AllView OTT - A Data-Driven Streaming Platform Powered by Market Intelligence and Visual Insights

1st Surya Chandra Raju Kurapati
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x23396920@student.ncirl.ie

2nd Nisarga Holenarasipur Ramesh
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x24101028@student.ncirl.ie

3rd Saranya Munikannaiah
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x23434821@student.ncirl.ie

***Abstract*—The entertainment industry is experiencing rapid global expansion, fueled by a surge in the production of movies, TV shows, and digital content. Simultaneously, the fast-paced lifestyles of modern consumers have driven a growing demand for flexible, on-the-go content consumption—making OTT platforms more essential than ever. With an increasing number of platforms competing for viewers' attention, launching a new product in this market requires more than just a compelling content library. It demands data-driven insights to understand shifting market dynamics and evolving consumer behavior in order to gain a competitive edge.**

This project explores the ETL pipeline design and analytical foundation required for processing and visualizing competitors' data to support the successful launch of *AllView*—an upcoming OTT platform. The core research of this paper is rooted in deriving insights from the data analysis of leading OTT platforms in the industry, such as Amazon Prime Video, Netflix, Hulu, and Disney+.

The project encompasses a custom end-to-end pipeline developed using the LIFE framework—covering data ingestion and cleaning of structured and semi-structured data, the development of a unified data model, and the construction of materialized views and dashboards for analysis and visualization. We employed a custom Python-based ETL framework called LIFE Framework, along with PostgreSQL and MongoDB for data storage, and custom dashboards built using Streamlit and Plotly to present findings that support strategic decision-making. The results are a well-rounded technical prototype and evidence-backed strategic insights that align to launch a scalable and data-driven OTT platform AllView to achieve launch target and 1 year goal.

***Index Terms*—AllView OTT, LIFE Framework, ETL, SCD, Data-driven, Analytics Dashboard, Competitors, Structured and Semi-Structured data, PostgreSQL, MongoDB, Streamlit**

I. Introduction

AllView isn't just another streaming service—it's a data-driven OTT platform designed to optimize content, guide global expansion, and target audience engagement, all backed by actionable insights to smash launch and Year 1 goals.

This research paper projects the implementation of the LIFE Framework and the strategic development of *AllView*, an upcoming OTT platform guided by in-depth analysis of competitors such as Netflix, Disney+, Amazon Prime, and Hulu. The framework delivers a scalable ETL pipeline and real-time analytics dashboard that tracks key performance indicators—from content performance to subscriber churn—providing actionable insights. This data-driven approach forms the foundation of our market strategy, transforming raw analytics into a competitive playbook for a successful platform launch and sustained growth.

Using a custom-built ETL pipeline powered by the LIFE framework (Load, Integrate, Filter, Export), this research paper consolidates diverse data sources—storing structured data in PostgreSQL and semi-structured data in MongoDB—to enable unified analytics. Interactive dashboards developed with Streamlit and Plotly provide real-time visibility into key performance metrics such as content volume, platform benchmarking, growth trends, and audience segmentation, enabling rational and data-informed decisions. To support this analytical foundation, a scalable data model and layered architecture have been implemented, along with historical tracking through Slowly Changing Dimensions (SCD) to ensure accuracy, traceability, and adaptability. The insights derived from this framework directly inform AllView's strategic launch targets, year-one objectives, and long-term roadmap for global expansion.

This report outlines the motivation, objective and methodologies we used for building the AllView OTT platform, setting the stage for a technology-powered data-driven entry into the global OTT market.

II. Literature Review

Several studies have examined the OTT market through various analytical lenses. Agarwal, Reeti, et al. [1] analyzed various content among all streaming platforms and its impact on subscriber withholding. These findings suggest that different genres of content significantly control the subscribers' decisions. Bokstr¨om, Victor, and Elin Eriksson. [2] explored user engagement metrics on OTT platforms, highlighting the correlation between personalized recommendations and screening time. They highlight the significance of algorithmic content curation in enhancing user experience. Regarding technical approaches, Gupta, Indranil, Vaijayanth Raghavan, and Mainak Ghosh. [3] demonstrated the efficient way of storing, transforming, and loading semi-structured media metadata in MongoDB for better results. Their work shows that NoSQL databases offer superior flexibility for handling heterogeneous data sources compared to traditional relational databases. Ge,

Shihao, et al. [4] proposed an ETL framework designed especially for streaming service data, combining batch and real-time processing experience. Their work achieved significant improvements in processing data efficiently and performing analytical queries. Although the authors provide valuable insights, there remains a gap in research that needs to be addressed, where technology combines ETL pipeline implementation with broad cross-platform analysis to bring an innovative entry strategy to market. Our paper fills this gap by integrating the customized LIFE Framework ETL pipeline and data derived from cross-platform analysis and visualization of user consumption of data based on the behavior, screening time, genre, and director's content.

III. Methodology

A. Data Collection

For the purpose of this study, data was collected from four of the most popular OTT platforms in the global entertainment industry: Netflix, Disney+, Amazon Prime, and Hulu. We chose these platforms' data strategically due to their worldwide user base and market dominance, which gives a collective presentation of international streaming trends. The datasets serves the purpose of it's collection by providing valuable information for understanding global audience behavior, including genre preferences, content engagement patterns, and viewing tendencies. By analyzing these platforms, the study aims to reveal meaningful insights into current entertainment consumption trends and inform strategic decisions for the proposed AllView OTT platform.

Dataset specific insights: The datasets were sourced from Kaggle.com and can verified in the below link. (https://www.kaggle.com/discussions/general/281540). The dataset has subscriber metrics, content ratings, viewing statistics, user reviews, platform interactions and financial data. Structured Data: Netflix, Amazon Prime and Hulu platforms data was obtained in a structured format and provided as CSV files. These structured datasets were ingested and stored in a PostgreSQL database for further processing and analysis. Semi-structured Data: Disney+ Hotstar's data was available in a semi-structured JSON format. To ensure efficient handling and scalability, this data was initially stored in MongoDB. Subsequently, it was transformed to align with a traditional relational database structure and then migrated into PostgreSQL to support consistent data integration and downstream analytics within the project framework.

B. ETL Pipeline Architecture

The Architecture begins with ingesting data from both structured and semi-structured sources of competitor data. This raw data enters an Extract, Load, Transfer (ETL) pipeline built using a custom Python framework called LIFE (Load Integrate Filter Export). The pipeline consists of three layers:.

Stage Layer: "Fig. 1" illustrates the extraction of data from diverse source types using the framework that leverage libraries such as pandas and json. The extracted data is staged according to its format—structured data (CSV files) is loaded into PostgreSQL, while semi-structured data (JSON files) is stored in MongoDB. The diagram depicts the corresponding database schemas and collections. In this layer, Slowly Changing Dimensions Type 1 (SCD Type 1) is implemented to support scalability and accommodate future enhancements, including frequent or daily data loads.

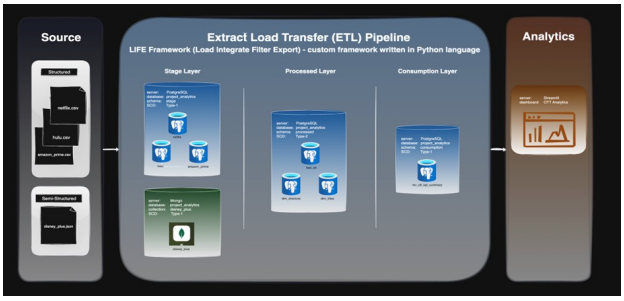

Fig. 1. Custom ETL Pipeline Architecture Overview

Processed Layer: Upon staged data, business logic is applied along with historical tracking. The refined data is then stored in a structured format within the PostgreSQL database using SCD Type 2 methodology. This approach enables the tracking of historical changes over time by maintaining versioned records in the form of fact and dimension tables.

Consumption Layer: As shown in "Fig. 1", the consumption layer serves as the final stage in the data pipeline, where modelled data from the processed layer—comprising facts and dimensions—is used to create materialized views. These views are auto-refreshed at defined intervals and stored in a dedicated PostgreSQL instance or schema, containing the refined data essential for analytics. This layer supports flexible analytical queries across multiple features, leveraging the historical context maintained in the processed layer. It also acts as the primary data source for the analytics module, powering dynamic dashboards and interactive visualizations that track Key Performance Indicators (KPIs) and other strategic metrics.

C. Data Model

"Fig. 2" illustrates data model that integrates information from separate staging tables for Netflix, Disney Plus, Hulu, and Amazon Prime into a central processed fact table (processed.fact_ott) for AllView OTT platform analysis.

D. LIFE Framework (Load Integrate Filter Export)

This project leveraged a custom ETL framework named LIFE "Fig. 3", developed in Python, to build a robust data pipeline for the launch of the AllView OTT platform. The framework comprises two main pipelines: stage and processed loader pipelines.

The execution of either pipeline is determined by the 'run_layer' parameter specified in a configuration file. The stage pipeline is the initial phase of the ETL process. It begins by reading source data, which can be in CSV or JSON format. Upon ingestion, the pipeline performs schema

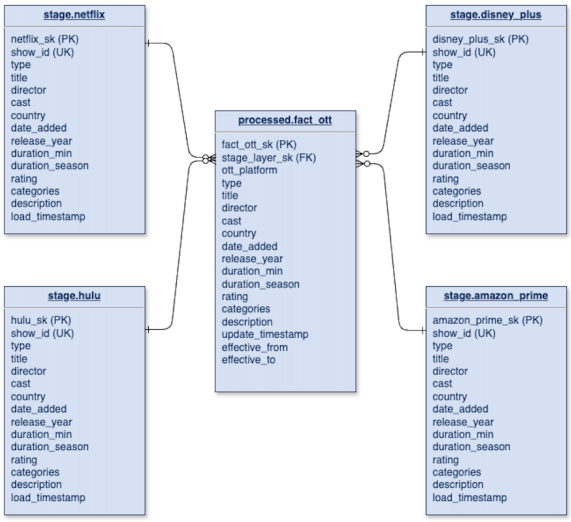

Fig. 2. Data Model

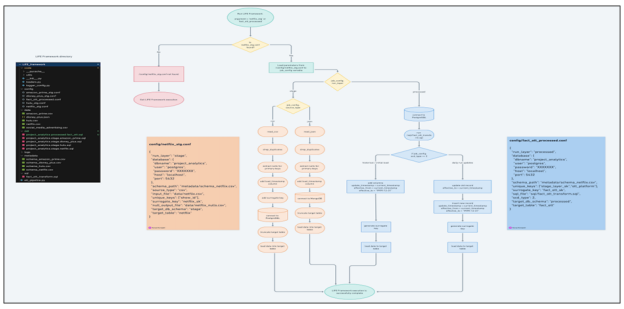

Fig. 3. LIFE Framework flowchart

validation to ensure data consistency. It then proceeds with data deduplication and null checks based on unique keys defined within the configuration file. Any records that fail these checks are flagged and written to a separate CSV file for further business review and potential correction. To ensure data traceability throughout the pipeline, a surrogate key is added to each record. Finally, the cleansed and prepared data is loaded into the appropriate database based on its structure: structured data is loaded into PostgreSQL using an SCD Type 1 (truncate and load) strategy, while semi-structured data is loaded into MongoDB, also using an SCD Type 1 strategy.

The processed pipeline focuses on applying business logic and managing historical data. It takes the enriched data from the staging area and applies SQL queries that encapsulate the necessary business rules to transform the data into meaningful facts and dimensions. This pipeline implements SCD Type 2 to track historical changes in the data. For each record, an 'effective_to' timestamp is maintained. The current, active record is marked with an 'effective_to' timestamp set to a high-end date ('9999-12-31'). When an update to a record occurs, the existing record's 'effective_to' timestamp is updated to the current date, effectively closing that historical version, and a new record with the updated information and a new 'effective_from' timestamp (the current date) is inserted. To guarantee uniqueness in the Type 2 table, a surrogate key is generated based on all the columns of the record before loading the data into the final destination.

The LIFE Framework is designed for ease of execution. To run the entire pipeline, the user simply provides the name of the configuration file as a command-line argument. This triggers the framework to execute the defined stage and processed pipelines in sequence, ultimately loading the transformed data into the final data warehouse. Furthermore, a well-defined logging strategy is integrated into the framework to meticulously track any errors or issues that may arise during any stage of the pipeline, facilitating debugging and monitoring.

IV. Data Visualization Methodology

An interactive dashboard is developed using Streamlit as the front-end framework and Plotly for dynamic and aesthetically pleasing visualizations, providing a consolidated view of key performance indicators (KPIs) derived from the data processed through the custom-built ETL LIFE framework. This visualization layer sources its data from the Consumption Layer, which utilizes materialized views created from the final refined data model in the processed layer. These curated metrics allow stakeholders to interact with high-level summaries of data while also offering the flexibility to drill down into specific analytical dimensions for more granular insights with filters to allow data slicing according to the OTT platform and release year.

Key visual components implemented in the dashboard include:

- **Total Titles and Platform Comparison:** Bar charts and summary panels compare the total volume of content across Netflix, Hulu, Amazon Prime, and Disney+, helping to assess the scale and diversity of each platform.

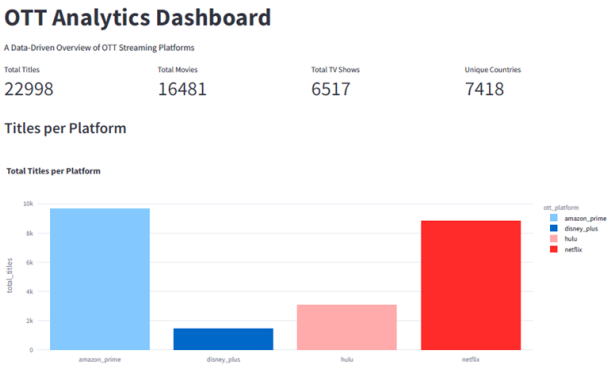

Fig. 4. Total Titles and Platform Comparison

- **Movies vs. TV Shows Distribution:** Dual bar visualizations present the proportion of movies and TV shows per platform, aiding in content mix analysis and guiding platform positioning strategies for AllView.

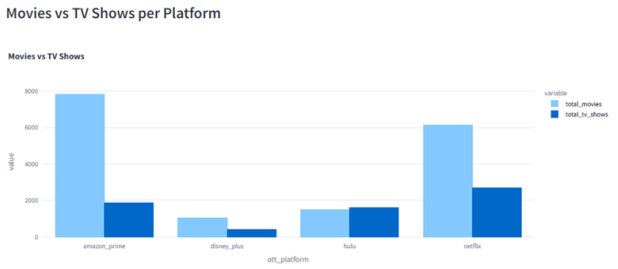

Fig. 5. Movies vs. TV Shows Distribution

- **Average Movie Duration per Platform:** This metric is visualized to identify content pacing preferences across platforms, supporting AllView's editorial and acquisition strategies.

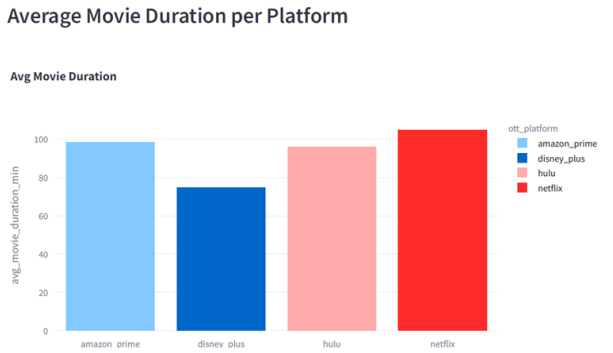

Fig. 6. Average Movie Duration per Platform

- **Monthly Content Addition Trends:** Time-series visualizations reflect how frequently new titles are added, helping to model potential growth strategies and benchmark AllView's rollout plans.

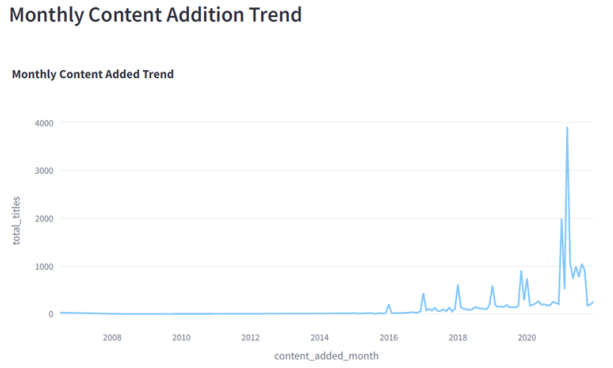

Fig. 7. Monthly Content Addition Trends

- **Unique Director Contributions:** Comparative analysis of directorial diversity across platforms informs the project's talent acquisition strategy, particularly emphasizing AllView's focus on promoting indie and first-time creators.

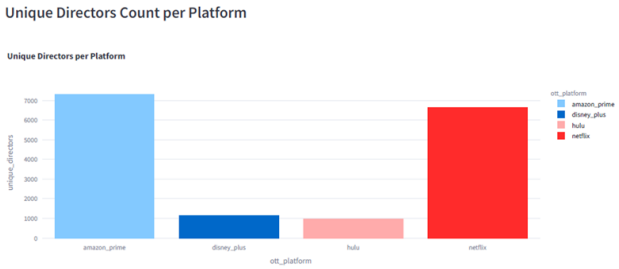

Fig. 8. Unique Director Contributions

- **Average Seasons per TV Show:** This measure, visualized as platform-wise comparisons, supports AllView's plan to emphasize short-format series and mini-series with high viewer retention.

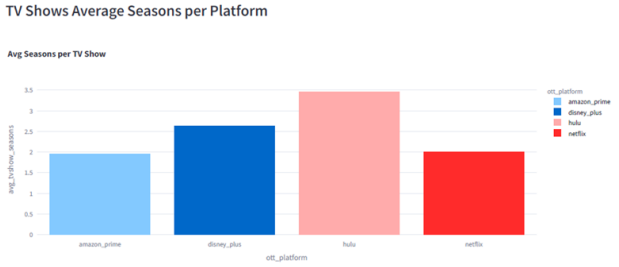

Fig. 9. Average Seasons per TV Show

Each chart is interactive, offering filtering "Fig. 10" and hovering functions, which make the dashboard not just a static report but a decision support system. The selection of visualization types adheres to best practices in data storytelling, ensuring readability, relevance, and usability for both technical and non-technical stakeholders.

Dashboard Layout Availability

The wireframe and layout of the end-to-end dashboard are available in our GitHub repository: https://github.com/suryakurapati-edu/project-Analytics/blob/main/OTT_Analytics_Dashboard/dashboard_layout.pdf

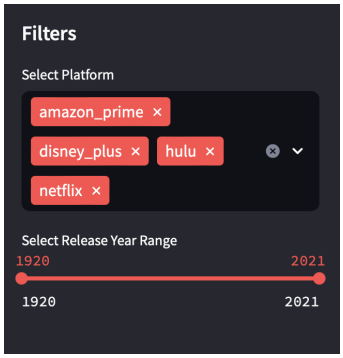

Fig. 10. Dashboard Filters

V. Results and Analysis

The analytical component of this project was designed to extract actionable insights from aggregated OTT datasets across four dominant competitor platforms. Netflix, Amazon Prime, Hulu, and Disney+. The results presented here are based on the coordinated data model developed through the custom LIFE (Load, Integrate, Filter, Export) ETL framework and visualized through an interactive dashboard powered by Streamlit and Plotly.

A. Content Volume and Distribution

Netflix leads with the largest content library, exceeding 10,000 titles, followed by Amazon Prime. Hulu and Disney+ maintain smaller, but curated libraries. This disparity shows the need for AllView to position its content strategy competitively, with an initial launch target of 5,000 titles and a projected first year expansion to 10,000-15,000 titles. The breakdown between movies and TV shows further reveals that most platforms maintain a 70:30 ratio, emphasizing feature films. This insight aligns with AllView's strategy to mirror industry distribution while also promoting unique formats such as short films and limited series to differentiate its offering.

B. Content Duration and Consumption Trends

The average movie duration was analyzed across platforms to assess pacing preferences. Results indicate that the majority of titles fall within the 80–100 minute range, which correlates with typical consumer viewing patterns favoring concise, engaging content. Disney+, in contrast, exhibited a slightly lower average duration, likely reflecting its focus on family-oriented and animated productions. These findings affirm AllView's focus on short-to-mid-length content for higher completion rates and engagement.

C. Growth Patterns and Geographic Reach

Time-series analysis of content additions shows that Netflix and Amazon Prime maintain steady monthly growth, averaging 100–150 new titles, scaling up to over 200 titles per month in some cases. This informed AllView's content acquisition roadmap, which targets a similar growth curve. Additionally, the geographic distribution of content highlights that while existing platforms are heavily concentrated in North American markets, there is a significant opportunity for AllView to expand into underrepresented regions by curating multilingual and culturally diverse libraries.

D. Talent and Programming Insights

The project also examined creative diversity through director-level analysis. Over 300 unique directors were identified, with approximately 30% being debut or indie filmmakers. This supports AllView's proposed strategy of nurturing emerging talent and highlights a market segment not fully leveraged by incumbent platforms. Moreover, the average number of seasons per show revealed that mini-series formats (6–10 episodes) dominate across platforms, suggesting high audience receptivity to limited, high-quality story arcs — a key format in AllView's planned programming.

E. Strategic Implications

The cross-platform analysis not only enabled benchmarking but also provided forward-looking guidance for AllView's launch. The results informed several strategic pillars:

- Content acquisition should emphasize quantity without sacrificing diversity, targeting the 70:30 movie-to-show ratio.
- Geographic content planning should prioritize multilingual availability and regional inclusivity.
- Talent onboarding strategies should promote indie creators to enhance platform identity and attract niche audiences.
- Growth planning must align with observed trends, scaling content consistently across the first operational year.

The analytical process demonstrated the value of consolidating structured and semi-structured data into a unified schema, enabling reliable comparisons and deep insights. By leveraging historical and comparative data, the AllView platform is strategically positioned to launch with a data-informed content portfolio and a differentiated market presence.

VI. Conclusion

This paper is establishing the data-driven base for the strategic launch of AllView, a proposed Over-The-Top (OTT) platform designed to enter a competitive and rapidly evolving digital entertainment market. By integrating of structured and semi-structured data from leading industry players - Netflix, Amazon Prime, Hulu, and Disney+ — and the implementation of a custom Python-based ETL pipeline under the LIFE (Load, Integrate, Filter, Export) framework, the project successfully consolidated disparate datasets into a unified analytics environment. The application of Slowly Changing Dimensions (SCD) methodologies enabled robust historical tracking and enriched the data's temporal value. A multi-layered architecture facilitated the processing of large volumes of content-related data, while interactive dashboards delivered actionable insights into user behavior, content preferences, platform growth strategies, and competitive positioning. These results collectively informed AllView's initial launch targets, content acquisition strategy, and expansion roadmap. From a technical perspective, the project demonstrated the effective use of PostgreSQL and MongoDB in tandem, leveraging their respective strengths in handling structured and semi-structured data. The use of tools such as Streamlit and Plotly further enabled intuitive visualization of complex metrics, ensuring accessibility for both technical and business stakeholders.

TABLE I
STRATEGY TARGETS AND YEAR 1 GOALS FOR ALLVIEW OTT PLATFORM

Strategy Area	**Metric / Insight**	**Launch Target**	**Year 1 Goal**
Content Volume	Total Titles	5,000 titles	10,000–15,000 titles
	Movies	3,500 titles	6,000–8,000 titles
	TV Shows	1,500 titles	3,000–5,000 titles
Content Mix	Distribution	70% Movies / 30% TV Shows	Maintain optimal mix
	Avg Movie Duration	80–100 mins	Same
	Short Films	Include 10–15%	Optional increase
Platform Benchmarking	Netflix	10,000 titles (8,000 movies, 2,000 shows)	Benchmark
	Amazon Prime	9,000 titles (8,000 movies, 2,000+ shows)	Benchmark
	Hulu	2,000 titles	Match or exceed
	Disney+	1,500 titles	Match or exceed
Monthly Growth	New Titles	100–150 titles/month	150–250 titles/month
Global Strategy	Geographic Coverage	Content from 20+ countries	Expand to 50+ countries
	Localization Strategy	Emphasis on regional/dubbed content	Continue to scale
Talent Strategy	Directors	300+ unique directors	Maintain or increase
	Indie Representation	30% debut/indie filmmakers	Continue strategy
	Avg Seasons	1–2 seasons/show	Increase if retention supports
TV Show Strategy	Series Format	Favor mini-series (6–10 episodes)	Continue with performance review
	Episode Volume	25–40 new episodes/month	Increase with demand

VII. FUTURE WORK

While the current implementation provides a strong baseline for AllView's launch, several areas have been identified for future enhancement:

- Real-time Data Integration: Future iterations may incorporate API-based ingestion of real-time or streaming data to enable dynamic tracking of user engagement, trending content, and social media reactions.
- Sentiment and Text Analytics: Integrating natural language processing (NLP) on reviews, social media comments, and user feedback can provide deeper qualitative insights into content reception and platform sentiment.
- Geospatial Analytics: Enhancing the regional rollout strategy with detailed geospatial data visualization could support better localization of content and marketing efforts.

In conclusion, this project has successfully laid the groundwork for a data-informed OTT platform launch. The architecture, tools, and insights developed here not only serve the immediate needs of the AllView initiative but also create a scalable foundation for continuous innovation and strategic decision-making in the digital entertainment domain.

ACKNOWLEDGMENT

The author would like to thank "Namita Agarwal" for supporting this research.

TEAM ALLVIEW

This project was driven by a collaborative team with specialized roles to ensure its success. Surya as a Data Engineer architectured and built the end-to-end ETL pipeline using a custom framework, integrating structured and semi-structured data sources while implementing SCD Type 2 for historical tracking. Nisarga as a Business Analyst derived strategic insights, identifying high-ROI content, target demographics, and regional opportunities to guide the platform's launch strategy. Saranya as a Data Analyst developed an interactive analytics dashboard, transforming complex data into actionable KPIs and uncovering trends like optimal release timing and audience engagement patterns. Together, the team combined technical expertise with business acumen to deliver a data-driven foundation for the AllView's success.

CODE AVAILABILITY

The complete source code and implementation details for this project are available at our GitHub repository: https://github.com/suryakurapati-edu/project-Analytics.git

REFERENCES

[1] Agarwal, Reeti, et al. "Over-the-top (OTT) retailing in the post pandemic world. Unveiling consumer drivers and barriers using a qualitative study." Journal of Retailing and Consumer Services 75 (2023): 103529.

[2] Bokström, Victor, and Elin Eriksson. "Customer Retention in OTT Subscription Services: Beyond the Content, Toward Improved Strategies for Enhancing Customer Satisfaction." (2023).

[3] Gupta, Indranil, Vaijayanth Raghavan, and Mainak Ghosh. "Leveraging metadata in no SQL storage systems." 2015 IEEE 8th International Conference on Cloud Computing. IEEE, 2015.

[4] Ge, Shihao, et al. "A scalable framework for multilevel streaming data analytics using deep learning." 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). Vol. 2. IEEE, 2019.