International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

"Bulid a Data-pipeline to Process & Analyze Data"

Mr. Rohan Shivaji Bhosale
Fabtach Collage Of Engineering Sangola
(Department of Artificial Intelligence And Data Science Engg)

Mr. Nikhil Navnath Shinde
Fabtach Collage Of Engineering Sangola
(Department of Artificial Intelligence And Data Science Engg)

Mr. Ajay Gorakh Jagtap
Fabtach Collage Of Engineering Sangola
(Department of Artificial Intelligence And Data Science Engg)

Prof. Manoj Sakharam Patil
Fabtach Collage Of Engineering Sangola
(Department of Artificial Intelligence And Data Science Engg)

This project focuses on building a robust data pipeline for processing movie and TV show data stored as CSV files on GitHub. The data is first ingested into **Azure Data Lake Storage**, forming the base for further processing. Using **Databricks Auto Loader**, the pipeline supports **incremental data loading**, allowing it to handle new data efficiently.

The pipeline adopts a **multi-layered architecture**:

Bronze Layer: Stores raw data.

Silver Layer: Contains cleaned and transformed data.

Gold Layer: Hosts refined datasets ready for reporting and analytics.

Delta Live Tables (DLT) are used in the Gold layer to ensure **streamlined data management** and maintain **data quality** throughout the pipeline.

I. INTRODUCTION

The project begins with the ingestion of data from GitHub, where CSV files containing information about movies and TV shows are stored. This data will be loaded into Azure Data Lake Storage, serving as the foundation for further processing. By employing Databricks, we will implement Auto Loader for incremental data loading, ensuring that our pipeline can efficiently handle new data as it becomes available.

The architecture of the pipeline is designed to follow a multi-layered approach: the Bronze layer will store raw data, the Silver layer will contain transformed data, and the Gold layer will provide refined datasets suitable for reporting and analysis. Delta Live Tables will be utilized to manage the Gold layer, enabling streamlined data management and ensuring data quality.

II. OVERVIEW

Data Ingestion: Build a system to ingest Netflix data from various sources, including user interactions, content metadata, and ratings.

Data Transformation: Cleanse and transform the data for further analysis, ensuring that it is in a usable format.

Data Storage: Store large volumes of processed data in a scalable and accessible format (e.g., using cloud storage solutions like AWS S3).

Data Analysis: Implement analytical models to provide actionable insights, such as user behavior patterns and content performance metrics.

Recommendation System: Build a basic recommendation engine based on historical data to suggest personalized content to users.

Automation: Ensure that the data pipeline is automated to handle new incoming data seamlessly.

Or **send alerts** to the user through IoT modules like Wi-Fi or GSM (if implemented).

The project aims to provide a **low-cost, reliable, and scalable solution** for farmers, gardeners, and agricultural researchers. It not only reduces the manual effort of checking soil conditions but also helps **prevent over- or under-watering**, ultimately contributing to better crop health and water conservation.This is best solution.
The architecture follows a **multi-layered approach**:

Bronze Layer for raw data,

Silver Layer for cleaned and transformed data,

Gold Layer, managed using **Delta Live Tables**, for analytics-ready datasets.

III. SCOPE AND OBJECTIVE

1 **Environment Setup**

Resource Group Creation: Set up a centralized resource group in Azure to manage all project-related services.

Azure Services Configuration: Deploy key Azure services including:

Azure Data Factory (ADF) for orchestrating data pipelines.

Azure Databricks for scalable data processing.

Azure Data Lake Storage for storing raw and processed data.

Azure Synapse Analytics for data exploration and visualization.

1. **Data Ingestion**

Use **Azure Data Factory** to extract data from multiple sources including on-premises databases.

Employ **Databricks Autoloader** for **incremental data ingestion**, allowing real-time updates.

Include **data validation checks** to ensure accuracy and integrity during the ingestion phase.

2. **Data Transformation**

Process and enrich the ingested data using **Apache Spark/PySpark in Azure Databricks**.

Apply **Medallion Architecture**:

Bronze Layer: Store raw, unprocessed data.

Silver Layer: Clean and transform the data to a usable format.

Gold Layer: Aggregate and refine the data for advanced analytics and reporting.

1.

IV. RELATED WORK

A. ***1. Project Setup***

Azure Account Creation: Set up Azure accounts and subscriptions.

Resource Provisioning:

-

-

Azure Data Lake Storage: To store raw and processed data.

Azure Data Factory (ADF): For orchestrating data movement and workflows.

Azure Databricks: For scalable data processing using Apache Spark.

Power BI: For data visualization and reporting.

B. ***2. Data Ingestion***

Source: Pull data from **GitHub (CSV files)** and other sources.

Tool: Use **Azure Data Factory** to move data into the **Azure Data Lake**.

Purpose: Ensure scalable and reliable data loading into the pipeline.

C. ***3. Data Processing***

Incremental Loading: Use **Databricks Auto Loader** to process only new data.

Structured Layers:

Bronze: Raw data.

Silver: Cleaned and validated data.

Gold: Aggregated and analytics-ready data using **Delta Live Tables (DLT)**.

D. ***4. Dynamic Notebooks***

Create **parameterized notebooks** in Databricks for reusable and flexible transformations.

Use **Databricks Workflows** to schedule and automate the execution of notebooks.

4.1. ***Data Validation***

Add **validation checks** at different stages (ADF and Databricks) to ensure:

Data consistency

Schema conformity

Null and duplicate checks

E. ***6. Reporting***

F. Connect the ***Gold layer*** *data to* ***Power BI*** *for building:*

G. Interactive dashboards

H. Reports showing content performance, user behavior, trends, etc

I.

V. METHODOLOGY

*J. 1. **Real-time Insights and Personalization***

*K. The pipeline is designed to **ingest and process data in near real-time**, enabling Netflix to track user behavior, streaming patterns, and content interactions.*

Real-time data supports **personalized recommendations**, adaptive content delivery, and improved user engagement.

---

*L. 2. **Scalability with Big Data Technologies***

*M. Technologies such as **Apache Spark** (via **Azure Databricks**) and cloud storage (Azure Data Lake, AWS S3) are used to process and store **large-scale datasets**.*

This allows the pipeline to **scale horizontally**, handling millions of records generated by user interactions without performance degradation.

---

*N. 3. **Automation of the Data Pipeline***

Azure Data Factory is used to orchestrate data ingestion workflows automatically.

Databricks Auto Loader enables **incremental data processing** as new data arrives, while **Workflows and Parameterized Notebooks** manage scheduled and triggered executions.

This minimizes manual intervention and ensures **continuous data flow**.

---

*O. 4. **Handling Complexity***

The project follows a **layered architecture** (Bronze, Silver, Gold) to organize data by processing stage, which simplifies pipeline complexity.

Modular development using **parameterized notebooks** and **Delta Live Tables (DLT)** ensures easier maintenance and debugging.

---

5. **Minimizing Latency**

While real-time systems may introduce some latency, the use of **optimized Spark jobs**, **Delta Lake**, and **streaming data sources** helps maintain low processing times.

Proper resource provisioning in Databricks ensures **timely delivery** of insights and recommendations.

6. **Ensuring Data Quality**

Data validation is embedded throughout the pipeline:

Schema checks

Null/duplicate handling

Logging and error handling

This helps prevent **data corruption** and ensures **accurate analytics output**.

VI. DOMAIN OVERVIEW

Azure Data Factory (ADF)

Orchestrates the data movement from GitHub and other sources into the Azure Data Lake.

Automates ETL/ELT processes for streamlined and reliable ingestion.

Azure Data Lake Storage

Acts as the **central storage layer** for raw and processed data.

Supports hierarchical storage to efficiently manage data across the Medallion Architecture.

Azure Databricks with Apache Spark

Performs high-performance data processing, cleansing, and transformation.

Implements the **Medallion Architecture**:

Bronze Layer: Raw data.

Silver Layer: Cleaned and validated data.

Gold Layer: Aggregated, analytics-ready datasets.

Delta Live Tables (DLT)

Manages the Gold layer with built-in **data quality checks**, **versioning**, and **incremental updates**.

Enhances pipeline reliability and simplifies operations.

Azure Synapse Analytics

Enables **big data querying** and advanced analytics.

Integrates seamlessly with the data lake and Databricks for deeper exploration.

Power BI

Connects to the Gold layer for building **interactive dashboards**.

Supports **real-time visualization** of key metrics like user behavior, trending content, and platform performance.

VII. SYSTEM ARCHITECTURE

The system uses Azure's cloud services to build a scalable and automated data pipeline.

Data is ingested from GitHub and other sources using Azure Data Factory.

Raw data is stored in Azure Data Lake under the **Bronze layer**.Azure Databricks processes and transforms data into **Silver and Gold layers**.

Delta Live Tables ensure data quality and manage incremental updates.

Final data is visualized using **Power BI** for reporting and insights.

The pipeline follows the **Medallion Architecture**: Bronze (raw), Silver (cleaned), Gold (analytics-ready).

This setup supports real-time analytics, automation, and personalized recommendations.

VIII. LITERATURE REVIEW

The development of real-time data engineering pipelines has been widely explored in academic and industry research due to the rapid growth of big data and the demand for scalable analytics platforms. This literature review outlines key studies and technologies that have influenced the design and implementation of this project.

.

IX. RESULTS AND DISCUSSION

I. ***1. Data Ingestion***

II. ***Result****: CSV files from GitHub and other sources were successfully ingested using Azure Data Factory (ADF).*

III. ***Discussion****: ADF pipelines performed scheduled extractions without failure, validating the system's ability to handle real-time and batch ingestion efficiently.*

IV. ***2. Incremental Data Loading***

Result: Auto Loader in Azure Databricks enabled seamless incremental loading of new data into the Bronze layer.

Discussion: This eliminated the need for reprocessing the entire dataset, reducing compute time and cost while ensuring pipeline efficiency.

V. ***3. Data Transformation & Cleaning***

Result: Raw data was transformed into structured formats using PySpark and stored in Silver and Gold layers.

Discussion: The use of Delta Lake ensured schema enforcement and data reliability, preventing data corruption or duplication.

4. Delta Live Tables (DLT) Usage

Result: DLT was used to automate the Gold layer with built-in data validation and monitoring.

Discussion: DLT helped in maintaining **data lineage**, improving **transparency**, and reducing manual transformation logic.

VI. ***5. Data Visualization***

Result: Processed Gold layer data was successfully connected to **Power BI** dashboards.

Discussion: Reports provided real-time insights into user interaction trends, content ratings, and performance metrics, enabling actionable business decisions.

6. Automation & Workflow Orchestration

VII. ***Result****: End-to-end data flow was automated using ADF and Databricks Workflows with parameterized notebooks.*

Discussion: Automation minimized human intervention, increased reliability, and made the pipeline production-ready.

XI. REFERENCES

· ***Zaharia, M., et al.*** *(2016). Apache Spark: A Unified Engine for Big Data Processing. Communications of the ACM, 59(11), 56–65.* *https://doi.org/10.1145/2934664*

· · **Armbrust, M., et al.** (2020). *Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores*. Proceedings of the VLDB Endowment, 13(12), 3411–3424. https://www.vldb.org/pvldb/vol13/p3411-armbrust.pdf

· · **Databricks**. (2022). *What is the Medallion Architecture?* https://www.databricks.com/glossary/medallion-architecture

· · **Microsoft Azure Documentation**. *Azure Data Factory, Data Lake Storage, Synapse Analytics & Databricks*. https://learn.microsoft.com/en-us/azure/

· · **Delta Live Tables**. Databricks Documentation. https://docs.databricks.com/workflows/delta-live-tables/index.html