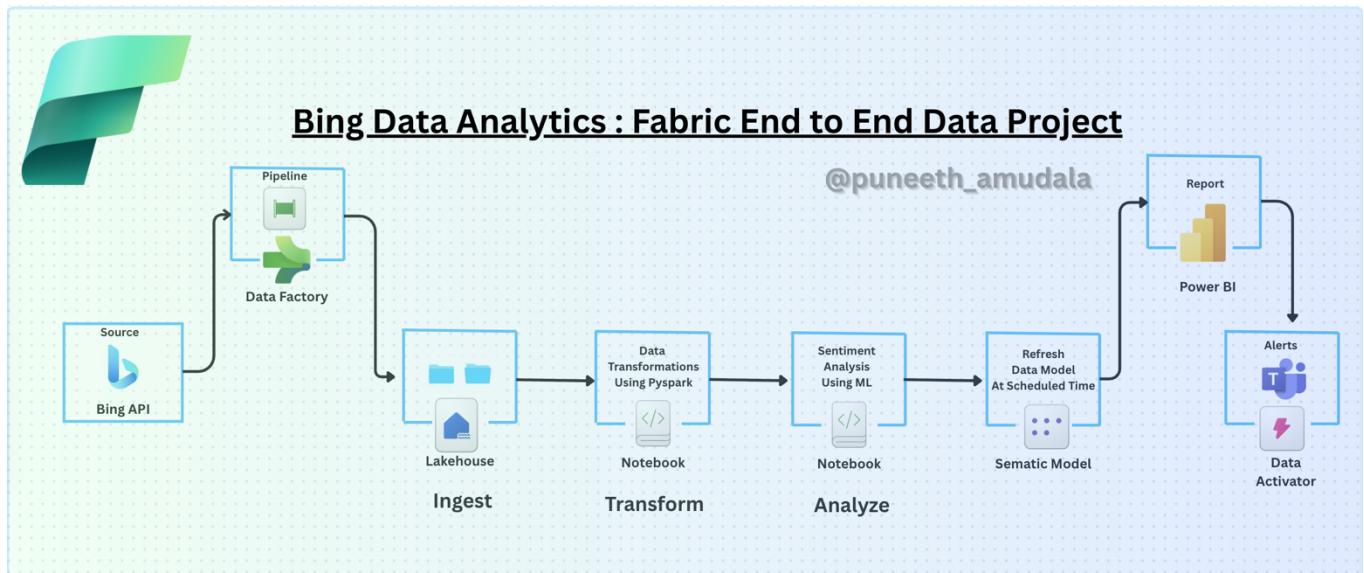


Fabric Data Project



1. Project Overview

This project demonstrates a complete data engineering workflow using **Microsoft Fabric** by ingesting real-time data from the **Bing News Search API**, transforming and applying **Sentiment detection using Machine Learning** with **Pyspark Notebooks**, and visualizing insights using **Power BI**. A pipeline was also developed to **schedule daily runs** and send **real-time alerts** for specific triggers.

2. Workspace and Lakehouse Setup

- Created a new Microsoft Fabric **Workspace**
- Created a **Lakehouse** to store both raw files and transformed tables

The screenshot shows the Microsoft Fabric workspace interface for the **bing-data-analytics-ws** workspace. The left sidebar includes links for Home, Workspaces, OneLake, Monitor, Real-Time, Workloads, and the current workspace. The main area displays a task flow builder with the following elements:

- Task Flow Title:** **bing-data-analytics-ws**
- Actions:** + New item, New folder, Import, Migrate
- Top Bar Buttons:** Create deployment pipeline, Create app, Manage access, Workflows, List view, Trial: 10 days left, Help, Filter by keyword, Filter, List view, Grid view.
- Central Area:** A placeholder for task flows with the text "Choose from predesigned task flows or add a task to build one".
- Buttons:** Select a predesigned task flow, Add a task, Import a task flow.

3. Data Ingestion (Bronze Layer)

3.1 Pipeline Creation

- Created a new **Data Factory pipeline**
- Used the **REST API connector** to connect to Bing News API
- Added required headers (e.g., Ocp-Apim-Subscription-Key)
- Configured parameters like:
- q=latest+news&count=100&freshness=Day&mkt=en-IN

3.2 Saving Raw Data

- Target location: Lakehouse > Files section
- Saved data in **JSON** format
- Verified pipeline run status and inspected saved files

4. 🖌️ Data Transformation (Silver Layer)

- Created a new PySpark Notebook
- Connected the Lakehouse and loaded JSON data
- Performed Data Cleaning and Data Transformations like flattening,
- Incrementally Loaded the transformed data into a Delta Table named today_news_in_India

The screenshot shows the Databricks workspace interface. The left sidebar lists workspaces: Home, OneLake, bing-data-analytics-ws, data_transformation..., bing_data_lakehouse, pi-end-to-end-bing..., Bing-News-Report, and daily_new_dashboard. The main area is titled "data_transformation_notebook | Saved". The top navigation bar includes Home, Edit, AI tools, Run, View, Connect (PySpark (Python)), Environment (Workspace default), Data Wrangler, and Copilot. A search bar at the top right says "Search". A message in the center states: "Other people in your organization may have access to notebooks and Spark job definitions in this workspace. Carefully review this item before running it." Below this, the "Extract" section is expanded, showing 3 cells hidden. The "Transform" section is collapsed, and the "Load" section is collapsed. There are buttons for "Edit", "Run all", and "Share" on the right.

This screenshot continues the notebook's execution. The "Importing functions and types" section is expanded, showing the following code in cell [1]:

```
1 from pyspark.sql.types import *
2 from pyspark.sql.functions import *
```

A status message indicates: "- Session ready in 10 sec 368 ms. Command executed in 364 ms by PuneethFabrics on 2:54:27 PM, 7/25/25". The "PySpark (Python)" tab is selected. The "Transform" section is collapsed below.

The "Exploding the Json column" section is expanded, showing the following code in cell [3]:

```
1 explodeddf= df.select(explode(df["value"]).alias('jsonobject'))
```

A status message indicates: "- Command executed in 366 ms by PuneethFabrics on 2:54:30 PM, 7/25/25". The "PySpark (Python)" tab is selected. The "Transform" section is collapsed below.

Home Edit AI tools Run View Comments History Develop Share

Workspaces OneLake bing-data-analytics-ws data_transformation... bing_data_lakehouse

Explorer

- Data items**
- Resources
- + Add data items

Flattening Data Frame

```

1 flattened_df= explodeddf.select(col('jsonobject.name').alias('title'), \
2 col('jsonobject.description').alias('description'), \
3 col('jsonobject.category').alias('category'), \
4 col('jsonobject.image.thumbnail.contentUrl').alias('image'), \
5 col('jsonobject.url').alias('url'), \
6 explode_outer(col('jsonobject.provider')).alias('provider'), \
7 col('jsonobject.datePublished').alias('datePublished'))

```

[5] ✓ - Command executed in 350 ms by PuneethFabrics on 2:54:31 PM, 7/25/25 PySpark (Python) ↴

Workspaces

OneLake bing-data-analytics-ws data_transformation... bing_data_lakehouse il-end-to-end-bing... [!]

Explorer

- Data items**
- Resources
- + Add data items

Items

- > bing_data_lakehouse

flattened_df.printSchema()

[8] ✓ - Command executed in 365 ms by PuneethFabrics on 2:54:34 PM, 7/25/25 PySpark (Python) ↴

```

root
|-- title: string (nullable = true)
|-- description: string (nullable = true)
|-- category: string (nullable = true)
|-- image: string (nullable = true)
|-- url: string (nullable = true)
|-- provider: struct (nullable = true)
|   |-- _type: string (nullable = true)
|   |-- image: struct (nullable = true)
|       |-- thumbnail: struct (nullable = true)
|           |-- contentUrl: string (nullable = true)
|       |-- name: string (nullable = true)
|-- datePublished: string (nullable = true)

```

Workspaces

OneLake bing-data-analytics-ws data_transformation... bing_data_lakehouse il-end-to-end-bing... Bing-News-Report daily_new_dashboard ...

Explorer

- Data items**
- Resources
- + Add data items

Items

- > bing_data_lakehouse

Load

```

1 # Incremental Load
2
3 from pyspark.sql.utils import AnalysisException
4
5 try:
6
7     table_name = 'bing_data_lakehouse.Today_news_in_india'
8
9     df.write.format("delta").saveAsTable(table_name)
10
11 except AnalysisException:
12
13     print("Table Already Exists")
14
15     df.createOrReplaceTempView("vw_df_final")
16     spark.sql(f"""
17         Merge into {table_name} t
18             using vw_df_final s
19                 on t.url= s.url
20                     when matched and
21                         s.title <> t.title or
22                             s.description <> t.description OR
23                                 s.category <> t.category
24                                     then update set *
25                                         when not matched then
26                                         INSERT *""")
```

[18] ✓ - Canceled by PuneethFabrics on 11:20:46 AM, 7/26/25 PySpark (Python) ↴

bing_data_lakehouse

Home Get data New semantic model Open notebook Manage OneLake data access (preview) Update all variables

10 days left

Lakehouse Share

Explorer

today_news_in_india

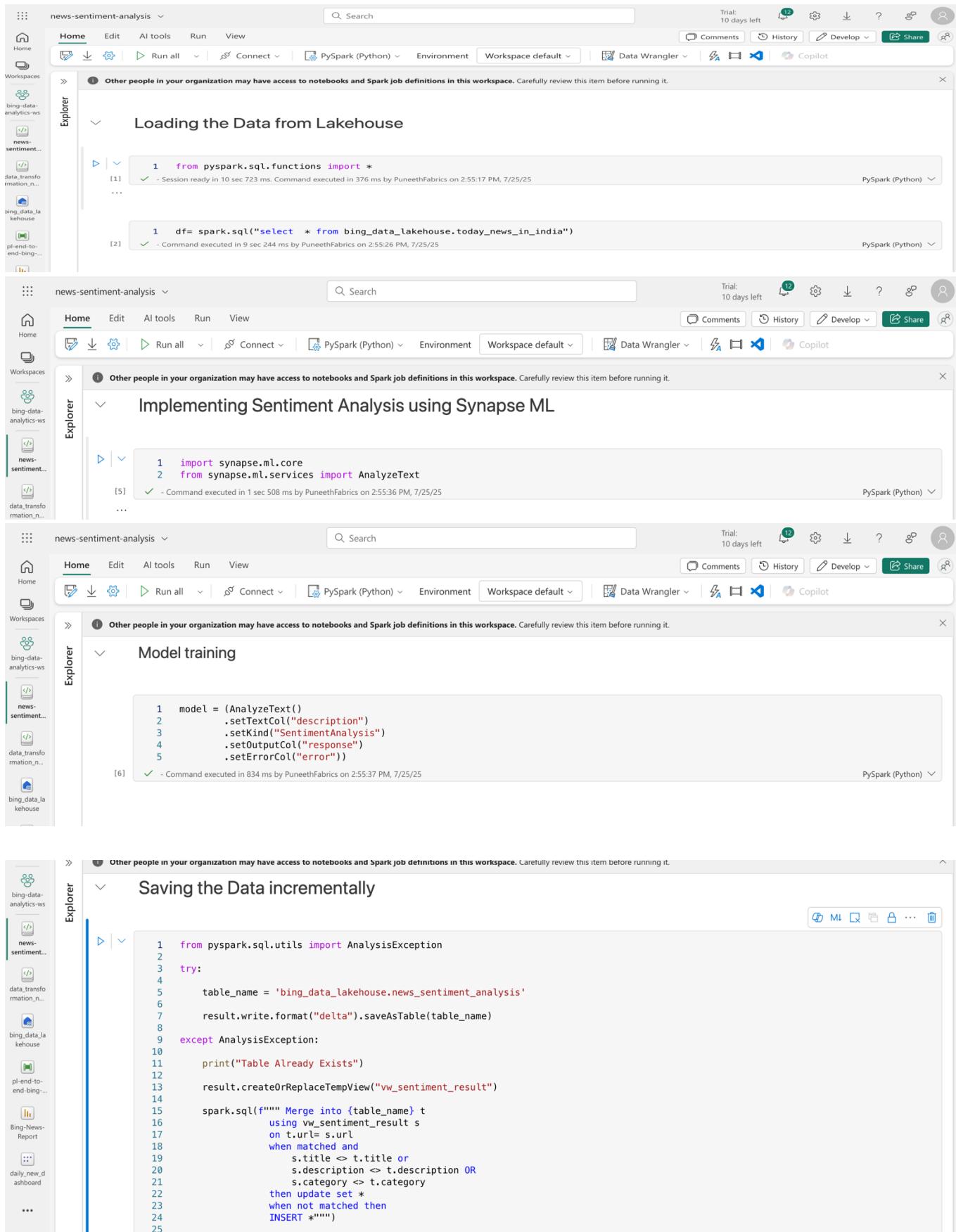
Showing 92 rows

Table view

	ABC title	ABC description	ABC category	ABC image	ABC url	ABC providername	date
1	India committed to sup...	Prime Minister Narendra...	World	https://www.bing.com/t...	https://www.hindustant...	Hindustan Times	7/26/2025 12:00:00
2	Latest news bulletin Ju...	Catch up with the most...	Entertainment	https://www.bing.com/t...	https://www.msn.com/...	Euronews on MSN.com	7/25/2025 12:00:00
3	Breaking News LIVE: 'In...	Kargil Vijay Diwas is ob...	World	https://www.bing.com/t...	https://www.timesnown...	Times Now	7/26/2025 12:00:00
4	2025 NFL training camp...	Welcome to The Athleti...	Sports	https://www.bing.com/t...	https://www.nytimes.co...	The New York Times	7/26/2025 12:00:00
5	Tired Of Your Office Vie...	Seychelles, the postcar...	LifeStyle	https://www.bing.com/t...	https://www.timesnown...	Times Now	7/26/2025 12:00:00
6	Saiyara Box Office Coll...	If the momentum holds...	Entertainment	https://www.bing.com/t...	https://www.livemint.co...	Mint	7/26/2025 12:00:00
7	Thailand-Cambodia Bor...	The UN Security Council...	World	https://www.bing.com/t...	https://www.indianexpress.c...	The Indian Express	7/26/2025 12:00:00
8	F1 news: Two drivers u...	Max Verstappen has ad...	Sports	https://www.bing.com/t...	https://www.express.co...	Daily Express	7/25/2025 12:00:00
9	Bajaj Finance vs Jio Fina...	The quarterly results fo...	Business	https://www.bing.com/t...	https://www.msn.com/...	Mint on MSN.com	7/26/2025 12:00:00
10	Prada heels or Rajastha...	Prada faces criticism ag...	LifeStyle	https://www.bing.com/t...	https://www.msn.com/...	The Times of India on ...	7/26/2025 12:00:00

5. Sentiment Analysis (Gold Layer)

- Created another notebook for **Machine Learning-based sentiment detection analysis**
- Loaded the cleaned data table `today_news_in_India`
- Classified the articles as *Positive*, *Negative*, or *Neutral*
- Saved the results as a new Delta Table named `news_sentiment_analysis`



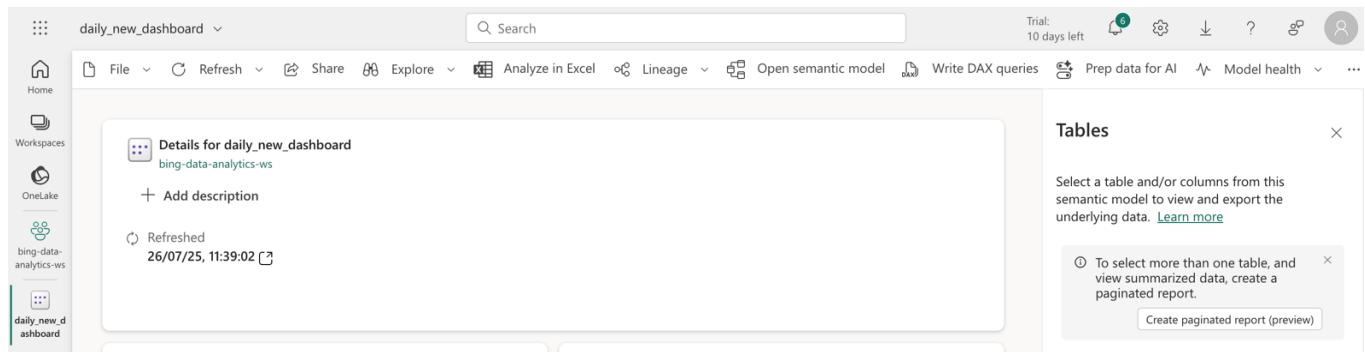
The image shows four stacked screenshots of the Azure Data Studio interface, each displaying a PySpark notebook. The notebooks are titled:

- Loading the Data from Lakehouse**: Contains code to import functions and select data from the `today_news_in_India` table.
- Implementing Sentiment Analysis using Synapse ML**: Contains code to import `synapse.ml.core` and `AnalyzeText`.
- Model training**: Contains code to define a `AnalyzeText` model with specific parameters.
- Saving the Data incrementally**: Contains code to handle `AnalysisException`, write the data to a Delta table, and merge it with an existing table using a temporary view.

The notebooks are part of a workspace named `news-sentiment-analysis`. The code snippets include line numbers and execution logs indicating successful command execution.

6. Power BI Reporting

- Created a **Semantic Model** called daily_news_dashboard on top of the gold layer
- Built a Power BI report using:
 - News Title, Sentiment, Source, Publish Time
 - Filters like date and keyword search



The screenshot shows the Power BI service interface. On the left, there's a navigation bar with icons for Home, Workspaces, OneLake, and a specific workspace named 'bing-data-analytics-ws'. Below these are two items: 'Details for daily_new_dashboard' and 'daily_new_dashboard'. The 'daily_new_dashboard' item is selected, indicated by a blue border. The main area displays the 'Details for daily_new_dashboard' page, which includes a 'Search' bar at the top right, a 'Trial: 10 days left' message, and various navigation links like File, Refresh, Share, Explore, Analyze in Excel, Lineage, Open semantic model, Write DAX queries, Prep data for AI, and Model health. Below the navigation is a section titled 'Details for daily_new_dashboard' with a 'bing-data-analytics-ws' icon. It shows a 'Refreshed' status from '26/07/25, 11:39:02'. To the right is a 'Tables' pane with instructions to 'Select a table and/or columns from this semantic model to view and export the underlying data.' It also includes a note about creating paginated reports and a 'Create paginated report (preview)' button.



The screenshot shows the 'Bing News Report' Power BI dashboard. At the top, there's a header with the Bing logo and the title 'Bing News Report'. Below the header, there are four large numerical cards: '0.00' (Positive Sentiment %), '32.00' (Negative Sentiment %), and '68.00' (Neutral Sentiment %). Each card has a corresponding color-coded background: green for positive, red for negative, and light blue for neutral. Below these cards is a table with the following columns: Image, News Title, News Provider, News URL, Category, Date Published, and Sentiment. The table contains five rows of news items:

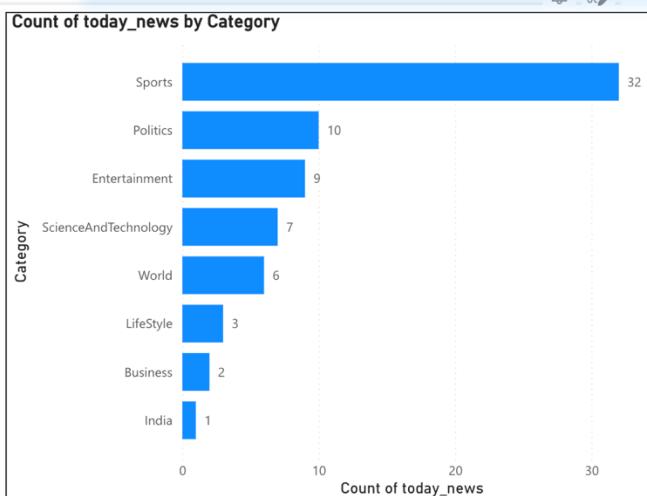
Image	News Title	News Provider	News URL	Category	Date Published	Sentiment
	India Live News Updates: CBI busts major international cyber fraud racket in Pune and Mumbai	The Economic Times	https://economictimes.indiatimes.com/news/newsblogs/india-news-live-updates-july-25-2025-weather-today-rains-himachal-bus-accident-parliament-monsoon-session-pm-modi-maldives-visit-uk-india-fta-trade-deal-congress-air-india-plane-crash-donald-trump-india-pakistan-ukraine-russia-plane-crash-thailand-cambodia-conflict/liveblog/122892046.cms	India	Friday, July 25, 2025	negative
	Why India-UK deal on critical minerals points to New Delhi's anxiety over Chinese grip	The Indian Express	https://indianexpress.com/article/business/india-uk-free-trade-agreement-critical-minerals-10148304/	ScienceAndTechnology	Friday, July 25, 2025	neutral
	Trump latest: Trump boards Air Force One for Scotland ahead of five-day visit	Sky News	https://news.sky.com/story/trump-latest-ghislaine-maxwell-meets-with-us-justice-department-nbc-reports-13398295	Politics	Friday, July 25, 2025	neutral
	Transfer news: Latest on Isak, Gyokeres and Trafford	BBC	https://www.bbc.com/sport/football/live/c80p0vdz289t	Sports	Friday, July 25, 2025	neutral
	Transfer news: Isak considering move away from Newcastle - latest	BBC	https://www.bbc.com/sport/football/live/e8e4zej8y0rt?page=4	Sports	Friday, July 25, 2025	neutral
	Anritsu Announced Its Latest 5G Mobile-Satellite	Electronics For You	https://www.electronicsforu.com/news/	ScienceAndTechnology	Friday, July 25, 2025	neutral



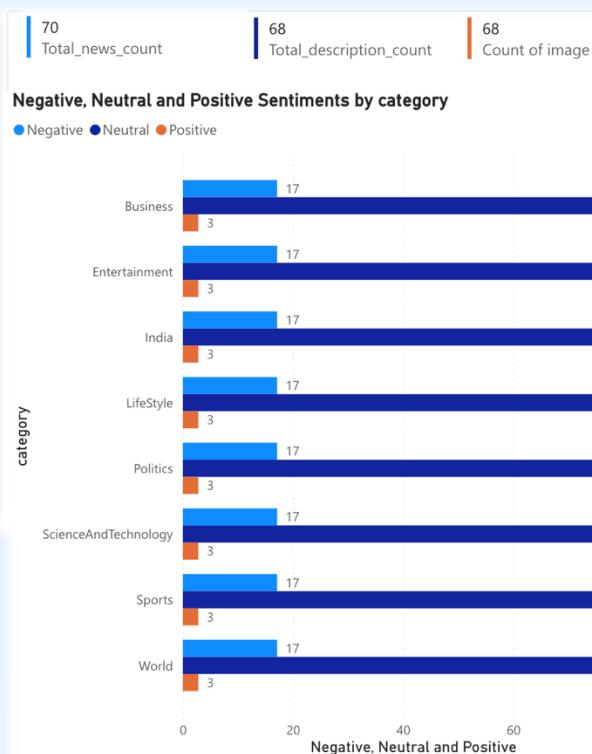
Bing News Report

Analysis

Summary



- At 32, Sports had the highest Count of today_news_in_india and was 3,100.00% higher than India, which had the lowest Count of today_news_in_india at 1.
- Sports accounted for 45.71% of Count of today_news_in_india.
- Across all 8 category, Count of today_news_in_india ranged from 1 to 32.



6.1 Scheduled Refresh

- Turned off Direct Lake refresh from the semantic model settings
- Refreshing the semantic model from Pipeline

Fabric bing-data-analytics-ws

General Dashboards Semantic models Workbooks Reports Dataflows

bing_data_lakehouse

daily_new_dashboard

Settings for daily_new_dashboard

[View semantic model](#)

This semantic model has been configured by Fabrics700@puneethkumaramudala7@gmail.onmicrosoft.com.

Last refresh succeeded: 26/07/2025, 11:39:02
[Refresh history](#)

▷ Semantic model description

▷ Gateway and cloud connections

▷ Data source credentials

▷ Parameters

▷ Query Caching

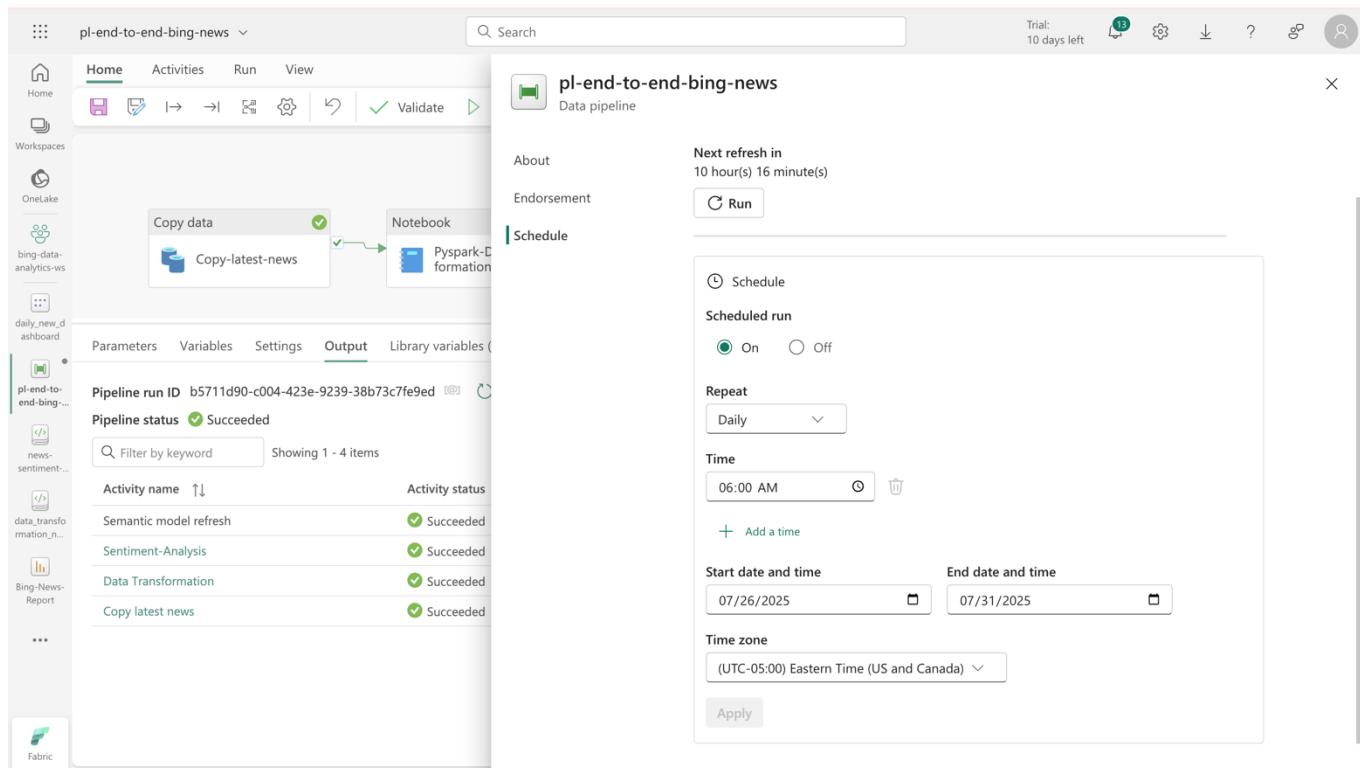
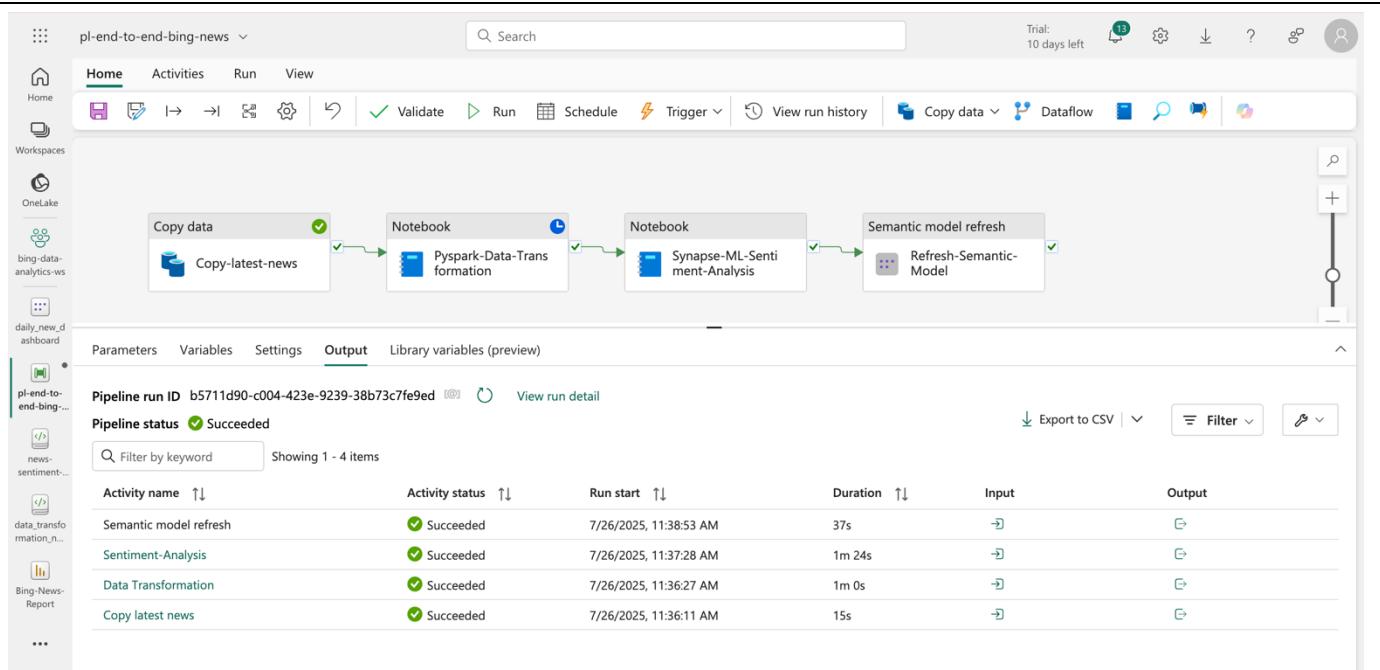
▷ Refresh

Keep your Direct Lake data up to date
Configure Power BI to detect changes to the data in OneLake and automatically update the Direct Lake tables that are included in this semantic model. [Learn more](#)

Off

7. Pipeline Orchestration

- Connected ingestion, transformation, and ML notebooks in a **single orchestrated pipeline**
- Scheduled to run **daily at 6 am for a period of time**



8. ⚡ Real-Time Alerts with Data Activator

- Enabled alerts via **Power BI Data Activator**
- Added an alert on visual elements like the number of negative news items
- Configured alert actions and schedules

9. End-to-End Testing

- Ran the full pipeline manually
- Verified successful refresh and updated visualizations

Activity name ↑↓	Activity status ↑↓	Run start ↑↓	Duration ↑↓	Input	Output
Semantic model refresh	Succeeded	7/26/2025, 11:38:53 AM	37s		
Sentiment-Analysis	Succeeded	7/26/2025, 11:37:28 AM	1m 24s		
Data Transformation	Succeeded	7/26/2025, 11:36:27 AM	1m 0s		
Copy latest news	Succeeded	7/26/2025, 11:36:11 AM	15s		

10. Resources

- Code Notebooks, Power BI Report File (PBIX) are available in : [Github-Link](#)

11. Conclusion

This project showcases how Microsoft Fabric can be used to design a full-scale data pipeline—from ingesting REST API data to transforming and analyzing it, to reporting and setting up actionable alerts. This solution is scalable, dynamic, and efficient for real-time data analytics.