IMDb Team Project - GROUP 3

Architecture Design

BRONZE LAYER (Raw Ingestion)

- Technology: Auto Loader (cloudFiles) + Delta Tables
- Schema evolution enabled
- Streaming ingestion from TSV files
- Metadata tracking: _source_file, _loaded_at
- Change Data Feed enabled
- Basic validation: ID format checks

SILVER LAYER (Cleansed Data)

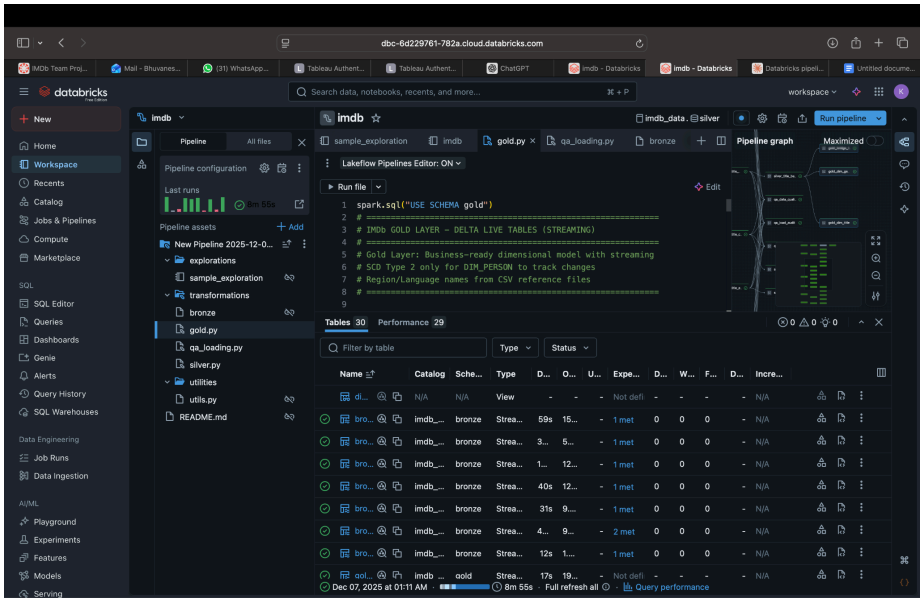
- Technology: Delta Live Tables (Streaming)
- Data type conversions (String → Int/Double/Boolean)
- Data cleansing and standardization
- Range validations (years, ratings, runtime)
- Array transformations (split multi-valued fields)
- @dlt.expect and @dlt.expect_or_drop for quality
- Timestamp tracking: source_loaded_at, silver_processed_at

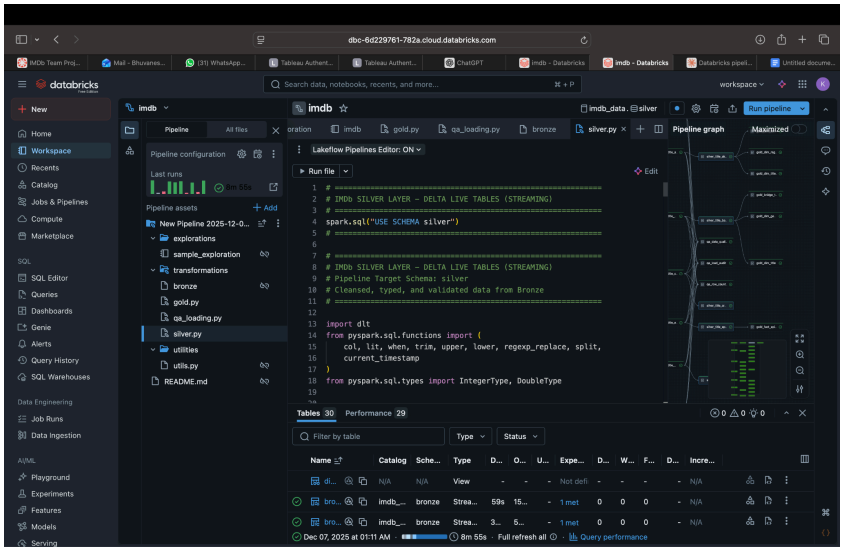
GOLD LAYER (Business-Ready Dimensional)

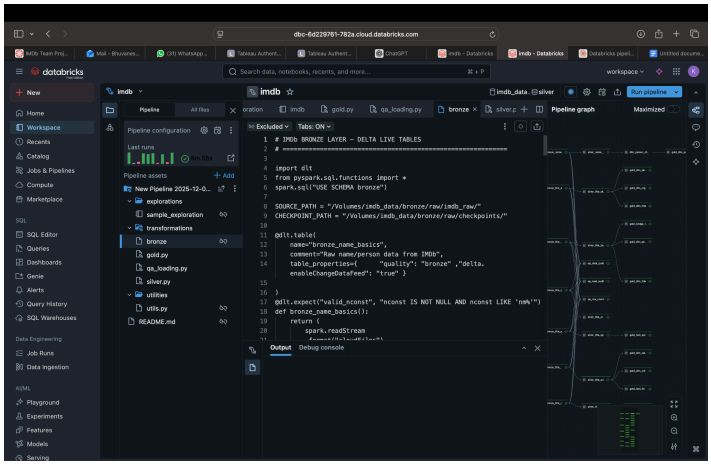
- Technology: Delta Live Tables (Streaming)
- Star schema: Dimensions + Facts + Bridge tables
- SCD Type 2 for DIM_PERSON
- Reference data enrichment (ISO codes)
- Business logic applied
- Optimized for analytics

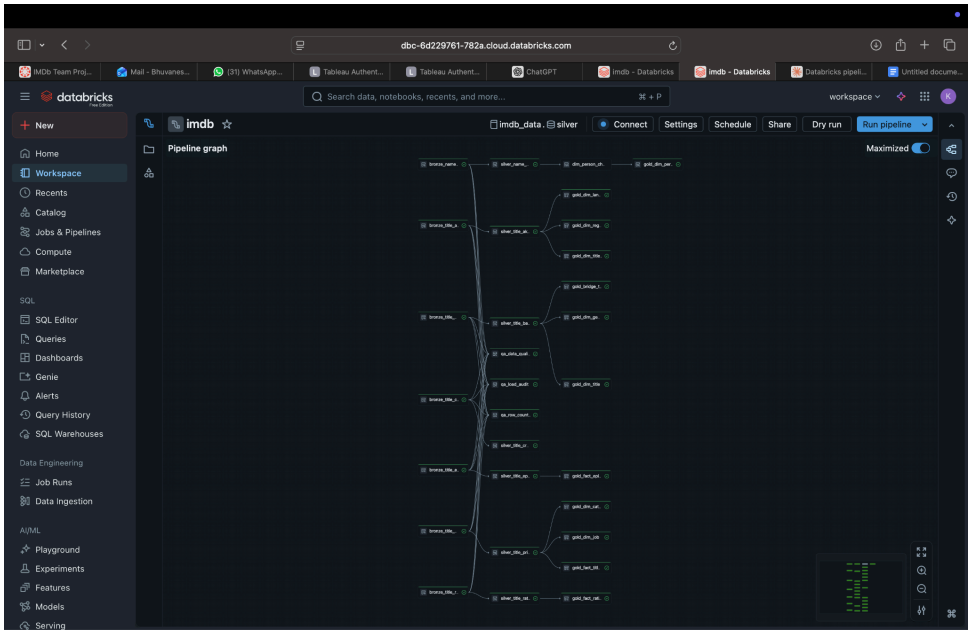
QA LAYER (Quality Assurance)

- Load audit tracking (qa_load_audit)
- Row count history (qa_row_count_history)
- Data quality metrics (qa_data_quality)
- Schema drift detection
- Quality scoring system


dbc-6d229761-782a.cloud.databricks.com
databricks
Search data, notebooks, recents, and more...
workspace
imdb
imdb_data . silver
Run pipeline
Pipeline graph
Maximized
Lakeflow Pipelines Editor: ON
Run file
Edit
spark.sql("USE SCHEMA gold")
# IMDb GOLD LAYER – DELTA LIVE TABLES (STREAMING)
# Gold Layer: Business-ready dimensional model with streaming
# SCD Type 2 only for DIM_PERSON to track changes
# Region/Language names from CSV reference files
Pipeline configuration
Last runs
Pipeline assets
Add
New Pipeline 2025-12-0...
explorations
sample_exploration
transformations
bronze
gold.py
qa_loading.py
silver.py
utilities
utils.py
README.md
Tables 30
Performance 29
Filter by table
Type
Status
Name
Catalog
Sche...
Type
Expe...
Incre...
Dec 07, 2025 at 01:11 AM
8m 55s
Full refresh all
Query performance


dbc-6d229761-782a.cloud.databricks.com
databricks
Search data, notebooks, recents, and more...
workspace
imdb
imdb_data . silver
Run pipeline
Pipeline graph
Maximized
Lakeflow Pipelines Editor: ON
Run file
Edit
# IMDb SILVER LAYER – DELTA LIVE TABLES (STREAMING)
spark.sql("USE SCHEMA silver")
# IMDb SILVER LAYER – DELTA LIVE TABLES (STREAMING)
# Pipeline Target Schema: silver
# Cleansed, typed, and validated data from Bronze
import dlt
from pyspark.sql.functions import (
col, lit, when, trim, upper, lower, regexp_replace, split,
current_timestamp
)
from pyspark.sql.types import IntegerType, DoubleType
Pipeline configuration
Last runs
Pipeline assets
Add
New Pipeline 2025-12-0...
explorations
sample_exploration
transformations
bronze
gold.py
qa_loading.py
silver.py
utilities
utils.py
README.md
Tables 30
Performance 29
Filter by table
Type
Status
Name
Catalog
Sche...
Type
Expe...
Incre...
Dec 07, 2025 at 01:11 AM
8m 55s
Full refresh all
Query performance


dbc-6d229761-782a.cloud.databricks.com
databricks
imdb
imdb_data . silver
Run pipeline
Pipeline graph
Maximized
Excluded
Tabs: ON
# IMDb BRONZE LAYER – DELTA LIVE TABLES
import dlt
from pyspark.sql.functions import *
spark.sql("USE SCHEMA bronze")
SOURCE_PATH = "/Volumes/imdb_data/bronze/raw/imdb_raw/"
CHECKPOINT_PATH = "/Volumes/imdb_data/bronze/raw/checkpoints/"
@dlt.table(
name="bronze_name_basics",
comment="Raw name/person data from IMDb",
table_properties={ "quality": "bronze" ,"delta.
enableChangeDataFeed": "true" }
)
@dlt.expect("valid_nconst", "nconst IS NOT NULL AND nconst LIKE 'nm%'")
def bronze_name_basics():
return (
spark.readStream
Output
Debug console



DATAFLOW DIAGRAM

TSV Files → Auto Loader → Bronze Tables → DLT Streaming→ Silver Tables

↓

QA Layer

↓

Quality Validation

↓

Silver Tables → DLT Transformations → Gold Dimensions & Facts

↓

Dashboards

Schema Architecture

imdb_data (Catalog)

├── bronze (Schema)

│ ├── bronze_name_basics

│ ├── bronze_title_basics

│ ├── bronze_title_ratings

│ ├── bronze_title_episode

│ ├── bronze_title_akas

│ ├── bronze_title_principals

│ └── bronze_title_crew

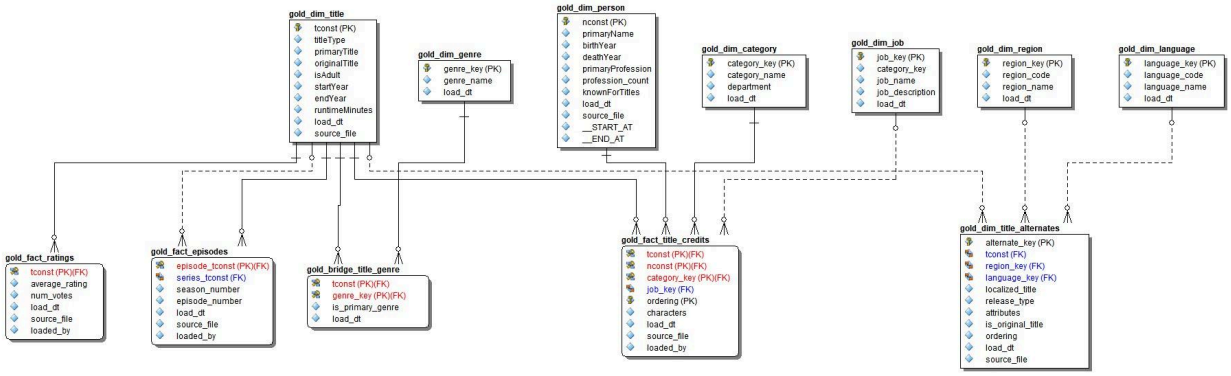
├── silver (Schema)

│ ├── silver_name_basics

│ ├── silver_title_basics

│ ├── silver_title_ratings

```
│   ├── silver_title_episode
│   ├── silver_title_akas
│   ├── silver_title_principals
│   └── silver_title_crew
├── gold (Schema)
│   ├── gold_dim_person (SCD Type 2)
│   ├── gold_dim_title
│   ├── gold_dim_genre
│   ├── gold_dim_category
│   ├── gold_dim_job
│   ├── gold_dim_region
│   ├── gold_dim_language
│   ├── gold_dim_title_alternates
│   ├── gold_bridge_title_genre
│   ├── gold_fact_ratings
│   ├── gold_fact_episodes
│   └── gold_fact_title_credits
└── qa (Schema)
    ├── qa_load_audit
    ├── qa_row_count_history
    └── qa_data_quality
```



Standard metadata columns-

Column Name	Data Type	Purpose	Layer
source_file	STRING	Original file path	Bronze
_loaded_at	TIMESTAMP	Bronze ingestion time	Bronze
source_loaded_at	TIMESTAMP	Bronze timestamp reference	Silver/Gold
silver_processed_at	TIMESTAMP	Silver processing time	Silver
load_dt	TIMESTAMP	Gold load time	Gold
Source_file	STRING	Source identifier	Gold
loaded_by	STRING	ETL job identifier	Gold

ETL Pipeline Implementation

Pipeline Architecture

Pipeline Components:

1. **explorations** - Setup and initial data quality checks
2. **transformations/bronze.py** - Raw data ingestion
3. **transformations/silver.py** - Data cleansing and validation
4. **transformations/gold.py** - Dimensional modeling
5. **transformations/qa_loading.py** - Quality assurance
6. **utilities/utils.py** - Helper functions

Bronze Layer Implementation

Key Features:

- Streaming ingestion from Volume paths
- Schema hints for all columns
- Automatic schema evolution (addNewColumns)
- Change Data Feed enabled
- Checkpoint management for incremental loads
- IMDb null handling ('\\N' → NULL)

Bronze Tables Created:

1. bronze_name_basics
2. bronze_title_basics
3. bronze_title_ratings
4. bronze_title_episode
5. bronze_title_akas
6. bronze_title_principals
7. bronze_title_crew

Validations:

- ID format checks (tt% for titles, nm% for persons)
- NOT NULL constraints on primary keys
- Metadata tracking for lineage

Silver Layer Implementation

Key Transformations:

Data Type Conversions

```
# String → Integer
when(col("birthYear").rlike("^[0-9]{1,4}$"), col("birthYear").cast(IntegerType()))

# String → Double
col("averageRating").cast(DoubleType())
# String → Boolean
when(col("isAdult") == "1", lit(True)).otherwise(lit(False))
```

Data Standardization

```
# Region codes: UPPER case
upper(trim(col("region")))
# Language codes: lower case
lower(trim(col("language")))
# Category standardization
lower(trim(col("category")))
```

Array Transformations

```
# Split comma-separated values into arrays
when(col("primaryProfession").isNotNull(),
   split(trim(col("primaryProfession")), ","))
# Split genres
when(col("genres").isNotNull(),
   split(trim(col("genres")), ","))
```

Data Cleansing

```
# Remove JSON brackets from characters field
regexp_replace(regexp_replace(trim(col("characters")), "^\\[\"", ""), "\"\\]$", "")
```

Silver Tables Created:

1. silver_name_basics
2. silver_title_basics
3. silver_title_ratings
4. silver_title_episode
5. silver_title_akas
6. silver_title_principals
7. Silver_title_crew

Gold Layer Implementation

Dimension Tables

DIM_PERSON (SCD Type 2)

- Tracks historical changes in: primaryName, deathYear, primaryProfession
- Uses dlt.apply_changes() for automatic SCD Type 2 management
- Includes __start_at, __end_at, __current columns

gold_dim_title - Title/movie details
gold_dim_genre - Genre lookup (exploded from arrays)
gold_dim_category - Role category with department classification
gold_dim_job- Job titles by category
gold_dim_region- Region codes enriched with ISO country names
gold_dim_language - Language codes enriched with ISO language names
gold_dim_title_alternates - Alternative titles by region/language

Reference Data Integration:

Bridge Tables

gold_bridge_title_genre
- Handles many-to-many relationship between titles and genres
- Explodes genre arrays into individual records
- Includes is_primary_genre flag

Fact Tables

Gold_fact_ratings
- Granularity: One record per title
- Measures: average_rating, num_votes
- Foreign Keys: tconst → dim_title

Gold_fact_episodes
- Granularity: One record per episode
- Links episodes to their parent series
- Includes season_number, episode_number

gold_fact_title_credits
- Granularity: One record per person per title per role
- Links cast/crew to titles
- Includes ordering, job, characters
- Foreign Keys: tconst → dim_title, nconst → dim_person

QA Layer Implementation

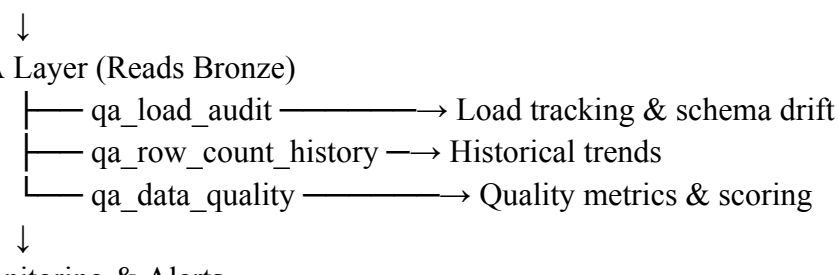
Data Quality & Validation

QA Layer Implementation

- Purpose: Automated quality assurance, monitoring, and validation framework
- Technology:Delta Live Tables (Batch mode)
- Schema: imdb_data.qa
- Trigger: Executes after Bronze layer completes

Architecture Overview

The QA layer provides automated data quality monitoring through three complementary tables:

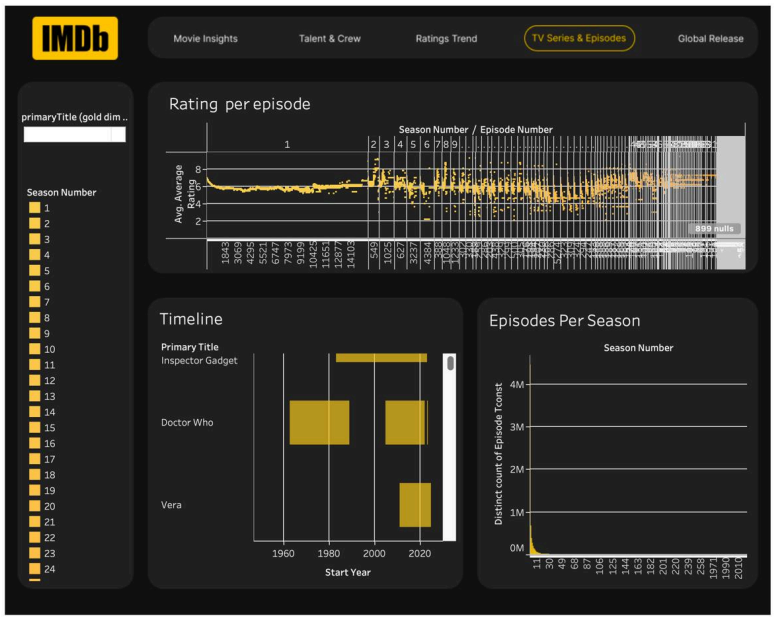


Key Features:

- Dynamic configuration-driven processing
- Schema drift detection
- Automated quality scoring
- Historical trend tracking

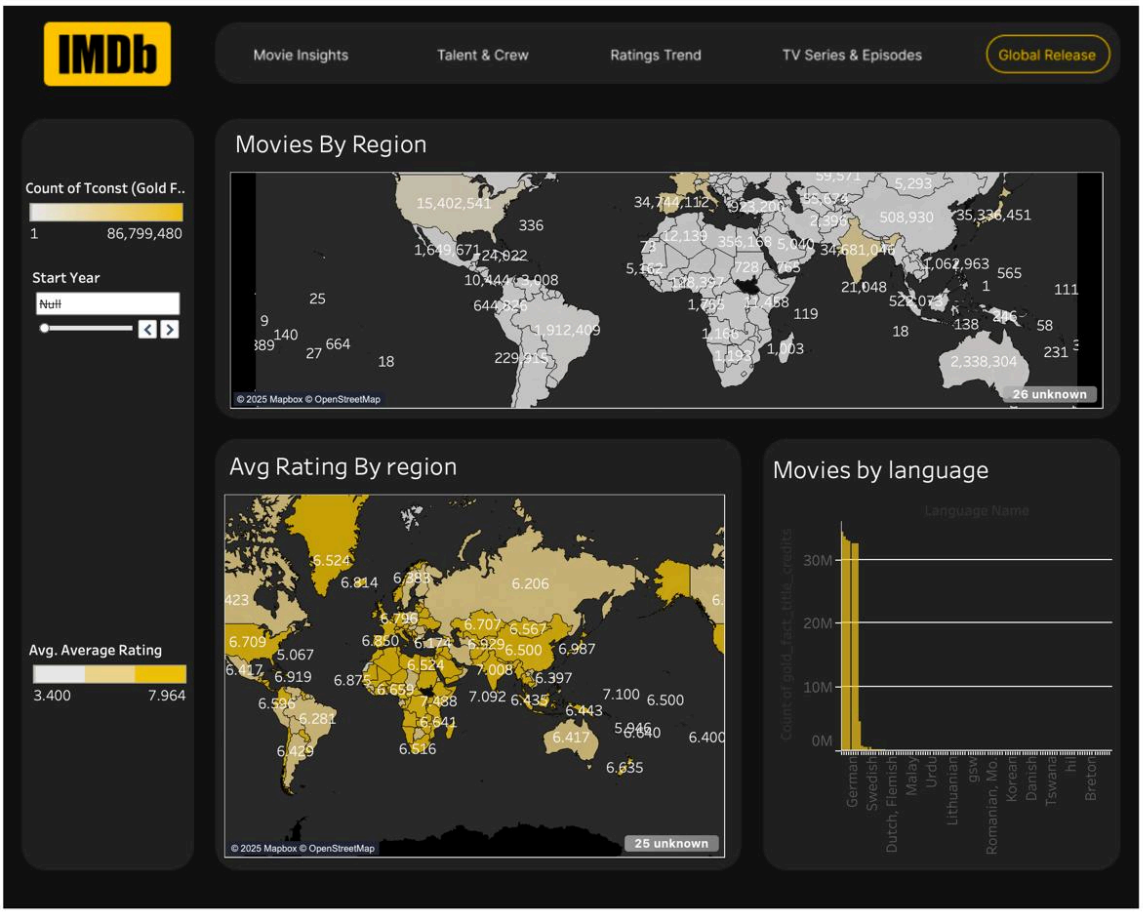
- Duplicate detection
- Null value analysis
- No hardcoded table references

DASHBOARD




IMDb
Movie Insights
Talent & Crew
Ratings Trend
TV Series & Episodes
Global Release
Title Type
(All)
movie
short
tvEpisode
tvMiniSeries
tvMovie
tvPilot
tvSeries
tvShort
tvSpecial
videoGame
multiacheivers
Primary Name
Primary Profession
50 Cent
music_artist,actor,produ..
98 Degrees
actor,soundtrack,archive..
100 Kila
actor,writer,composer
100dcx
animation_department,a..
112
music_artist,actor,compo..
311
actor,soundtrack,archive..
808 State
actor,music_department,..
911
actor,soundtrack,archive..
01099
music_artist,actor,compo..
TheGhostFace
director,actor,writer
'Big' James Wrobl..
writer,actor,cinematogra..
People by no of professions
profession count
Count of Nconst
1000K
800K
600K
400K
200K
0K
1
2
3
Top Directors
Primary Name
Megumi Ishitani
Tucker Gates
0
1
2
Count of gold_dim_title
Top Writers
Primary Name
Ann Druyan
Arunabh Kumar
B.J. Novak
Bob Kane
Bruce Timm
Bryan Konietzko
Carl Sagan
Chandan Kumar
Clyde Phillips
0
1
2
3
4
Count of gold_dim_title


IMDb
Movie Insights
Talent & Crew
Ratings Trend
TV Series & Episodes
Global Release
Genre Name
(All)
Avg rating by years
Avg. Average Rating
10
5
0
1920
1940
1960
1980
2000
2020
Start Year
Avg Rating by Genre
Genre Name
Null
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Top rated movies
Primary Title
The Shawshank Redempti..
The Godfather
The Chaos Class
Attack on Titan the Movie:..
The Dark Knight
Ramayana: The Legend of ..
The Lord of the Rings: The ..
The Godfather Part II
Schindler's List
8.6
8.8
9.0
9.2
9.4
Avg. Average Rating
Rating Distribution
10K
5K
0K
2
4
6
8
10
Average Rating


IMDb
Movie Insights
Talent & Crew
Ratings Trend
TV Series & Episodes
Global Release
Count of Tconst (Gold F..
1
86,799,480
Start Year
Null
Movies By Region
15,402,541
336
1,649,671
24,922
10,441,008
644,
1,912,409
25
140
27
664
18
34,
508,930
35,336,451
34,681,045
1,062,963
565
1
111
21,048
119
18
138
58
231
2,338,304
© 2025 Mapbox © OpenStreetMap
26 unknown
Avg. Average Rating
3.400
7.964
Avg Rating By region
6.524
6.814
6.206
6.709
5.067
6.919
6.875
6.707
6.567
6.500
6.987
6.524
7.008
6.397
7.092
7.100
6.500
6.443
6.417
6.400
6.635
6.516
© 2025 Mapbox © OpenStreetMap
25 unknown
Movies by language
Language Name
Count of gold_fact_title_credits
30M
20M
10M
0M
German
Swedish
Dutch, Flemish
Malay
Urdu
Lithuanian
gsw
Romanian, Mo
Korean
Danish
Tswana
hil
Breton