Guía

October 19, 2025

1 Guía y reflexiones

Objetivo: dejar un paquete "listo para producción" que permita analizar el catálogo y desempeño (IMDb) de títulos, directores, actores y países, con calidad de datos trazable y un dashboard ejecutivo en **Tableau Public**.

1.1 1) Resumen ejecutivo (qué logramos)

- **Modelo analítico consistente**: esquema estrella con **dimensiones** de Título, Persona y País; **hechos** de ratings; puentes M:N para Personas y Países.
- **Calidad medida y gobernada**: sin duplicados exactos; validaciones de rango, coherencias (p. ej., `seasons` solo en `SHOW`), y **gates** en ETL.
- **Imputación responsable**: no forzamos temporadas; **cruce por `title_norm` + `release_year`** cuando hay datos en otras fuentes; si no, **mediana por género** para `imdb_score`/`imdb_votes`.
- **Métricas definidas y reutilizables**: conteos únicos, promedios, **rating ponderado bayesiano** y segmentaciones (tipo, género, país, año).
- **Dashboard** en Tableau: selector `MOVIE/SHOW`, **Top 5 géneros**, matriz género×año, **Top 5 & emergentes** (directores/actores), **donut de clasificación IMDb** (etiquetas traducidas), **mapa** por país y tarjetas clave.

1.2 2) Alcance y fuentes

- **Fuentes base** (CSV): `dim_title.csv`, `dim_person.csv`, `dim_country.csv`, `fact_ratings.csv`, `bridge_title_person.csv`, `bridge_title_country.csv`.
- **Cobertura temporal**: últimos 20 años (filtro listo para usar en las vistas; el modelo conserva histórico).
- **Limitaciones conocidas**: `date_added` no disponible; parte del `imdb_id` ausente; `age_certification` incompleto.
- **Lineaje**: EDA → ETL (v3 countries) → salidas dimensionales → Tableau.

1.3 3) Hallazgos del EDA (lo que importa para el ETL)

- **Integridad**: 0 duplicados exactos por archivo; llaves `title + release_year` con **>99%** de cobertura.

- **Faltantes relevantes**: `imdb_score` y `imdb_votes` ~**9%** cada uno; `imdb_id` ~**7–8%**; `age_certification` alta ausencia.
- **No hubo match cruzado** (título+año) suficiente para completar `imdb_score/imdb_votes` desde auxiliares → **se justifica el fallback por género**.
- **Distribución**: `imdb_votes` con cola pesada (necesita log/winsor); **SHOW** suele puntuar un poco mejor que **MOVIE**; outliers manejables.
- **Coherencias**: `seasons` **solo** para `SHOW`; patrón `imdb_id` tipo `tt\d+`.

  Implicación directa: el **ETL** debe (a) conservar el valor original, (b) intentar cruce por `title_norm + release_year`, y (c) si falla, imputar por **mediana del género**. `seasons_final` **no se imputa**.

1.4 4) Modelo de datos (estrella)

Hecho

- `fact_ratings`
  - Claves: `title_id` (FK → `dim_title`).
  - Métricas: `imdb_score_original`, `imdb_votes_original`, `imdb_score_final`, `imdb_votes_final`, `rating_ponderado`, etc.

Dimensiones

- `dim_title` (`title_id`, `imdb_id`, `title`, `title_norm`, `type` {MOVIE, SHOW}, `release_year`, `main_genre`, `age_certification`, `runtime_final`, `seasons_final`).
- `dim_person` (`person_id`, `name`).
- `dim_country` (`country_id`, `country_code`).

Puentes M:N

- `bridge_title_person` (`title_id`, `person_id`, `role`).
- `bridge_title_country` (`title_id`, `country_id`).

Notas de modelado

- Las tablas **bridge** evitan duplicar títulos/personas y permiten **contar distintos** en BI.
- Reglas de conteo:
  - **Títulos por director**: `COUNTD(title_id)` filtrando `role="DIRECTOR"`.
  - **Títulos por actor**: `COUNTD(title_id)` filtrando `role="ACTOR"`.
  - Al cruzar con países: usar `COUNTD(title_id)` para no inflar por coproducciones.

1.5 5) Decisiones de ETL y *quality gates*

Limpieza & normalización

- `title_norm`: minúsculas, sin acentos, sin signos → mejor *join* por título+año.
- Tipificación numérica segura (`to_numeric(errors="coerce")`); patrón `imdb_id` válido.
- `character` vacío en actores → **"desconocido"** (no se pierde el registro).

- **No eliminar** `imdb_id` en actores/títulos (necesario para BI y trazabilidad).

Imputación

1. **Cruce**: si existe la misma (título_norm, año) en otro dataset con `score/votes` → usar ese valor.
2. **Fallback**: si no hay donante, **mediana por** `main_genre`.
3. `seasons_final` = `seasons` (sin imputar; coherente con `type=SHOW`).

Quality gates (bloquean el pipeline)

- `imdb_score_final` fuera de [0..10] → **fail**.
- `runtime_final` fuera de [1..600] → **warn/fix** (cap/winsor).
- `seasons_final` no nulo con `type` `SHOW` → **fail**.
- `release_year` en futuro → **fail**.

Performance

- Categorización para `type`, `main_genre`, `country_code` → menor memoria.
- *Seeds* y logs para reproducibilidad.

1.6 6) Métricas y KPIs (definiciones)

- **Títulos**: `COUNTD(title_id)`.
- **Prom IMDb**: `AVG(imdb_score_final)`.
- **Votos IMDb**: `SUM(imdb_votes_final)`.
- **Rating ponderado (Bayes)**, evitando sesgo por pocos votos:

  ```
  rating_ponderado = (v / (v + m)) * s + (m / (v + m)) * C
  ```

  - `s`: `imdb_score_final` del título
  - `v`: `imdb_votes_final` del título
  - `C`: promedio global de `imdb_score_final` (o por filtro actual)
  - `m`: umbral de estabilización (p. ej., p90 de votos o un valor de negocio)

  En Tableau: implementarlo como calculadas anidadas todas a nivel de **fila del título** o como **LOD FIXED** para `C`/`m`.

1.7 7) Dashboard — diseño y uso

Controles

- **Botones MOVIE/SHOW** (parámetro + acción o filtro de dimensión).
- Filtros: Años (últimos 20 por defecto), Género, País.

Vistas

1. **Top 5 géneros** (barra horizontal, `COUNTD(title_id)`).
2. **Matriz género × año** (círculos; tamaño por `COUNTD(title_id)` y color por tendencia).

3. **Top 5 directores** y **emergentes (6–10)**
    - Cálculo: `COUNTD(title_id)` filtrando `role="DIRECTOR"`, ordenar y limitar (Index/Rank).
4. **Top 5 actores** y **emergentes (6–10)** (igual, `role="ACTOR"`).
5. **Donut IMDb** por `age_certification`
    - **Traducciones** (ejemplos): `G` Apta para todo público · `PG` Guía paterna · `PG-13` 13+ · `R` Restringida (menores de 17 con adulto) · `NC-17` Solo adultos · `TV-Y` Niños · `TV-G` General · `TV-PG` Guía paterna · `TV-14` 14+ · `TV-MA` Adultos (17+) · `NR/UR` Sin clasificación.
    - Mostrar %: formato de número **Porcentaje** con 0–1 decimales.
6. **Mapa** por país (de `dim_country` vía `bridge_title_country`), medida `COUNTD(title_id)`.

Interacción

- Clic en **Top 5 géneros** filtra el resto; botones `MOVIE/SHOW` influyen en todas las vistas; tooltip con `rating_ponderado`, votos y año.

1.8 8) Riesgos, supuestos y mitigación

- **M:N y doble conteo**: siempre usar `COUNTD(title_id)` al cruzar Personas y Países.
- **Valores faltantes**: la mediana por género evita sesgo; documentar tasa de imputación.
- **Sesgo por pocos votos**: usar `rating_ponderado` en listados "Top".
- **Campos no disponibles**: `date_added` fuera de alcance; si se incorpora, añadir **tabla de fechas** y `snapshot`.

1.9 9) Recomendaciones de siguiente iteración

1. **Enriquecer** `imdb_id` y `age_certification` con fuentes externas (OMDb/IMDb datasets).
2. **Matching difuso** (fuzzy) para mejorar el cruce por título.
3. **Segmentar** `m` (umbral bayesiano) por tipo o género.
4. **Pruebas automatizadas** del ETL (pytest + great_expectations) y **CI** en Git.
5. **Semántica común** (definiciones de KPI) en un **Diccionario de Métricas** compartido.

1.10 10) Glosario (variables clave del ETL)

- **`imdb_score_final`**: puntuación consolidada (original, o imputada por **mediana del género** si faltante).
- **`imdb_votes_final`**: votos consolidados (original, o **mediana del género** si faltante).
- **`rating_ponderado`**: score estabilizado mediante promedio global `C` y umbral `m`.
- **`title_norm`**: normalización de `title` (minúsculas, sin acentos/ruido) para cruces.
- **`main_genre`**: género dominante del título (base para imputación segmentada).
- **`seasons_final`**: temporadas; **no imputada**; solo aplica a `SHOW`.
- **`bridge_title_person`**: puente M:N entre título y persona (roles: ACTOR/DIRECTOR/…).

- `bridge_title_country`: puente M:N entre título y país (coproducciones).

1.11 Reflexiones finales

El proyecto expone un balance entre **rigor analítico** y **practicidad de negocio**. Del lado de datos, priorizamos una imputación **parsimoniosa**: primero intentamos recuperar información por cruce (`title_norm` + `release_year`) y solo en ausencia de donantes aplicamos **mediana por género**. Esta decisión limita el sesgo y mantiene la interpretabilidad, dos rasgos críticos cuando las métricas alimentan decisiones visibles (e.g., rankings de contenido, mejores directores/actores).

En modelado, la elección de un **esquema estrella** con puentes M:N, más el uso sistemático de `COUNTD(title_id)`, reduce el riesgo de **doble conteo** al cruzar personas y países. La experiencia mostró que pequeños desalineamientos en granularidad (episodios vs. obras) pueden inflar indicadores; por ello consolidamos claves (`imdb_id`, `title_id`, `work_key` cuando aplique) y verificamos con hojas de control en BI. Esta disciplina —sumada a *quality gates* explícitos— fortalece la **confiabilidad del insight** y habilita auditoría.

Del lado de consumo, evitamos sobre-ingeniería: el dashboard comunica con **pocas vistas potentes** (matriz género×año, Top & Rising, donut de clasificación, mapa) y controles mínimos (botones `MOVIE/SHOW`, filtros de año/género/país). El objetivo es que un usuario de negocio **entienda y actúe** en segundos, sin leer un manual. Cuando haya que profundizar, el **rating ponderado** ofrece una métrica robusta para comparaciones "justas" frente a volúmenes de votos muy dispares.

Finalmente, dejamos un camino claro de **evolución**: enriquecer `imdb_id` y `age_certification`, introducir *matching* difuso, y automatizar validaciones con *data tests*. Esto permite escalar el caso de uso (más catálogos, más mercados) manteniendo la **coherencia semántica** y la **trazabilidad.**