Pranoy Chakraborty | Bangalore, India | +917996915434 | pcslg1998@gmail.com |LinkedIn | GitHub

🎬 End-to-End Streaming Analytics with Microsoft Azure, Databricks and Microsoft Power BI

Domain: Media & Entertainment

1. Introduction

The Media & Entertainment (M&E) industry is undergoing a profound transformation, driven by rapid digitization, increased competition among OTT platforms, and the explosion of content consumption across geographies. In this context, data analytics has become pivotal for content strategy, viewer engagement, and operational efficiency.

This project presents a comprehensive, end-to-end analytics platform that simulates how a media tech company like Netflix can leverage modern cloud-native tools (Microsoft Azure, Databricks, Power BI) to transform raw content metadata into strategic business insights.

The dataset used is inspired by publicly available Netflix title metadata, capturing details of over 6,000 titles including movies and TV shows, across multiple countries, genres, and contributors.

---

2. Problem Statement

With thousands of titles released across various countries, languages, and genres, media companies struggle to answer questions like:

- Which genres perform best in which countries?
- How is content consumption evolving over time?
- Who are the top directors or actors contributing to globally?
- How does the platform's content mix compare across years and regions?

To address these, we built a **data intelligence pipeline** that transforms raw content metadata into **Gold-standard business metrics** for consumption via Power BI dashboards.

---

3. Architecture and Technologies Used

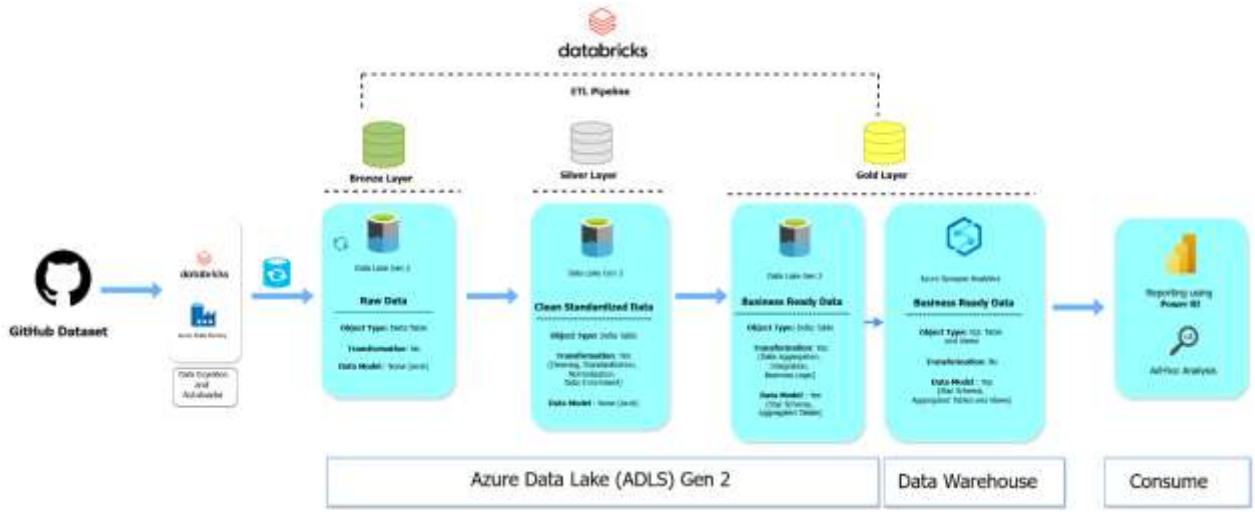
This project uses the **Microsoft Azure data stack** with Databricks and Power BI, following the **Medallion Architecture** (Bronze–Silver–Gold) to ensure data quality, modularity, and performance.

3.1 Tools & Technologies

Layer	Technology
Ingestion	Azure Data Factory, Databricks Auto Loader
Storage	Azure Data Lake Storage Gen2 (ADLS Gen2)
Transformation	PySpark in Azure Databricks (Delta Lake)
Data Warehouse	Azure Synapse Analytics (SQL Pools)
Visualization	Power BI (Connected to Synapse)

3.2 Medallion Architecture Overview

- **Bronze Layer**: Raw ingested data in Delta format. No schema enforcement or transformation.
- **Silver Layer**: Cleaned and standardized data. Null handling, date parsing, categorical normalization.
- **Gold Layer**: Business-ready aggregated data. Star-schema-like outputs (fact + dimension tables).



4. Dataset Overview

The source dataset comprises five normalized schemas:

- **netflix_titles**: Title-level metadata (type, director, cast, release year, rating, etc.)
- **netflix_cast**: Exploded cast list per title
- **netflix_directors**: Exploded director list per title
- **netflix_countries**: Country-level availability per title
- **netflix_category**: Extracted genre tags from combined genre column

After transformation, we created materialized views and aggregated fact tables used in Power BI dashboards.

---

5. ETL Implementation in Databricks

5.1 Ingestion Pipeline

- Used **Databricks Auto Loader** for ingesting CSVs from the mounted ADLS path.
- Schema inference enabled for auto-adapting to changes.
- Raw data stored in bronze tables.

5.2 Transformation Logic

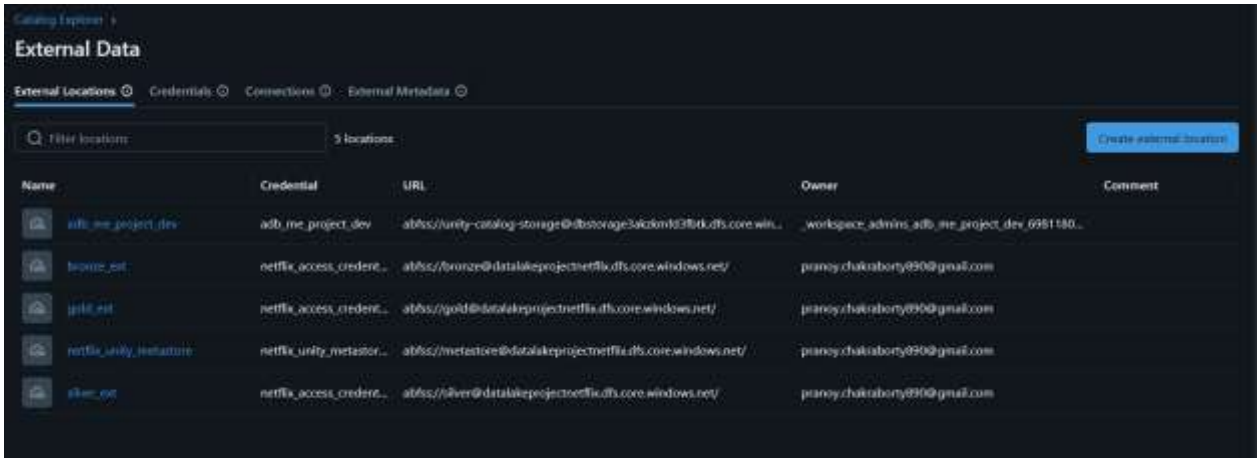
- Null value handling (NA, empty strings)
- Standardizing date formats (e.g., 8/15/2018 → 2018-08-15)
- Splitting multi-value columns (cast, country, listed_in) into exploded rows
- Deduplicating entries and removing noise (e.g., removing 'Unknown' cast entries)

All cleaned data is stored in Silver Delta tables with enforced schemas.

5.3 Gold Layer Outputs

- fact_titles (title-level facts)
- dim_country, dim_genre, dim_cast, dim_director
- agg_content_trends: titles added per year, per country
- agg_duration_by_rating: average duration grouped by content rating

Pranoy Chakraborty | Bangalore, India | +917996915434 | pcslg1998@gmail.com |LinkedIn | GitHub



6. Insights from Power BI Dashboards

The Gold-layer tables are queried using Synapse SQL endpoints and visualized in Power BI. Three main dashboards were created:

6.1 Content Overview Dashboard

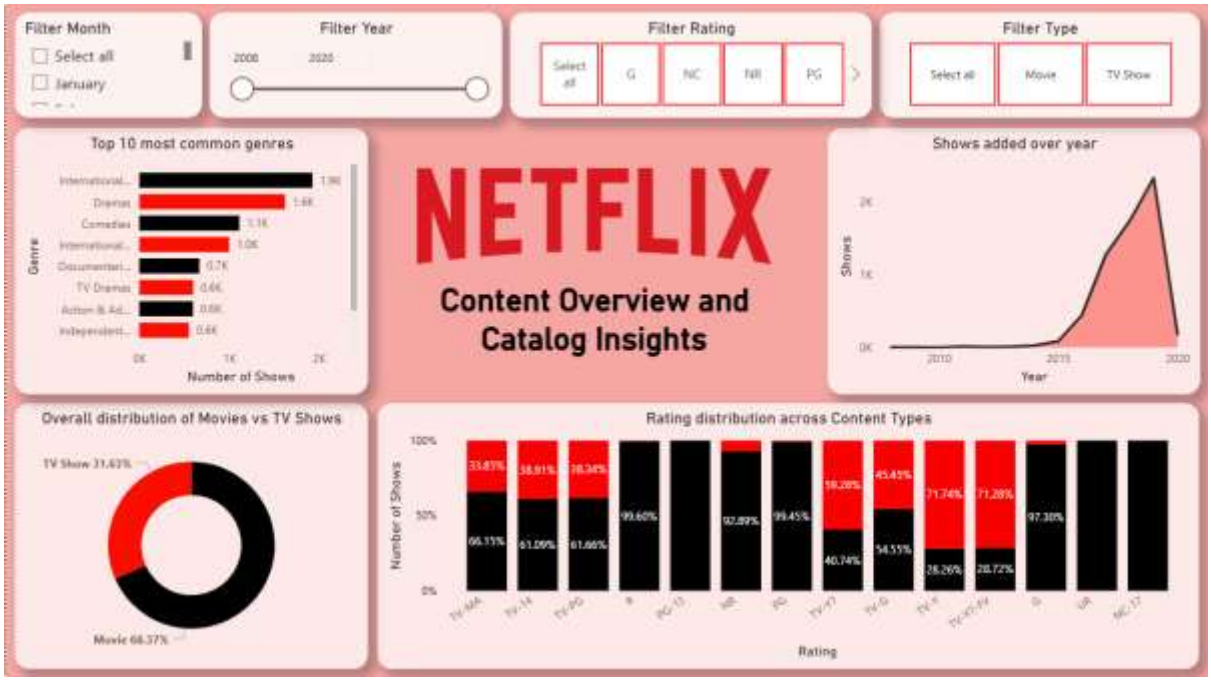
Purpose: Provide a bird's-eye view of the Netflix catalog.

Insights:

- Titles peaked in **2019** with **2,323** additions.
- Top genres: **International Movies**, **Dramas**, and **Comedies**.
- Movie to TV show split: **68.37% vs 31.63%**
- Dominant ratings: **TV-MA**, **TV-14**, **TV-PG**.

Charts:

- Bar chart: Titles per year
- Pie chart: Content type distribution
- Tree map: Genre popularity



6.2 Geography & Talent Insights Dashboard

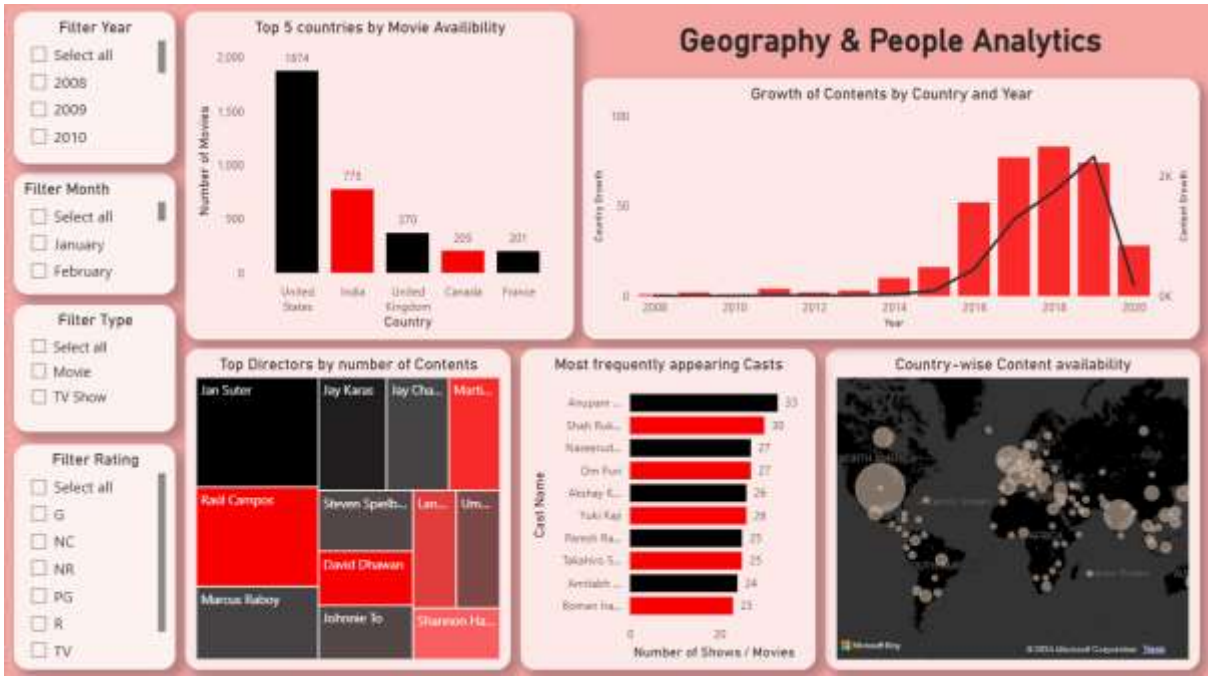
Purpose: Understand Netflix's global presence and content contributors.

Insights:

- Top contributing countries: **USA**, **India**, **UK**.
- Actors with most titles: **Anupam Kher**, **Shah Rukh Khan**, **Naseeruddin Shah**.
- Most frequent directors: **Jan Suter**, **Jay Karas**, **Raúl Campos**.
- Year-wise country content heatmap shows global expansion post-2015.

Charts:

- World map: Title count by country
- Line chart: Country growth over time
- Leaderboards: Top 10 actors and directors



6.3 Duration & Release Patterns Dashboard

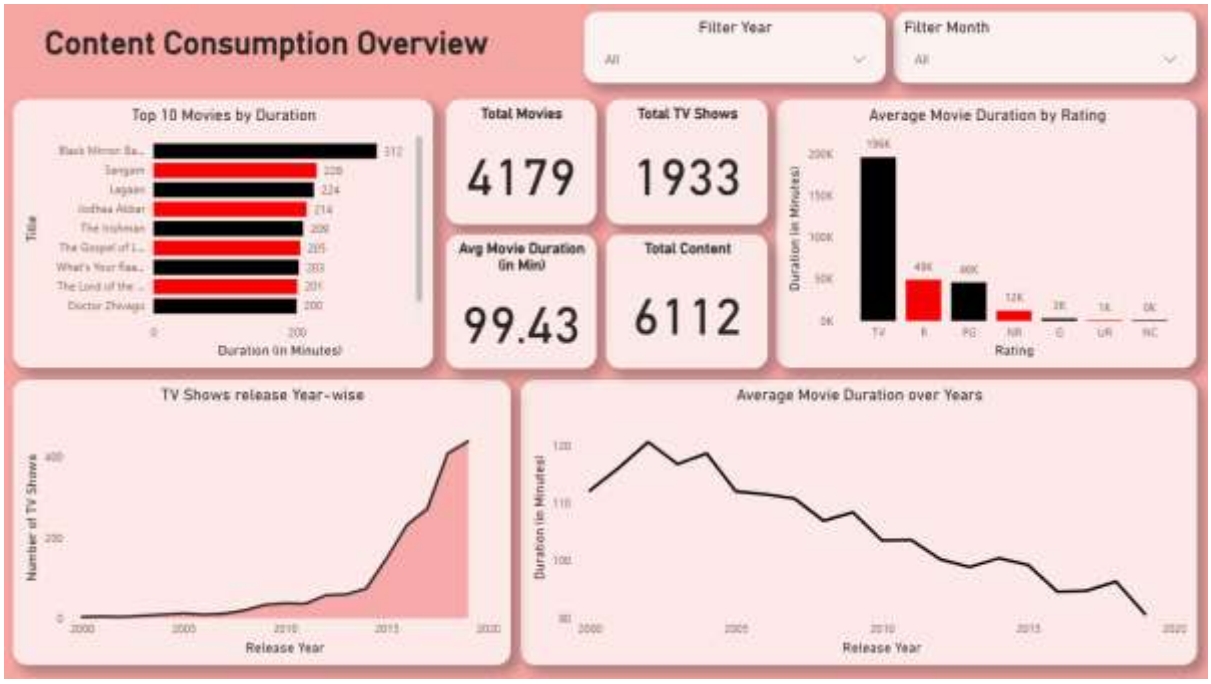
Purpose: Analyze content length and release evolution.

Insights:

- **Average movie duration**: **99.43 mins**
- Longest titles: *Black Mirror: Bandersnatch* (312 mins), *Sangam* (228 mins)
- TV shows exhibit steep growth post-2015
- Increasing TV-MA content in recent years
- Shift towards shorter movie durations in 2020s

Charts:

- Histogram: Duration distribution
- Line chart: Duration trend by year
- Rating vs. Duration boxplot



7. Business Value Delivered

💼 **Strategic Benefits:**

- **Content Strategy**: Identify underrepresented genres or regions to expand into.
- **Localization Planning**: Gauge which countries lack dubbed/subtitled content.
- **Talent Analytics**: Assess actor/director contributions by content popularity or region.
- **Catalog Optimization**: Trim low-duration or underperforming titles.

⚙️ **Technical Outcomes:**

- **Delta Lake Storage**: ACID transactions, time travel for audit & rollback
- **Schema Evolution**: Supports dynamic metadata without code changes
- **Scalable**: Built to scale for real-world multi-OTT analytics
- **Unified Stack**: All Azure-native tools, production-ready deployment pipeline

---

8. Future Scope & Enhancements

1. **Recommendation System**:
    - Collaborative filtering using user viewing history (if available)
2. **ML Integration**:
    - Predict content success using features like genre, duration, cast
3. **Cross-OTT Benchmarking**:
    - Integrate data from Prime Video, Disney+, etc.
4. **Sentiment Analysis**:
    - Analyze audience feedback or reviews via NLP
5. **Dynamic Alerts**:
    - Real-time monitoring of title additions or drops via Event Grid + Logic Apps

---

9. Conclusion

This project simulates a real-world, enterprise-scale analytics solution for the Media & Entertainment industry. Using Microsoft Azure technologies, we have demonstrated how structured and semi-structured content data can be transformed into rich visual insights with real business value. This project not only showcases technical proficiency but also reflects the data-driven culture that companies like Netflix are built on.

By leveraging Azure Databricks, Delta Lake, Synapse, and Power BI, this pipeline is modular, cloud-native, and enterprise-ready — suitable for consulting use cases, product analytics, or portfolio presentation.

✅ **Project Summary Table**

Component	Implementation
Data Source	Netflix Metadata (CSV)
Ingestion Tool	Azure Data Factory, Databricks Auto Loader
Storage Format	Delta Lake on ADLS Gen2
ETL Engine	PySpark in Azure Databricks
Gold Output Tables	fact_titles, dim_genre, agg_trends, etc.
Visualization Tool	Power BI (Synapse SQL Connector)
Total Content	6,112 (4,179 Movies + 1,933 TV Shows)