Guía Exposición TP7

Presentación: [Click aquí para ver la presentación]

Introducción y Etapas 1-2

Objetivo del Proyecto

Analizar un dataset real de películas de Hollywood para responder preguntas de negocio:

- ¿Qué géneros son más rentables?
- ¿Existe relación entre calidad (score) y éxito comercial?
- ¿Conviene invertir mucho o poco presupuesto?
- ¿Cómo evolucionó la industria del cine en 40 años?

¿Por qué usamos estas tecnologías?

- **Apache Spark**: Framework que procesa grandes volúmenes de datos (big data) de forma distribuida (procesamiento paralelo).
- **Databricks**: Plataforma en la nube que facilita el uso de Spark. Brinda un entorno tipo "notebook" para escribir código, visualizar datos y compartir resultados. Nos evitó tener que instalar Spark localmente.
- **PySpark**: API de Python que permite usar Spark (lenguaje más accesible). Pudimos escribir código de Big Data usando Python

Etapa 1: Preparación del entorno

Configuramos Databricks y verificamos que Spark funcionara.

Etapa 2: Carga y Exploración

¿De dónde vienen los datos? Dataset público de TMDb (The Movie Database) alojado en GitHub con información de películas de 1980 a 2020.

¿Qué cargamos?

- 7,668 películas
- 15 columnas: nombre, año, score, género, director, presupuesto, recaudación, clasificación, etc.

Exploramos antes de transformar para conocer los datos antes de trabajarlos:

- ¿Qué tipo de información tenemos?
- ¿Hay datos faltantes?
- ¿Qué rangos tienen los valores?

Hallazgos de la exploración:

- 14 géneros distintos (Drama, Comedy, Action, etc.)
- Drama es el más producido (1,103 películas)
- Score promedio general: 6.5/10
- El dataset abarca 40 años de cine

Etapa 3 - Transformaciones

3.1 Análisis de Calidad de los Datos y Tratamiento de nulos

Verificamos los valores nulos de cada columna. Valores nulos: Datos faltantes o vacíos. Por ejemplo, una película sin presupuesto registrado (cantidad por columna y % de completo).

¿Qué encontramos?

- **name, year, score (puntaje):** 3 películas no los tienen.
- **director, genre:** Casi completos. Solo 70 películas no tienen director o género registrado.
- **budget, gross (presupuesto y recaudación):** El 30% de las películas NO tiene datos financieros. Son 2,300 películas sin presupuesto o recaudación registrada.

3.2 Columnas calculadas

¿Qué hicimos con los nulos y POR QUÉ?

1. **Columnas críticas (name, year, score)**: Eliminar las 3 películas sin estos datos.
2. **Budget y gross (~30% nulos)**: NO los eliminamos porque:
    - Perderíamos demasiados datos (2,300 películas)
    - Estos nulos no afectan otros análisis (géneros, scores)

¿Por qué eliminar duplicados? Evitamos contar la misma película dos veces, lo que falsearía estadísticas.

Resultado: 7,665 películas limpias (99.96% de integridad)

3.3 Columnas Calculadas

¿Por qué crear nuevas columnas? Los datos originales no responden directamente nuestras preguntas. Necesitamos métricas derivadas.

¿Qué columnas creamos?

💰 **Financieras:**

- **profit**: recaudación - presupuesto → Ganancia neta
- **roi_percentage**: (ganancia / presupuesto) × 100 → Rentabilidad en %
    - Ejemplo: Si costó $10M y ganó $50M → ROI = 400% ✅
    - Ejemplo: Si costó $100M y ganó $50M → ROI = -50% ❌
- **budget_millions / gross_millions**: Convertir a millones (legibilidad)

🏷️ **Categorías:**

- categoria_score: ≥8.0 = Excelente | ≥7.0 = Muy buena | ≥6.0 = Buena | <6.0 = Regular/Mala
- categoria_recaudacion: ≥$500M = Blockbuster | ≥$100M = Muy exitosa | <$10M = Baja

¿Por qué el ROI es tan importante? Una película que recauda $50M parece exitosa, pero:

- Si costó $10M → ROI = 400% ✅ (muy rentable)
- Si costó $100M → ROI = -50% ❌ (pérdida)

El ROI nos dice la eficiencia de la inversión.

3.4 Filtros

¿Por qué filtramos? Para responder preguntas específicas.

Filtro 1: Películas exitosas (buena calificación Y alta recaudación)

- score > 7.5 Y gross > $100M
- Resultado: 248 películas

Filtro 2: Películas modernas

- año ≥ 2000
- Para analizar tendencias recientes

Filtro 3: Blockbusters rentables (alta recaudación Y alto ROI)

- gross > $500M Y roi_percentage > 100%
- Resultado: 18 películas

3.5 Agregaciones

Agregación: Agrupar datos y calcular estadísticas sobre esos grupos.

Antes de agregar: Teníamos 7,665 filas (una por película)
Después de agregar por género: Tenemos 14 filas (una película por cada género)

¿Por qué agregamos (groupBy)?
Para encontrar patrones por categorías:

- ¿Qué género recauda más en total?
- ¿Qué director es más exitoso?
- ¿Cómo varía el score por década?

Agregaciones realizadas:

Agregación 1: Estadísticas por género

- groupBy("genre")
- Calculamos: cantidad, score promedio, recaudación total
- Hallazgo: Animation más taquillero ($15,843M), Biography mejor score (7.1/10)

Agregación 2: Ranking de directores (con ≥3 películas)

- groupBy("director")
- Calculamos: cantidad de películas, recaudación total, ROI promedio
- Hallazgo: Steven Spielberg más exitoso ($4,223M)

Agregación 3: Tendencias anuales

- groupBy("year")
- Calculamos: cantidad por año, recaudación anual, score promedio
- Para ver evolución de la industria

Agregación 4: Análisis por clasificación (PG, PG-13, R, etc.)

- groupBy("rating")
- Calculamos: cantidad por clasificación, calificación promedio, recaudación
- Para entender audiencias

3.6 JOINS

¿Qué es un JOIN? Combinar dos tablas usando una columna común (como una llave).

¿Para qué lo usamos? Enriquecimos nuestros datos de películas con información contextual de géneros que NO estaba en el dataset original.

Información agregada:
Creamos una tabla auxiliar con:

- Presupuesto típico del género (Alto, Medio, Bajo)
- Público objetivo (Familiar, Adulto, Joven adulto)

Ejemplos:

- Action → Presupuesto Alto, Público Joven adulto
- Horror → Presupuesto Bajo, Público Joven adulto
- Animation → Presupuesto Alto, Público Familiar

¿Cómo lo hicimos?
df_completo = df_transformado.join(df_info_generos, on="genre", how="left")
Columna común: genre (género)

Resultado: 7,665 películas enriquecidas con contexto adicional que permitió análisis cruzados como:

- ¿Los géneros de presupuesto alto recaudan más?
- ¿Qué público objetivo genera más ingresos?

Etapas 4-5 Visualización

Etapa 4: Almacenamiento con Data Lake

Se buscó almacenar los nuevos datos del DataSet completo y transformado:

- 7,665 películas limpias
- 24 columnas:
    - 15 columnas originales (name, year, score, genre, etc.)
    - 9 columnas calculadas (profit, roi_percentage, budget_millions, etc.)
    - 2 columnas del JOIN (presupuesto_tipico, publico_objetivo)

Delta Lake: Formato de almacenamiento optimizado con:

- Versionado (time travel)
- Transacciones ACID
- Mejor rendimiento

¿Qué problema encontramos? Databricks Community Edition tiene DBFS deshabilitado → No podemos escribir en disco.

¿Cómo lo solucionamos? Usamos vistas temporales que guardan datos en memoria. Si bien no es permanente (sólo sesión), tiene funcionalidades SQL.

Etapa 5: Visualizaciones

5.1 - Preparación de Datos para Visualización

Usamos tablas

- Dataset 1: Preparar datos para visualizar los géneros más taquilleros
- Dataset 2: Analizar la evolución del cine desde 1990 hasta la actualidad

5.2 - Gráficos

Los gráficos comunican patrones que las tablas no muestran claramente.

- **Gráfico 1**: Score vs Recaudación - Sin correlación fuerte
- **Gráfico 2**: Evolución Temporal → Presupuestos y recaudación crecen juntos
- **Gráfico 3**: Top ROI
    - Paranormal Activity: 19,758%
    - Blair Witch Project: 10,931%
- **Gráfico 4**: Rating - R: 46% | PG-13: 35% | PG: 12%

Análisis estadístico y conclusiones

5.3 - Análisis de Estadísticas

Correlación: Es un número entre -1 y +1 que mide si dos cosas están relacionadas:

- **Cerca de +1**: Cuando una sube, la otra también sube
- **Cerca de 0**: No tienen relación
- **Cerca de -1**: Cuando una sube, la otra baja

¿Qué encontramos?

Variables	Correlaciones	Interpretación
budget ↔ gross	0.740 🟢	Muy fuerte: más presupuesto = más recaudación
votes ↔ gross	0.631 🟢	Fuerte: más popularidad = más recaudación
score ↔ gross	0.211 🔴	Débil: calidad NO garantiza éxito

Conclusión estadística clave: El éxito comercial depende del presupuesto y la popularidad, NO de la calidad artística.

Análisis de Outliers

Outliers: Valores extremos que se alejan del promedio (usamos regla de 3 desviaciones estándar).
¿Qué encontramos? 23 películas con recaudación >$478M (extremadamente taquilleras):

- Avengers Endgame: $858M
- Avatar: $760M
- Star Wars VII: $936M

¿Por qué importa? Estas películas distorsionan los promedios. Por eso también analizamos medianas.

Rentabilidad del Presupuesto (Análisis más importante del proyecto)

Rango de Presupuesto	Cantidad de Películas	ROI	Recaudación
Bajo (<$10M)	1,237	1,933% 🟢	$18.93M
Medio ($10-50M)	2,948	168%	$61.06M
Alto ($50-100M)	811	157%	$175.48M
Muy Alto (>$100M)	440	215% 🟢	$486.83M

¿Qué significa esto?
Estrategia A: Bajo presupuesto

- ✅ Menos riesgo, más rentable porcentualmente
- ❌ Menor ganancia absoluta

Estrategia B: Alto presupuesto

- ✅ Mayor ganancia absoluta
- ❌ Más riesgo, menos eficiente

CONCLUSIONES FINALES

Técnicas:

1. Spark es eficaz para análisis de Big Data
2. La limpieza de datos consume ~30% del tiempo pero es fundamental
3. Las vistas temporales son alternativa funcional a Delta Lake

De Negocio:

1. Calidad ≠ Éxito comercial
   - Correlación score-recaudación: solo 0.211
   - Películas "malas" pueden ser muy taquilleras
   - El marketing importa más que las críticas
2. El presupuesto predice recaudación
   - Correlación: 0.740
   - Pero mayor presupuesto = menor eficiencia (ROI)
3. Horror es el género más inteligente
   - Bajo costo de producción
   - ROI promedio: 278%
   - Ejemplos: Paranormal Activity, Blair Witch
4. Animation es el rey de la taquilla
   - Alto presupuesto pero alta recaudación
   - Público familiar = múltiples entradas (padres + hijos)

Recomendaciones para un estudio:

- Maximizar recaudación: Animation/Action con >$100M
- Maximizar ROI: Horror/Thriller con <$10M
- Estrategia óptima: Diversificar portafolio

Cierre:

"El cine es un negocio donde el presupuesto y la popularidad determinan el éxito, no la calidad artística. Las películas de bajo presupuesto son las más rentables, pero los blockbusters dominan la taquilla."