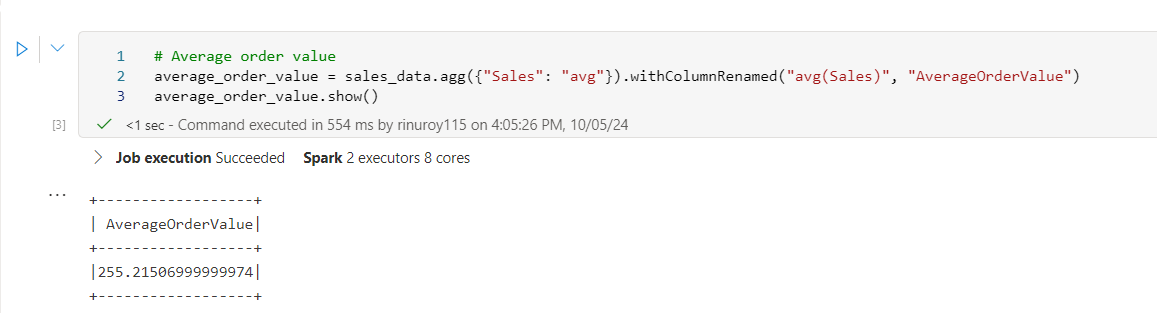
* + - Write Python scripts to read the sales data into data frame(sales\_data) by creating a notebook in spark pool:

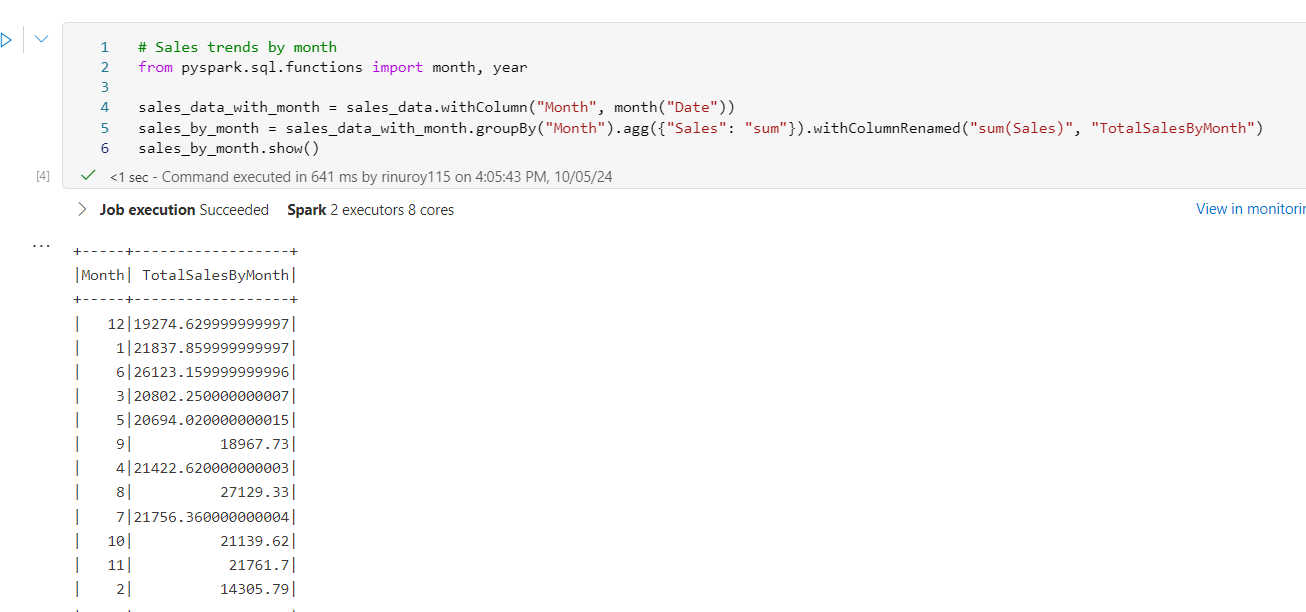


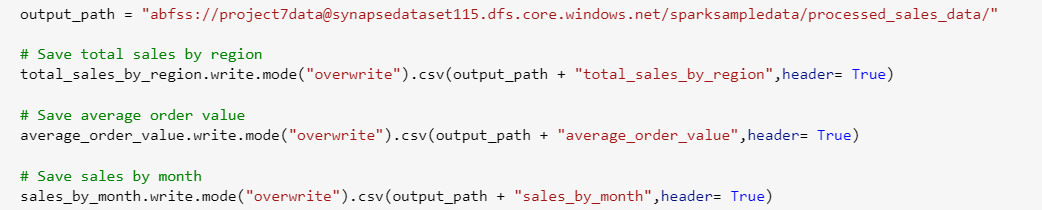


* Performing some transformations on data frame(sales\_data) in spark pool:

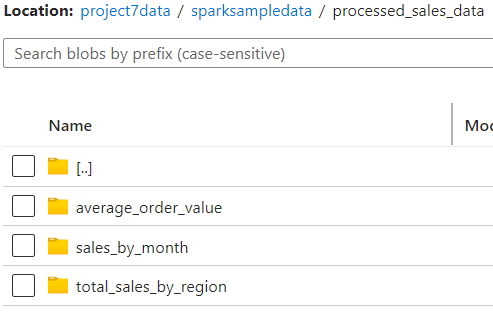




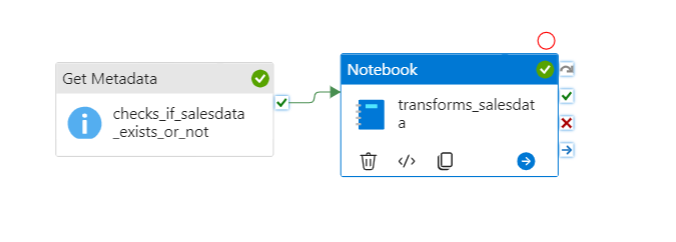


* Write/save the output of transformations to ADLS location mentioned:

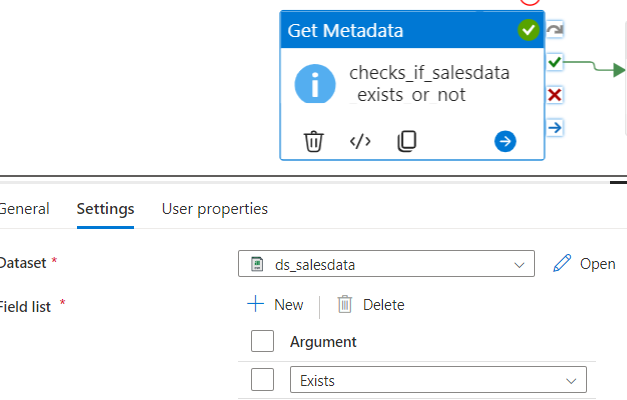




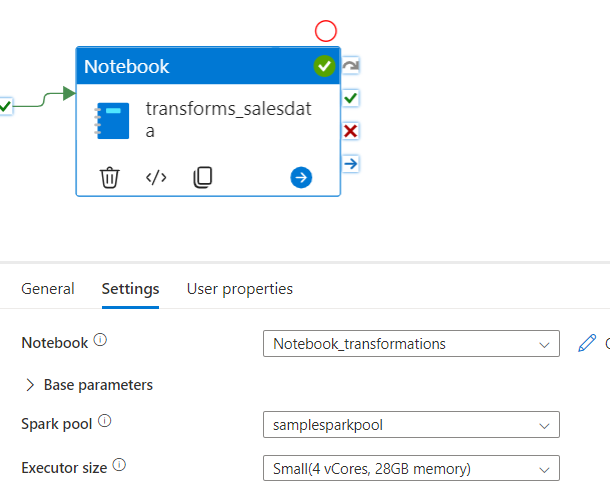
* Integrate the notebook with pipeline as an activity: pipeline has 2 activities



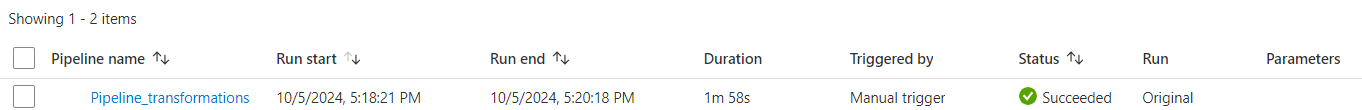
Get metadata activity: to check if file exists or not in ADLS location



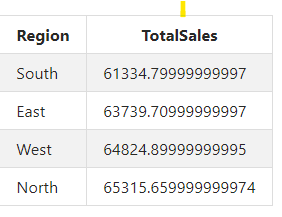
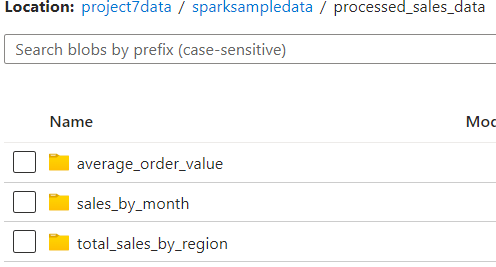
Notebook activity: to perform transformations on the data using spark pool



Triggered the pipeline and it got completed successfully:

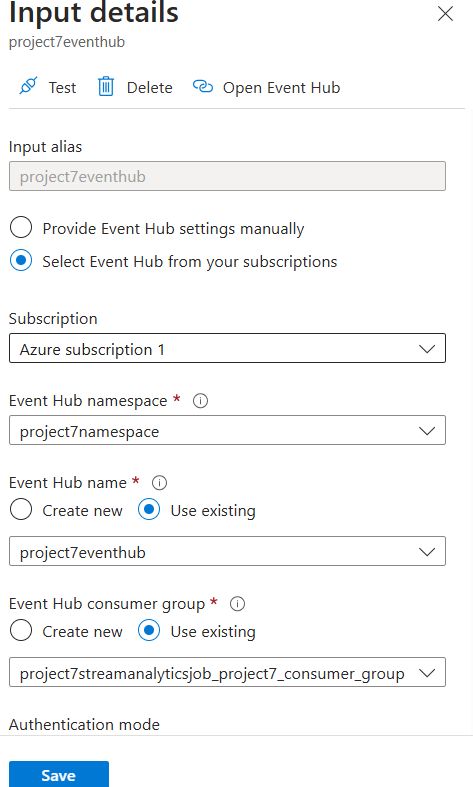
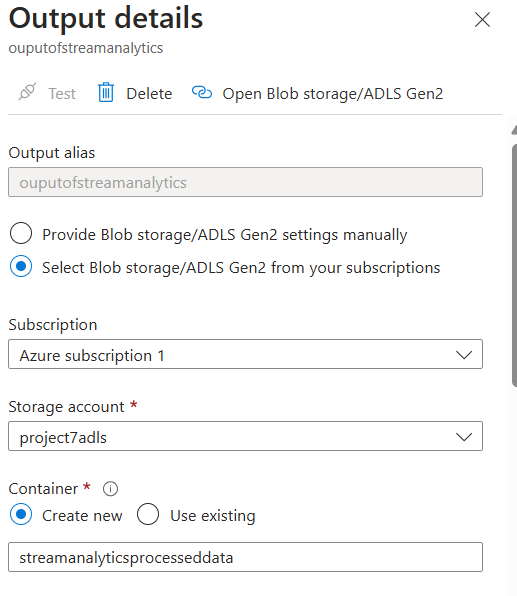


Output: data got saved to mentioned ADLS location

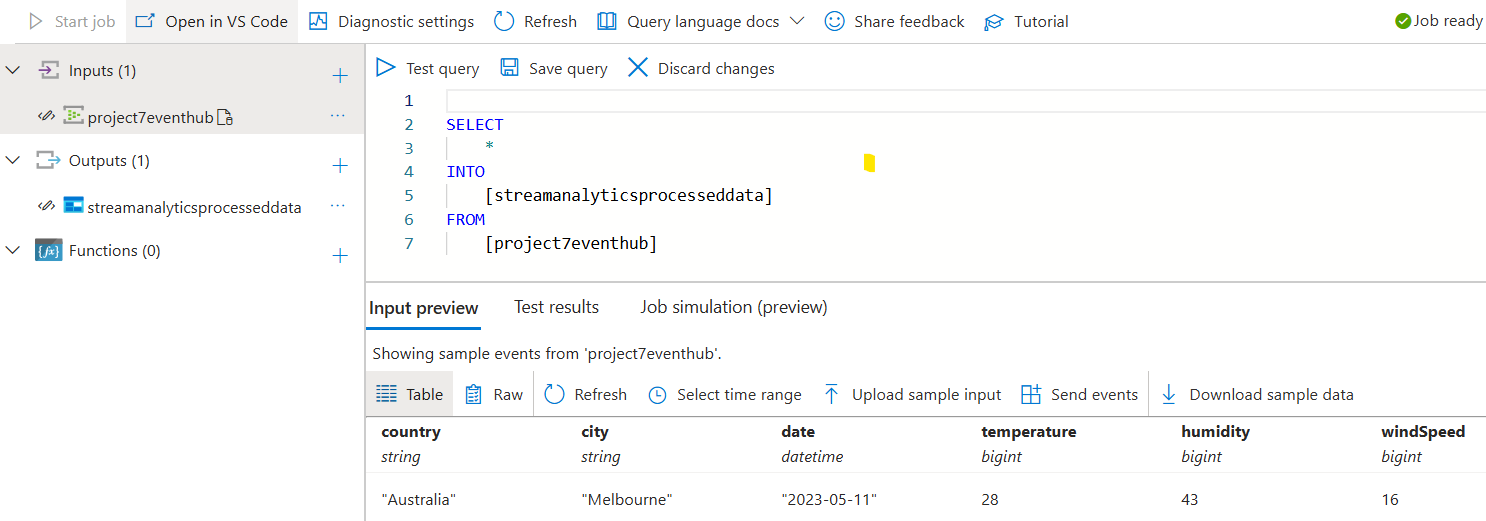


3. Using azure event hub and stream analytics:

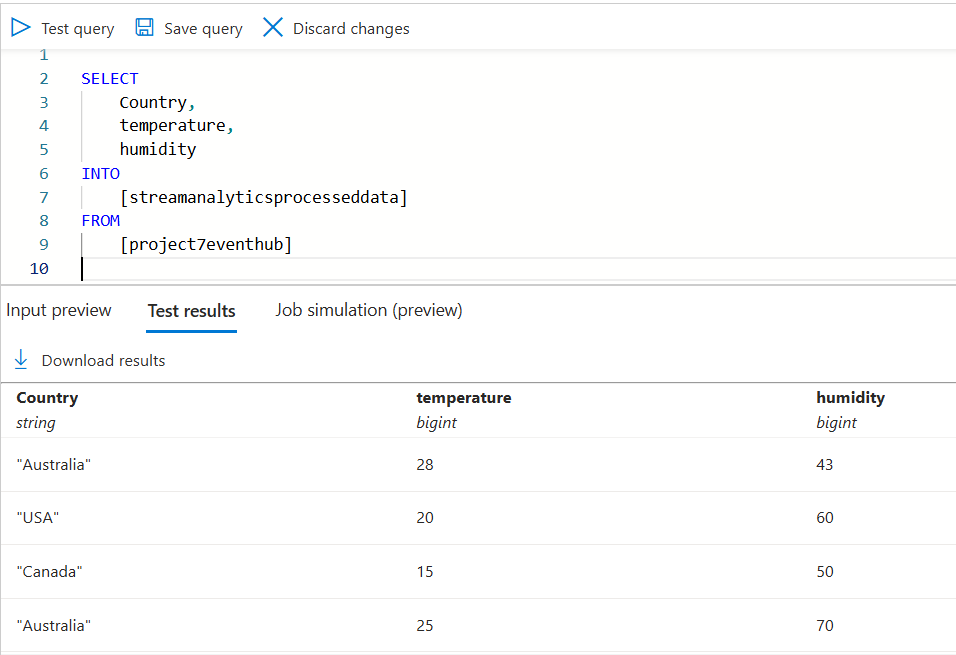
Send data to Azure Event Hub, processing it with Azure Stream Analytics, and output the stream analytics data to Azure Data Lake Storage (ADLS).

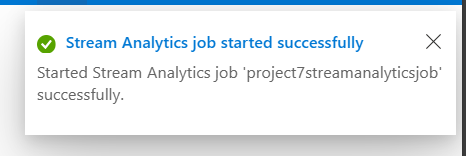
The input, ouput and query:



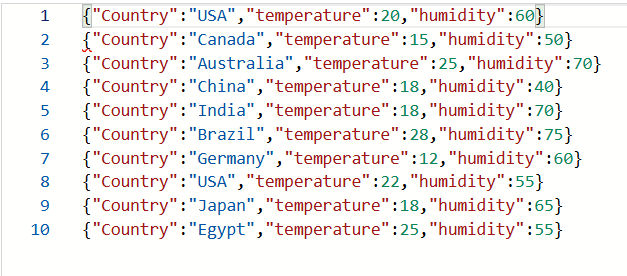
Test the query:



Run the stream analytics job:



Data gets loaded into ADLS:



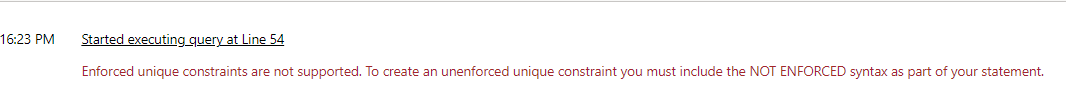
ERRORS FACED:

1. Mentioned Incorrect container name and got this error – corrected the name and it worked

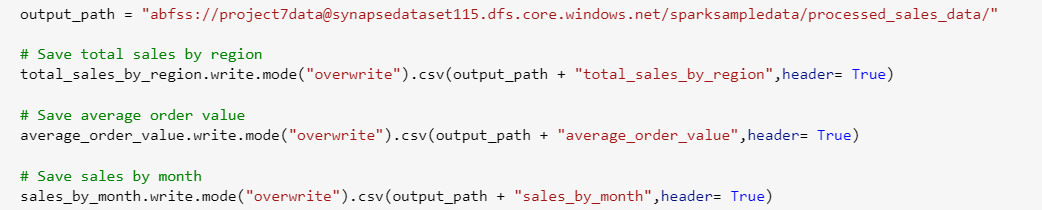


1. Got this error while adding alternate key to SalesData table: Added the syntax ‘NOT ENFORCED’ and it worked.





1. Forgot to mention Header = True in a statement and data was saved without column names in ADLS: corrected it and worked



1. Created Event hub to integrate with stream analytics for real time data analytics, but could not create dedicated pool.

