# News Analytics on Microsoft Azure & Fabric

This project showcases an end-to-end News Analytics solution designed and implemented in Microsoft Azure and Microsoft Fabric. The workflow demonstrates data engineering, data science, and business intelligence capabilities by automating the ingestion, transformation, analysis, and visualization of news data. The solution is orchestrated using Data Factory pipelines and delivers actionable insights via Power BI dashboards with real-time alerts configured through Data Activator.

## 🔄 End-to-End Workflow

The solution is divided into five major steps that ensure a smooth data pipeline from ingestion to actionable insights:

* 1. **Data Ingestion** – Collect latest news data (World & Canada) using News API and ingest into Fabric Lakehouse in JSON format.
* 2. **Data Transformation** – Use Synapse Data Engineering (Spark Notebooks) to process raw JSON, flatten nested structures, and convert into clean Delta Tables. Implement incremental loading for efficiency.
* 3. **Sentiment Analysis** – Leverage SynapseML pre-trained text analytics models to classify articles into Positive, Negative, or Neutral sentiments.
* 4. **Data Reporting** – Build a Power BI dashboard connected to Lakehouse semantic models to visualize sentiment distribution and trends for the last 7 days.
* 5. **Alerts & Automation** – Configure Data Activator to trigger Teams/Email alerts when negative sentiment exceeds a threshold.

## 📋 Agenda

- Environment Setup

- Data Ingestion

- Data Transformation (Incremental Load)

- Sentiment Analysis (Incremental Load)

- Data Reporting

- Building Pipelines

- Setting up Alerts (Data Activator)

- End-to-End Testing

## ⚙️ Environment Setup

• Ensure a Microsoft Fabric license is available.  
• Create a resource group in Azure for API access (NewsAPI).  
• Set up Microsoft Fabric capacity via Power BI Service.  
• All subsequent steps (ingestion, transformation, ML, dashboards) are implemented in Microsoft Fabric.

## 📥 Data Ingestion

• Created a dedicated workspace in Power BI.  
• Built a Lakehouse Database to store ingested raw data in JSON format.  
• Configured Data Factory pipeline with REST API connector to fetch news data from NewsAPI (https://newsapi.org/v2/top-headlines).  
• Stored API results as JSON files in Lakehouse storage.  
• Orchestration included parameters for World News and Canada News queries.  
• Scheduled pipelines for continuous ingestion.

## 🛠️ Data Transformation

• Used Synapse Data Engineering notebook to read JSON files.  
• Selected and flattened required fields: title, description, content, author, url, urlToImage, source, publishedAt.  
• Exploded nested 'articles' JSON objects into tabular rows.  
• Converted 'publishedAt' into proper date format.  
• Implemented incremental loading using SQL Merge (Type 1 SCD) to handle duplicates and updates.  
• Final output stored as clean Delta Tables in Lakehouse.

## 💡 Sentiment Analysis

• Used Synapse Data Science notebooks with SynapseML for machine learning.  
• Applied pre-trained sentiment analysis model on the 'description' column.  
• Generated sentiment labels (Positive, Negative, Neutral) for each article.  
• Cleaned output saved into Lakehouse with sentiment column included.

## 📊 Data Reporting & Dashboard

• Created a Semantic Model in Power BI for clean data governance.  
• Built measures for Positive %, Negative %, and Neutral % sentiments.  
• Designed interactive dashboard showing sentiment distribution across days, countries, and news categories.  
• Insights focus on last 7 days of news data.

## 🚨 Alerts & Automation

• Implemented Data Activator in Power BI to set up automated triggers.  
• Example use case: Send Teams/Email alert when Negative sentiment exceeds threshold (e.g., >5%).  
• Ensured real-time monitoring capability.

## 🚀 Key Learnings & Impact

• Designed a scalable, serverless pipeline in Microsoft Fabric.  
• Applied data engineering best practices (incremental loading, structured Delta tables).  
• Integrated ML seamlessly into a data pipeline with SynapseML.  
• Delivered real-time insights and alerts for business decision-making.  
• Project demonstrates full lifecycle of Data Engineering + Data Science + BI.

## 🛠️ Tech Stack

- Microsoft Azure & Microsoft Fabric

- Azure Data Factory (Ingestion & Orchestration)

- Synapse Data Engineering (Apache Spark)

- SynapseML (Sentiment Analysis)

- Fabric Lakehouse (Delta Tables)

- Power BI (Semantic Models & Dashboards)

- Data Activator (Real-time Alerts)