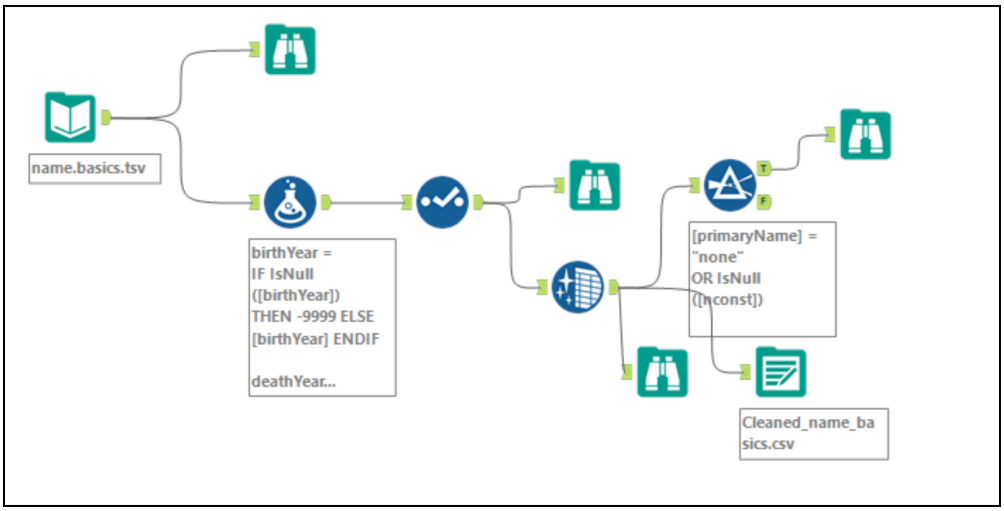
# IMDb Team Project

## Team 7

**ALTERYX**

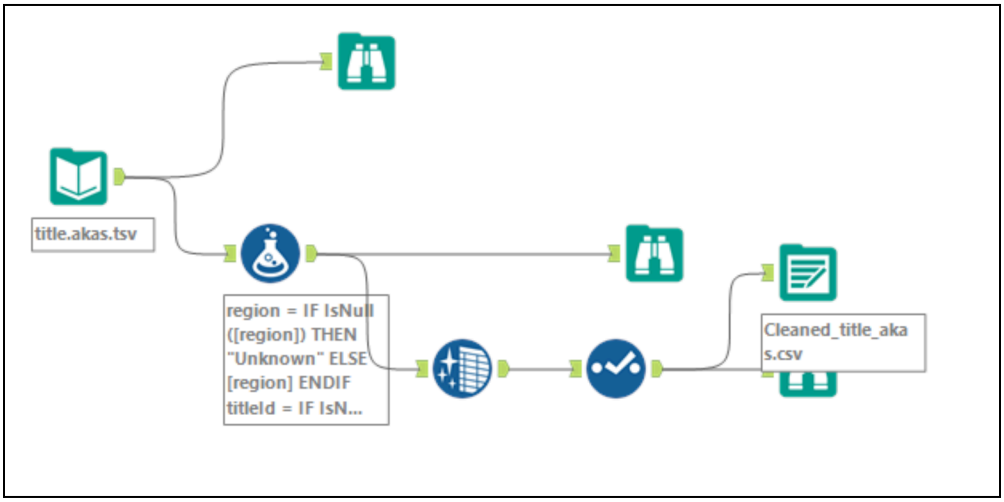
### 1. name.basics

- Replaced NULL birthYear and deathYear → -9999.
- Replaced rows where primaryName = "none" or nconst is NULL to “unknown” (invalid PK).
- Standardized text fields (trim, cleaned blanks).
- Output: Cleaned_name_basics.csv.



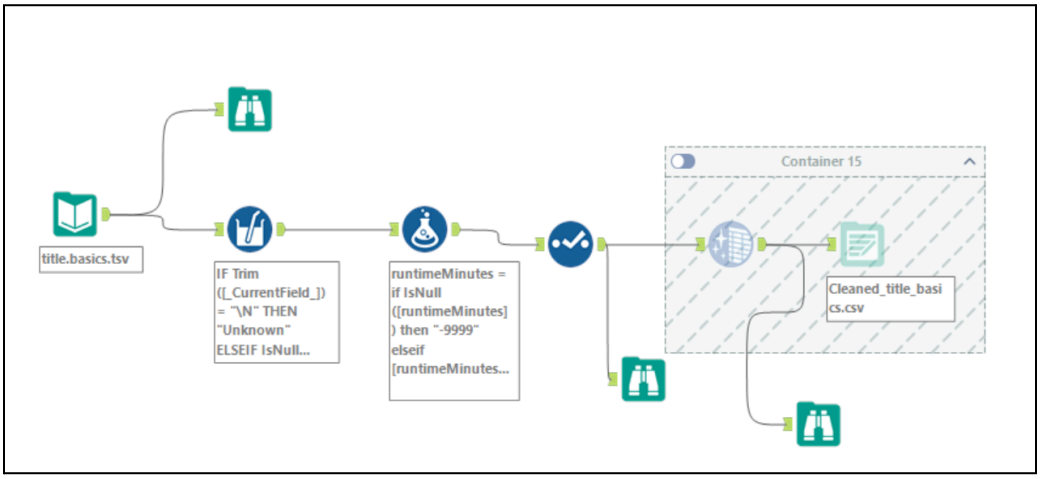

### 2. title.akas

- Replaced NULL region and language → "Unknown".
- Ensured titleId is not null (FK requirement).
- Trimmed all text fields.
- Output: Cleaned_title_akas.csv.



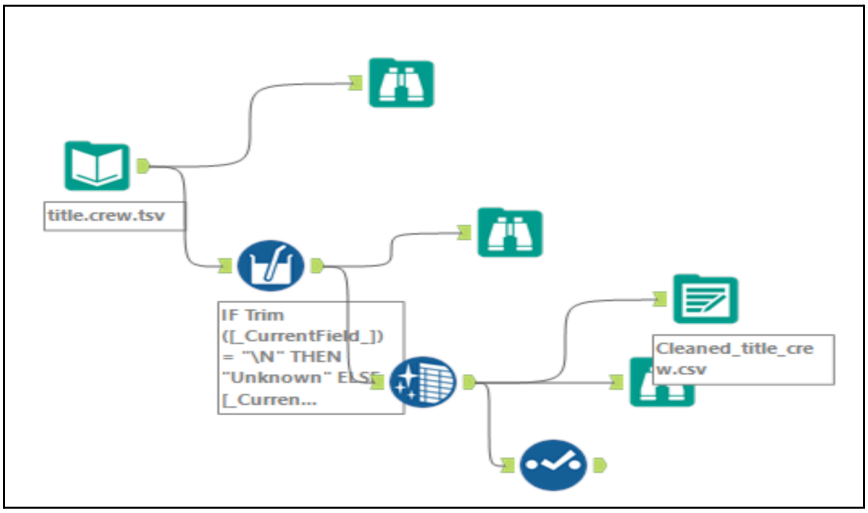
**3. title.basics**

- Replaced all \N values → "Unknown".
- runtimeMinutes NULL → -9999.
- Cleaned whitespace in text columns.
- Output: Cleaned_title_basics.csv.



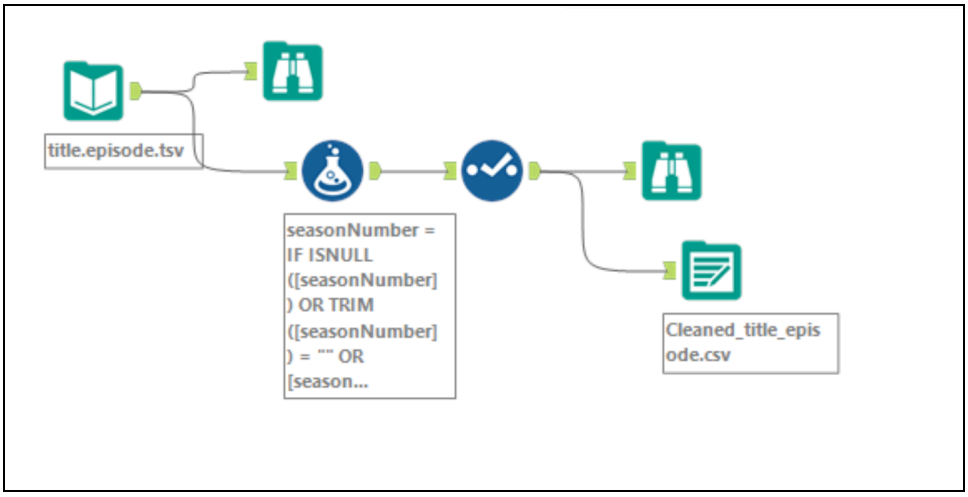

**4. title.crew**

- Replaced \N in directors and writers → "Unknown".
- Trimmed fields for consistency.
- Output: Cleaned_title_crew.csv.



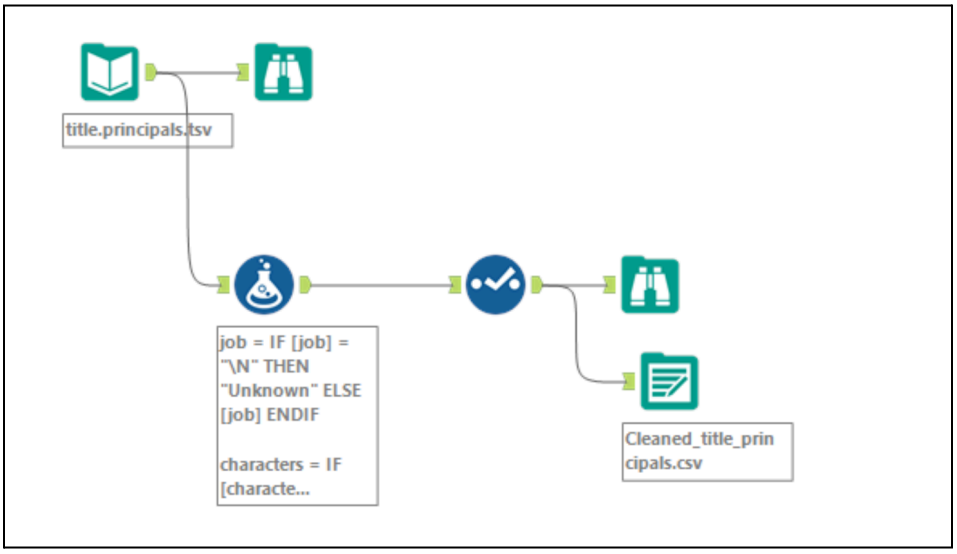

**5. title.episode**

- seasonNumber and episodeNumber:
  If NULL, empty, or invalid → "Unknown".
- Ensured tconst and parentTconst are not null.
- Output: Cleaned_title_episode.csv.



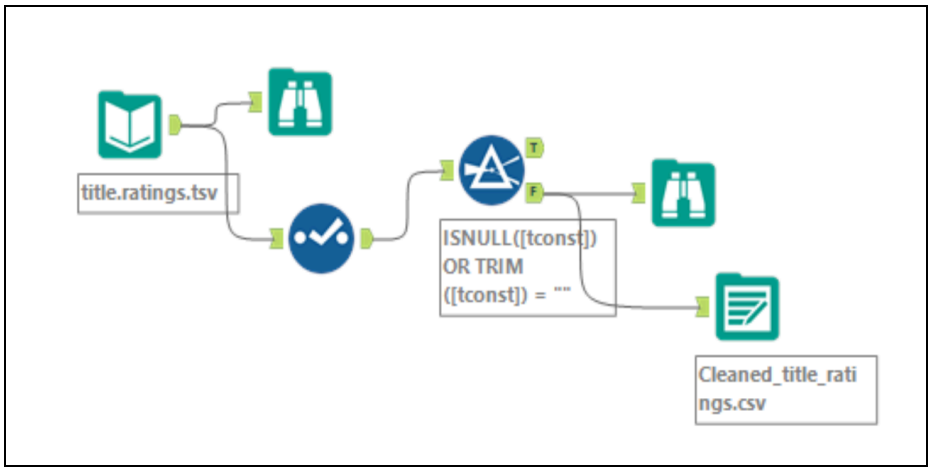

**6. title.principals**

- Replaced \N in job and characters → "Unknown".
- Preserved all rows (sparse data expected).
- Output: Cleaned_title_principals.csv.

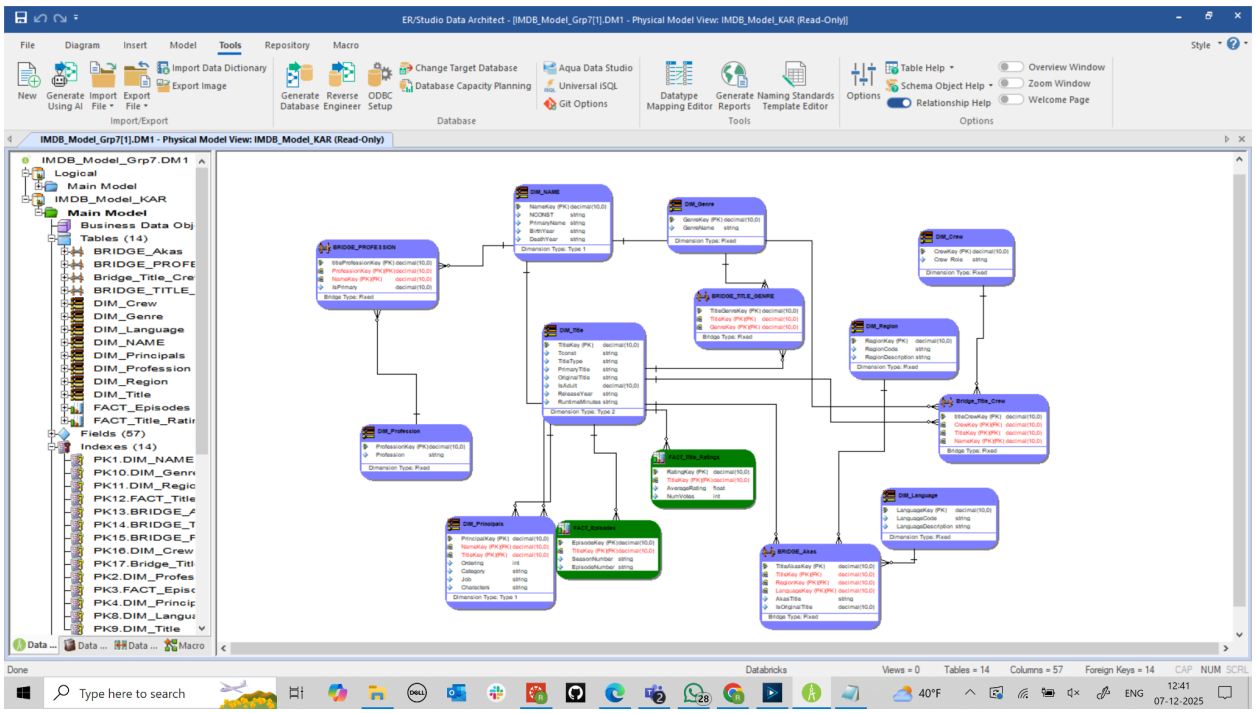



**7. title.ratings**

- Removed rows where tconst is NULL or empty (invalid FK).
- Kept averageRating and numVotes as provided.
- Output: Cleaned_title_ratings.csv.




## ER DIMENSIONAL MODEL




### 1. Overall Design

- Implemented a Kimball-style dimensional warehouse for IMDb data.
- Modeled using ER/Studio, generated DDL for Databricks Delta.
- Architecture includes:
  7 Dimensions
  2 Facts
  4 Bridge Tables for multi-valued relationships.
- Surrogate keys (decimal(10,0)) used for consistency and performance.
- All relationships enforced with foreign keys for referential integrity.

## 2. Dimension Tables

### 1. DIM_Name

- Attributes: NameKey (PK), NCONST, PrimaryName, BirthYear, DeathYear.
- Purpose: Represents unique personnel.
- SCD Type: Type-1 (overwrite changes).
- FK Usage: Linked from Principals, Profession, Crew bridges.

### 2. DIM_Title

- Attributes: TitleKey (PK), Tconst, TitleType, PrimaryTitle, OriginalTitle, IsAdult, ReleaseYear, RuntimeMinutes.
- SCD Type: Type-2 (keeps history).
- Central hub dimension referenced by almost all facts and bridges.

### 3. DIM_Genre

- Attributes: GenreKey (PK), GenreName.
- Fixed list of genres; no SCD.

### 4. DIM_Region

- Attributes: RegionKey (PK), RegionCode, RegionDescription.
- Used primarily by Akas.

### 5. DIM_Language

- Attributes: LanguageKey (PK), LanguageCode, LanguageDescription.
- Used in title localization.

### 6. DIM_Profession

- Attributes: ProfessionKey (PK), Profession.
- Supports multi-profession relationships via BRIDGE_PROFESSION.

### 7. DIM_Crew

- Attributes: CrewKey (PK), Crew_Role.
- Supports Director/Writer relationships via Bridge_Title_Crew.

## 3. Fact Tables

### 1. FACT_Title_Ratings

- Attributes: RatingKey (PK), TitleKey (FK), AverageRating, NumVotes.
- Grain: One record per title rating.
- Represents measurable KPIs (ratings).

2. **FACT_Episodes**

- Attributes: EpisodeKey (PK), TitleKey (FK), SeasonNumber, EpisodeNumber.
- Grain: One record per episode per title.
- Captures series structure.

**4. Bridge Tables (Multi-Valued Dimensions)**

1. **BRIDGE_PROFESSION**

- Links Name ↔ Profession.
- Allows multiple professions per person.
- Columns: titleProfessionKey, ProfessionKey, NameKey, IsPrimary.
- Bridge Type: Fixed.

2. **BRIDGE_TITLE_GENRE**

- Links Title ↔ Genre.
- Supports multi-genre classification.

3. **BRIDGE_Akas**

- Links Title ↔ Region ↔ Language for alternate titles.
- Columns: TitleAkasKey, TitleKey, RegionKey, LanguageKey, AkasTitle, IsOriginalTitle.

4. **Bridge_Title_Crew**

- Links Title ↔ Crew ↔ Name.
- Supports directors and writers for each movie/episode.

**5. Relationship Design**

- DIM_Title is the central dimension for almost all facts and bridges.
- BRIDGE_PROFESSION connects multiple professions to each Name.
- BRIDGE_TITLE_GENRE supports multi-genre titles.
- Bridge_Title_Crew models many-to-many relationships between titles and crew members.
- BRIDGE_Akas connects title localization details (region + language).
- FACT tables reference the surrogate TitleKey for reliable joins.

**6. Why This Schema Works**

Supports your stated queries:

- Find professions per person → BRIDGE_PROFESSION + DIM_Profession
- Count personnel with multiple professions → Bridge table structure
- Get genres per title → BRIDGE_TITLE_GENRE
- Movie length metrics → DIM_Title.RuntimeMinutes
- Adult vs non-adult → DIM_Title.IsAdult
- Languages per title → BRIDGE_Akas + DIM_Language

- Regions of release → BRIDGE_Akas + DIM_Region
- Directors and writers → Bridge_Title_Crew + DIM_Crew + DIM_Name
- Episode counts → FACT_Episodes

This confirms the ERD meets every business requirement.

**DATABRICKS PIPELINE (BRONZE-SILVER-GOLD)**

Databricks workspace shows 28 materialized views/tables representing the full Medallion architecture:

- 7 Bronze (raw cleaned inputs from Alteryx)
- 7 Silver (validated, standardized integration layer)
- 14 Gold (dimensions, facts, bridges based on ERD)

**1. Bronze Layer – Raw Ingest Layer**

Purpose:

- Store cleaned Alteryx outputs exactly as received.
- Maintain source traceability.
- No business transformations.

Bronze Tables:

- bronze_name_basics
- bronze_title_basics
- bronze_title_akas
- bronze_title_crew
- bronze_title_episode
- bronze_title_principals
- bronze_title_ratings

Actions Performed:

- Loaded cleaned CSVs into Delta tables.
- Schema preserved as strings (IMDb raw format).
- No joins, no filtering, no datatype changes.

**2. Silver Layer – Transformed, Standardized, Strongly-Typed Data**

Purpose:

- Convert Bronze data into validated, typed, conformed datasets.
- Apply datatype conversions, explode arrays, standardize nulls, prepare FK mappings.
- Make data Gold-ready.

**2.1 Dataset-Wise Silver Transformations (with datatype fixes)**

**SILVER – name.basics**
Source Characteristics

- birthYear, deathYear as strings
- primaryProfession as array
- knownForTitles as array

Transformations (Silver)
Converted:

- birthYear → INT (NULL → -9999)
- deathYear → INT (NULL → -9999)
- Exploded primaryProfession → 1 row/profession.
- Ensured nconst NOT NULL for FK mapping.

Output
silver_name_basics → feeds DIM_NAME, BRIDGE_PROFESSION

**SILVER – title.basics**
Source Characteristics

- startYear, endYear, runtimeMinutes are strings
- isAdult boolean (0/1 as string)
- genres array

Transformations
Converted:

- startYear → INT
- endYear → INT
- runtimeMinutes → INT (NULL → -9999)
- isAdult → INT
- Exploded genres array → 1 row per genre.
- Cleaned unknown/missing titles → “Unknown”

Output
silver_title_basics → feeds DIM_Title, BRIDGE_TITLE_GENRE

**SILVER – title.akas**
Source Characteristics

- ordering integer
- region, language strings
- isOriginalTitle boolean

Transformations
Converted:

- ordering → INT
- isOriginalTitle → INT

Standardized:

- NULL region → “Unknown”
- NULL language → “Unknown”
- Prepared region/language codes for dimension mapping.

Output
silver_title_akas → feeds BRIDGE_AKAS

**SILVER – title.crew**
Source Characteristics

- directors, writers as arrays of nconst

Transformations

- Exploded directors → 1 row per director.
- Exploded writers → 1 row per writer.
- Added Crew_Role = Director / Writer.
- Removed null/empty IDs.

Output
silver_title_crew → feeds BRIDGE_TITLE_CREW

**SILVER – title.episode**
Source Characteristics

- seasonNumber, episodeNumber may be empty strings

Transformations
Converted:

- seasonNumber → INT (NULL/empty → -1)
- episodeNumber → INT (NULL/empty → -1)
- Ensured tconst NOT NULL.

Output
silver_title_episode → feeds FACT_Episodes

**SILVER – title.principals**

Source Characteristics

- ordering integer
- job, characters may be '\N'

Transformations
Converted:

- ordering → INT

Standardized:

- job = "Unknown" if NULL
- characters = "Unknown" if NULL
- Ensured tconst and nconst NOT NULL.

Output
silver_title_principals → feeds DIM_Principals

**SILVER – title.ratings**
Transformations
Converted:

- averageRating → FLOAT
- numVotes → INT
- Removed invalid rows where tconst is NULL or empty.

Output
silver_title_ratings → feeds FACT_Title_Ratings

**3. Gold Layer – Dimensional Model Implementation (ERD)**
Purpose

- Build final Dimensions, Facts, and Bridge Tables using transformed Silver data.
- Apply SCD rules, surrogate keys, and enforce referential integrity.

**Gold Dimensions**

- gold_dim_name
- gold_dim_title
- gold_dim_genre
- gold_dim_region
- gold_dim_language
- gold_dim_profession
- gold_dim_principals
- gold_dim_crew

**Gold Facts**

- gold_fact_episodes
- gold_fact_title_ratings

**Gold Bridges**

- gold_bridge_profession
- gold_bridge_title_genre
- gold_bridge_title_crew
- gold_bridge_akas

**3.1. Gold Layer – Actions Performed (UPDATED with SCD Types)**

Surrogate Keys Assigned

Every dimension uses decimal(10,0) surrogate keys:

- NameKey, TitleKey, GenreKey, RegionKey, LanguageKey, ProfessionKey, PrincipalKey, CrewKey.

**Fact Tables Built**

Correct grain established:

- Title-level ratings
- Episode-level structure
- All facts reference surrogate TitleKey for consistency.

**Bridge Tables Built for M:N Relationships**

- BRIDGE_PROFESSION → person can have multiple professions
- BRIDGE_TITLE_GENRE → title can have multiple genres
- BRIDGE_TITLE_CREW → multiple directors/writers per title
- BRIDGE_AKAS → multiple regional/language title variations

Exploded values from Silver used to populate these.

**SCD TYPES (Updated as requested)**

Type-1 Dimensions (Overwrite old data)

**1. DIM_Name**

(BirthYear, DeathYear, PrimaryName can be overwritten; history not required)

**2. DIM_Principals**

(Job, Characters, Category change frequently and do not require historical tracking)

**Type-2 Dimensions (Maintain history)**

**1. DIM_Title**

(ReleaseYear, TitleType, IsAdult, RuntimeMinutes may evolve → history needed)

**Fixed Dimensions (Reference values; no SCD)**

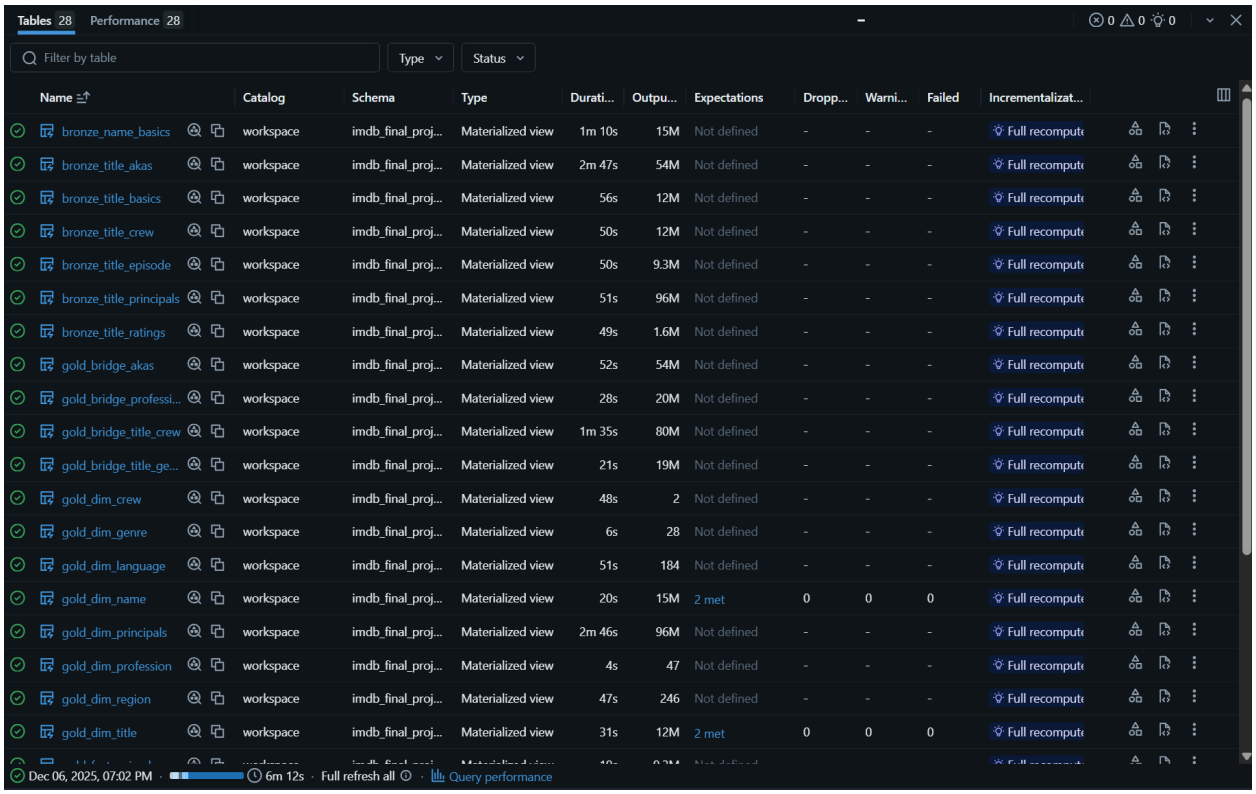
1. DIM_Genre
2. DIM_Region
3. DIM_Language
4. DIM_Profession
5. DIM_Crew

**Referential Integrity + Consistency Checks**

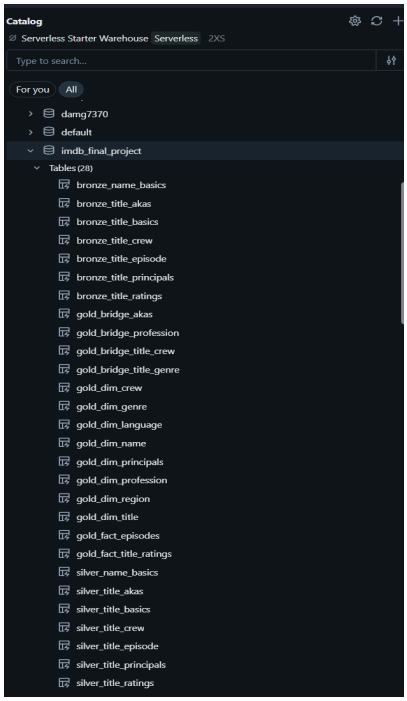
- All Gold tables joined using surrogate keys.
- Null keys handled in Silver → ensures no broken FK relationships.
- Fully enforced model matches ERD + DDL.

**SCREENSHOTS**

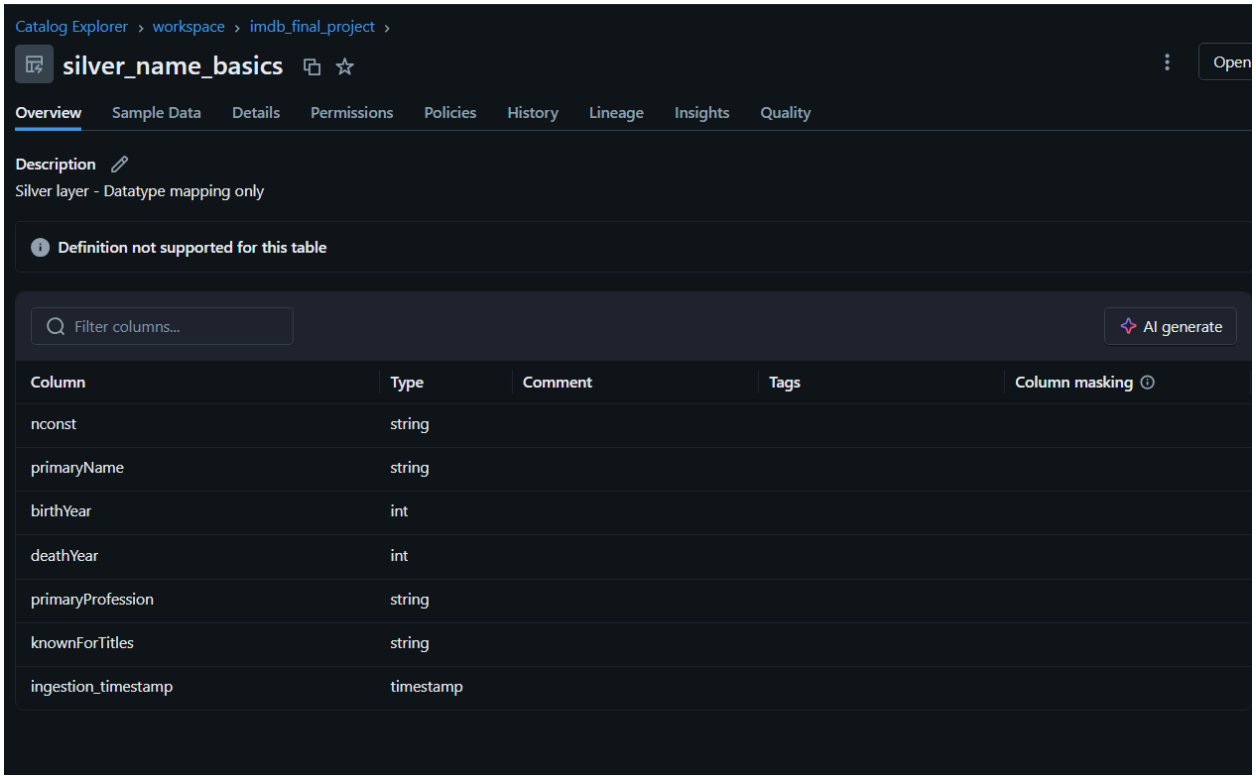
**Pipeline**

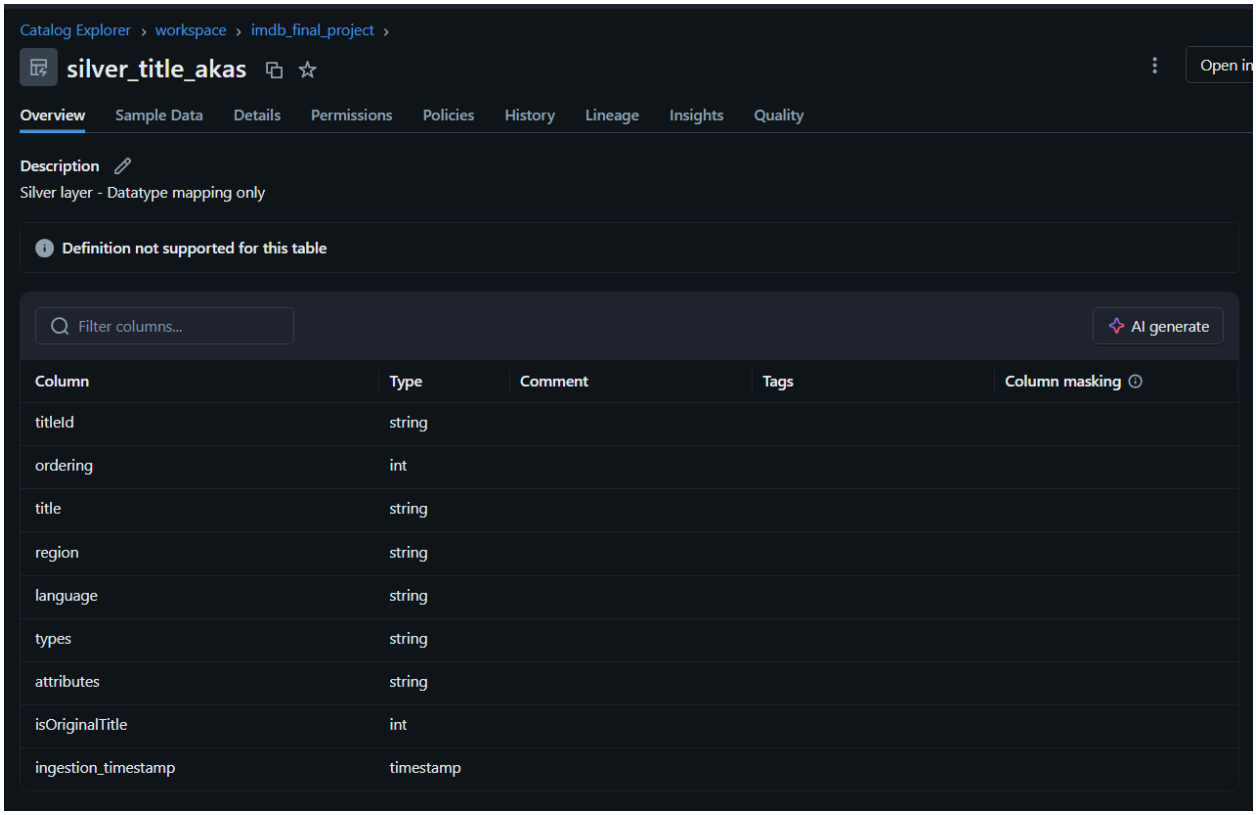


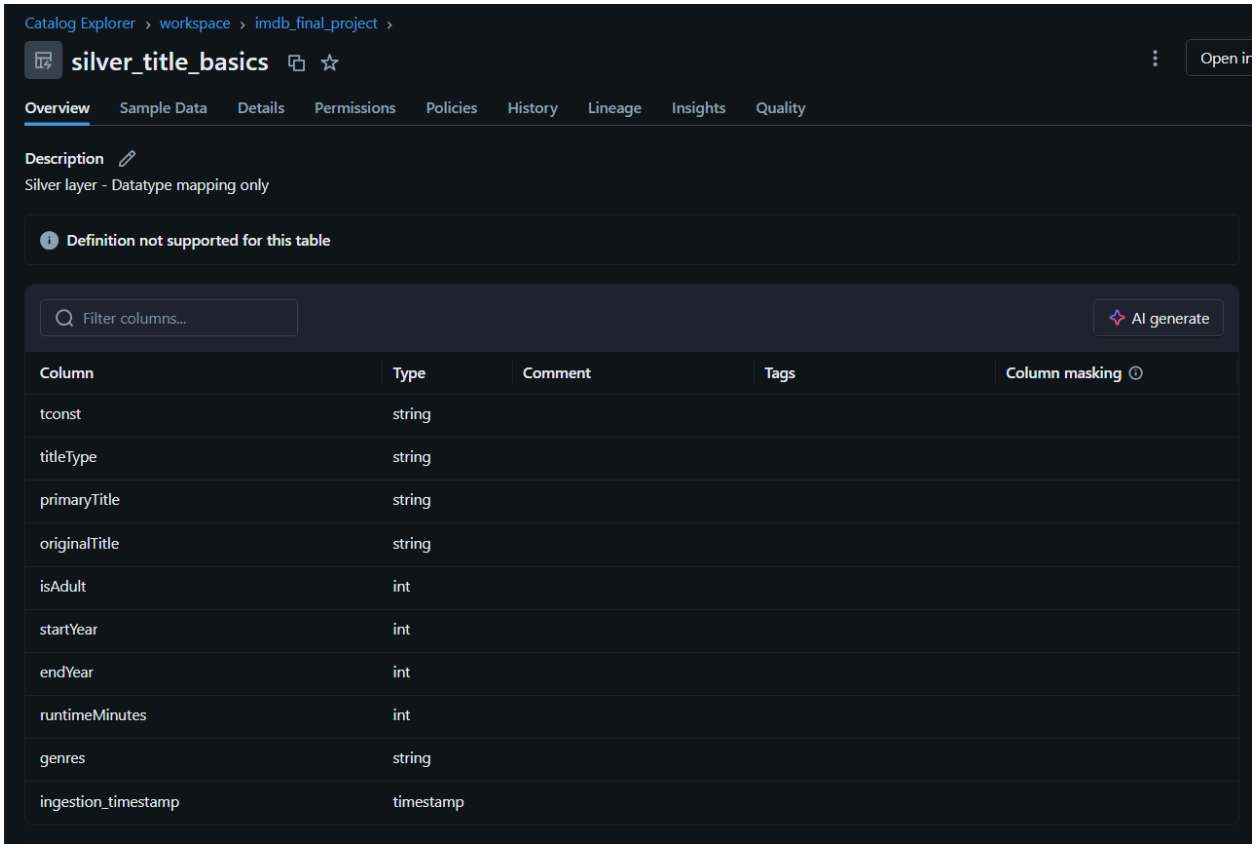

# Catalog Tables

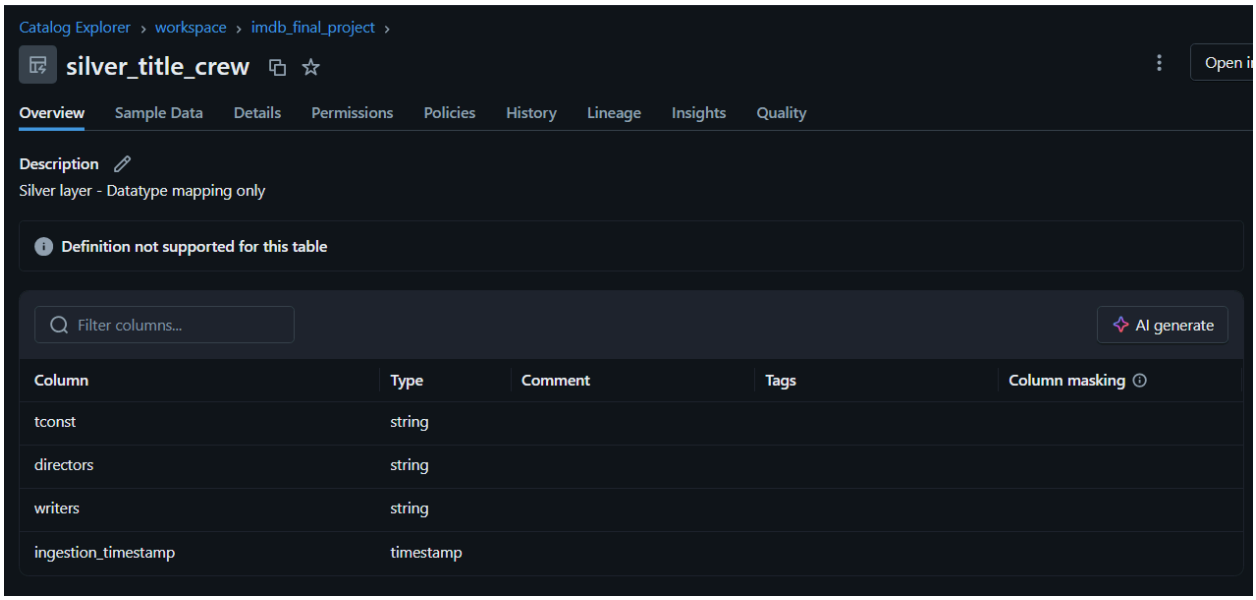


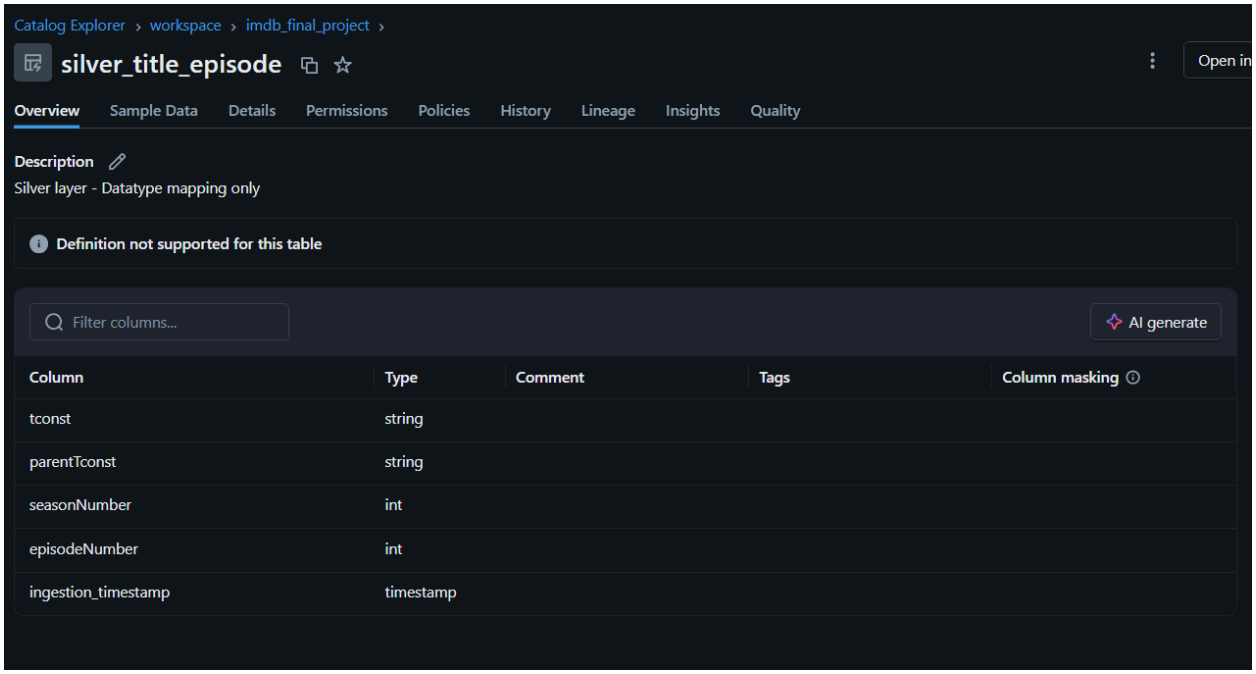
## Silver Table Updated Datatypes

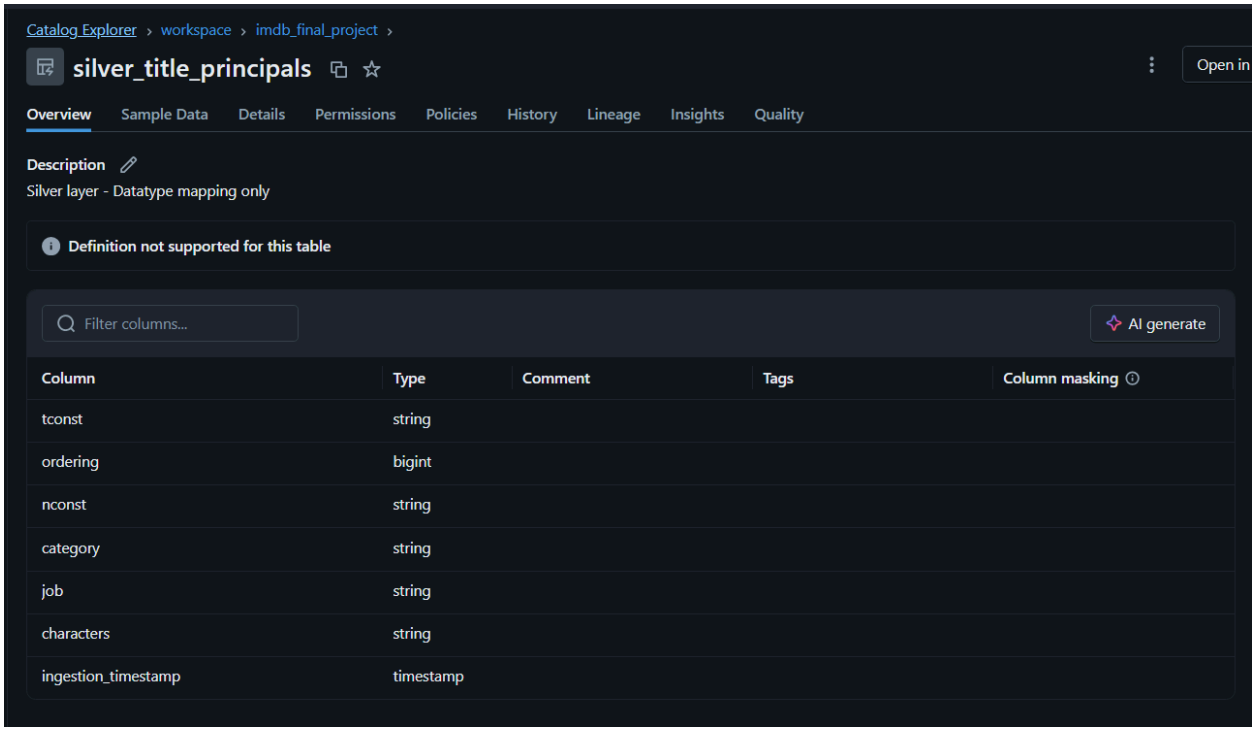


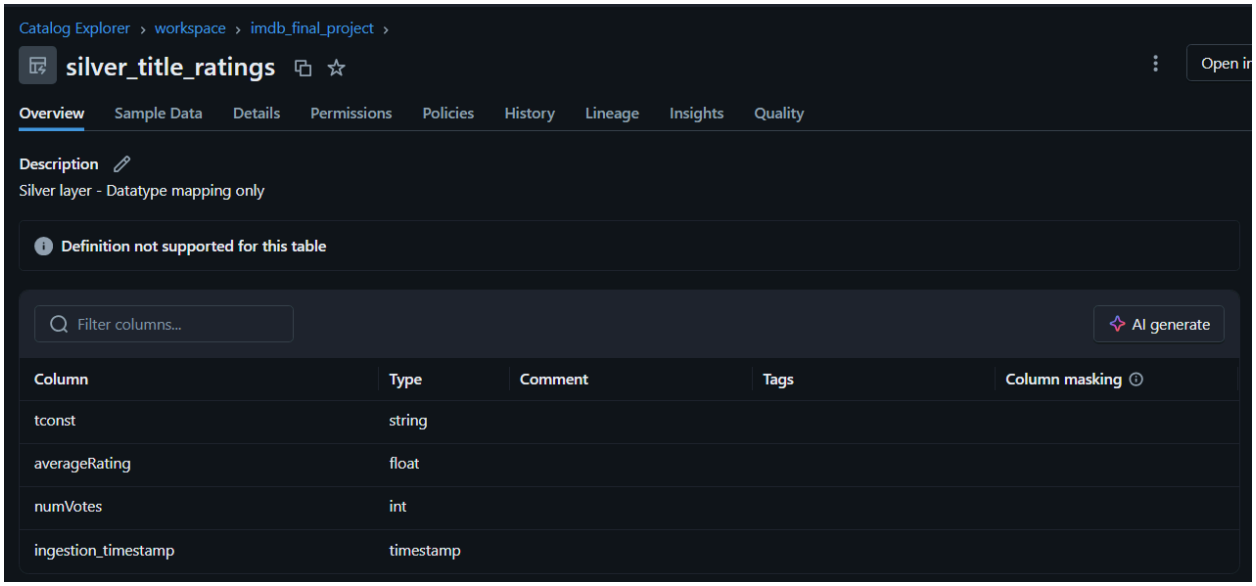




Catalog Explorer › workspace › imdb_final_project ›
silver_title_basics
Open in
Overview
Sample Data
Details
Permissions
Policies
History
Lineage
Insights
Quality
Description
Silver layer - Datatype mapping only
Definition not supported for this table
Filter columns...
AI generate
Column
Type
Comment
Tags
Column masking
tconst
string
titleType
string
primaryTitle
string
originalTitle
string
isAdult
int
startYear
int
endYear
int
runtimeMinutes
int
genres
string
ingestion_timestamp
timestamp


Catalog Explorer › workspace › imdb_final_project ›
silver_title_crew
Open in
Overview
Sample Data
Details
Permissions
Policies
History
Lineage
Insights
Quality
Description
Silver layer - Datatype mapping only
Definition not supported for this table
Filter columns...
AI generate
Column
Type
Comment
Tags
Column masking
tconst
string
directors
string
writers
string
ingestion_timestamp
timestamp


Catalog Explorer › workspace › imdb_final_project ›
silver_title_episode
Open in
Overview Sample Data Details Permissions Policies History Lineage Insights Quality
Description
Silver layer - Datatype mapping only
Definition not supported for this table
Filter columns...
AI generate
Column Type Comment Tags Column masking
tconst string
parentTconst string
seasonNumber int
episodeNumber int
ingestion_timestamp timestamp


Catalog Explorer › workspace › imdb_final_project ›
silver_title_principals
Open in
Overview Sample Data Details Permissions Policies History Lineage Insights Quality
Description
Silver layer - Datatype mapping only
Definition not supported for this table
Filter columns...
AI generate
Column Type Comment Tags Column masking
tconst string
ordering bigint
nconst string
category string
job string
characters string
ingestion_timestamp timestamp




**Bronze Table Row Count**


**Silver Table Row Count**

Removed 5 rows from bronze_title_principals table due to nulls

Table +

| | table_name | row_count |
|---|---|---|
| 1 | silver_name_basics | 14909809 |
| 2 | silver_title_akas | 54180283 |
| 3 | silver_title_basics | 12101466 |
| 4 | silver_title_crew | 12105185 |
| 5 | silver_title_episode | 9329138 |
| 6 | silver_title_principals | 96206760 |
| 7 | silver_title_ratings | 1606282 |

7 rows | 7.93s runtime

## Gold Bridge Tables

20 hours ago (4s) 14

```
%sql
select * from imdb_final_project.gold_bridge_akas
```

See performance (1)

Table +

| | TitleAkasKey | TitleKey | RegionKey | LanguageKey | AkasTitle | IsOriginalTitle |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | -9999 | -9999 | carmencita | 1 |
| 2 | 2 | 1 | 55 | -9999 | carmencita | 0 |
| 3 | 3 | 1 | 229 | -9999 | carmencita | 0 |
| 4 | 4 | 1 | 99 | -9999 | carmencita - spanyol tánc | 0 |
| 5 | 5 | 1 | 88 | -9999 | καρμενσίτα | 0 |
| 6 | 6 | 1 | 189 | -9999 | карменсита | 0 |
| 7 | 7 | 1 | 226 | -9999 | карменсіта | 0 |
| 8 | 8 | 1 | 113 | 75 | カルメンチータ | 0 |
| 9 | 9 | 2 | -9999 | -9999 | le clown et ses chiens | 1 |
| 10 | 10 | 2 | 99 | -9999 | a bohóc és kutyái | 0 |
| 11 | 11 | 2 | 187 | -9999 | clovnul si cainii sai | 0 |
| 12 | 12 | 2 | 55 | -9999 | der clown und seine hunde | 0 |
| 13 | 13 | 2 | 73 | -9999 | le clown et ses chiens | 0 |
| 14 | 14 | 2 | 229 | -9999 | the clown and his dogs | 0 |
| 15 | 15 | 2 | 189 | -9999 | клоун и его собаки | 0 |

10,000+ rows | Truncated data | 4.38s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

20 hours ago (3s) 15

```
%sql
select * from imdb_final_project.gold_bridge_profession
```

See performance (1)

Table +

| | titleProfessionKey | ProfessionKey | NameKey | IsPrimary |
|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 1 |
| 2 | 2 | 28 | 1 | 0 |
| 3 | 3 | 32 | 1 | 0 |
| 4 | 4 | 3 | 2 | 1 |
| 5 | 5 | 28 | 2 | 0 |
| 6 | 6 | 40 | 2 | 0 |
| 7 | 7 | 3 | 3 | 1 |
| 8 | 8 | 30 | 3 | 0 |
| 9 | 9 | 32 | 3 | 0 |
| 10 | 10 | 2 | 4 | 1 |
| 11 | 11 | 30 | 4 | 0 |
| 12 | 12 | 47 | 4 | 0 |
| 13 | 13 | 2 | 5 | 0 |
| 14 | 14 | 19 | 5 | 0 |
| 15 | 15 | 47 | 5 | 1 |

10,000+ rows | Truncated data | 3.49s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

20 hours ago (3s)

```
%sql
select * from imdb_final_project.gold_bridge_title_crew
```

See performance (1)

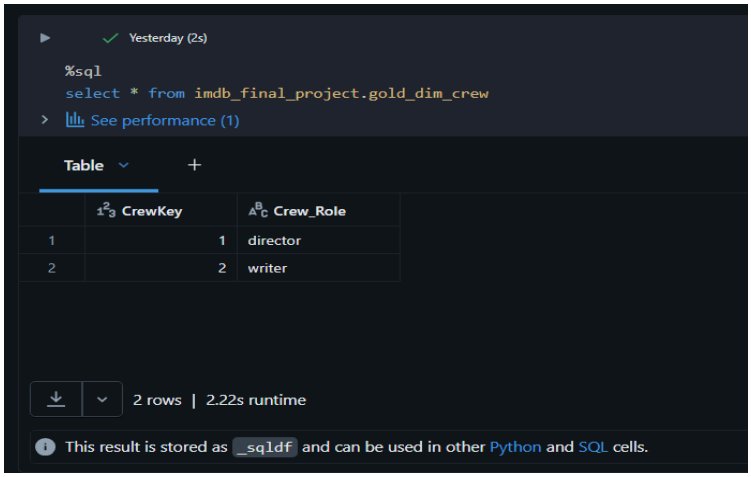
Table +

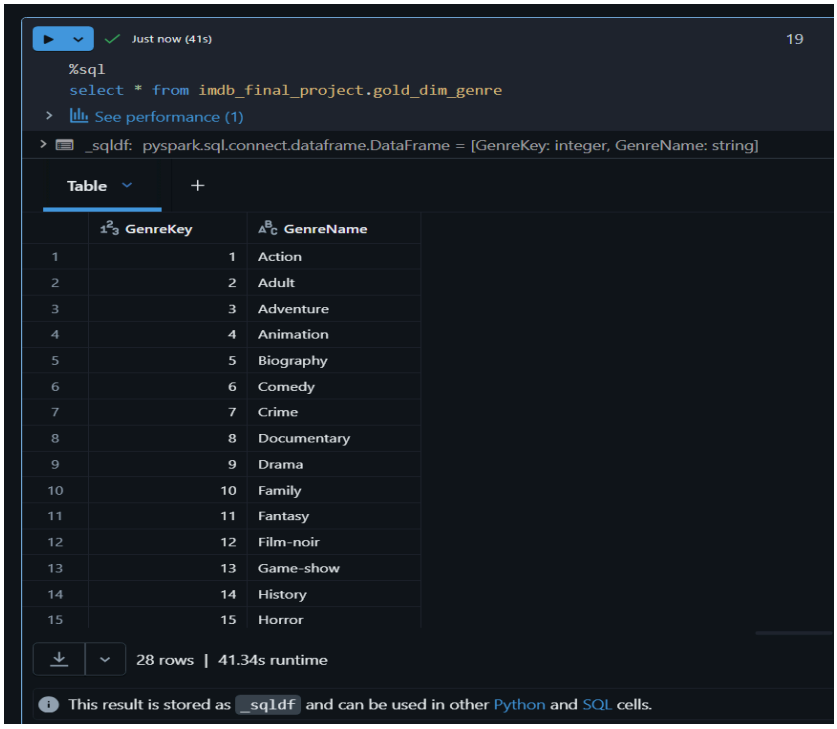
| | titleCrewKey | CrewKey | TitleKey | NameKey |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 5665 |
| 2 | 9 | 2 | 2 | 14909806 |
| 3 | 55 | 2 | 11 | 14909805 |
| 4 | 60 | 1 | 12 | 496116 |
| 5 | 67 | 1 | 13 | 496118 |
| 6 | 84 | 2 | 16 | 14909807 |
| 7 | 102 | 2 | 19 | 14909806 |
| 8 | 114 | 2 | 21 | 14909806 |
| 9 | 123 | 2 | 22 | 14909809 |
| 10 | 125 | 2 | 23 | 14909805 |
| 11 | 126 | 2 | 23 | 14909806 |
| 12 | 127 | 2 | 23 | 14909807 |
| 13 | 135 | 2 | 24 | 14909809 |
| 14 | 138 | 2 | 25 | 14909806 |
| 15 | 150 | 2 | 27 | 14909806 |

10,000+ rows | Truncated data | 3.29s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.


## Gold Dim Tables





Just now (41s) 19

```
%sql
select * from imdb_final_project.gold_dim_genre
```

See performance (1)

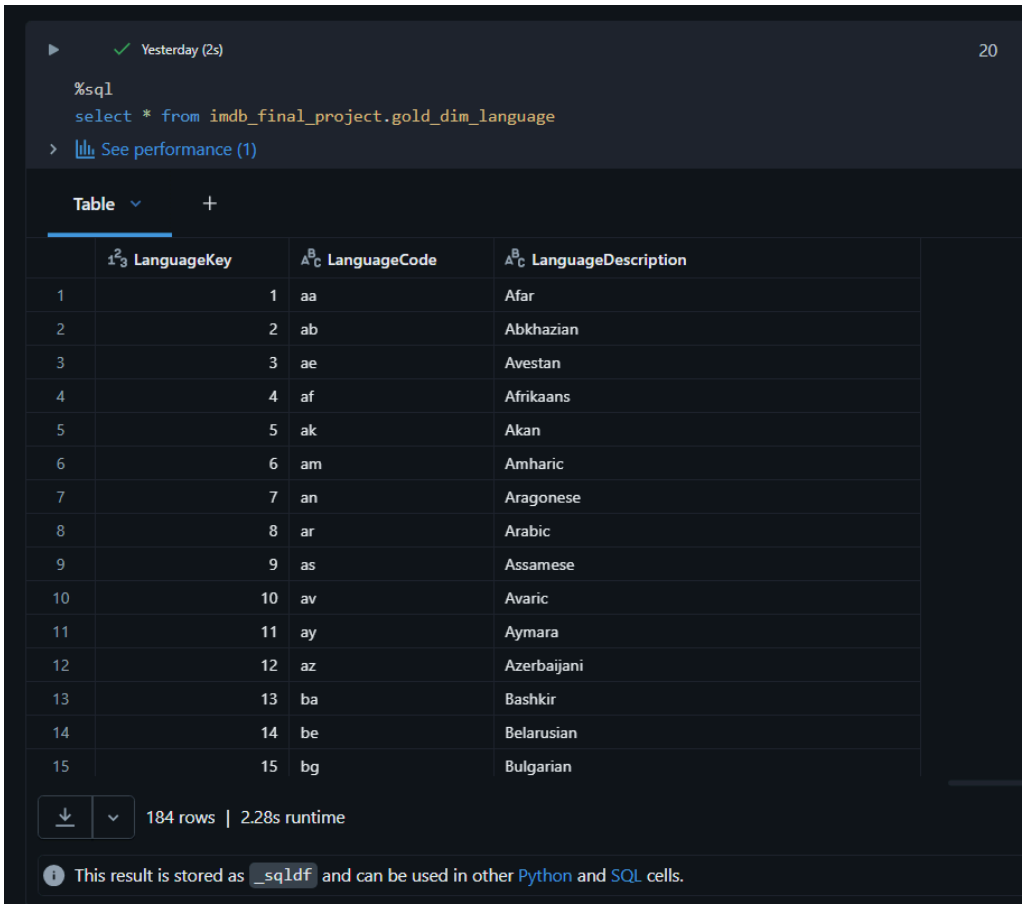
_sqldf: pyspark.sql.connect.dataframe.DataFrame = [GenreKey: integer, GenreName: string]

Table +

| | GenreKey | GenreName |
|---|---|---|
| 1 | 1 | Action |
| 2 | 2 | Adult |
| 3 | 3 | Adventure |
| 4 | 4 | Animation |
| 5 | 5 | Biography |
| 6 | 6 | Comedy |
| 7 | 7 | Crime |
| 8 | 8 | Documentary |
| 9 | 9 | Drama |
| 10 | 10 | Family |
| 11 | 11 | Fantasy |
| 12 | 12 | Film-noir |
| 13 | 13 | Game-show |
| 14 | 14 | History |
| 15 | 15 | Horror |

28 rows | 41.34s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.



Yesterday (2s) 20

```
%sql
select * from imdb_final_project.gold_dim_language
```

See performance (1)

Table +

| | LanguageKey | LanguageCode | LanguageDescription |
|---|---|---|---|
| 1 | 1 | aa | Afar |
| 2 | 2 | ab | Abkhazian |
| 3 | 3 | ae | Avestan |
| 4 | 4 | af | Afrikaans |
| 5 | 5 | ak | Akan |
| 6 | 6 | am | Amharic |
| 7 | 7 | an | Aragonese |
| 8 | 8 | ar | Arabic |
| 9 | 9 | as | Assamese |
| 10 | 10 | av | Avaric |
| 11 | 11 | ay | Aymara |
| 12 | 12 | az | Azerbaijani |
| 13 | 13 | ba | Bashkir |
| 14 | 14 | be | Belarusian |
| 15 | 15 | bg | Bulgarian |

184 rows | 2.28s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Yesterday (2s) 21

```
%sql
select * from imdb_final_project.gold_dim_region
```

See performance (1)

Table +

| | RegionKey | RegionCode | RegionDescription |
|---|---|---|---|
| 1 | 1 | ad | Andorra |
| 2 | 2 | ae | United Arab Emirates |
| 3 | 3 | af | Afghanistan |
| 4 | 4 | ag | Antigua and Barbuda |
| 5 | 5 | ai | Anguilla |
| 6 | 6 | al | Albania |
| 7 | 7 | am | Armenia |
| 8 | 8 | an | Netherlands Antilles |
| 9 | 9 | ao | Angola |
| 10 | 10 | aq | Antarctica |
| 11 | 11 | ar | Argentina |
| 12 | 12 | as | American Samoa |
| 13 | 13 | at | Austria |
| 14 | 14 | au | Australia |
| 15 | 15 | aw | Aruba |

246 rows | 2.24s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

20 hours ago (39s) 22

```
%sql
select * from imdb_final_project.gold_dim_name
```

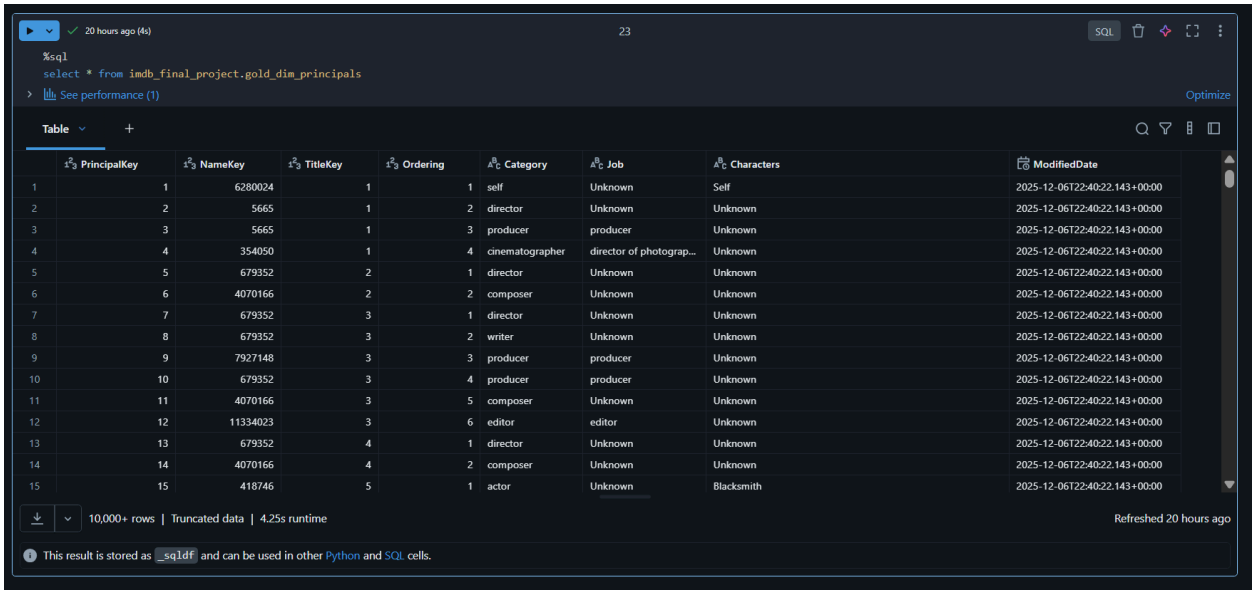
See performance (1)

Table +

| | NameKey | NCONST | PrimaryName | BirthYear | DeathYear | ModifiedDate |
|---|---|---|---|---|---|---|
| 1 | 1 | nm0000001 | fred astaire | 1899 | 1987 | 2025-12-06T22:39:54.403+00:00 |
| 2 | 2 | nm0000002 | lauren bacall | 1924 | 2014 | 2025-12-06T22:39:54.403+00:00 |
| 3 | 3 | nm0000003 | brigitte bardot | 1934 | -9999 | 2025-12-06T22:39:54.403+00:00 |
| 4 | 4 | nm0000004 | john belushi | 1949 | 1982 | 2025-12-06T22:39:54.403+00:00 |
| 5 | 5 | nm0000005 | ingmar bergman | 1918 | 2007 | 2025-12-06T22:39:54.403+00:00 |
| 6 | 6 | nm0000006 | ingrid bergman | 1915 | 1982 | 2025-12-06T22:39:54.403+00:00 |
| 7 | 7 | nm0000007 | humphrey bogart | 1899 | 1957 | 2025-12-06T22:39:54.403+00:00 |
| 8 | 8 | nm0000008 | marlon brando | 1924 | 2004 | 2025-12-06T22:39:54.403+00:00 |
| 9 | 9 | nm0000009 | richard burton | 1925 | 1984 | 2025-12-06T22:39:54.403+00:00 |
| 10 | 10 | nm0000010 | james cagney | 1899 | 1986 | 2025-12-06T22:39:54.403+00:00 |
| 11 | 11 | nm0000011 | gary cooper | 1901 | 1961 | 2025-12-06T22:39:54.403+00:00 |
| 12 | 12 | nm0000012 | bette davis | 1908 | 1989 | 2025-12-06T22:39:54.403+00:00 |
| 13 | 13 | nm0000013 | doris day | 1922 | 2019 | 2025-12-06T22:39:54.403+00:00 |
| 14 | 14 | nm0000014 | olivia de havilland | 1916 | 2020 | 2025-12-06T22:39:54.403+00:00 |
| 15 | 15 | nm0000015 | james dean | 1931 | 1955 | 2025-12-06T22:39:54.403+00:00 |

10,000+ rows | Truncated data | 38.80s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.



20 hours ago (4s)

23

SQL

```
%sql
select * from imdb_final_project.gold_dim_principals
```

See performance (1)

Optimize

Table

| | PrincipalKey | NameKey | TitleKey | Ordering | Category | Job | Characters | ModifiedDate |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 6280024 | 1 | 1 | self | Unknown | Self | 2025-12-06T22:40:22.143+00:00 |
| 2 | 2 | 5665 | 1 | 2 | director | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 3 | 3 | 5665 | 1 | 3 | producer | producer | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 4 | 4 | 354050 | 1 | 4 | cinematographer | director of photograp... | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 5 | 5 | 679352 | 2 | 1 | director | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 6 | 6 | 4070166 | 2 | 2 | composer | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 7 | 7 | 679352 | 3 | 1 | director | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 8 | 8 | 679352 | 3 | 2 | writer | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 9 | 9 | 7927148 | 3 | 3 | producer | producer | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 10 | 10 | 679352 | 3 | 4 | producer | producer | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 11 | 11 | 4070166 | 3 | 5 | composer | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 12 | 12 | 11334023 | 3 | 6 | editor | editor | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 13 | 13 | 679352 | 4 | 1 | director | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 14 | 14 | 4070166 | 4 | 2 | composer | Unknown | Unknown | 2025-12-06T22:40:22.143+00:00 |
| 15 | 15 | 418746 | 5 | 1 | actor | Unknown | Blacksmith | 2025-12-06T22:40:22.143+00:00 |

10,000+ rows | Truncated data | 4.25s runtime

Refreshed 20 hours ago

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Yesterday (2s)

```
%sql
select * from imdb_final_project.gold_dim_profession
```

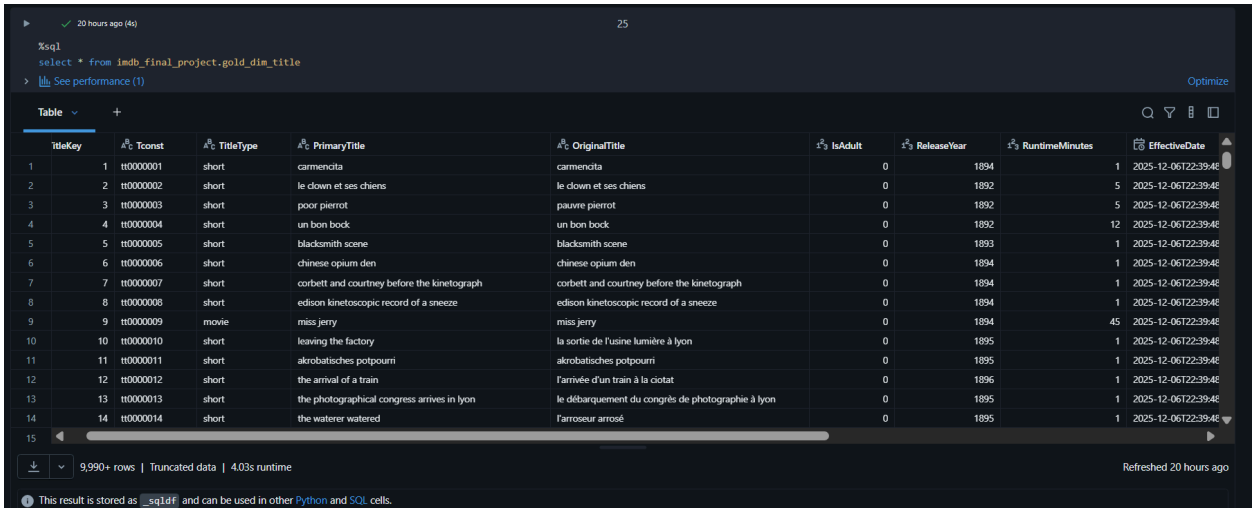
See performance (1)

Table

| | ProfessionKey | Profession |
|---|---|---|
| 1 | 1 | accountant |
| 2 | 2 | actor |
| 3 | 3 | actress |
| 4 | 4 | animation_department |
| 5 | 5 | archive_footage |
| 6 | 6 | archive_sound |
| 7 | 7 | art_department |
| 8 | 8 | art_director |
| 9 | 9 | assistant |
| 10 | 10 | assistant_director |
| 11 | 11 | camera_department |
| 12 | 12 | casting_department |
| 13 | 13 | casting_director |
| 14 | 14 | choreographer |
| 15 | 15 | cinematographer |

47 rows | 2.18s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.



```
%sql
select * from imdb_final_project.gold_dim_title
```

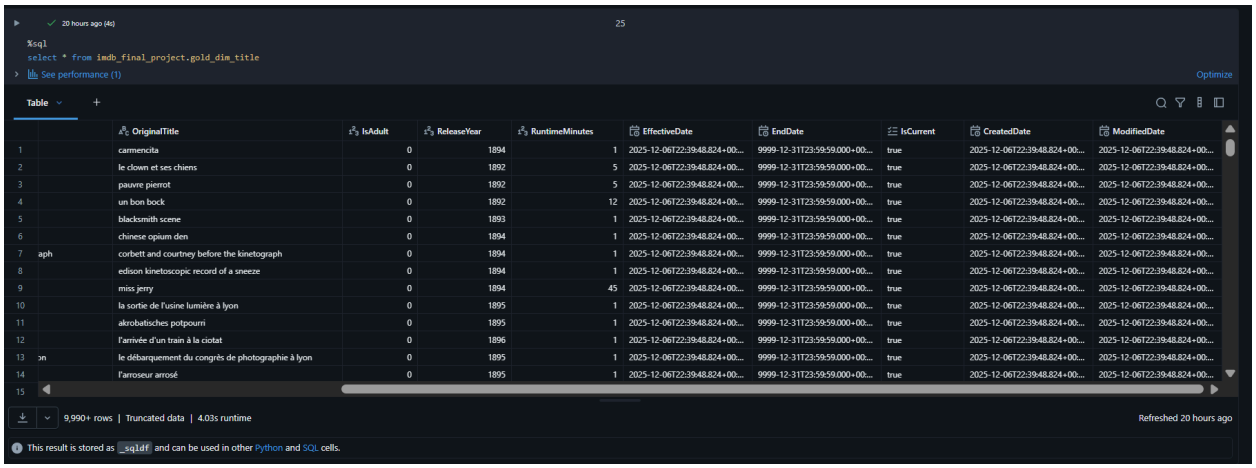
See performance (1) — Optimize

Table

| | titleKey | Tconst | TitleType | PrimaryTitle | OriginalTitle | IsAdult | ReleaseYear | RuntimeMinutes | EffectiveDate |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | tt0000001 | short | carmencita | carmencita | 0 | 1894 | 1 | 2025-12-06T22:39:48 |
| 2 | 2 | tt0000002 | short | le clown et ses chiens | le clown et ses chiens | 0 | 1892 | 5 | 2025-12-06T22:39:48 |
| 3 | 3 | tt0000003 | short | poor pierrot | pauvre pierrot | 0 | 1892 | 5 | 2025-12-06T22:39:48 |
| 4 | 4 | tt0000004 | short | un bon bock | un bon bock | 0 | 1892 | 12 | 2025-12-06T22:39:48 |
| 5 | 5 | tt0000005 | short | blacksmith scene | blacksmith scene | 0 | 1893 | 1 | 2025-12-06T22:39:48 |
| 6 | 6 | tt0000006 | short | chinese opium den | chinese opium den | 0 | 1894 | 1 | 2025-12-06T22:39:48 |
| 7 | 7 | tt0000007 | short | corbett and courtney before the kinetograph | corbett and courtney before the kinetograph | 0 | 1894 | 1 | 2025-12-06T22:39:48 |
| 8 | 8 | tt0000008 | short | edison kinetoscopic record of a sneeze | edison kinetoscopic record of a sneeze | 0 | 1894 | 1 | 2025-12-06T22:39:48 |
| 9 | 9 | tt0000009 | movie | miss jerry | miss jerry | 0 | 1894 | 45 | 2025-12-06T22:39:48 |
| 10 | 10 | tt0000010 | short | leaving the factory | la sortie de l'usine lumière à lyon | 0 | 1895 | 1 | 2025-12-06T22:39:48 |
| 11 | 11 | tt0000011 | short | akrobatisches potpourri | akrobatisches potpourri | 0 | 1895 | 1 | 2025-12-06T22:39:48 |
| 12 | 12 | tt0000012 | short | the arrival of a train | l'arrivée d'un train à la ciotat | 0 | 1896 | 1 | 2025-12-06T22:39:48 |
| 13 | 13 | tt0000013 | short | the photographical congress arrives in lyon | le débarquement du congrès de photographie à lyon | 0 | 1895 | 1 | 2025-12-06T22:39:48 |
| 14 | 14 | tt0000014 | short | the waterer watered | l'arroseur arrosé | 0 | 1895 | 1 | 2025-12-06T22:39:48 |
| 15 | | | | | | | | | |

9,990+ rows | Truncated data | 4.03s runtime — Refreshed 20 hours ago

This result is stored as `_sqldf` and can be used in other Python and SQL cells.



```
%sql
select * from imdb_final_project.gold_dim_title
```

See performance (1) — Optimize

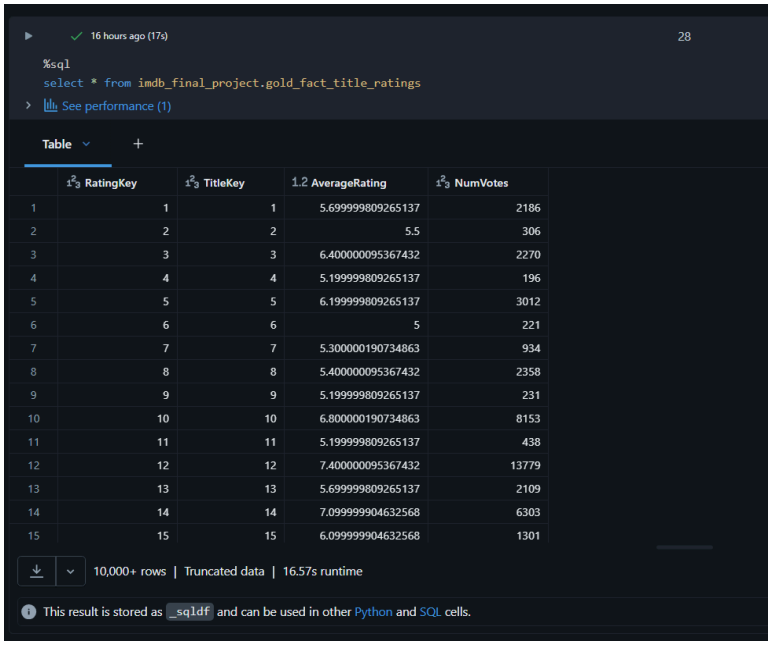
Table

| | OriginalTitle | IsAdult | ReleaseYear | RuntimeMinutes | EffectiveDate | EndDate | IsCurrent | CreatedDate | ModifiedDate |
|---|---|---|---|---|---|---|---|---|---|
| 1 | carmencita | 0 | 1894 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 2 | le clown et ses chiens | 0 | 1892 | 5 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 3 | pauvre pierrot | 0 | 1892 | 5 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 4 | un bon bock | 0 | 1892 | 12 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 5 | blacksmith scene | 0 | 1893 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 6 | chinese opium den | 0 | 1894 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 7 | aph corbett and courtney before the kinetograph | 0 | 1894 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 8 | edison kinetoscopic record of a sneeze | 0 | 1894 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 9 | miss jerry | 0 | 1894 | 45 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 10 | la sortie de l'usine lumière à lyon | 0 | 1895 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 11 | akrobatisches potpourri | 0 | 1895 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 12 | l'arrivée d'un train à la ciotat | 0 | 1896 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 13 | on le débarquement du congrès de photographie à lyon | 0 | 1895 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 14 | l'arroseur arrosé | 0 | 1895 | 1 | 2025-12-06T22:39:48.824+00:... | 9999-12-31T23:59:59.000+00:... | true | 2025-12-06T22:39:48.824+00:... | 2025-12-06T22:39:48.824+00:... |
| 15 | | | | | | | | | |

9,990+ rows | Truncated data | 4.03s runtime — Refreshed 20 hours ago

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

# Gold Fact Tables



```
%sql
select * from imdb_final_project.gold_fact_title_ratings
```

See performance (1)

Table

| | RatingKey | TitleKey | AverageRating | NumVotes |
|---|---|---|---|---|
| 1 | 1 | 1 | 5.699999809265137 | 2186 |
| 2 | 2 | 2 | 5.5 | 306 |
| 3 | 3 | 3 | 6.400000095367432 | 2270 |
| 4 | 4 | 4 | 5.199999809265137 | 196 |
| 5 | 5 | 5 | 6.199999809265137 | 3012 |
| 6 | 6 | 6 | 5 | 221 |
| 7 | 7 | 7 | 5.300000190734863 | 934 |
| 8 | 8 | 8 | 5.400000095367432 | 2358 |
| 9 | 9 | 9 | 5.199999809265137 | 231 |
| 10 | 10 | 10 | 6.800000190734863 | 8153 |
| 11 | 11 | 11 | 5.199999809265137 | 438 |
| 12 | 12 | 12 | 7.400000095367432 | 13779 |
| 13 | 13 | 13 | 5.699999809265137 | 2109 |
| 14 | 14 | 14 | 7.099999904632568 | 6303 |
| 15 | 15 | 15 | 6.099999904632568 | 1301 |

10,000+ rows | Truncated data | 16.57s runtime

This result is stored as `_sqldf` and can be used in other Python and SQL cells.

Yesterday (3s) 27

```
%sql
select * from imdb_final_project.gold_fact_episodes
```

See performance (1)

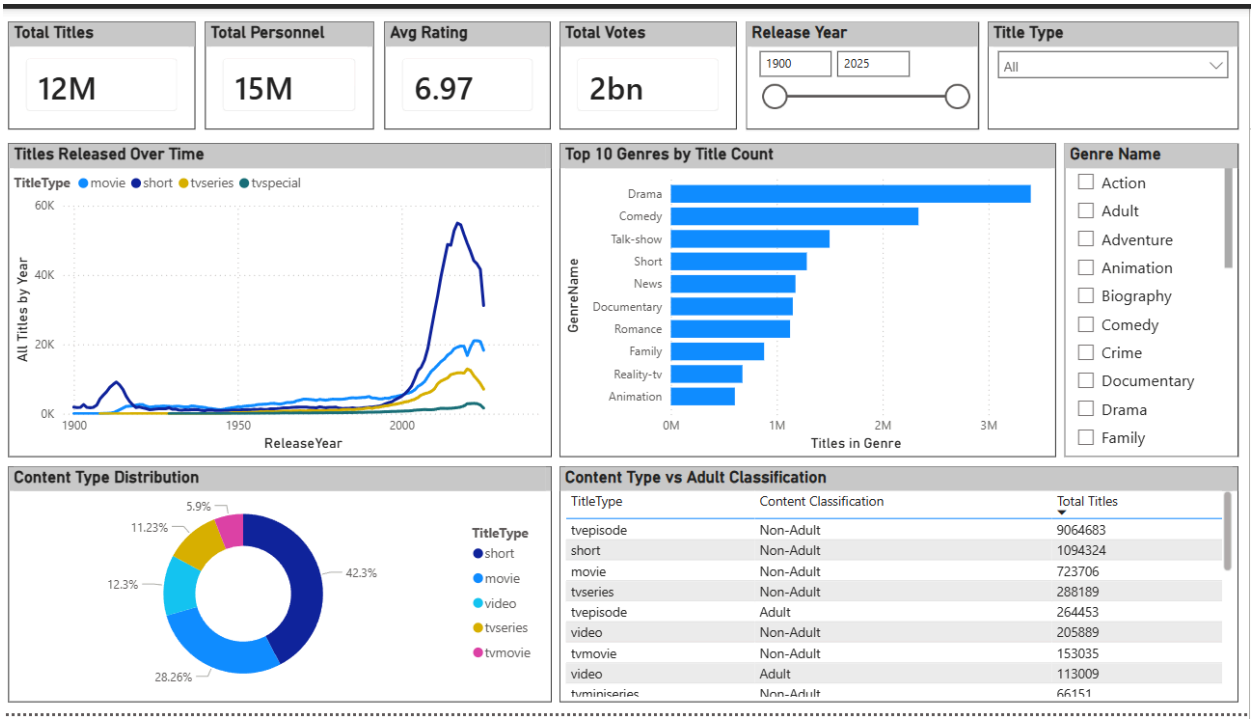
Table +

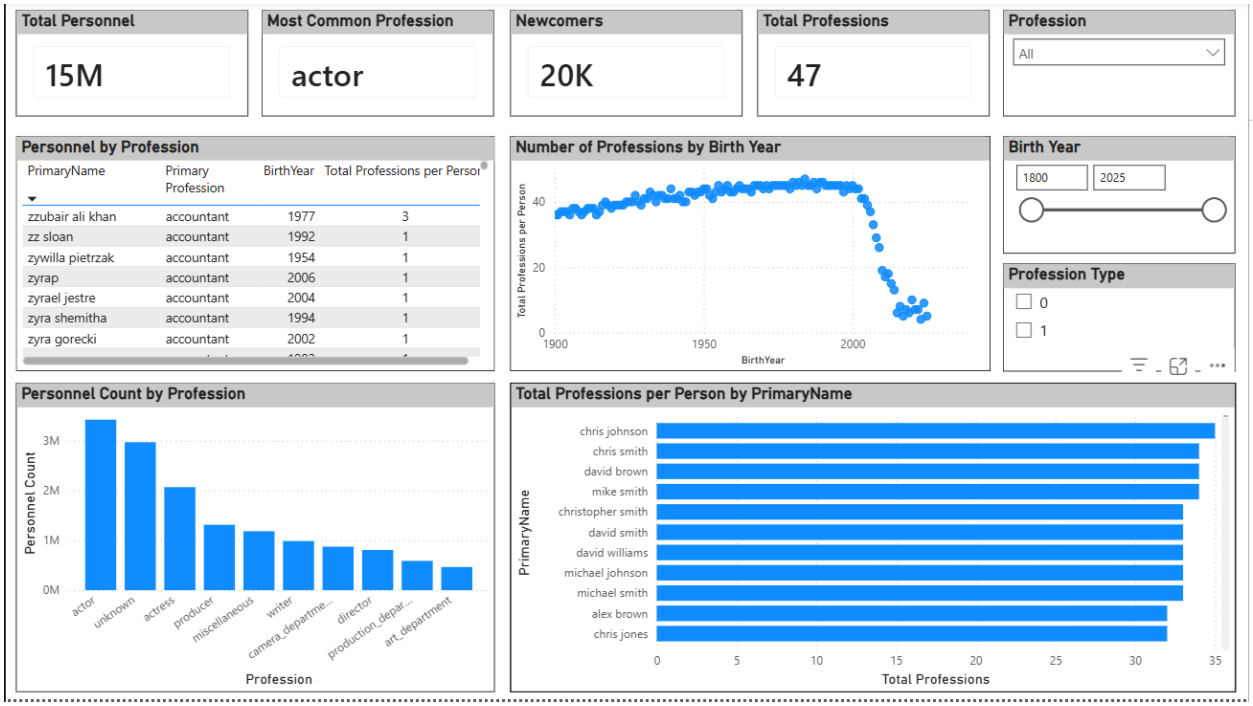
| | EpisodeKey | TitleKey | SeasonNumber | EpisodeNumber |
|---|---|---|---|---|
| 1 | 1 | 30892 | -9999 | -9999 |
| 2 | 2 | 41210 | 1 | 9 |
| 3 | 3 | 42058 | 1 | 17 |
| 4 | 4 | 42130 | -9999 | -9999 |
| 5 | 5 | 42654 | 3 | 42 |
| 6 | 6 | 42859 | 2 | 16 |
| 7 | 7 | 42920 | 2 | 8 |
| 8 | 8 | 42937 | 3 | 3 |
| 9 | 9 | 43309 | 1 | 6 |
| 10 | 10 | 43871 | 2 | 16 |
| 11 | 11 | 44100 | 3 | 46 |
| 12 | 12 | 44705 | 4 | 11 |
| 13 | 13 | 45137 | 2 | 3 |
| 14 | 14 | 45309 | 4 | 5 |
| 15 | 15 | 45323 | -9999 | -9999 |

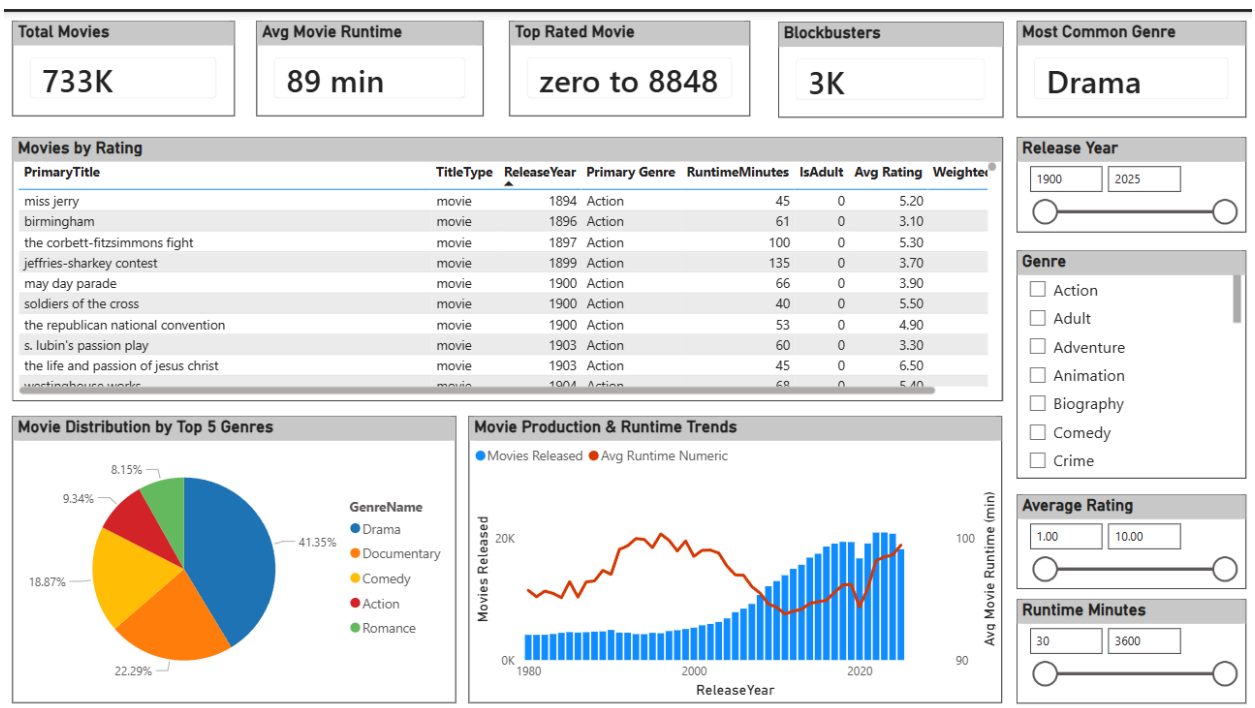
10,000+ rows | Truncated data | 3.05s runtime

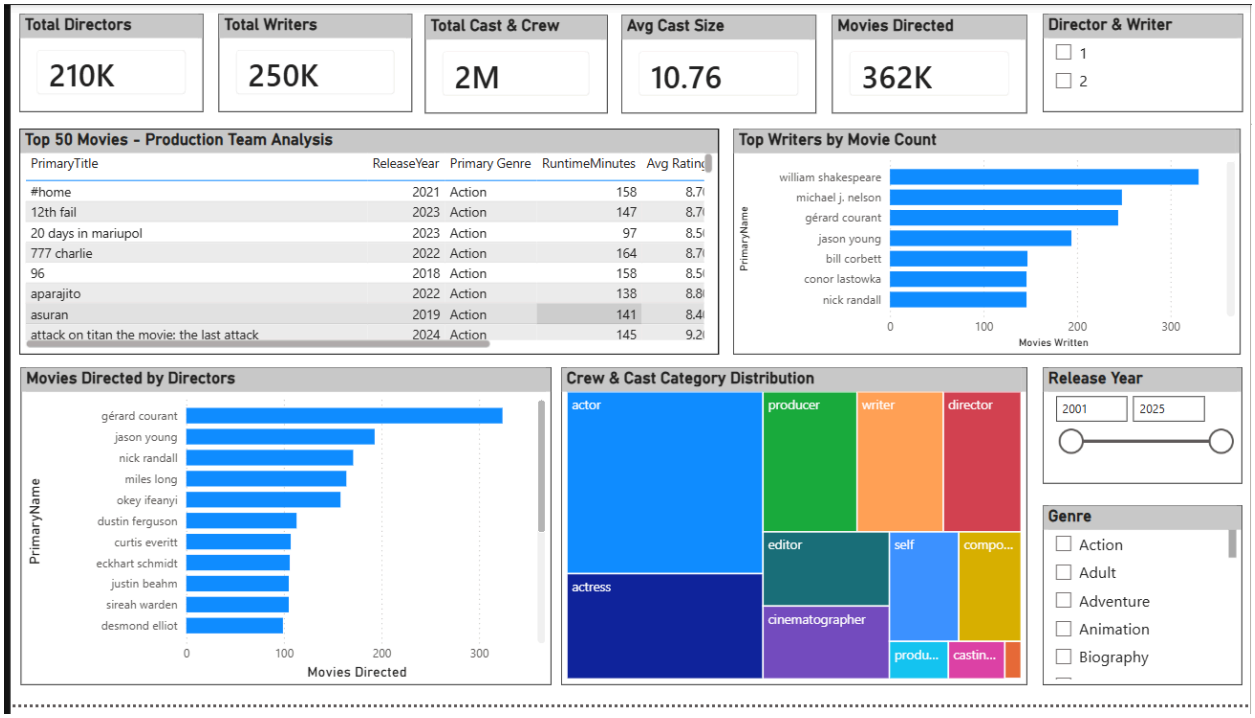
This result is stored as `_sqldf` and can be used in other Python and SQL cells.

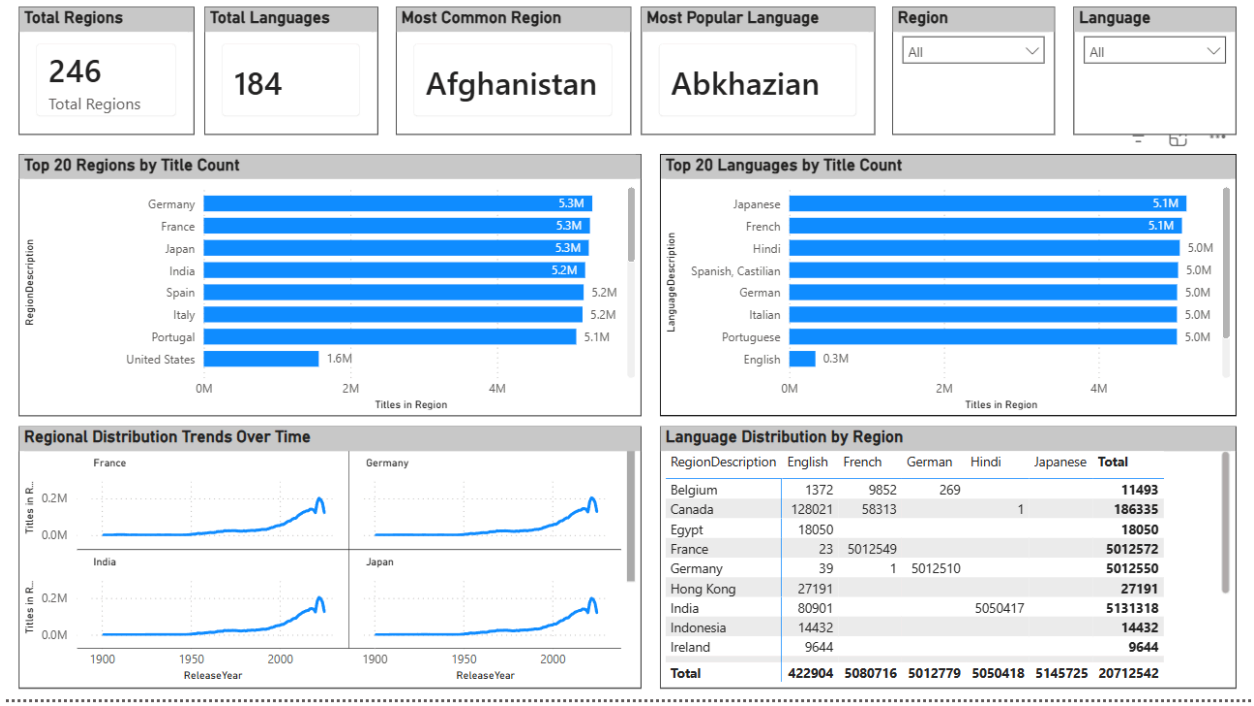
# POWERBI DASHBOARDS

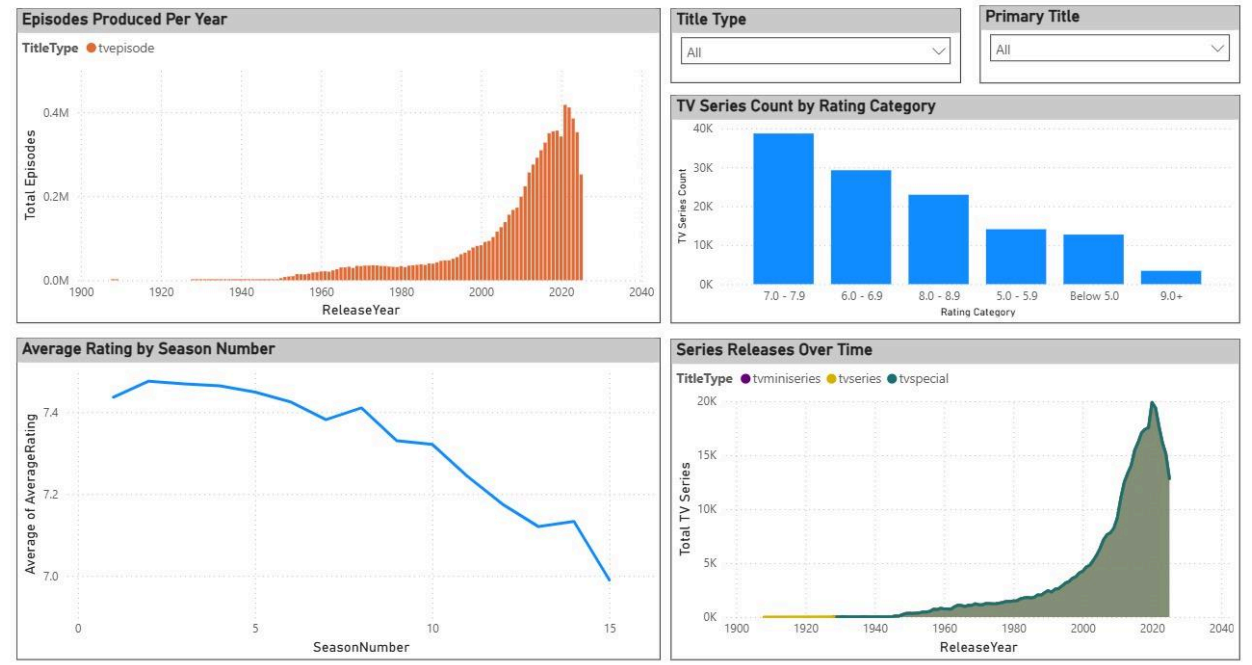



Total Personnel
15M
Most Common Profession
actor
Newcomers
20K
Total Professions
47
Profession
All
Personnel by Profession
PrimaryName
Primary Profession
BirthYear
Total Professions per Perso
zzubair ali khan accountant 1977 3
zz sloan accountant 1992 1
zywilla pietrzak accountant 1954 1
zyrap accountant 2006 1
zyrael jestre accountant 2004 1
zyra shemitha accountant 1994 1
zyra gorecki accountant 2002 1
Number of Professions by Birth Year
Total Professions per Person
BirthYear
Birth Year
1800
2025
Profession Type
0
1
Personnel Count by Profession
Personnel Count
Profession
actor
unknown
actress
producer
miscellaneous
writer
director
art_department
Total Professions per Person by PrimaryName
PrimaryName
chris johnson
chris smith
david brown
mike smith
christopher smith
david smith
david williams
michael johnson
michael smith
alex brown
chris jones
Total Professions


Total Movies
733K
Avg Movie Runtime
89 min
Top Rated Movie
zero to 8848
Blockbusters
3K
Most Common Genre
Drama
Movies by Rating
PrimaryTitle
TitleType
ReleaseYear
Primary Genre
RuntimeMinutes
IsAdult
Avg Rating
Weighte
miss jerry movie 1894 Action 45 0 5.20
birmingham movie 1896 Action 61 0 3.10
the corbett-fitzsimmons fight movie 1897 Action 100 0 5.30
jeffries-sharkey contest movie 1899 Action 135 0 3.70
may day parade movie 1900 Action 66 0 3.90
soldiers of the cross movie 1900 Action 40 0 5.50
the republican national convention movie 1900 Action 53 0 4.90
s. lubin's passion play movie 1903 Action 60 0 3.30
the life and passion of jesus christ movie 1903 Action 45 0 6.50
Release Year
1900
2025
Genre
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Movie Distribution by Top 5 Genres
8.15%
9.34%
41.35%
18.87%
22.29%
GenreName
Drama
Documentary
Comedy
Action
Romance
Movie Production & Runtime Trends
Movies Released
Avg Runtime Numeric
Movies Released
Avg Movie Runtime (min)
ReleaseYear
Average Rating
1.00
10.00
Runtime Minutes
30
3600


Total Directors
210K
Total Writers
250K
Total Cast & Crew
2M
Avg Cast Size
10.76
Movies Directed
362K
Director & Writer
1
2
Top 50 Movies - Production Team Analysis
PrimaryTitle ReleaseYear Primary Genre RuntimeMinutes Avg Ratin
#home 2021 Action 158 8.7
12th fail 2023 Action 147 8.7
20 days in mariupol 2023 Action 97 8.5
777 charlie 2022 Action 164 8.7
96 2018 Action 158 8.5
aparajito 2022 Action 138 8.8
asuran 2019 Action 141 8.4
attack on titan the movie: the last attack 2024 Action 145 9.2
Top Writers by Movie Count
william shakespeare
michael j. nelson
gérard courant
jason young
bill corbett
conor lastowka
nick randall
PrimaryName
0
100
200
300
Movies Written
Movies Directed by Directors
gérard courant
jason young
nick randall
miles long
okey ifeanyi
dustin ferguson
curtis everitt
eckhart schmidt
justin beahm
sireah warden
desmond elliot
PrimaryName
0
100
200
300
Movies Directed
Crew & Cast Category Distribution
actor
producer
writer
director
actress
editor
self
compo...
cinematographer
produ...
castin...
Release Year
2001
2025
Genre
Action
Adult
Adventure
Animation
Biography


Total Regions
246
Total Regions
Total Languages
184
Most Common Region
Afghanistan
Most Popular Language
Abkhazian
Region
All
Language
All
Top 20 Regions by Title Count
Germany 5.3M
France 5.3M
Japan 5.3M
India 5.2M
Spain 5.2M
Italy 5.2M
Portugal 5.1M
United States 1.6M
RegionDescription
0M
2M
4M
Titles in Region
Top 20 Languages by Title Count
Japanese 5.1M
French 5.1M
Hindi 5.0M
Spanish, Castilian 5.0M
German 5.0M
Italian 5.0M
Portuguese 5.0M
English 0.3M
LanguageDescription
0M
2M
4M
Titles in Region
Regional Distribution Trends Over Time
France
Germany
India
Japan
Titles in R..
0.2M
0.0M
1900
1950
2000
ReleaseYear
Language Distribution by Region
RegionDescription English French German Hindi Japanese Total
Belgium 1372 9852 269 11493
Canada 128021 58313 1 186335
Egypt 18050 18050
France 23 5012549 5012572
Germany 39 1 5012510 5012550
Hong Kong 27191 27191
India 80901 5050417 5131318
Indonesia 14432 14432
Ireland 9644 9644
Total 422904 5080716 5012779 5050418 5145725 20712542




## BUSINESS REQUIREMENTS

**1.Get different types of professions for any personnel**

Tables Used: DIM_Name, DIM_Profession, BRIDGE_Profession

Dashboards:

- Personnel by Profession
- Personnel Count by Profession
- Total Professions per Person

Explanation: Displays each person's primary + secondary professions through bridge table relationships.

**2. Identify personnel with multiple professions**

Tables Used: BRIDGE_Profession

Dashboards:

- Total Professions per Person (Bar chart)

Explanation: Helps highlight individuals with diverse roles across media.

**3. Get list of genres for a title & movies by genre**

Tables Used: DIM_Genre, BRIDGE_Title_Genre, DIM_Title

Dashboards:

- Top Genres by Title Count
- Genre filter panels
- Movie Distribution by Genre (Pie Chart)

Explanation: Enables selecting a genre and viewing corresponding titles.

**4. Find movies released in a given year**
Tables Used: DIM_Title (ReleaseYear)
Dashboards:

- Titles Released Over Time
- Movie Production Trends

Explanation: User can filter movies across any year range.

**5. Track movie length over release year**
Tables Used: DIM_Title (RuntimeMinutes)
Dashboards:

- Movie Production & Runtime Trends

Explanation: Correlates runtime patterns with release year.

**6. List adult vs non-adult movies**
Tables Used: DIM_Title (IsAdult)
Dashboards:

- Content Type vs Adult Classification

Explanation: Categories movies into Adult / Non-Adult buckets.

**7. Show languages each title is associated with**
Tables Used: DIM_Language, BRIDGE_Akas
Dashboards:

- Top Languages by Title Count
- Language Distribution by Region

Explanation: Surfaces multilingual titles and regional language trends.

**8. List regions where a movie is released**
Tables Used: DIM_Region, BRIDGE_Akas
Dashboards:

- Top Regions by Title Count
- Regional Distribution Trends Over Time

Explanation: Tracks country/region-specific film releases.

**9. Directors & writers involved in a movie**
Tables Used: BRIDGE_Title_Crew, DIM_Name
Dashboards:

- Movies Directed by Directors
- Top Writers by Movie Count

- Crew & Cast Category Distribution

Explanation: Helps identify the most active creators.

**10. Number of episodes per season**

Tables Used: FACT_Episodes

Dashboards:

- Episodes Produced Per Year
- Average Rating by Season
- Series Releases Over Time

Explanation: Shows episode volumes and their impact on ratings.

**11. Crew/cast types for any title**

Tables Used: DIM_Principals

Dashboards:

- Crew & Cast Category Distribution

Explanation: Displays categories such as actor, director, editor, producer, etc.

**12. List of jobs involved in a title**

Tables Used: DIM_Principals (Job)

Dashboards:

- Crew & Cast Category Distribution

Explanation: Helps understand job roles behind each production.

**13. List of characters involved in a title**

Tables Used: DIM_Principals (Characters)

Dashboards:

- Cast Analysis (part of personnel dashboards)

Explanation: Shows key characters portrayed.

**14. Top-rated movies by year or genre**

Tables Used: FACT_Title_Ratings, DIM_Title

Dashboards:

- Movies by Rating
- Top Rated Movie KPI
- Movies by Rating Category

Explanation: Easily compare best movies across time or genre.

**15. Movie Insights (General Performance Dashboard)**

Dashboards:

- Total Movies KPI

- Avg Movie Runtime
- Movie Distribution by Genre
- Movie Trends Over Time

Explanation: Provides end-to-end movie performance analytics.

**16. Crew Member Analysis (Actor, Director, Producer, Writers, etc.)**

Dashboards:

- Movies Directed by Directors
- Top Writers
- Crew & Cast Category Tree Map
- Personnel Count by Profession

Explanation: Deep dive into human resource contributions in film production.

**17. Movie Trend Analysis Based on Ratings**

Dashboards:

- Movies by Rating
- Average Rating by Season Number
- Rating-based filters & visuals

Explanation: Understand user preferences and high-rating patterns.

**18. Non-movie Title Analysis (Seasons, Episodes & Ratings)**

Dashboards:

- Episodes Produced Per Year
- Average Rating by Season Number
- Series Releases Over Time

Explanation: Allows comparison of series, episodes, and rating behaviour.

**19. Region & Country-Based Movie Releases**

Dashboards:

- Top Regions by Title Count
- Regional Distribution Over Time
- Language Distribution by Region

Explanation: Helps understand geographical consumption & production.

This covers all business requirements, maps them to the correct Power BI visuals, and shows exactly which dimension/fact/bridge tables satisfy each requirement.