Extract

EXTERNAL DATA SOURCES

```
IMDb Non-Commercial Datasets
                                               Box Office Mojo
     https://datasets.imdbws.com/
                                               (Web Scraping)
     • name.basics.tsv.gz (291 MB)
                                               • Revenue Data
     • title.basics.tsv.gz (212 MB)
                                               (22 MB)
     • title.crew.tsv.gz (78 MB)
                                               • BoxOfficeMojo IDs
     • title.episode.tsv.gz (51 MB)
     • title.principals.tsv.gz (736)
                                                Mapping
     • title.ratings.tsv.gz (8 MB)
     Total: 1.37 GB compressed
               [Manual Download & Decompress]
            LOCAL FILE SYSTEM (MySQL Uploads Directory)
            C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/
                              • title.principals.tsv
      name.basics.tsv
     • title.basics.tsv
                              title.ratings.tsv
      • title.crew.tsv
                              • box_office_revenue.tsv
      • title.episode.tsv
               [LOAD DATA INFILE - Bulk Loading]
             (High-performance, bypasses SQL parsing)
LAYER 2: STAGING AREA
```

SCHEMA: boxofficemojo

(daily revenue records) BoxOfficeMojoIds

BoxOfficeRevenue

(tconst mapping)

{ STAGING SCHEMAS }

SCHEMA: imdb

- NameBasics (12M+ rows)
- TitleBasics (11.9M rows)
- TitleCrew (director/writer lists)
- TitleEpisode (9.18M rows)
- TitlePrincipals (94M+ rows - largest)
- TitleRatings (1.4M rows)

Transform

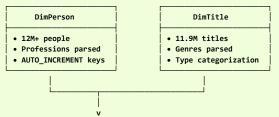
```
[1] STRING PARSING FOR MULTI-VALUED ATTRIBUTES
                                                                        [4] DATE DIMENSION POPULATION (ON-DEMAND STRATEGY)
Fixed-Width Attributes (SUBSTRING_INDEX):
                                                                            INSERT IGNORE INTO DimDate (...)
 genres = "Action,Drama,Thriller"
                                                                            SELECT DISTINCT revenue_date FROM BoxOfficeRevenue
    --> genre1 = "Action"
    --> genre2 = "Drama"
                                                                            SELECT CURDATE()
    --> genre3 = "Thriller"
                                                                            SELECT CURDATE() + INTERVAL 1 DAY;
 knownForTitles = "tt001,tt002,tt003,tt004"
   --> knownForTitle1 = "tt001"
                                                                            Approach: Only materialize dates referenced in fact tables
    --> knownForTitle2 = "tt002"
                                                                            Trade-off: Smaller storage vs. ETL complexity
    --> knownForTitle3 = "tt003"
    --> knownForTitle4 = "tt004"
                                                                        [5] DATA VALIDATION & FILTERING
Variable-Width Attributes (RECURSIVE CTES):
 directors = "nm0001.nm0002.nm0003.....nm0150"
                                                                        Defensive WHERE Clauses:
    WITH RECURSIVE DirectorSplit AS (
                                                                            • Filter NULL references before insertion
     -- Base case: Extract first director
                                                                            • Validate foreign key relationships
                                                                            • Remove orphaned records (e.g., episodes without parents)
      -- Recursive case: Extract next director
     -- Repeat until exhausted
                                                                         Temporary Constraint Disabling:
                                                                            SET FOREIGN_KEY_CHECKS = 0;
    --> Row 1: titleID=1, directorID=nm0001
                                                                            SET UNIQUE_CHECKS = 0;
    --> Row 2: titleID=1, directorID=nm0002
                                                                            -- Bulk insert operations --
    --> Row 3: titleID=1, directorID=nm0003
                                                                            SET FOREIGN_KEY_CHECKS = 1;
    ... (up to 150+ rows for single title)
                                                                            SET UNIQUE_CHECKS = 1;
  ⚠ Bottleneck: 30% of transformation time
  	riangle Recursion depth: Can exceed 1,000 iterations
 ♠ Config: cte_max_recursion_depth = 5000
[2] DATA TYPE CONVERSIONS
Boolean Flags:
 isAdult = '1' --> TRUE (BOOLEAN)
 isAdult = '0' --> FALSE (BOOLEAN)
NULL Markers:
  '\N' --> NULL (handled by LOAD DATA INFILE)
Numerical Precision:
 averageRating = "7.5" --> 7.5 (DECIMAL(3,1))
numVotes = "12345" --> 12345 (INT)
revenue = "12345678" --> 12345678.00 (DECIMAL(12,2))
[3] SURROGATE KEY GENERATION
Natural Keys (IMDb IDs):
                                  Surrogate Keys (Performance):
 tconst = "tt0123456"
                              --> titleID = 1 (AUTO_INCREMENT)
  nconst = "nm0987654"
                              --> personID = 1 (AUTO_INCREMENT)
  • 4-byte integer vs 20-character string (join performance)
  • Immutable (natural keys may change)
  • Sequential allocation (cache-friendly)
 • Natural keys stored as UNIQUE columns
  • Enables source data traceability
```

Supports incremental ETL

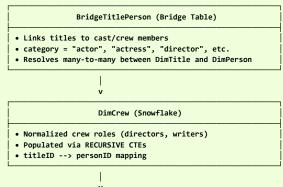
Load

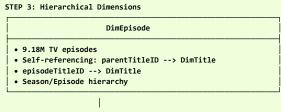
LOADING ORDER (Dependency-Aware):

STEP 1: Independent Dimensions



STEP 2: Bridge Tables (Many-to-Many Relationships)





STEP 4: Time Dimension

| DimDate | |
|---|--|
| On-demand population Year, Quarter, Month, Day hierarchy Supports temporal roll-up/drill-down | |
| l | |

STEP 5: Fact Tables (Transaction Grain)

| FactRatingSnapshot | FactBoxOfficeRevenue |
|----------------------------|----------------------------|
| • Periodic Snapshot | • Transaction Fact Table |
| • Grain: One row per title | • Grain: One row per title |
| per snapshot date | per revenue date |
| Foreign Keys: | Foreign Keys: |
| - titleID> DimTitle | - titleID> DimTitle |
| - snapshotDateID> DimDate | - dateID> DimDate |
| Measures: | Measures: |
| - averageRating (DECIMAL) | - revenue (DECIMAL) |
| - numVotes (INT) | - rank (INT) |