Media Content Analytics Platform

Abstract

This project presents a comparative media analytics platform designed to analyze, clean, and visualize content data from two major streaming platforms: **HBO** and **Paramount**.
The system follows a complete **ETL pipeline** using PySpark for data ingestion and preprocessing, ensuring scalable and efficient handling of large media datasets.
Cleaned datasets are then used to generate meaningful insights through Python-based visualizations, focusing on content trends, genre distribution, duration patterns, production regions, and year-by-year performance.

The platform ultimately provides a unified analytical view of both streaming services, enabling comparisons of content strategy, production trends, and market positioning

Project Overview

The **Media Analytics Platform** is a data-driven application that performs end-to-end processing of media catalog datasets from HBO and Paramount.
The project is structured into three major components:

1. Extraction & Cleaning (ETL)

- Raw CSV files are loaded through PySpark using a defined schema.
- Column names are normalized for consistency.
- Important fields such as genres, runtime, and release year are parsed and standardized.
- Unique content identifiers are generated.
- Final deduplicated and cleaned datasets are exported as CSV for visual analysis.

2. Data Transformation for Visualization

- Cleaned data is loaded using Pandas.
- Nested/array fields (e.g., genres) are processed into usable formats.

- Filtering, grouping, and statistical operations prepare the data for charts.

3. Visualization & Insights

Multiple visualizations compare HBO vs Paramount on:

- Release year trends
- Genre priorities
- Duration distribution
- Year-over-year performance
- Production country contributions

Each visualization produces actionable insights, revealing how both platforms shape their content strategy, global reach, and audience targeting.

Tools and Technologies Used

- **Programming Language:** Python
- **Data Processing:** PySpark (Spark SQL, DataFrames)
- **IDE:** PyCharm
- **Visualization:** Plotly, Matplotlib
- **Libraries:** Pandas, NumPy, Plotly Express, Matplotlib
- **Environment Setup:** Java JDK 17, Spark 3.4.5

ETL (Extract, Transform, Load)

Data Source: The raw financial datasets for HBO and Paramount were sourced from Kaggle, a publicly available data platform containing real-world industry datasets

1. Titles by Release Year (Trend Analysis)

What it Visualizes:
Count of titles released each year for HBO vs Paramount.

Insights:

- **Trend Analysis:** Shows historical growth or decline periods.
- **Platform Comparison:** Side-by-side release volume comparison.
- **Growth Patterns:** Identifies whether platforms are expanding content production.
- **Peak Years:** Highlights high-output years.
- **Market Strategy:** Indicates production strategy shifts.

Example Findings:

- If HBO shows a spike in 2019–2021 → increased investment in content.
- If Paramount has steady growth → stable long-term expansion.

2. Top Genres Comparison

What it Visualizes:
The most frequent genres for HBO and Paramount.

Insights:

- **Content Strategy:** Which genres each platform focuses on.
- **Market Positioning:** Differences in genre strengths.
- **Audience Targeting:** Type of content each platform pushes.
- **Competitive Analysis:** Points of overlap and differentiation.

Example Findings:

- HBO dominating **Drama** → preference for serious, narrative-driven content.
- Paramount leading in **Comedy** → lighter entertainment strategy.
- Overlapping genres → direct competition.

Visualization Summary:

A **bar plot** comparing top genres, showing genre dominance and overlap between platforms.

3. Duration Distribution (Content Length Strategy)

What it Visualizes:
Histogram of content duration (in minutes) for both platforms.

Insights:

- **Content Length Strategy:** Movie-length vs episode-length preferences.
- **Peak Durations:** Most common runtime ranges.
- **Distribution Shape:** Single peak or multi-peak patterns.
- **Platform Differences:** Who produces longer or shorter content generally.

Example Findings:

- Peak around **90–100 minutes** → movie-heavy catalog.
- Peak around **30–60 minutes** → episodes/series format.
- Long tail → diverse content lengths.

Visualization Summary:

An **overlaid histogram** showing duration patterns, helping identify platform formatting strengths.

4. Year-by-Year Comprehensive Comparison

What it Visualizes:
Production volume, ratings, and genre breakdown over time.

Insights:

- **Volume vs Quality:** Are platforms producing more but scoring less?
- **Trend Identification:** Whether score averages rise/fall.
- **Platform Strategy:** Quantity-first vs quality-first approaches.
- **Year-Specific Insights:** Which genres spike in recent years.

Example Findings:

- High volume + high IMDb scores → strong content strategy.
- Declining scores → possible quality issues.
- Genre spikes → changing content focus.

Visualization Summary:

A **multi-metric comparison** showing not just quantity but content score patterns over time.

5. Production Countries by Content Type (Movie vs Show)

What it Visualizes:
Countries contributing to movies and shows for each platform.

Insights:

- **Geographic Strategy:** Domestic vs global production footprint.
- **Content Type Distribution:** Countries specializing in movies or shows.
- **Global vs Local:** Whether platforms source globally.
- **Production Hubs:** Top producing nations.
- **Platform Preferences:** Who collaborates with which region.

Example Findings:

- Heavy US production → domestic-first content model.
- Multiple countries → global diversification.
- Movie vs show split → content-type specialization.

Visualization Summary:

A **stacked or grouped bar chart** showing origin-country patterns across content types.