12/22/2025

Movie Lens ELT

Report

[1. Introduction 2](#_Toc217469357)

[2. Project Objectives 2](#_Toc217469358)

[3. Dataset Overview 3](#_Toc217469359)

[3.1 Contents: 3](#_Toc217469360)

[3.2 Key Characteristics: 3](#_Toc217469361)

[4. Architecture Overview 4](#_Toc217469362)

[4.1 Ingestion Layer 4](#_Toc217469363)

[4.2 Raw Storage Layer 4](#_Toc217469364)

[4.3 Processing Layer 4](#_Toc217469365)

[4.4 Transformation Layer 4](#_Toc217469366)

[4.5 Analytics Storage Layer 5](#_Toc217469367)

[4.6 Query Engine 5](#_Toc217469368)

[5. Methodology 5](#_Toc217469369)

[5.1 Raw Ingestion (Local → S3) 5](#_Toc217469370)

[5.2 Bronze Layer (S3 → BigQuery) 5](#_Toc217469371)

[Bronze Characteristics 6](#_Toc217469372)

[5.3 Silver Layer (Data Cleaning & Standardization) 6](#_Toc217469373)

[Transformations Applied 6](#_Toc217469374)

[Output 6](#_Toc217469375)

[5.4 Gold Layer 6](#_Toc217469376)

[5.5 Lineage Graph 7](#_Toc217469377)

[6. Cross-Cloud Data 8](#_Toc217469378)

[7. S3 Tables & Iceberg 8](#_Toc217469379)

[8. Dashboard Overview 9](#_Toc217469380)

[8.1 Content Performance & Movie Analytics Dashboard 9](#_Toc217469381)

[8.2 Genre Intelligence Dashboard 10](#_Toc217469382)

[8.3 Tag & Content Discovery Dashboard 11](#_Toc217469383)

[8.4 User Engagement & Behavior Dashboard 11](#_Toc217469384)

[9. Conclusion 12](#_Toc217469385)

# 1. Introduction

Modern analytics platforms increasingly rely on cloud-native, decoupled architectures that separate storage, compute, and transformation layers. This project demonstrates the design and implementation of a production-style ELT pipeline using AWS, Google Cloud, dbt, and Apache Iceberg (S3 Tables).

The pipeline ingests raw MovieLens data, processes it through Bronze, Silver, and Gold layers, and publishes analytics-ready datasets to S3 Tables for scalable querying via Amazon Athena.

The solution emphasizes **modularity, scalability, and data governance**, enabling independent scaling of storage and compute while supporting multiple analytical workloads. By adopting open data formats such as **Parquet** and **Iceberg**, the platform avoids vendor lock-in and ensures long-term interoperability across cloud services.

Additionally, the project showcases **multi-cloud data integration**, where data is processed in Google BigQuery and seamlessly exported to AWS for downstream analytics. This approach reflects real-world enterprise architectures, where organizations leverage best-in-class services across cloud providers while maintaining a unified analytics layer.

# 2. Project Objectives

The primary objectives of this project are:

1. **Build a scalable ELT architecture** using cloud object storage
2. **Demonstrate multi-cloud interoperability** (AWS + GCP)
3. **Implement Medallion Architecture** (Bronze / Silver / Gold)
4. **Apply data modeling and transformation best practices** using dbt
5. **Enable analytics consumption** via Athena and Iceberg-based S3 Tables
6. **Automate schema extraction and metadata generation**

# 3. Dataset Overview

**Dataset:** MovieLens

## 3.1 Contents:

1. **Movies metadata**Movie titles, release years, and genre classifications  
   Supports dimensional modeling and content-based analysis
2. **User ratings**Explicit user–movie rating interactions  
   Enables behavioral analysis, aggregation, and recommendation metrics
3. **User tags**Free-text annotations applied by users to movies  
   Useful for sentiment analysis, thematic exploration, and metadata enrichment
4. **Genome scores and tags**High-dimensional relevance scores linking movies to standardized tags  
   Supports advanced similarity analysis and machine learning use cases
5. **External links (IMDB, TMDB)**References to external movie databases  
   Enables data enrichment and integration with third-party sources

## 3.2 Key Characteristics:

* **Relational data structure** with well-defined primary and foreign keys
* **Moderate data volume**, suitable for both batch processing and analytical workloads
* **Time-based attributes** enabling trend and behavioral analysis
* **High cardinality relationships**, ideal for fact and dimension modeling
* **Well-suited for Medallion Architecture**, supporting progressive data refinement
* **Optimized for analytical modeling**, including star schemas and KPI aggregation

# 4. Architecture Overview

The architecture is designed around **decoupled, cloud-native principles**, where ingestion, storage, processing, transformation, and analytics are handled by specialized services. This approach improves scalability, flexibility, and maintainability while allowing each layer to evolve independently.

## 4.1 Ingestion Layer

* **Technology:** Python, boto3
* Responsible for extracting raw data from local sources and loading it into cloud storage
* Handles bucket creation, folder structure, and file uploads
* Ensures reliable and repeatable ingestion through programmatic control
* Acts as the entry point of the ELT pipeline

## 4.2 Raw Storage Layer

* **Technology:** Amazon S3
* Serves as the durable landing zone for raw datasets
* Stores data in its original format without transformation
* Enables low-cost, highly available storage
* Decouples data ingestion from downstream processing
* Provides a single source of truth for raw data

## 4.3 Processing Layer

* **Technology:** Google BigQuery
* Used for large-scale data processing and transformations
* Handles raw data ingestion into the **Bronze layer**
* Provides high-performance SQL execution for analytical workloads
* Supports schema inference, partitioning, and scalable compute

## 4.4 Transformation Layer

* **Technology:** dbt
* Implements transformation logic following the **Medallion Architecture**
* Organizes transformations into:
  1. **Bronze:** Raw, minimally processed data
  2. **Silver:** Cleaned, standardized, and validated datasets
  3. **Gold:** Business-ready fact and dimension tables
* Enables version-controlled, testable, and documented transformations
* Encourages modular SQL development and data lineage tracking

## 4.5 Analytics Storage Layer

* **Technology:** S3 Tables (Apache Iceberg)
* Stores transformed Silver and Gold datasets in an open table format
* Supports ACID transactions, schema evolution, and time travel
* Optimized for analytical queries via Amazon Athena
* Eliminates tight coupling between storage and query engine

## 4.6 Query Engine

* **Technology:** Amazon Athena
* Provides serverless SQL querying over data stored in S3 Tables
* Enables fast, ad-hoc analytics without infrastructure management
* Integrates seamlessly with Iceberg tables and AWS Glue Catalog
* Supports BI and reporting workloads

# 5. Methodology

## 5.1 Raw Ingestion (Local → S3)

* Raw MovieLens CSV files are uploaded to Amazon S3
* Bucket is created programmatically if it does not exist
* Files are organized under a logical prefix (movielens/)

**Why S3?**

* Durable, cheap storage
* Ideal landing zone for raw data
* Decouples ingestion from processing

## 5.2 Bronze Layer (S3 → BigQuery)

**Purpose:** Preserve raw data with minimal transformation

**Process:**

1. CSV files are read directly from S3
2. Loaded into BigQuery Bronze dataset
3. Metadata columns added:

* \_ingestion\_timestamp
* \_source\_file
* \_source\_bucket

#### Bronze Characteristics

* One BigQuery table per source file
* Schema auto-detected
* Raw but traceable

## 5.3 Silver Layer (Data Cleaning & Standardization)

**Tool:** dbt  
 **Location:** BigQuery

#### Transformations Applied

* Column renaming and normalization
* Data type casting
* Deduplication using window functions
* Data quality flags:
  + Missing values
  + Invalid ranges
  + Malformed
* Derived columns:
  + Date breakdowns
  + Genre arrays
  + Text

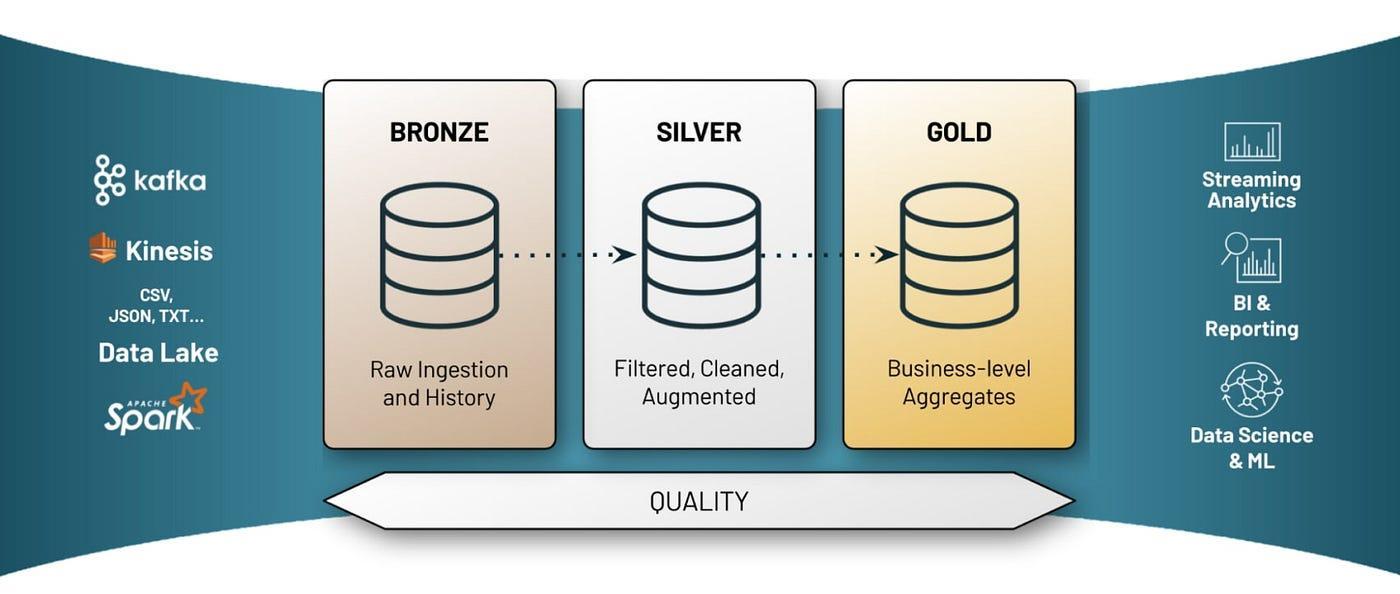
#### Output

* Clean, analytics-ready tables
* Still granular, not aggregated

## 5.4 Gold Layer

The Gold layer represents the final stage of the data pipeline and is designed to deliver **business-ready, analytics-optimized datasets**. At this stage, data is fully transformed, validated, and modeled to support reporting, dashboards, and advanced analytical use cases.

The Gold layer is organized into a structured hierarchy consisting of **fact tables, dimension tables, views, and KPIs**.. Overall, the Gold layer serves as the **single source of truth for analytics**, ensuring consistency, performance, and usability across all downstream consumers.



## 5.5 Lineage Graph

The lineage graph illustrates the **end-to-end data flow** across the Medallion Architecture, clearly showing how raw data evolves into analytics-ready insights. Data originates in the **Bronze layer**, where raw MovieLens source tables (movies, ratings, tags, genome data, and links) are ingested with minimal transformation. These Bronze tables feed the **Silver layer**, where data is cleaned, standardized, deduplicated, and enriched to create reliable, conformed datasets.

From the Silver layer, curated **dimension tables** (such as movies, genres, tags, users, and dates) and **fact tables** (including ratings, genome scores, and user interactions) are built. These core analytical models then power the **Gold layer**, which consists of analyst-friendly **views** and **pre-aggregated KPI tables**. Gold models aggregate metrics such as genre performance, tag intelligence, movie performance, user engagement, and time-series trends, ultimately converging into an executive summary KPI.

Overall, the lineage graph provides **full transparency and traceability**, demonstrating how each KPI and analytical output is derived from trusted upstream sources. This structure supports data governance, impact analysis, and confident decision-making by clearly mapping dependencies from raw ingestion to business-level metrics.



# 6. Cross-Cloud Data

Because BigQuery does not support direct exports to Amazon S3, a two-step export strategy is implemented. First, data is exported from BigQuery to Google Cloud Storage in **Parquet format with SNAPPY compression**, allowing sharded outputs for large tables. In the second step, the Parquet shards are read and merged using **PyArrow**, then uploaded as a **single optimized Parquet file per table** to Amazon S3. The exported data is systematically organized by layer and table name to ensure consistency and ease of downstream consumption.

# 7. S3 Tables & Iceberg

**Why S3 Tables?**

* Managed Iceberg tables on S3
* ACID transactions
* Schema evolution

**Process**

The final analytics layer is implemented using **S3 Tables with Apache Iceberg**. The process begins by creating an S3 Table Bucket and defining a namespace to logically organize tables. External tables are then registered in the AWS Glue Catalog to reference the underlying Parquet data stored in S3. Using **Athena CTAS (Create Table As Select)** statements, Iceberg tables are created and populated from the Silver and Gold datasets. This approach enables transactional, schema-evolvable, and high-performance analytical tables that are fully managed and queryable through Amazon Athena.

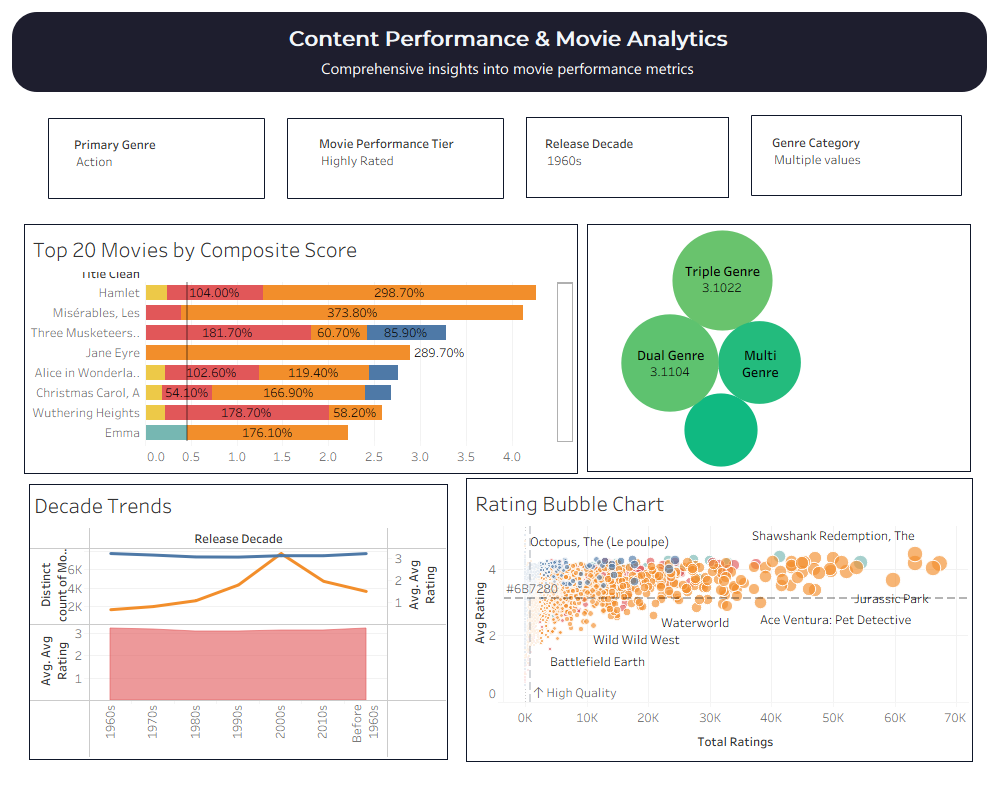
**Result:**

This enables fast, reliable access to analytics-ready datasets using Amazon Athena.

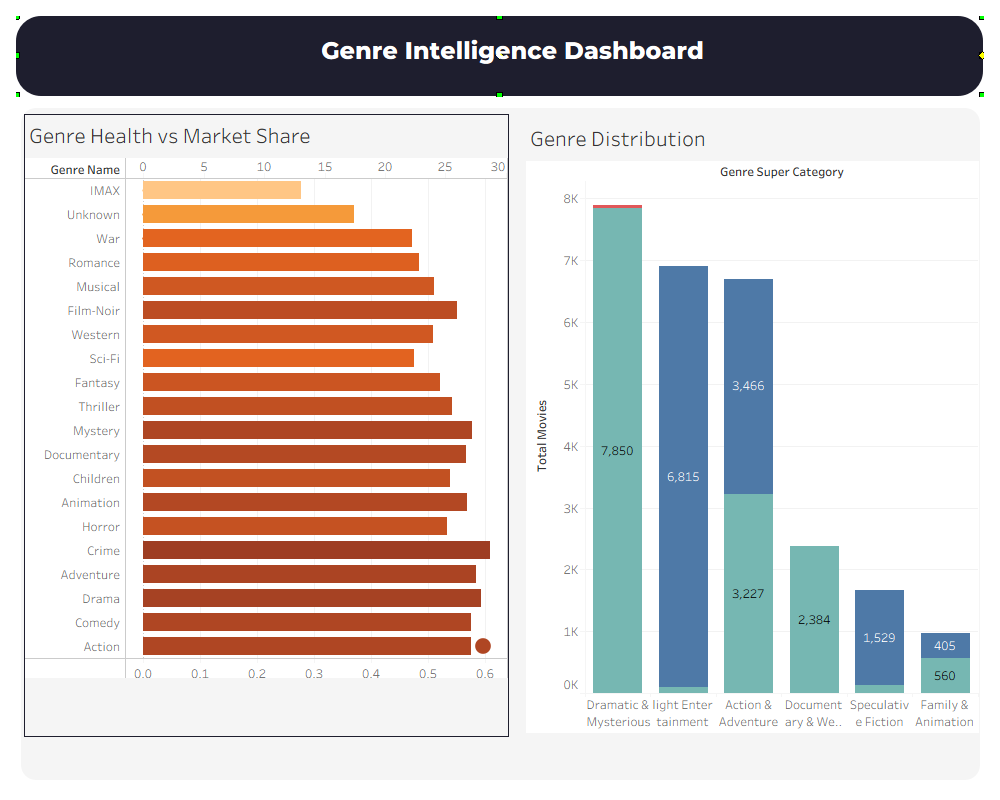
# 8. Dashboard Overview

## 8.1 Content Performance & Movie Analytics Dashboard

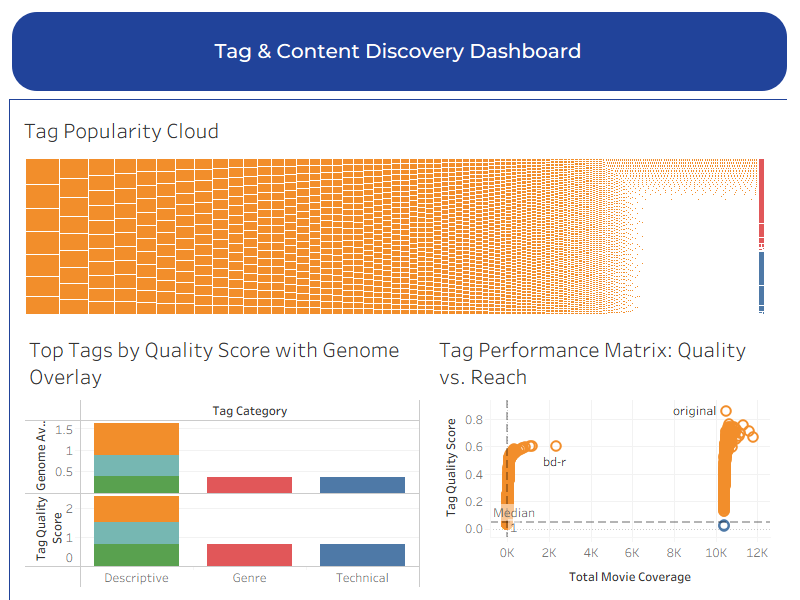
Provides an overview of **movie performance across genres and time**. It highlights top-performing movies using a composite score, compares genre structures (single vs. multi-genre), tracks decade-wise trends, and analyzes the relationship between rating quality and audience reach.



## 8.2 Genre Intelligence Dashboard

Focuses on **genre-level insights**, showing market share, performance health, and overall distribution across major genre categories. It helps identify dominant genres, emerging opportunities, and underrepresented segments.

## 8.3 Tag & Content Discovery Dashboard

Analyzes **content tags and metadata effectiveness**. It shows tag popularity, tag quality scores, and the balance between tag reach and relevance to support better content discovery and recommendation strategies.

## 

## 8.4 User Engagement & Behavior Dashboard

Summarizes **user activity and engagement patterns** across the platform. It includes key metrics, rating activity over time, engagement distribution, user tenure analysis, and monthly active user trends.

# 9. Conclusion

This project successfully demonstrates the design and implementation of a modern, cloud-native ELT data platform using a multi-cloud architecture. By integrating Amazon S3, Google BigQuery, dbt, and Apache Iceberg, the solution achieves a fully decoupled and scalable pipeline that follows the Medallion Architecture. The pipeline ensures reliable data ingestion, structured transformation, and the delivery of analytics-ready datasets through S3 Tables queryable via Amazon Athena. Overall, the project reflects real-world data engineering best practices, emphasizing scalability, interoperability, data governance, and performance optimization for analytical workloads.