Media Content Analytics Platform

Table of Contents

1. Introduction

In today's rapidly evolving digital ecosystem, social media platforms and online news portals have emerged as significant sources of information, engagement, and influence. Businesses, content creators, and decision-makers rely heavily on data-driven insights to understand audience behavior, content effectiveness, and emerging trends. The Media Content Analytics Platform is an end-to-end data engineering solution designed to collect, process, and analyze large volumes of data from **YouTube** and **News sources**.

This project is structured using the **Medallion Architecture** consisting of **bronze, silver, and gold layers**, which ensures high data quality, efficient processing, and actionable business insights. The pipeline transforms raw data into meaningful intelligence that can support **strategic planning**, **content optimization**, and **trend forecasting**. This synopsis provides a detailed overview of the project including objectives, architecture, methodology, technology stack, expected outcomes, and future enhancements.

2. Problem Statement

In the age of digital media, organizations face the challenge of extracting meaningful insights from large volumes of unstructured social media and news data. Manual analysis is time-consuming and prone to inaccuracies, making it difficult to monitor trends, measure content performance, and respond to public sentiment in real-time. There is a need for an automated and scalable solution that transforms raw media data into actionable intelligence.

This project addresses the following key problems: -

1. Difficulty in handling large, dynamic datasets from multiple online platforms.
2. Lack of standardized procedures for data cleaning and transformation.
3. Inability to quickly identify trending topics or content performance.
4. Limited accessibility of insights for non-technical teams.

3. Objectives of the Project

The main objectives of the Social Media Analytics Pipeline are: -

1. To build a scalable and automated pipeline to handle social media and news data.
2. To convert raw, unstructured data into clean and organized datasets.
3. To derive analytical insights related to content engagement, performance, and trends.
4. To support decision-making for businesses, content creators, and analysts through dashboards.
5. To establish a foundation for predictive analytics and machine learning integration.

4. System Architecture – Medallion Model

The project follows the Medallion Architecture, which organizes the data transformation workflow into three layers:

4.1 Bronze Layer – Raw Data Collection

- Stores raw CSV/JSON files exactly as received.
- Preserves original data for backup, audit, and integrity.
- Ingests YouTube and News data without any modifications.

4.2 Silver Layer – Data Cleaning and Standardization

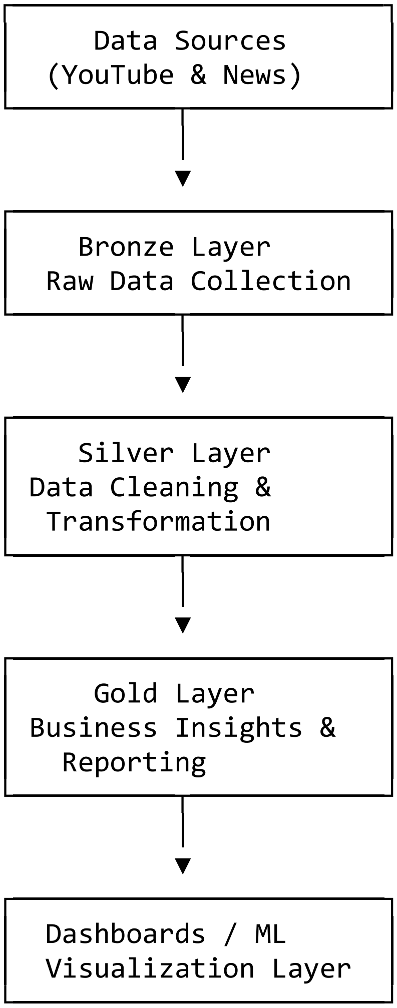
- Formats dates, text fields, and numerical values consistently.
- Handles missing values, duplicates, and structural inconsistencies.
- Converts raw data into clean, reliable, and well-structured datasets.

4.3 Gold Layer – Final Insights Layer

- Aggregates and computes business-relevant insights.
- Produces summarized, analytics-ready data for reporting and dashboards.
- Used directly by analysts, decision-makers, and data scientists.

This architecture ensures data accuracy, scalability, traceability, and easier maintenance.

System Architecture (Medallion Model)



5. Data Sources and Their Purpose

5.1 YouTube Dataset

Contains video-level metadata such as:

- Video titles
- Channel names
- Views, likes, and comments counts
- Content categories (e.g., Entertainment, Education, News)
- Publish dates and trending dates
- Basic engagement indicators

Purpose:

To analyze content performance, understand audience engagement patterns, and identify which topics or channels generate higher reach and interaction. It also helps in comparing performance across categories and tracking growth over time.

5.2 News Dataset

Includes text-based article information such as:

- Article headlines
- Short descriptions
- Publication dates
- News categories (Business, Politics, Technology, Sports, etc.)
- Text suitable for topic modeling and sentiment analysis

Purpose: Track trending news topics and evaluate public sentiment.

6. Technology Stack

The following technologies are used to build the pipeline:

Databricks: Main data processing platform.

Python: Python language for doing the operations.

API: Api is used for taking the news from google (Gnews).

Workspace File System: Stores ingested data locally.

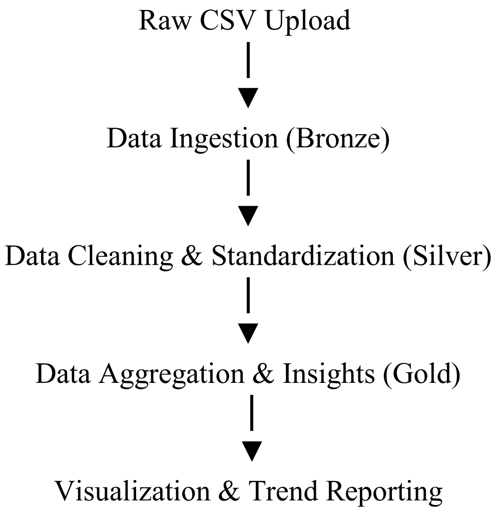
Visualization Tools: Used for dashboards and reporting.

7. Workflow Summary

1. **Data Upload** – Raw CSV files for YouTube and News.
2. **Ingestion to Bronze Layer** – Data stored without transformation.
3. **Transformation to Silver Layer** – Cleaning and structuring.
4. **Aggregation to Gold Layer** – Insight extraction and business-ready outputs.
5. **Visualization** – Dashboards highlight trends and analytics.

Each step builds upon the previous one, ensuring that data quality improves throughout the process.

Workflow Diagram


Raw CSV Upload
Data Ingestion (Bronze)
Data Cleaning & Standardization (Silver)
Data Aggregation & Insights (Gold)
Visualization & Trend Reporting

8. Coding & Technical Implementation Overview

This section outlines the core technical components and pseudocode used in the project.

8.1 Bronze Layer (Data Ingestion)

BRONZE RAW FILE PATH

```
bronze_path = "/Workspace/social media/Bronze/Youtube-dataset-sample.csv"
bronze_path1 = "/Workspace/social media/Bronze/news.csv"
```

LOAD RAW BRONZE DATA

```
df = pd.read_csv(bronze_path)
print ("Original Bronze shape:", df.shape)
```

8.2 Silver Layer (Transformation)

```
df_silver = df_bronze.copy()
```

Convert date column to datetime

```
df_silver["date"] = pd.to_datetime(df_silver["date"], errors="coerce")
```

Convert numeric columns properly

```
numeric_cols = ["likes", "replies", "replies_value"] for col in numeric_cols: df_silver[col] = pd.to_numeric(df_silver[col], errors="coerce")
```

Drop rows where required fields are missing

```
required_cols = ["comment_id", "comment_text", "username", "video_id"] df_silver = df_silver.dropna(subset=required_cols)
```

Add derived column: comment length

```
df_silver["comment_length"] = df_silver["comment_text"].astype(str).apply(len)
```

8.3 Gold Layer (Aggregations)

```
df_gold = df_silver.copy()
```

Engagement Score

```
df_gold["engagement_score"] = df_gold["likes"].fillna(0) + df_gold["replies"].fillna(0)
```

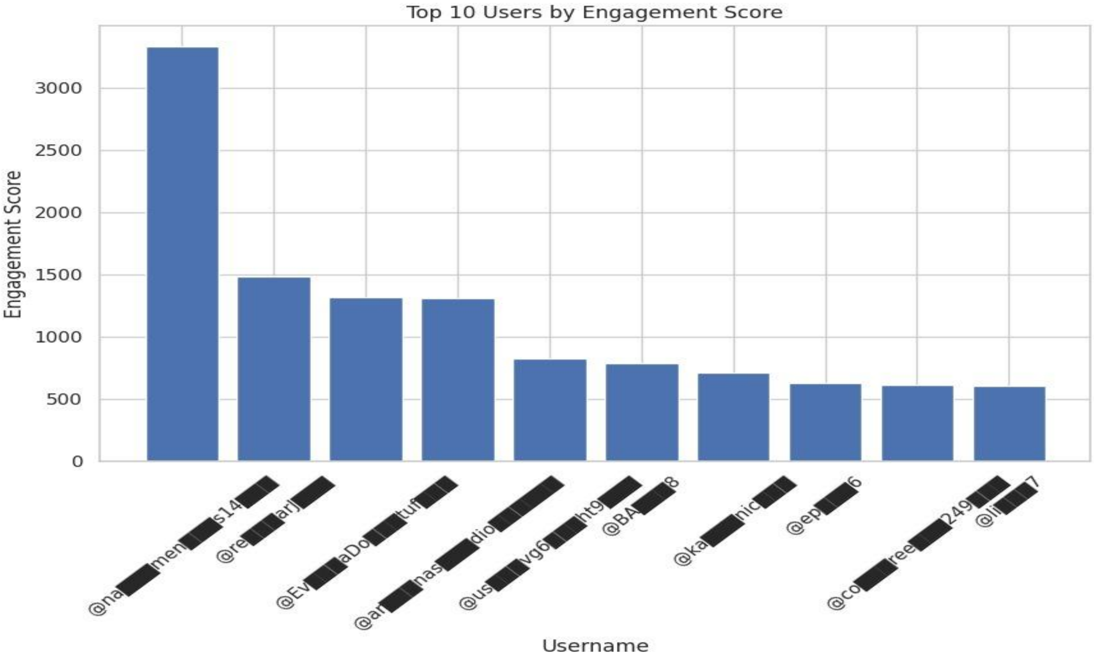
User-level aggregates

```
user_stats = df_gold.groupby("username").agg({ "comment_id": "count", "likes": "sum",
"replies": "sum", "engagement_score": "sum" }).reset_index()
user_stats.rename(columns={"comment_id": "total_comments"}, inplace=True)
user_stats.head()
```

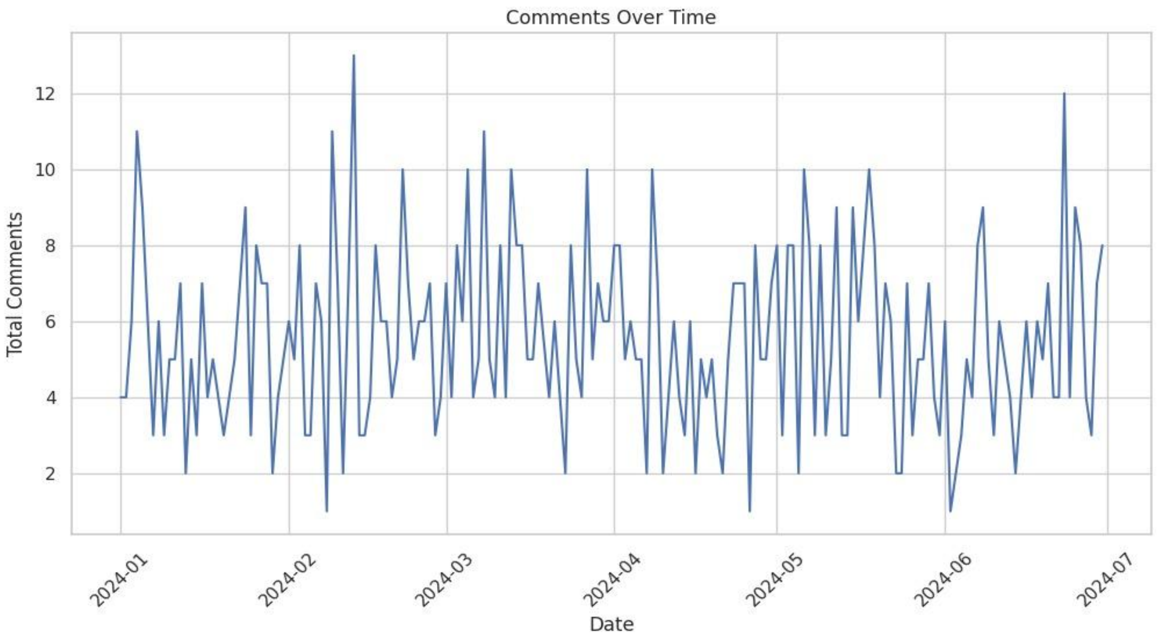
8.4 Visualization Layer

```
top_users = fact_user_stats.sort_values( "engagement_score", ascending=False ).head(10)

plt.figure(figsize=(10, 6)) plt.bar( top_users["username"], top_users["engagement_score"] )
plt.xticks(rotation=45) plt.title("Top 10 Users by Engagement Score") plt.xlabel("Username")
plt.ylabel("Engagement Score")

path = f"{viz_folder}/top_users.png" plt.savefig(path, bbox_inches="tight") plt.show()
```

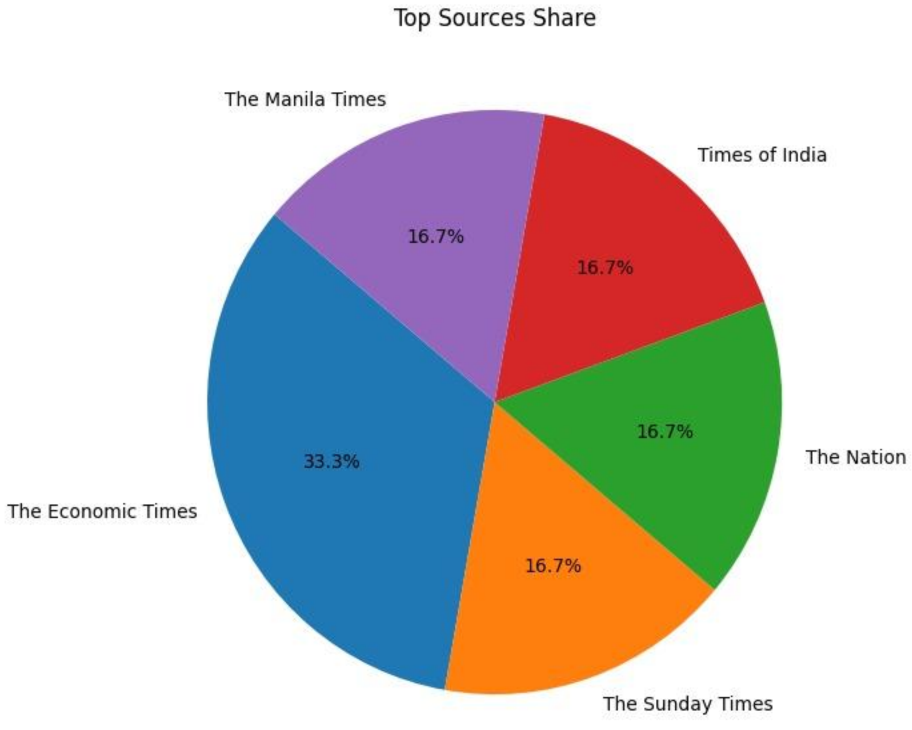
9. RESULTS



Comments over Time in a Youtube



News Sources shares on different news publishers



10. Expected Outcomes

The project delivers: Clean, standardized datasets.

1. Insightful metrics about content performance and engagement.
2. Trend detection and forecasting support.
3. Audience behavior analysis.
4. A strong foundation for predictive analytics and machine learning.
5. Enhanced decision-making capabilities for content planning and marketing.

11. Limitations of the Project

While the Social Media Analytics Pipeline provides structured data processing, it has certain limitations: -

1. **Manual data upload**: Currently depends on CSV uploads instead of real-time API ingestion.
2. **Limited sentiment analysis**: Only basic trends and engagement analysis are performed.
3. **Platform dependency**: Focused only on YouTube and News data; does not include platforms like Twitter, Instagram, or Facebook.
4. **Storage constraints**: Local Databricks storage may not scale for larger enterprise applications.
5. **No real-time processing**: Delayed insights due to batch processing instead of streaming.

12. Future Enhancements

Potential future improvements include: -

1. API-based automatic data loading.
2. Real-time streaming instead of manual uploads.
3. Sentiment analysis using NLP techniques.
4. Topic modeling for trend identification.
5. Integration with cloud storage (AWS, Azure, GCP).
6. Predictive analytics for forecasting engagement and viral trends.

13. Conclusion

The Media Content Analytics Platform is a powerful solution that transforms raw, unorganized data into meaningful insights. It follows a structured and efficient approach based on Medallion Architecture to ensure high data quality and scalability. By leveraging modern technologies such as Databricks, this system supports advanced analytics and real-time decision-making.

With expanding capabilities such as API integration and machine learning, the pipeline is adaptable for future enhancements. It serves as an asset for businesses, content creators, and analysts seeking to maximize digital presence and enhance strategic planning.