systems

Internship Report

Prepared by: Usama Yazdani

Submitted to : Sir Matteen Sajjad Daar

Date : 10/18/2024

Table of Contents:

Project 1: **Bing News Data Analytics - End-to-End Azure Data Engineering using Microsoft Fabric**

Overview Of Microsoft fabric

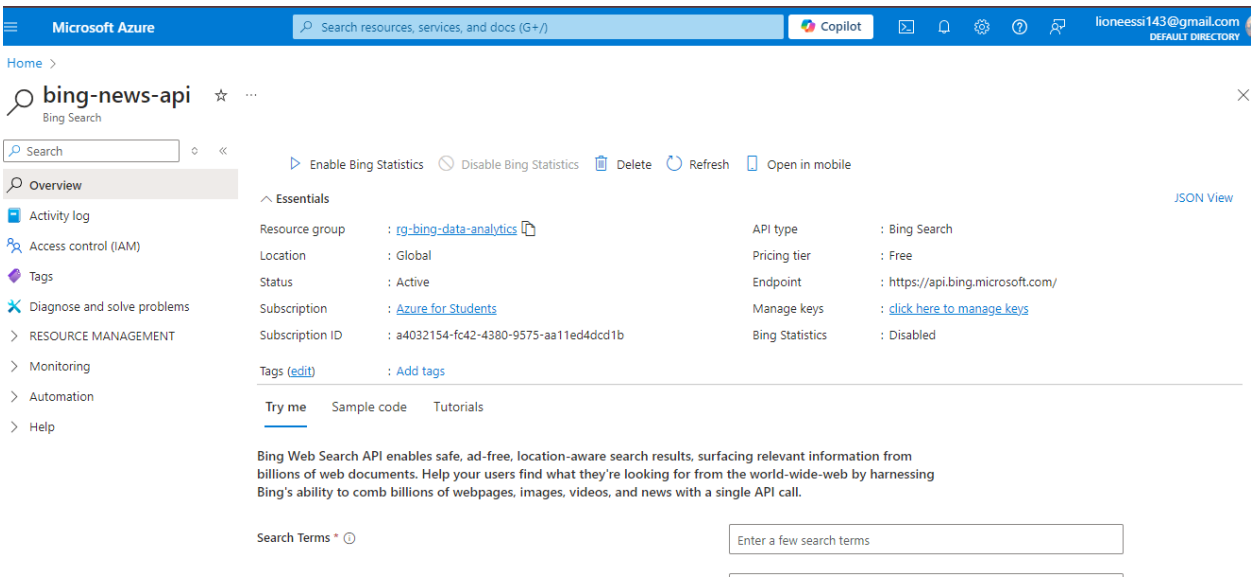
Microsoft Fabric is a comprehensive, unified analytics platform that brings together Azure Data Factory, Power BI, Synapse Analytics, and data lakes, making it an ideal solution for end-to-end data engineering and analytics workflows. It provides a single pane of glass for all data operations, allowing seamless integration between data ingestion, transformation, storage, and visualization.

Step 1: Data Gathering

To begin the project, I used **Microsoft Azure Marketplace** to source data. The marketplace offers a free **Bing News API**, which I used to retrieve real-time news articles. I connected to the API and configured it to gather news based on specific topics and keywords.

API Configuration: Used the Bing News API from Azure Marketplace, filtering data for relevant news categories.

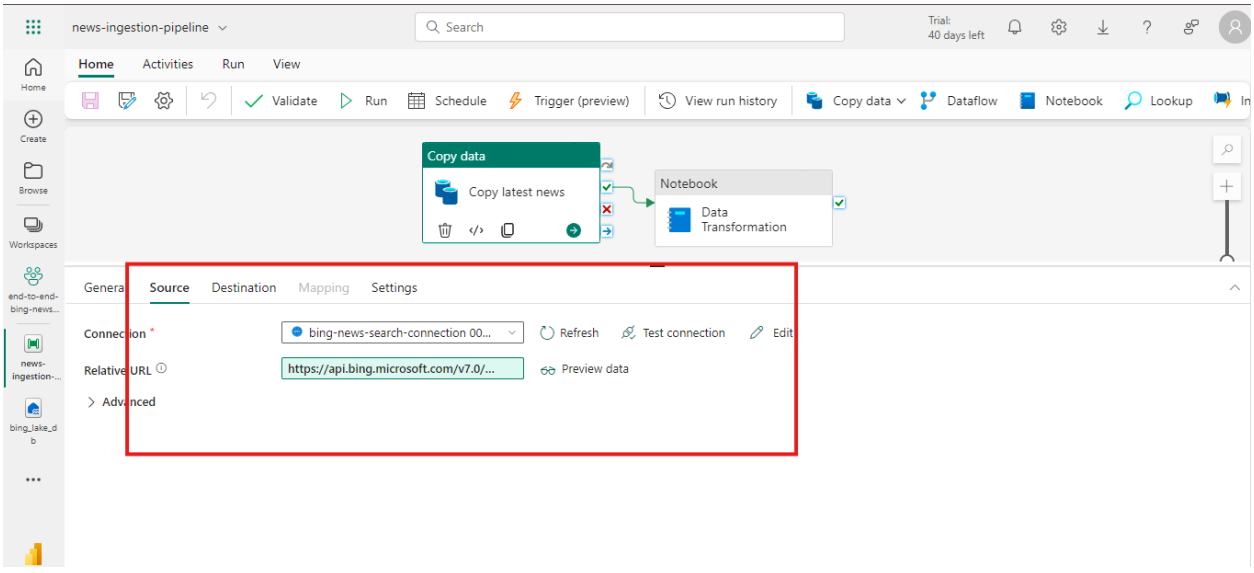
Data Storage Format: The retrieved data was stored in JSON format to ensure flexibility and ease of transformation.

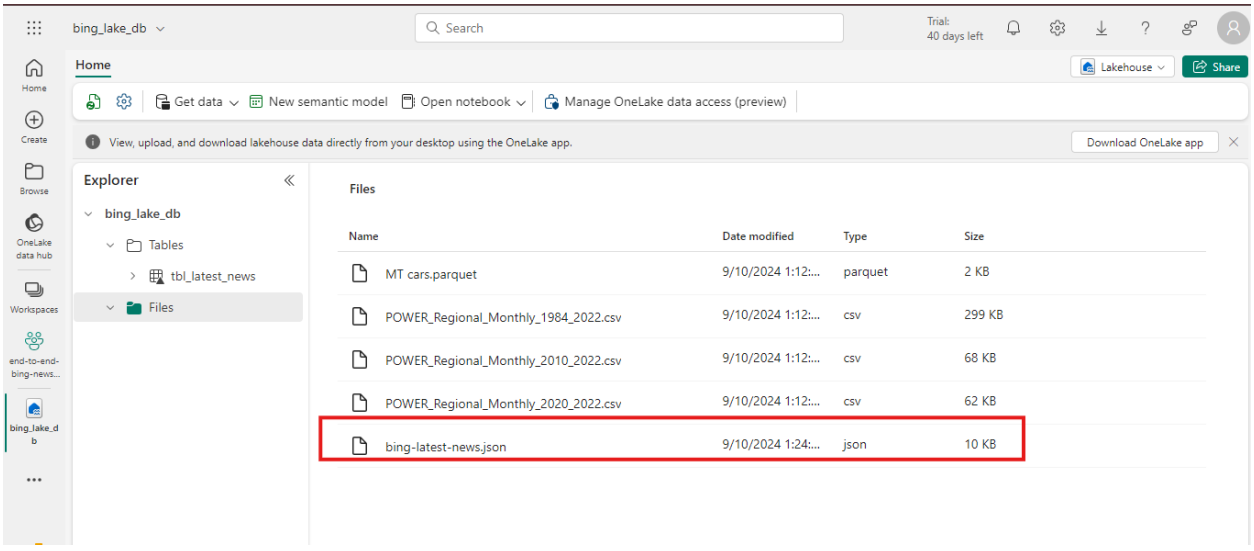


Step 2: Data Ingestion and Storage in Lakehouse

After fetching the data via the API, I utilized **Microsoft Fabric's Data Factory** to load the data into a **Lakehouse**. The Data Factory allowed for smooth orchestration of the pipeline, ensuring that the data was ingested in real-time.

- **Lakehouse Creation:** The JSON files were stored in **Microsoft Fabric's Lakehouse**, which served as the primary data storage location.
- **Data Organization:** The Lakehouse enabled structured storage, making it easier to perform downstream transformations.

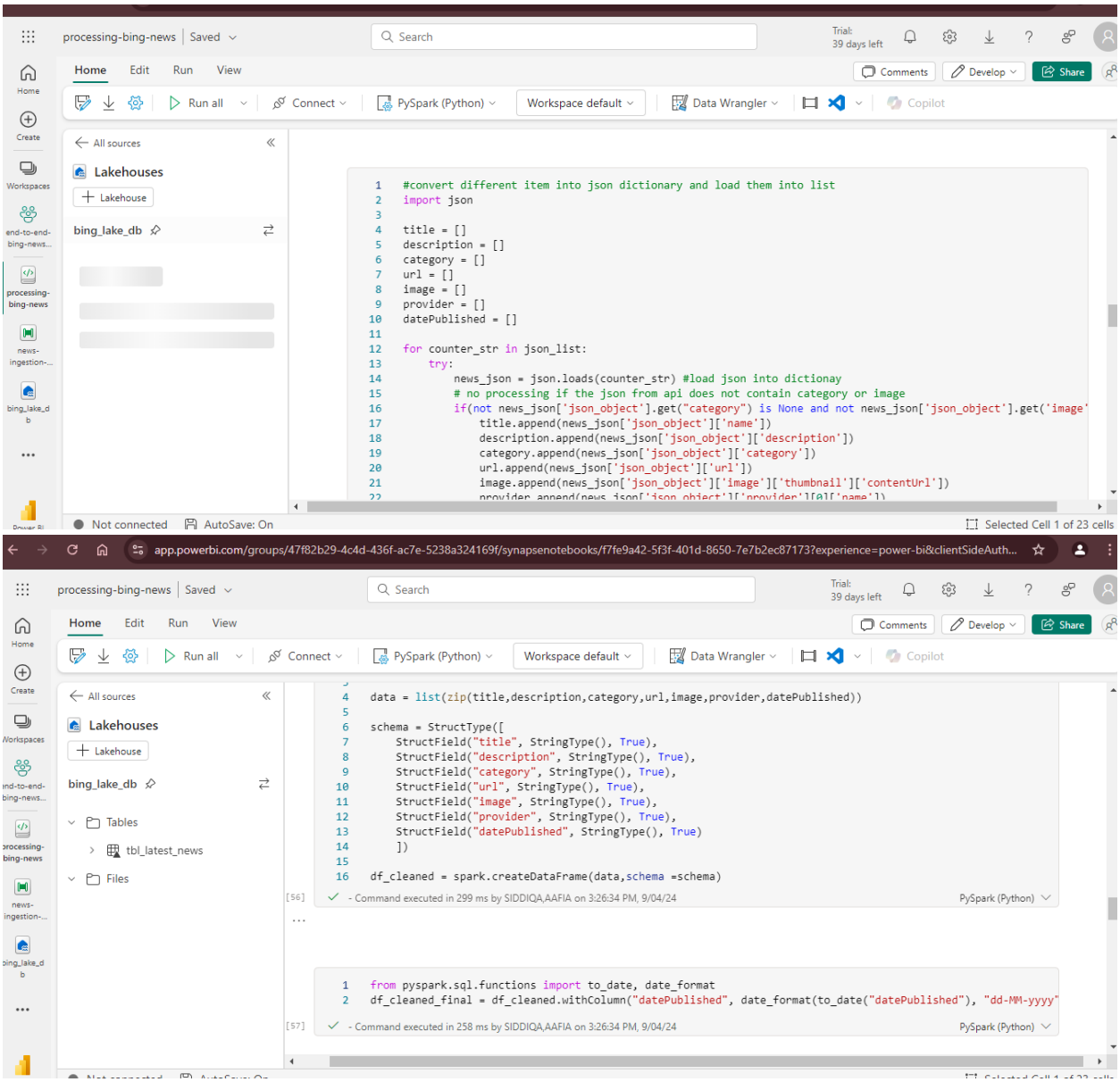




Step 3: Data Preprocessing using Synapse Spark Notebooks

Once the data was stored in the Lakehouse, I used **Azure Synapse Data Engineering** to preprocess and clean the data. This was done through a **Synapse Spark Notebook** where I implemented the necessary transformations in **Python** using Spark.

- **Data Cleaning:** Removed any irrelevant fields, handled missing data, and optimized the structure for further processing.
- **Preprocessing:** Applied transformations such as filtering based on certain criteria (e.g., news date range) and formatting fields.
- Design a schema : I design a schema related to news and also create a view there so the new news should not overwrite the existing one.

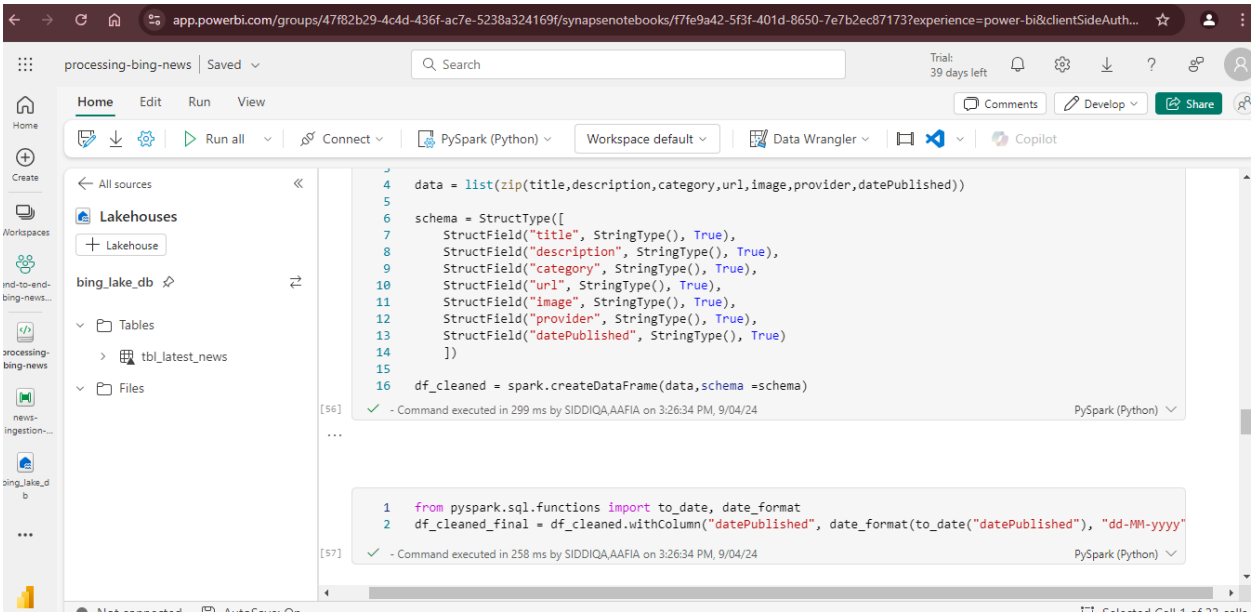


Step 4: Data Modeling – Creating Dimensions and Fact Tables

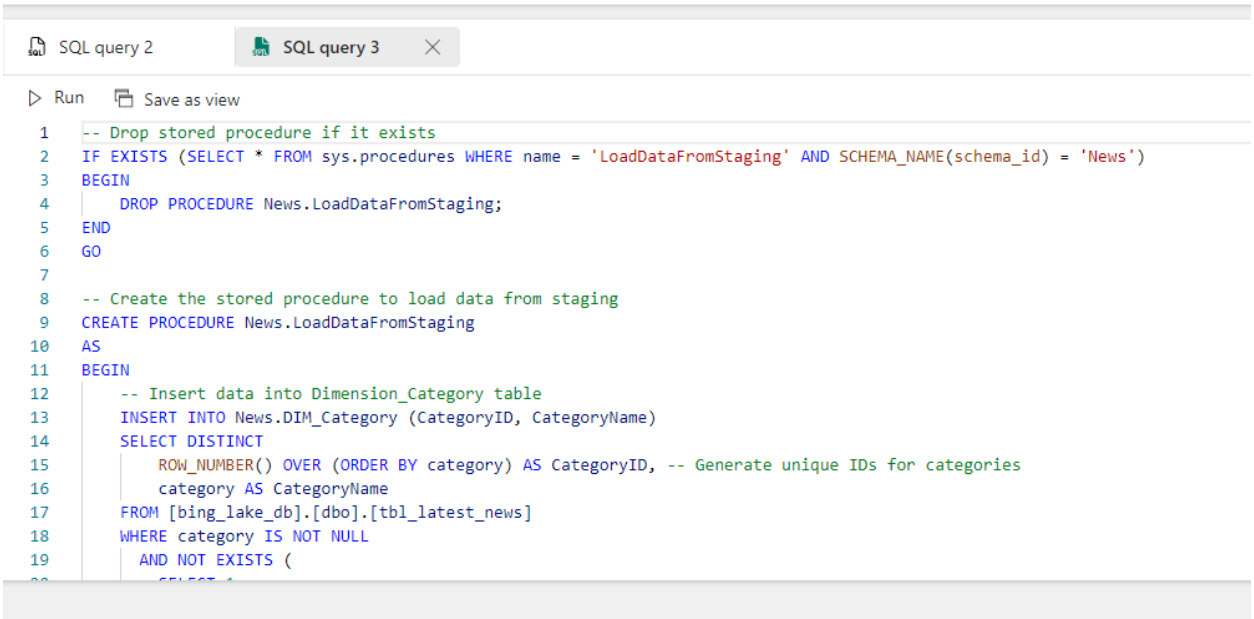
After preprocessing, the next step was to model the data in a structured form, suitable for reporting. I used **T-SQL** to define **dimension tables** and a **fact table**, ensuring the data was ready for analysis in the warehouse.

- **Fact Table:** Created a fact table that captures the key metrics (e.g., news image, Url, etc.).

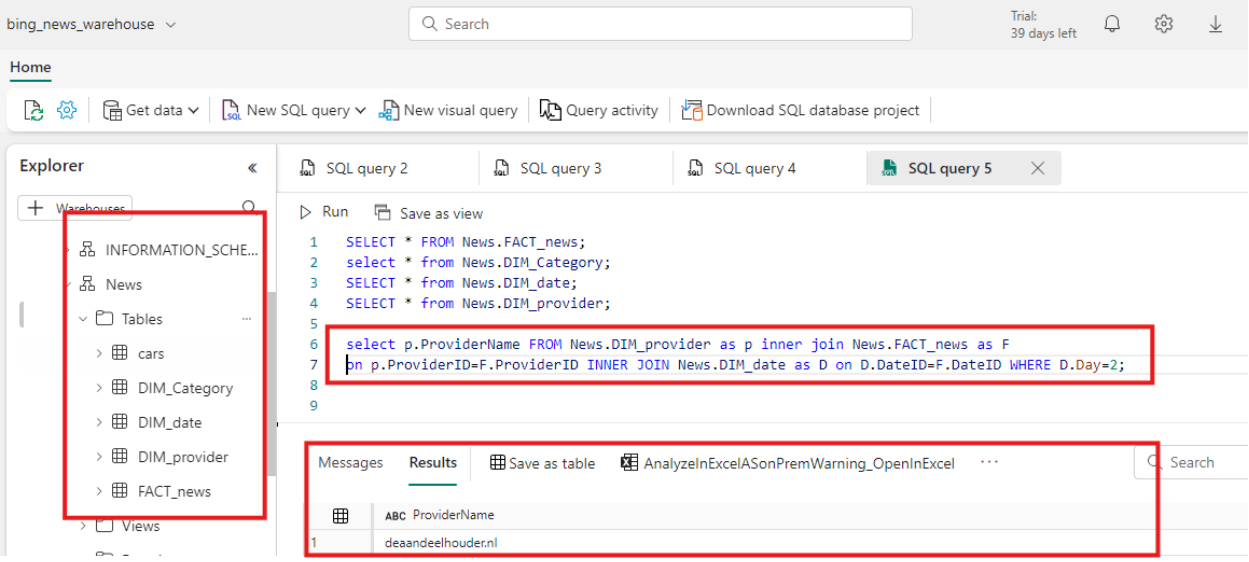
- **Dimension Tables:** Established dimension tables to categorize data (e.g., Categories, Provider, etc.).
- **Warehouse:** All the data was then loaded into a **Data Warehouse** using Microsoft Fabric.



After creating a table then insert data from lakehouse to warehouse :



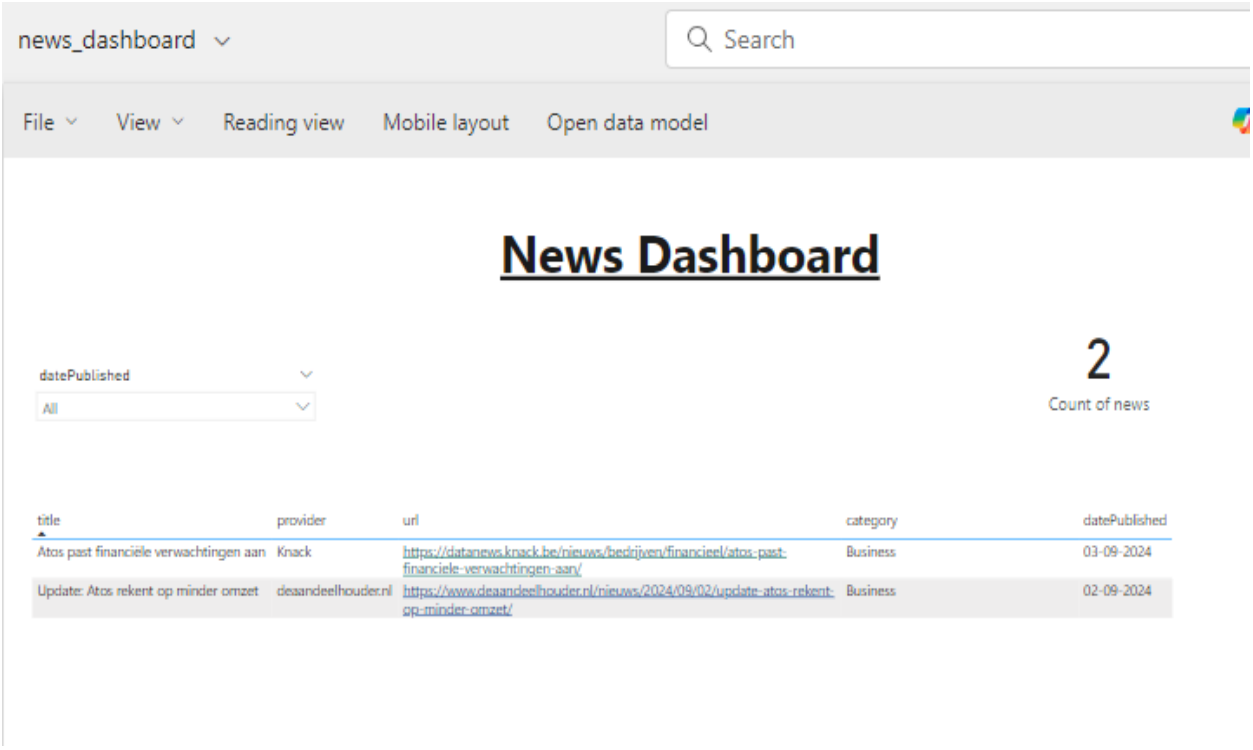
And running query on it :



Step 5: Creating the Semantic Model and Power BI Report

To visualize the data, I built a **semantic model** that defined relationships between fact and dimension tables, making it easier to interpret the data. Afterward, I connected this semantic model to **Power BI** for reporting.

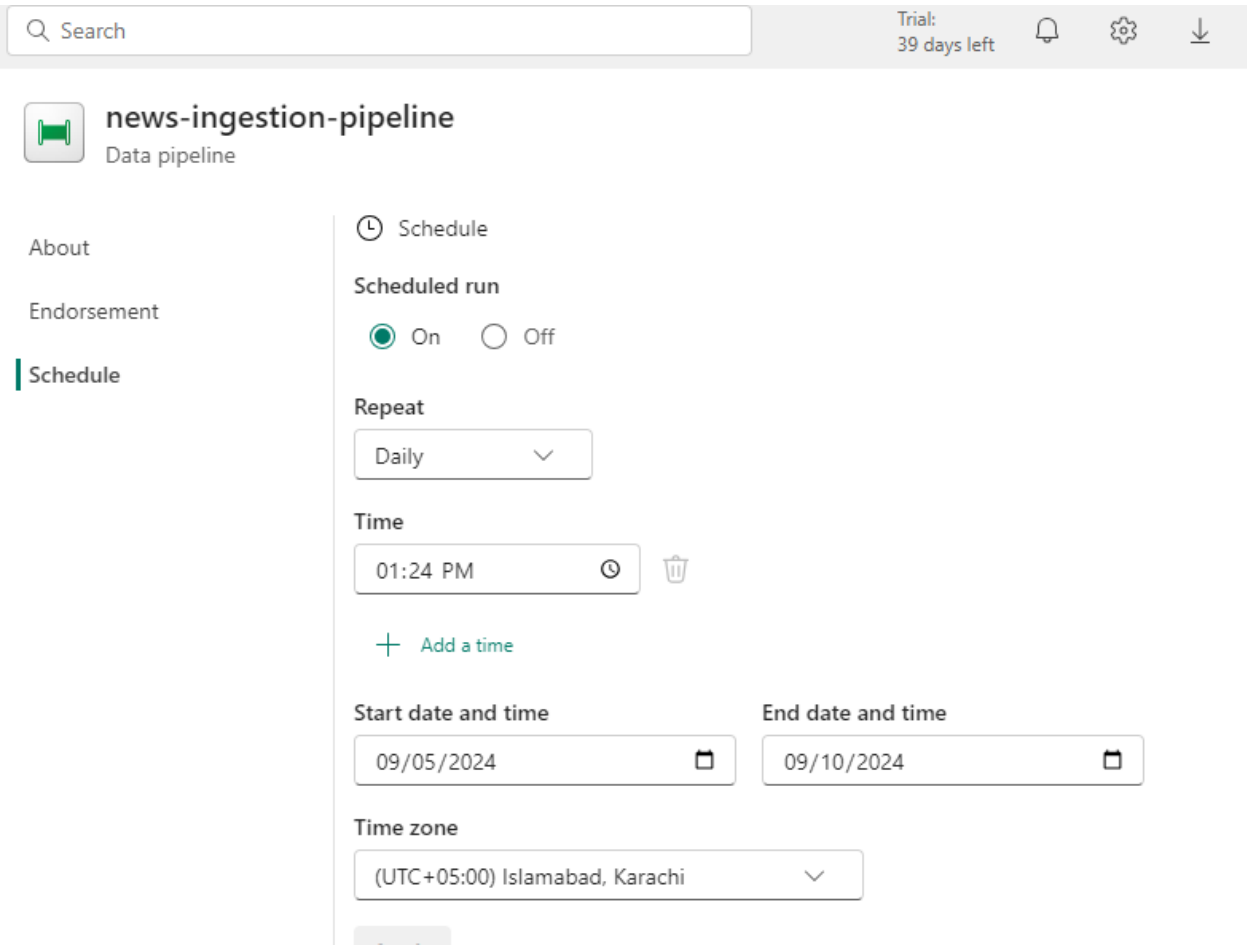
- **Semantic Model:** Created a model that mapped the relationships between the datasets.
- **Power BI Dashboard:** Used Power BI to generate insightful reports, including trending news topics, sentiment analysis, and category breakdowns.



Step 6: Automation and Orchestration

To ensure the data was always up-to-date, I implemented an **automated orchestration pipeline** using **Microsoft Fabric's Data Factory**. This pipeline ran every **1 hour** to fetch the latest news data from the API and update the existing reports and dashboards.

- **Orchestration:** The process was automated to run at regular intervals, ensuring that the latest news data was always available for analysis.
- **Scheduling:** Implemented a trigger that initiates the pipeline every hour to update the warehouse and refresh the Power BI reports.

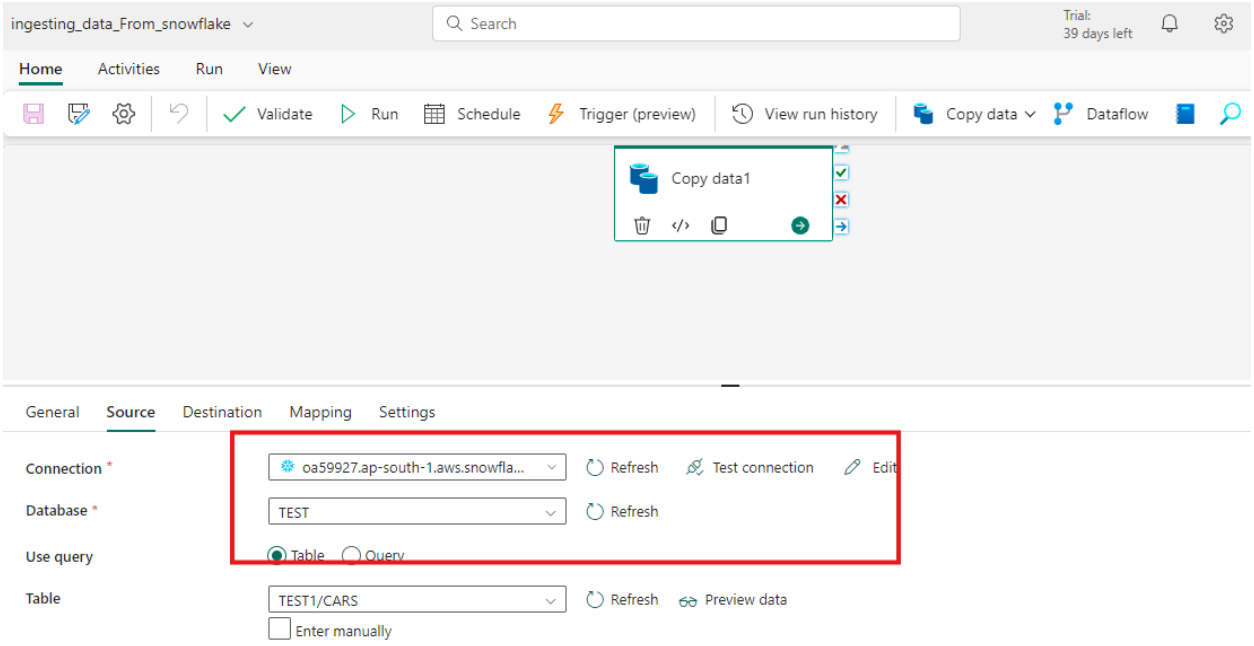


----------End ----------

Additional Tasks in Microsoft Fabric

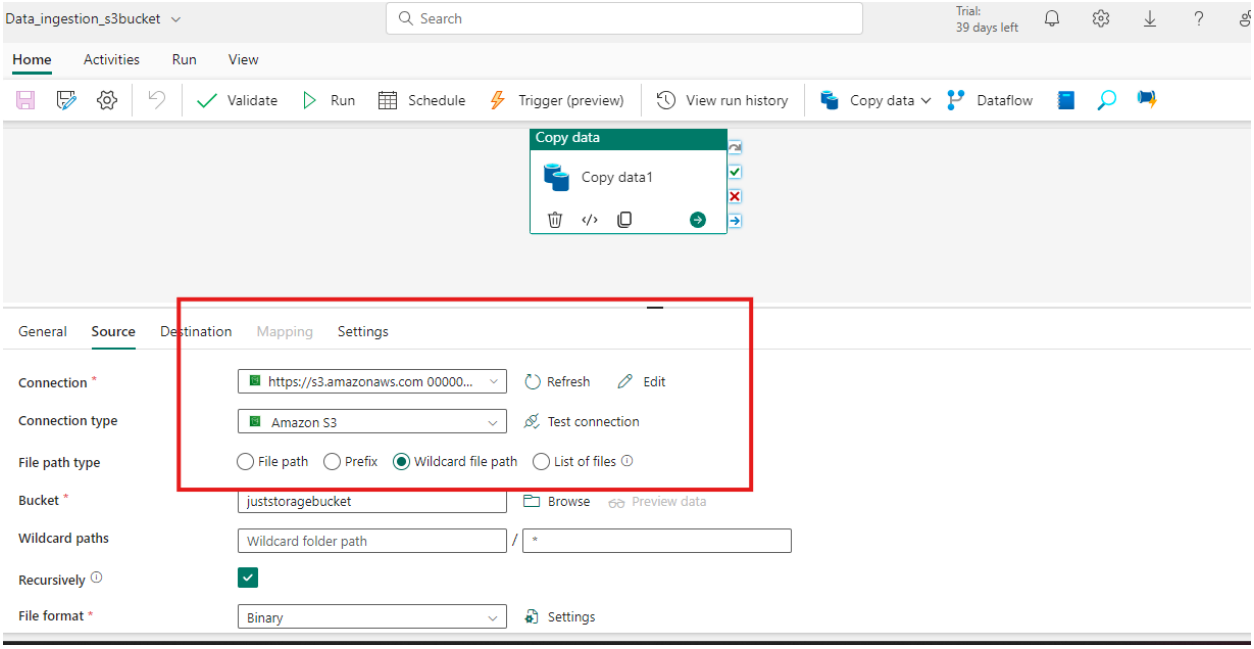
Besides the Bing News project, I worked on a few other tasks assigned by my manager, including:

1. **Data Ingestion from Snowflake:**
    a. I ingested data from Snowflake into Microsoft Fabric using the **Data Factory**, integrating Snowflake's structured data into the Lakehouse for further processing.



Data Ingestion from AWS S3:

- Similarly, I fetched data from **AWS S3 buckets** and ingested it into the Microsoft Fabric environment for analysis.



Conclusion and Final Notes

Through this project, I gained hands-on experience with **end-to-end data engineering** on Microsoft Fabric, working with tools like Azure Synapse, Power BI, and automated orchestration. The integration of real-time data sources and building of reusable data models provided valuable insights, and the automation ensured that the reports were consistently up-to-date without manual intervention.

Power BI - Data Professional Survey Breakdown

Introduction:

In this project,I work on survey dataset which is related to people job and their related things.I used power bi to create a interactive dashboard where put can have multiple functionality to do analysis.

The focus of the analysis was to break down key insights like demographics, salary trends, job satisfaction, and programming preferences.

. Data Source:

- The dataset contained fields like job titles, salary, country of origin, favorite programming languages, and happiness metrics.

Key Metrics and Visualizations:

- **Count of Survey Takers**
- **Count of Survey Takers**
- **Favorite Programming Languages**:
- **Difficulty to Break into the Data Profession**:
- **Happiness Metrics**:

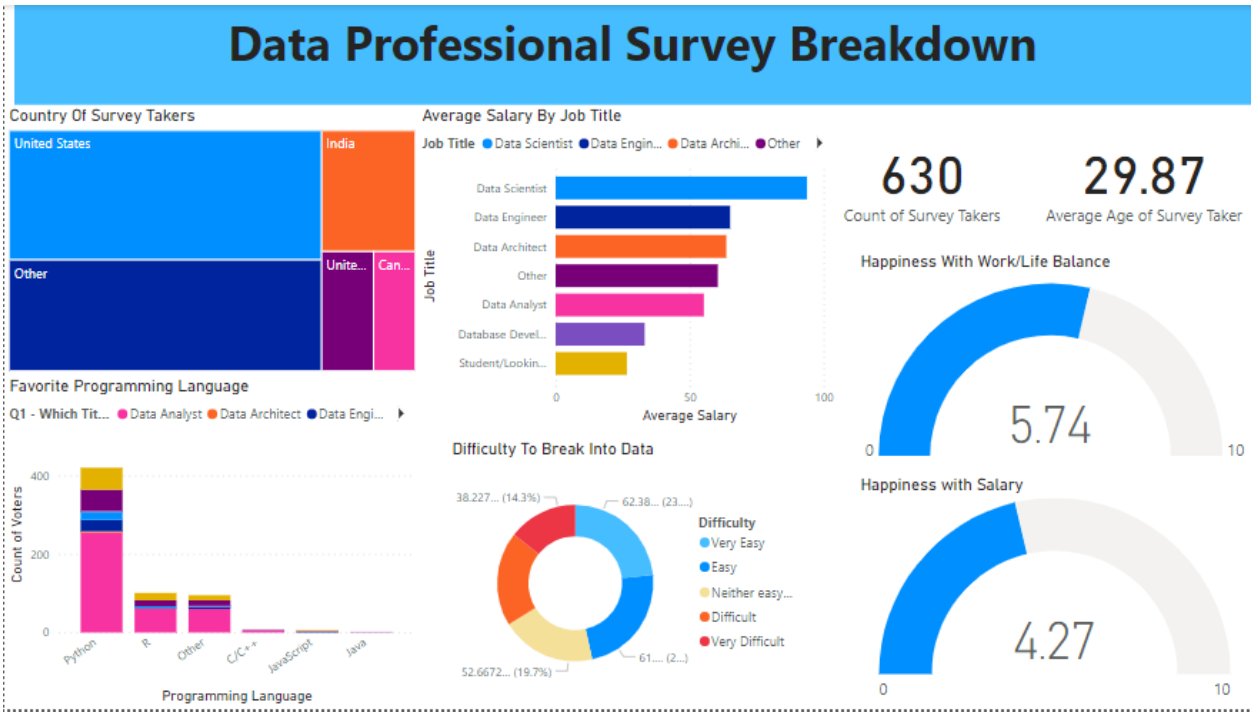
Insight :

- U.S. and India dominate the data profession in terms of participation.
- Data Scientists tend to have higher average salaries than Data Engineers.
- Python remains the top programming language across most job roles.
- A significant number of professionals find it moderately difficult to break into the field.

- Overall happiness with salary is moderate, with work/life balance ratings being relatively higher.

Conclusion:

State that this analysis can help organizations or job seekers get an overview of what to expect in terms of salary, challenges, and satisfaction within the industry.



----------> Thank You 😊 <----------