Project Report

TrendLens:
NextGen Short-Video Analysis

IE 6760 Data Warehousing & Integration

Group 4

Student 1: Neha Patil

Student 2: Medhavi Uday Pande

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Neha Patil

Signature of Student 2: Medhavi Uday Pande

1. Summary

The proliferation of short-form video platforms like TikTok and YouTube Shorts has fundamentally transformed digital content consumption. With billions of videos circulating monthly, creators and marketers face a significant challenge in deciphering the specific mechanics that drive virality and engagement. The TrendLens project addresses this gap by establishing a structured, scalable data warehousing framework designed to analyze engagement trends with precision.

To achieve this, the project implemented a Galaxy Schema data warehouse and a comprehensive ETL pipeline utilizing Talend and PostgreSQL. This pipeline extracts, transforms, and integrates fragmented video metadata and engagement metrics, processing nearly 70,000 records into a unified analytical environment. The system enables advanced analytical capabilities, including ROLAP operations, Type 3 Slowly Changing Dimensions (SCD) for tracking creator evolution, and cross-platform performance comparisons. By converting inconsistent, high-volume data into structured assets, TrendLens empowers stakeholders to identify optimal posting times, high-impact content types, and the underlying drivers of virality.

2. Background

By 2025, short-form video content has solidified its position as the core of modern digital interaction. These platforms are defined by rapid trend cycles, high creativity, and constant audience migration. However, despite the richness of this ecosystem, meaningful analysis remains elusive. Data structures vary drastically across platforms, engagement metrics evolve rapidly, and the velocity of content production often outpaces traditional tracking methods.

The TrendLens dataset aggregates critical information from TikTok and YouTube Shorts, including video attributes, creator profiles, hashtag usage, and regional identifiers. The raw data, however, was plagued by inconsistencies, varying formats, and platform-specific definitions. To overcome these hurdles, a centralized and structured data warehouse was essential. TrendLens provides this unified analytical space, allowing creators and analysts to move beyond intuition and make evidence-based decisions regarding content strategy and audience growth.

3. Problem Definition

The short-video analytics landscape presents several distinct, interconnected challenges that this project aims to resolve:

- **Data Fragmentation:** Platforms utilize disparate data structures, rendering direct cross-platform comparison difficult without substantial harmonization.
- **Inconsistent Data Quality:** The raw data suffers from missing fields, inconsistent timestamps, and varying naming conventions that limit immediate analysis.
- **Complex Derived Metrics:** Core performance indicators such as retention rates, engagement velocity, and interaction ratios are not natively available and must be calculated from raw logs.
- **High Content Velocity:** The rapid evolution of trends means that without an automated, structured pipeline, insights become obsolete almost immediately.

TrendLens addresses these issues through rigorous data standardization, intelligent dimensional modeling, and automated ETL processing.

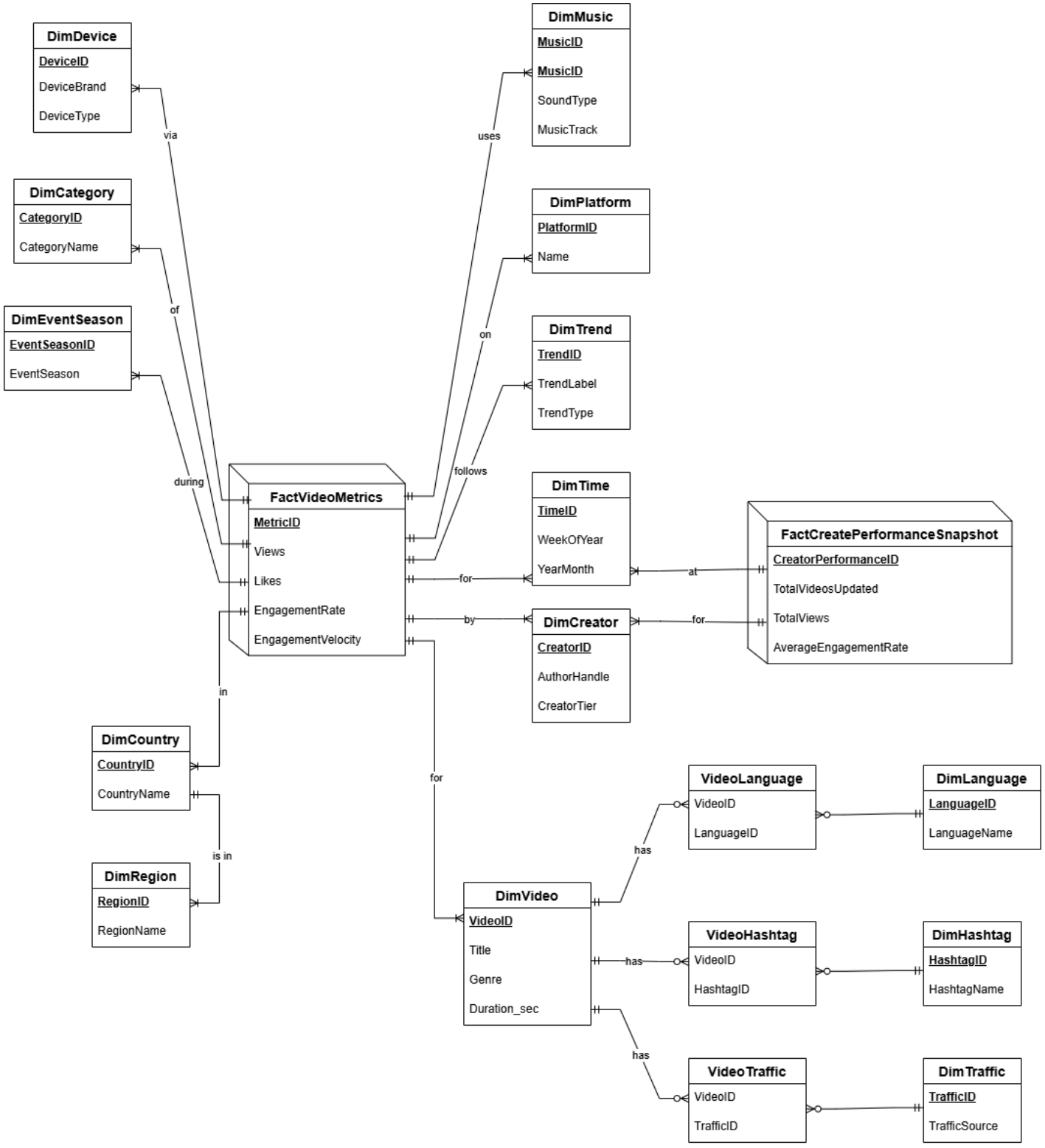
4. Objectives

The primary objectives of the TrendLens project were to:

- **Extract & Ingest:** Efficiently pull raw data from the 2025 dataset, covering diverse entities such as videos, creators, trends, and platform metrics.
- **Clean & Transform:** Standardize inconsistent fields, compute essential analytical metrics (e.g., Engagement Rate), and resolve schema differences between platforms.
- **Analyze & Compare:** Uncover variances in trend performance across different content categories, geographic regions, and creator tiers.
- **Load & Structure:** Construct a robust data warehouse supporting complex queries, including drill-down, roll-up, and time-series analysis.
- **Visualize Engagement:** Deploy geospatial maps and interactive dashboards to identify audience hotspots and peak activity windows.

5. Conceptual Model

The conceptual model serves as the blueprint for the data warehouse, defining the relationships between key entities in the short-video ecosystem.



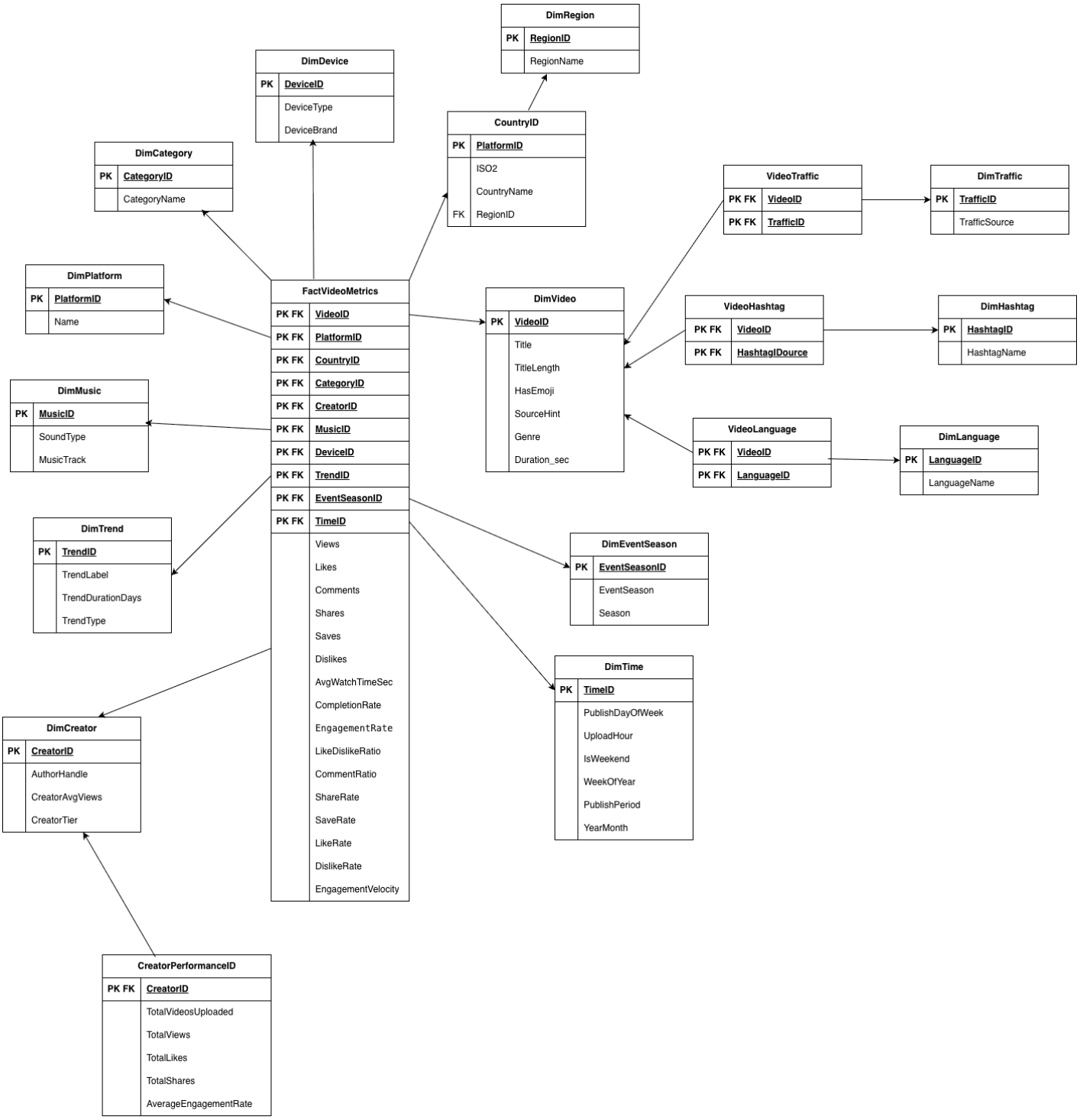
- **Core Entities:**
  - Video: The central unit of content.
  - Creator: The entity producing the content.
  - Platform: The distribution channel (TikTok, YouTube Shorts).
  - Region & Country: Geographic dimensions for spatial analysis.
  - Trend: Categorical identifiers for trend types and lifecycles.
- **Attributes:** The model captures both static metadata (upload time, language, genre) and dynamic engagement data (views, likes, shares, saves).
- **Relationships:** Complex many-to-many relationships (e.g., Video-to-Hashtag, Video-to-Language) are managed through bridge structures to ensure accurate associations.

This design ensures TrendLens supports multi-dimensional analytics while retaining the rich context of the source data.

6. Logical Model

The logical model translates the conceptual design into a Galaxy Schema optimized for analytical performance.

- **Fact Tables:**
  - FactVideoMetrics: Captures atomic, daily performance metrics for individual videos.
  - FactCreatorPerformanceSnapshot: Provides aggregated, creator-level performance summaries.
- **Dimension Tables:** A suite of 15 dimensions including DimCreator, DimTrend, DimCategory, DimRegion, and DimTime provides descriptive context.
- **Bridge Tables:** To maintain referential integrity in many-to-many relationships, the model includes bridge_video_hashtag, bridge_video_language, and bridge_video_trafficSource.

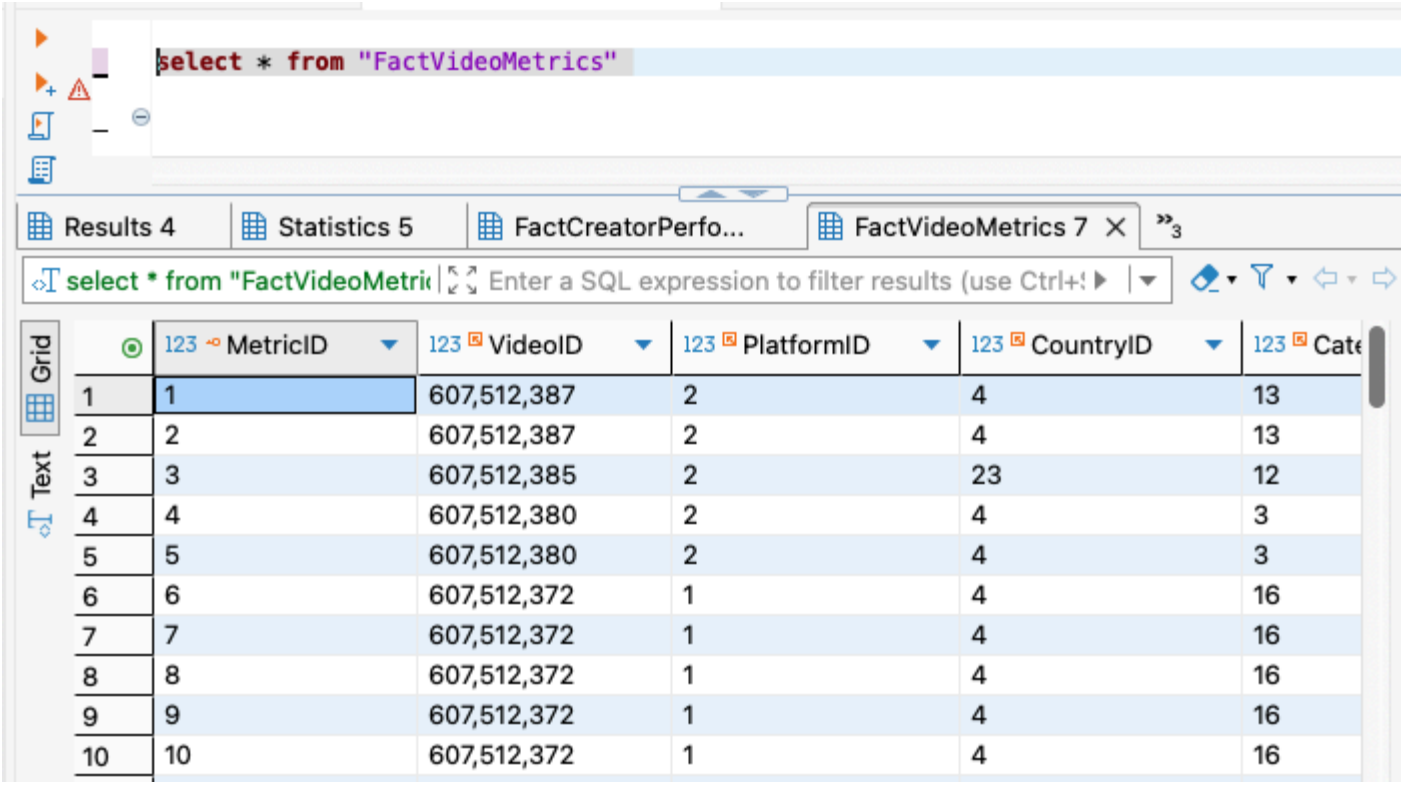


This structure facilitates robust cross-platform comparisons and deep dives into specific engagement behaviors.

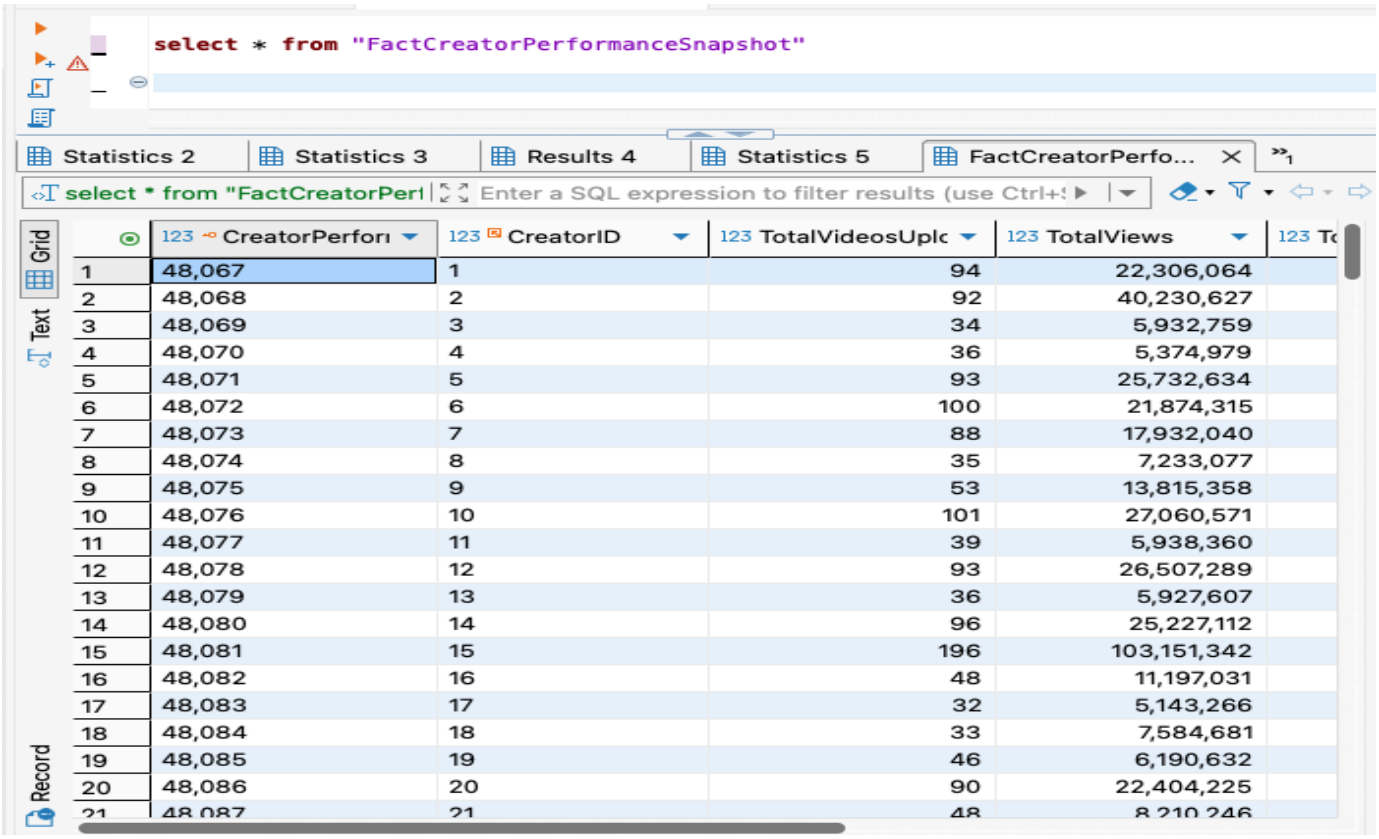
7. Loading Data into PostgreSQL Database

The process of loading data into the PostgreSQL database began with acquiring the YouTube Shorts and TikTok Trends 2025 dataset from Kaggle, which served as the foundation for the project's analytical framework. The raw dataset contained approximately 48,000 records but presented challenges such as inconsistent formats, unstructured metrics, and platform-specific variances that required significant preprocessing. To address these complexities, an automated ETL pipeline was engineered using Talend, utilizing components like tMap and tAggregateRow for data manipulation, cleaning, and the derivation of key metrics such as average engagement rates. A staging layer was also employed to temporarily store raw data, allowing for validation and structural adjustments before final processing.

Once the data transformation logic was established, a PostgreSQL database named TrendsLens was created to house the data warehouse. The database schema was meticulously designed based on a Galaxy Schema logical model, featuring two central fact tables and fifteen dimension tables to support multi-dimensional analysis. Primary and foreign keys were strictly enforced to maintain referential integrity, while bridge tables were implemented to handle complex many-to-many relationships. The loading process was orchestrated via a sequential control flow, populating dimension tables first to generate surrogate keys, followed by bridge tables, and finally the fact tables. Following the successful load of approximately 69,900 records, validation was conducted using comprehensive SQL queries to verify row counts, check constraint integrity, and ensure the accuracy of the imported data. This rigorous process resulted in a structured, high-quality data warehouse optimized for ROLAP operations and strategic reporting.



	MetricID	VideoID	PlatformID	CountryID	Cate
1	1	607,512,387	2	4	13
2	2	607,512,387	2	4	13
3	3	607,512,385	2	23	12
4	4	607,512,380	2	4	3
5	5	607,512,380	2	4	3
6	6	607,512,372	1	4	16
7	7	607,512,372	1	4	16
8	8	607,512,372	1	4	16
9	9	607,512,372	1	4	16
10	10	607,512,372	1	4	16



	CreatorPerfor	CreatorID	TotalVideosUplc	TotalViews	Tc
1	48,067	1	94	22,306,064	
2	48,068	2	92	40,230,627	
3	48,069	3	34	5,932,759	
4	48,070	4	36	5,374,979	
5	48,071	5	93	25,732,634	
6	48,072	6	100	21,874,315	
7	48,073	7	88	17,932,040	
8	48,074	8	35	7,233,077	
9	48,075	9	53	13,815,358	
10	48,076	10	101	27,060,571	
11	48,077	11	39	5,938,360	
12	48,078	12	93	26,507,289	
13	48,079	13	36	5,927,607	
14	48,080	14	96	25,227,112	
15	48,081	15	196	103,151,342	
16	48,082	16	48	11,197,031	
17	48,083	17	32	5,143,266	
18	48,084	18	33	7,584,681	
19	48,085	19	46	6,190,632	
20	48,086	20	90	22,404,225	
21	48,087	21	48	8,210,246	

8. OLAP (Online Analytical Processing)

OLAP is a powerful analytical paradigm used to transform raw, transactional data into meaningful business intelligence. In the TrendLens project, OLAP was critical for addressing the multi-dimensional nature of short-form video metrics from YouTube Shorts and TikTok . By integrating an OLAP system with our Galaxy Schema data warehouse, we moved beyond simple reporting to multidimensional analysis. This enabled us to efficiently query vast datasets across axes such as time, geography, content category, and creator demographics, transforming granular engagement logs into actionable insights regarding content strategy and regional performance.

1. **Total Views and Likes by Region (Roll-up)**

   RegionPerformance ← Rollup * (FactVideoMetrics, DimCountry, DimRegion, CountryName → RegionName, sum(Views), sum(Likes))

   Result ← RegionPerformance

2. **Performance of Creators within the "Gaming" Category (Drill-Down)**

   GamingCreators ← Drilldown * (FactVideoMetrics, DimCategory, DimCreator, CategoryName = 'Gaming' → AuthorHandle, avg(EngagementRate), sum(Views))

   Result ← GamingCreators

3. **Performance Metrics for the "YouTube" Platform (Slice)**

   YouTubeMetrics ← Slice * (FactVideoMetrics, DimPlatform, Name = 'YouTube' → sum(Views), sum(Likes), avg(EngagementRate))

   Result ← YouTubeMetrics

4. **Engagement for "Evergreen" Trends in the USA during "Fall" (Dice)**

   FallTrendAnalysis ← Dice * (FactVideoMetrics, DimTrend, DimCountry, DimEventSeason, TrendType = 'Evergreen', CountryName = 'USA', Season = 'Fall' → avg(EngagementRate))

   Result ← FallTrendAnalysis

This OLAP capability allows stakeholders to seamlessly toggle between high-level strategic views and granular performance details. By uncovering correlations obscured by the sheer volume of data—such as specific trend lifecycles or regional engagement hotspots—the system empowers content strategists and marketers with the evidence-based intelligence needed to optimize schedules and maximize audience reach.

9. ETL Implementation

The Extract, Transform, and Load (ETL) implementation for this project was designed and executed using Talend, forming the backbone of the data warehousing solution. The objective was to efficiently extract raw data from multiple operational sources, transform it into a structured analytical format, and load it into the PostgreSQL data warehouse with full referential integrity and reliability.

Extraction

The extraction phase focused on sourcing raw data from multiple operational datasets related to video content performance and creator analytics. The extracted data included:

- One of the Dimension - Creator Dimension: Containing attributes such as Creator ID, handle, and tier classification.
- Video Dimension: Including metadata such as title, genre, and duration.
- Video Metrics Source: Capturing raw engagement statistics like views, likes, shares, and comments used to populate fact tables.

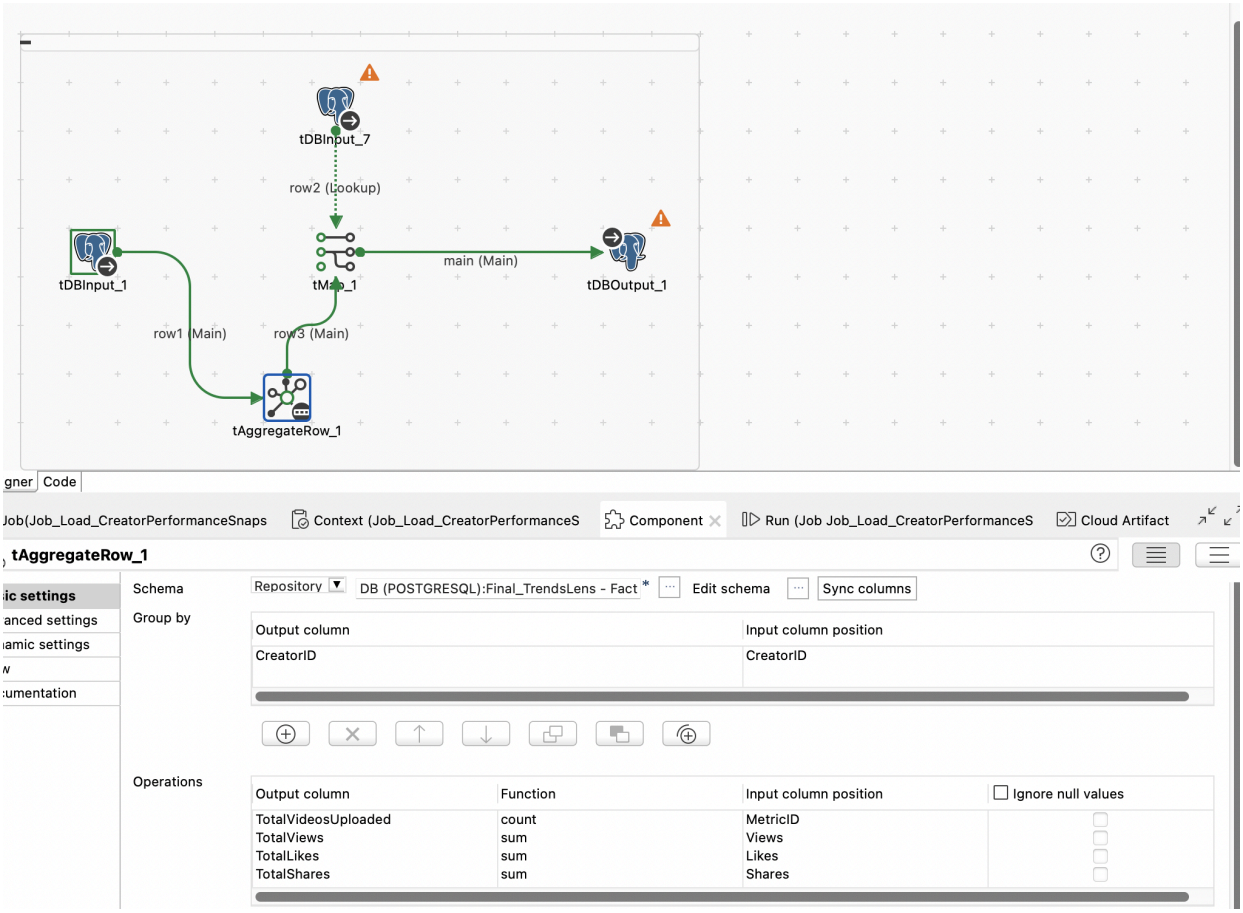
Each source was imported into Talend and connected to dedicated staging tables, serving as an intermediate data layer. This staging process ensured data consistency and readiness for subsequent transformation activities.

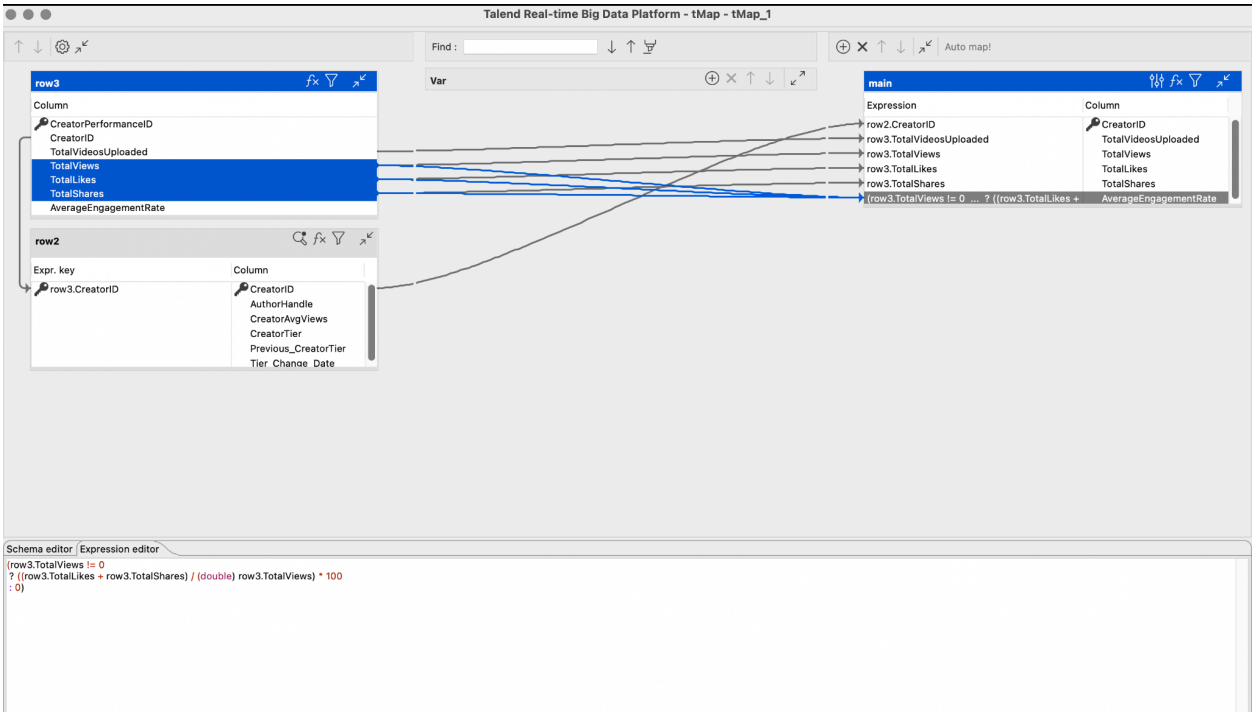
Transformation

The transformation phase was implemented using Talend's tMap and tAggregateRow components to clean, map, and enrich the source data according to the data warehouse schema.

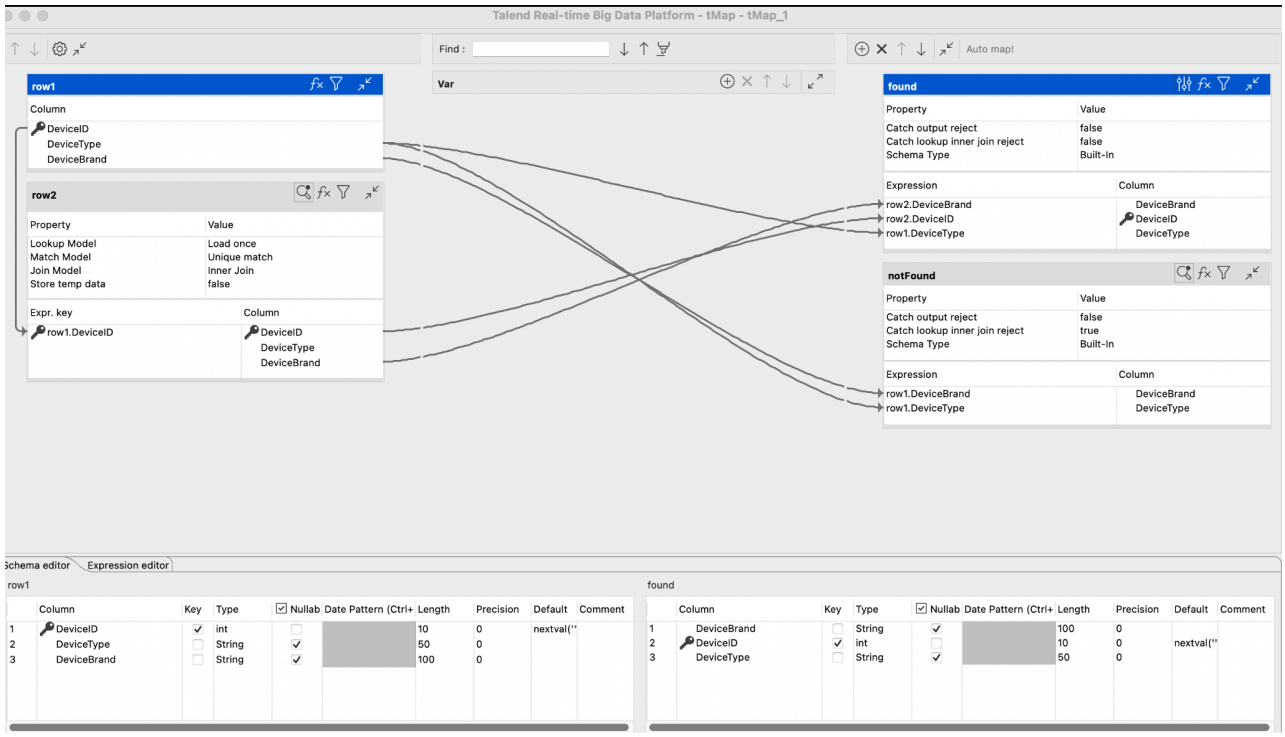
Key transformation activities included:

- **Data Aggregation (Fact Table Preparation):**
  Raw video-level data was aggregated at the Creator level using tAggregateRow. Summarized metrics such as total views, likes, and shares were computed for each CreatorID.
- **Derived Measures:**
  A new calculated metric, Average Engagement Rate, was derived within tMap using the formula: AverageEngagementRate=TotalViews(TotalLikes+TotalShares). Safety checks were applied to prevent division-by-zero errors, ensuring robustness in metric computation.





- **Data Cleansing and Standardization:**
  Transformation logic included handling missing values, normalizing inconsistent formats, and performing data type conversions (e.g., String to Integer/Long).



- **Lookup and Upsert Logic:**
    - Update Existing Records: When a matching natural key was found in the target dimension (e.g., existing CreatorID), the record was updated with the latest information.
    - Insert New Records: If no match existed, the record was inserted as new.

  This hybrid Upsert (Update–Insert) approach ensured all dimensions remained current while maintaining historical accuracy across related entities.

Loading

Following transformation, the processed data was loaded into the target PostgreSQL data warehouse:

- **Fact Tables:**
  The central fact table (FactVideoMetrics) was populated with aggregated measures (views, likes, shares, engagement rate) and foreign keys linking to corresponding dimension tables.

- **Dimension and Bridge Tables:**
  Dimension tables (Creator, Video, Platform, Device, etc.) were updated or extended with new and modified records. Bridge tables such as Bridge_Video_Hashtag and Bridge_Video_Language were populated using inner joins and lookups, ensuring referential integrity and valid foreign key relationships.

Control Flow in the ETL Process

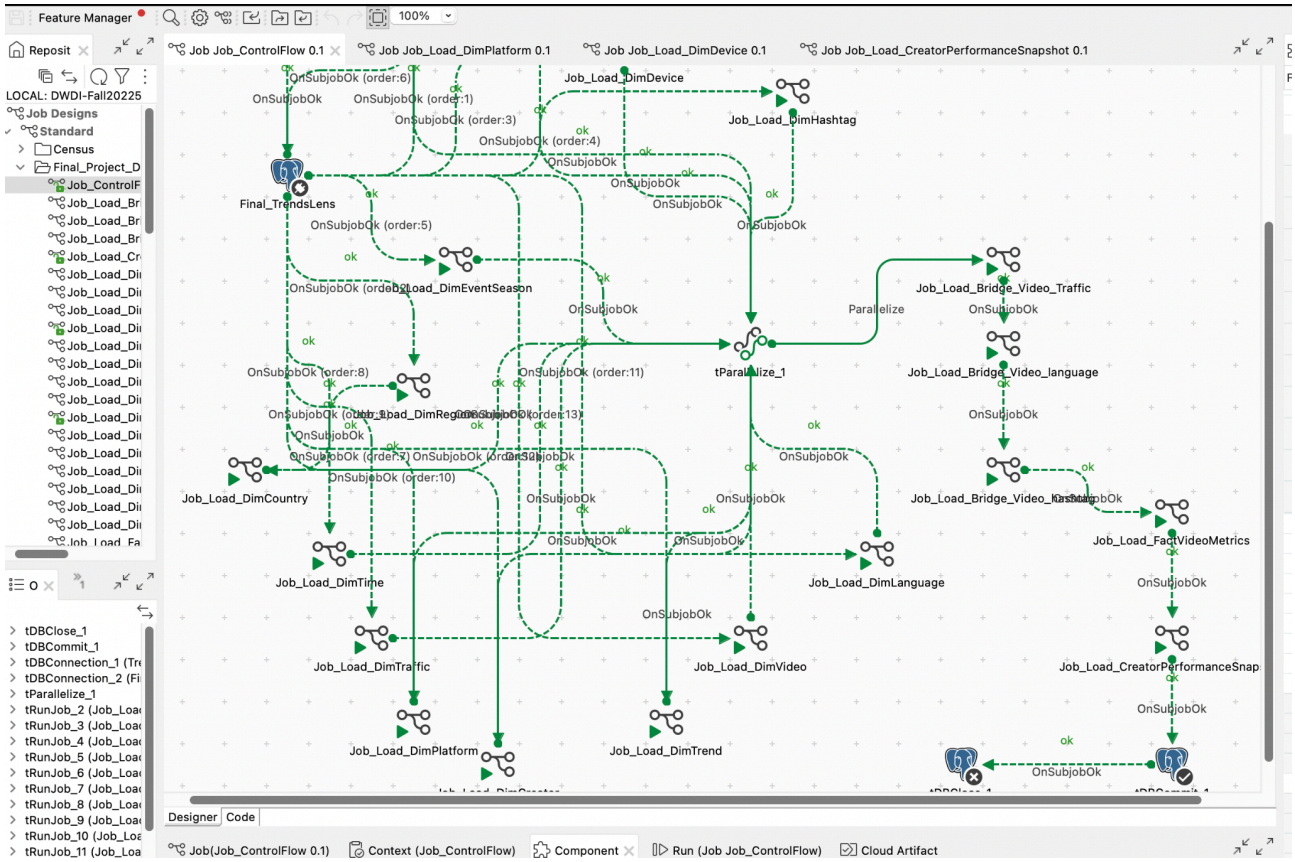
A comprehensive Control Flow Job was implemented in Talend to orchestrate execution sequencing, error handling, and performance optimization:

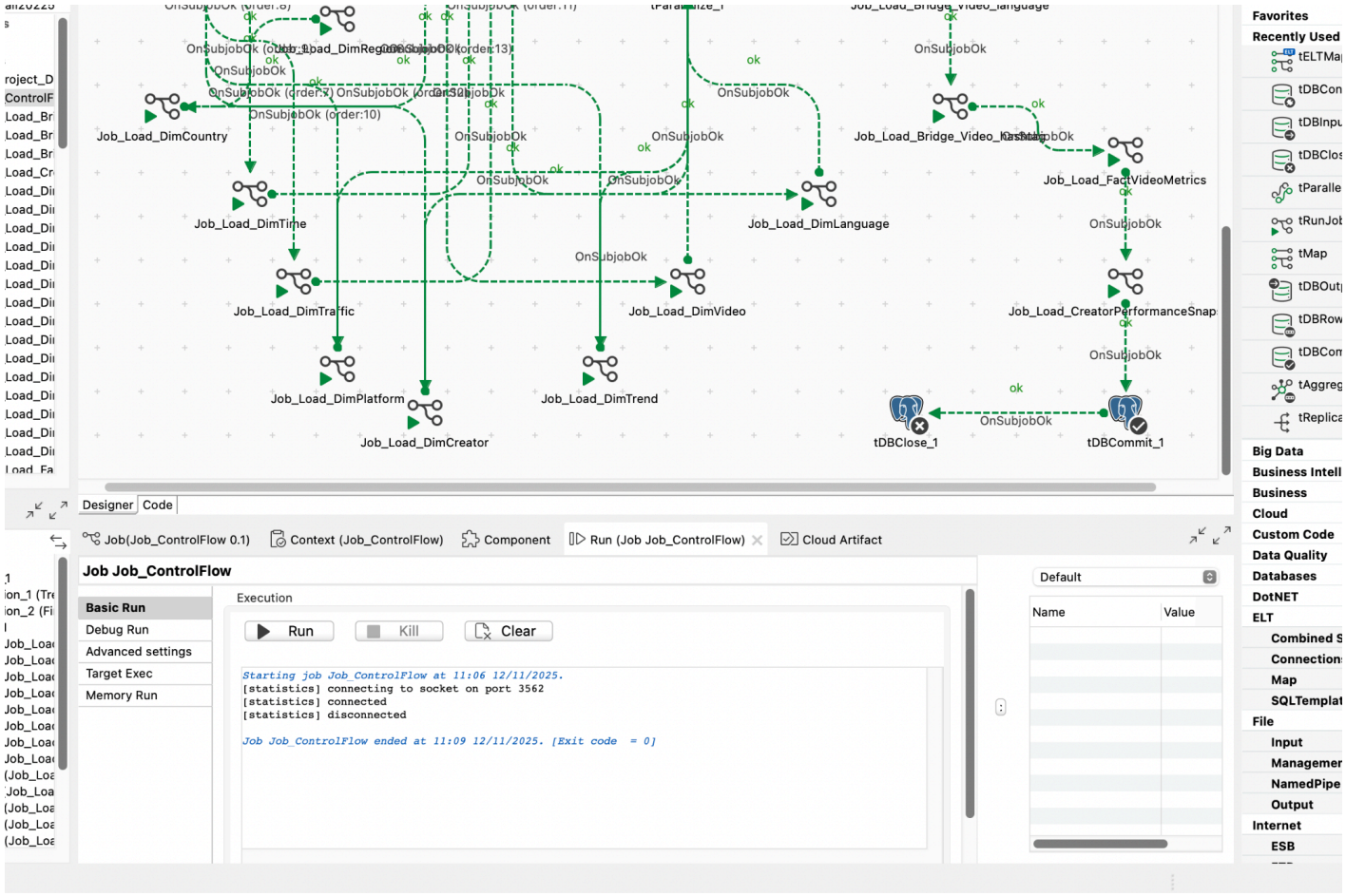
- **Sequential Execution:**
  The `Job_ControlFlow` managed the logical order of execution, loading all dimension tables first (e.g., `Job_Load_DimCountry`, `Job_Load_DimDevice`), followed by bridge tables, and finally the `FactVideoMetrics` table. This ensured all surrogate keys were available for foreign key resolution before loading facts.

- **Parallelization:**
  The tParallelize component was utilized to run independent dimension jobs (e.g., `Job_Load_DimMusic` and `Job_Load_DimDevice`) concurrently, improving overall batch performance.

- **Conditional Execution:**
  The OnSubjobOK trigger was configured across the flow to ensure each downstream job executed only upon successful completion of the preceding one, maintaining robustness and data integrity.

- **Post-Execution Validation:**
  Upon completion, validation queries in PostgreSQL confirmed record counts, referential integrity, and data accuracy.





Through this systematic control flow, the ETL pipeline achieved full automation, reliability, and repeatability, providing a consistent data refresh process for the analytical platform.

10. Handling Slowly Changing Dimensions (SCD)

Given the rapid evolution of content creators, tracking their growth trajectory is vital. TrendLens implemented Type 3 SCD for the Creator Dimension to monitor changes in Creator Tier (e.g., from Beginner to TopTier).

- Schema Enhancements: Columns for Previous_Creator_Tier, Current_Creator_Tier, and Tier_Change_Date were added.
- ETL Logic: When a tier change is detected, the current value is moved to the "Previous" column, the new value is updated in the "Current" column, and the timestamp is recorded.

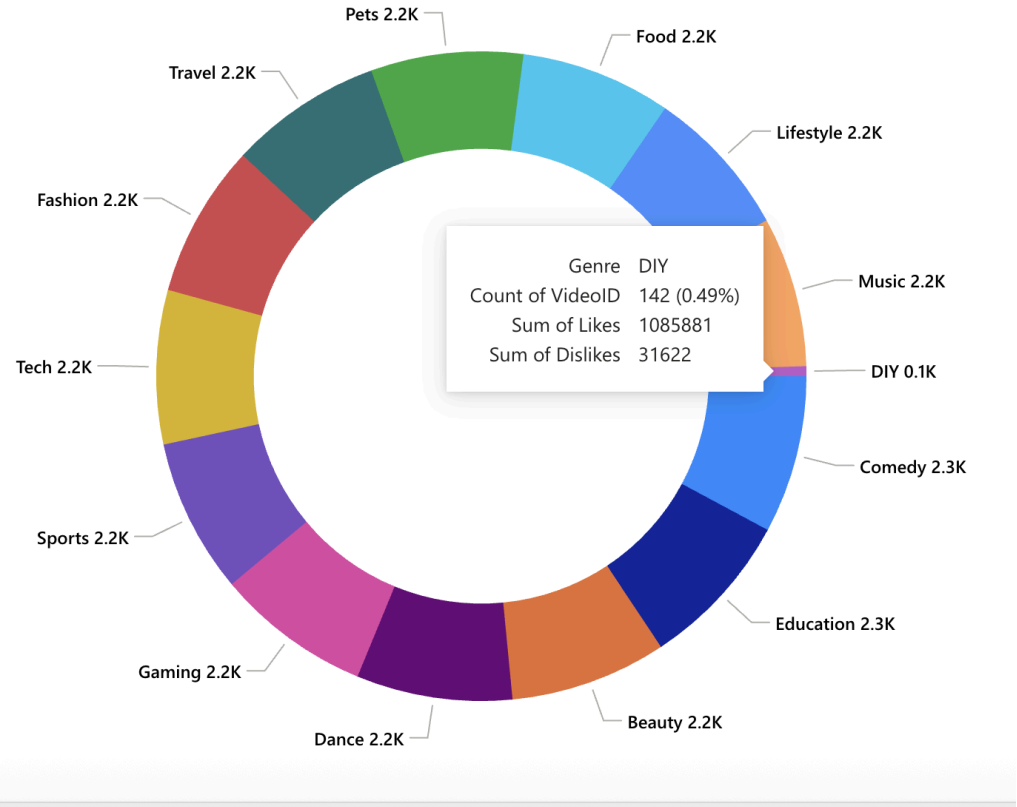
This approach preserves valuable historical context regarding creator evolution without the overhead of creating multiple row versions.

11. Insights & Recommendations

Data analysis within TrendLens revealed several critical insights:

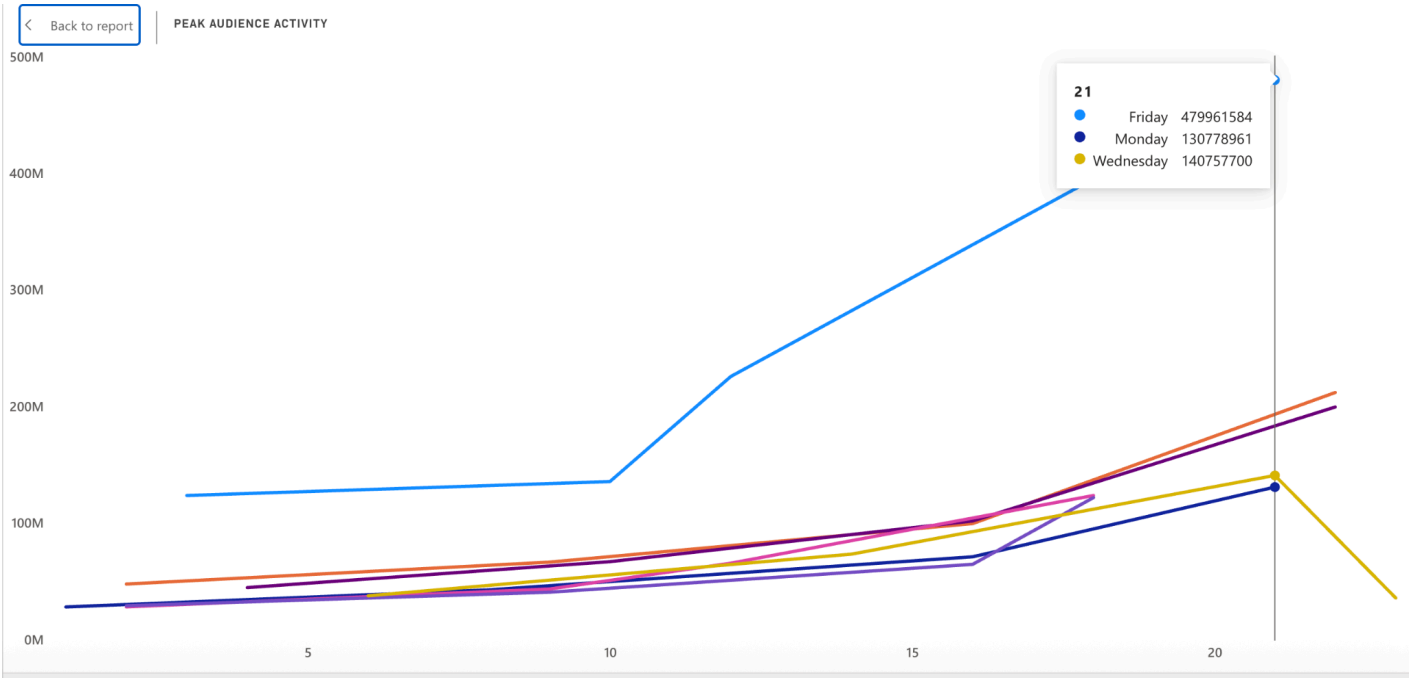
1. Genre Saturation vs. Opportunity

- Insight: Mainstream genres like Travel, Food, and Gaming are highly saturated, each containing ~2.2k videos.
- Opportunity: The DIY category, with only 0.1k videos, represents a significant "blue ocean" opportunity for creators to gain visibility with less competition.



2. Peak Audience Activity

- Insight: User engagement spikes dramatically on Fridays, reaching nearly 480M interactions, significantly outperforming other weekdays.
- Recommendation: Creators should schedule "Hero" or premium content releases for Fridays to maximize organic reach.

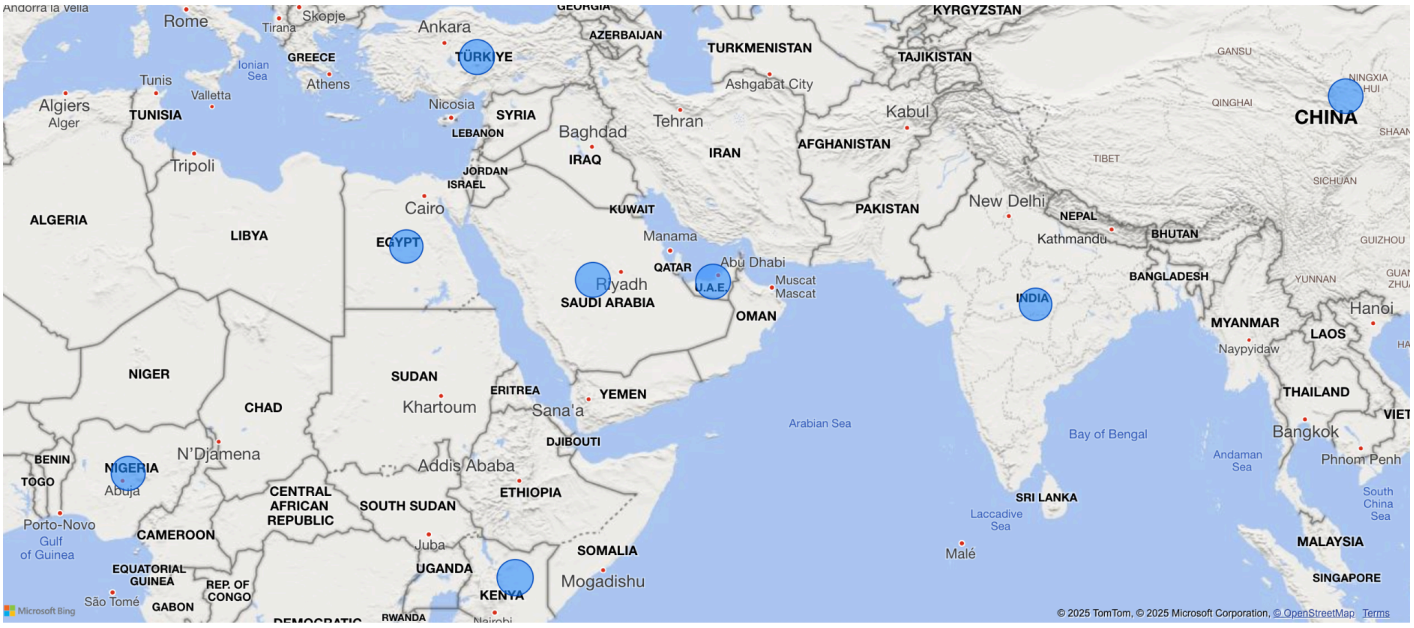


3. Trend Type Performance

- Insight: "Short" trends generated the highest activity (~8M), outperforming "Medium" and "Evergreen" content.
- Implication: Success on these platforms requires agility and the ability to capitalize on fast-paced content cycles.

4. Category-Specific Engagement

- Insight: Different categories drive different behaviors; Education earns the most Saves (utility), while Travel earns the most Likes (aspiration).
- Recommendation: Call-to-actions (CTAs) should be tailored to the category (e.g., "Save this for later" for tutorials).



12. Conclusion

TrendLens successfully transforms fragmented, platform-specific data into a unified, high-performance analytical environment. By leveraging a Galaxy Schema, a robust Talend ETL pipeline, and a PostgreSQL backend, the project processes nearly 70,000 records to support sophisticated OLAP analysis.

The integration of Type 3 SCDs allows for longitudinal analysis of creator growth, while the derived insights regarding posting times, genre saturation, and trend velocity offer immediate strategic value. TrendLens effectively bridges the gap between overwhelming raw data and actionable strategy, laying a strong foundation for future capabilities such as predictive virality modeling and automated trend detection.