Evaluación 2 - Big Data

**Internet Movie Database (IMDb):**

Gestión de Grandes Volúmenes de Datos

línea horizontal

# Marcador de posición de imagen

**Integrantes**:

Alex Pedreros - Daniel Vásquez - Gabriel Jiménez

**Sección:**

001D

# 

**Índice**

[**Introducción 3**](#_arolcxe0i15c)

[**Contexto del caso semestral 4**](#_la5jp5tnimjw)

[**Análisis de las fuentes de datos 4**](#_jkb9rzs2c6g0)

[**Desarrollo del caso 6**](#_hvbrahwo9vmm)

[Cloud Storage 7](#_i63rjv9iwxtc)

[Dataprep 8](#_rm23qa48mnom)

[BigQuery 9](#_iheyqlgi09rl)

[Looker Studio 10](#_imnrpiwq6q7d)

[**Reportes de Actividades Desarrolladas 11**](#_wag7k760sidn)

[Actividad 1: 11](#_r75y7hmqg5g6)

[Actividad 2: 12](#_mb6mo8kzxvrj)

[Actividad 3: 13](#_mc8g4r57c8br)

[Actividad 4: 14](#_vv3o8qtzvvz7)

[Actividad 5: 15](#_owwg9cz44qup)

[Actividad 6: 16](#_xyekvqlxrpym)

[Actividad 7: 17](#_ocrtkwsqzb7a)

[Actividad 8: 19](#_qnj8xhqa1hcr)

[Actividad 9: 20](#_fb5zpbkn59ut)

[Actividad 10: 20](#_owobv8814vbc)

[**Conclusión 22**](#_n5hm1mmh7069)

# Introducción

En este siguiente informe se presentará la utilización de herramientas de Big Data, con el objetivo de ejecutar tareas de gestión de grandes volúmenes de datos en formato Batch para facilitar su análisis y comprensión

En este caso se ocupará un conjunto de fuentes de datos de IMDb, la cuales una base de datos en constante expansión, que recopila información de una variedad de fuentes, incluyendo estudios cinematográficos, redes de televisión, festivales de cine, distribuidores de medios y, en algunos casos, los propios usuarios. La plataforma tiene como objetivo proporcionar información detallada y precisa sobre cada producción, incluyendo su título, fecha de lanzamiento, duración, género, país de origen, elenco y equipo, sinopsis y calificaciones de los usuarios. Además, IMDb ofrece una variedad de características adicionales, como noticias de la industria del entretenimiento, trailers, reseñas y listas de las películas y programas de televisión más populares.

Dentro de este informe se detalla cómo fueron utilizados los datos y fueron analizados en cada proceso y con las herramientas que se encargaban de hacer limpieza de los datos, además de realizar reportes en base a los datos ya en limpio.

Las herramientas empleadas fueron principalmente de GCP (Google Cloud Platform) la cual tendrá las herramientas necesarias para trabajar los datos principalmente ocupando datos para almacenar, limpieza, selección de información específica y visualización.

# Contexto del caso semestral

Como se mencionó anteriormente IMDb es una plataforma en línea esencial para cualquier persona interesada en la industria del entretenimiento. Al proporcionar información detallada sobre películas, programas de televisión y personas relacionadas con la industria, la plataforma se ha convertido en una herramienta indispensable para fans,periodistas, profesionales de la industria y analistas de datos. Su importancia radica en su capacidad para proporcionar información precisa y exhaustiva, y en su papel en la promoción y evaluación del éxito.

Por lo que se busca diferentes datos, comportamientos, relaciones y información relevante para ocupar este dataset por lo que se realizaron 10 actividades la cual se utilizará estos datos para sacar información específica ocupando la herramienta de GCP como Cloud Storage, Dataprep, BigQuery y Looker Studio, que serán herramientas claves para utilizar en este proyecto.

# Análisis de las fuentes de datos

Los datos fueron entregados en un archivo .rar, por lo que al descomprimirlos dejaron una carpeta la cual contiene 7 archivos que tenían los datos que trabajaremos. Cada conjunto de datos está contenido en un archivo con formato de valores separados por tabuladores (TSV) en el conjunto de caracteres UTF-8. La primera línea de cada archivo contiene encabezados que describen lo que hay en cada columna. Un '\N' se utiliza para indicar que un campo en particular falta o es nulo para ese título/nombre.

En primera instancia se realizó una lectura de las fuentes analizando la información que disponíamos en cada una de las fuentes, identificando errores en las columnas y sus datos poblados.

Los datos una vez subidos al almacenamiento se procedieron a tratar para posteriormente realizar el análisis, por lo que optamos realizar las siguientes acciones para poder resolverlo:

* Reemplazar campos con ‘\N’ por vacío
* Eliminar espacios en blanco antes y después de un valor
* Conversión de columnas (de un tipo a otro según corresponda)
* Eliminar simbolos especificos (en las columnas que corresponda)
* Eliminar filas duplicadas

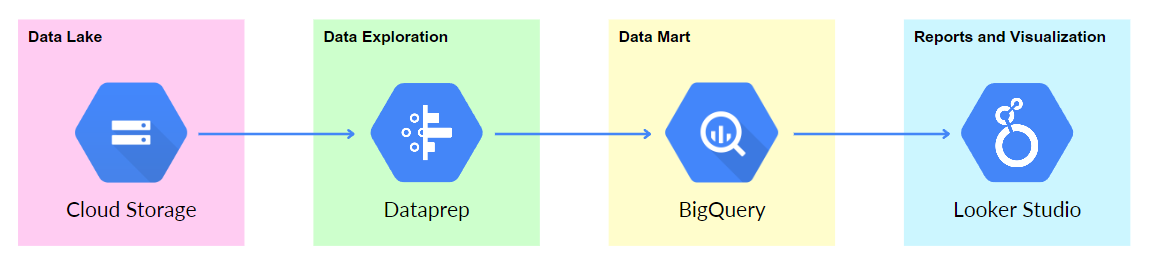
principalmente se busca las principales utilización de datos con la que nos adentraremos en el datasets para realizar nuestro análisis y preparar para el desarrollo de este caso:

Con estas actividades podremos desarrollar los datos y trabajarlos con los objetivos e información que necesitamos para realizar nuestro caso, utilizando estos datos bajo las reglas de Big Data.

# 

# Desarrollo del caso

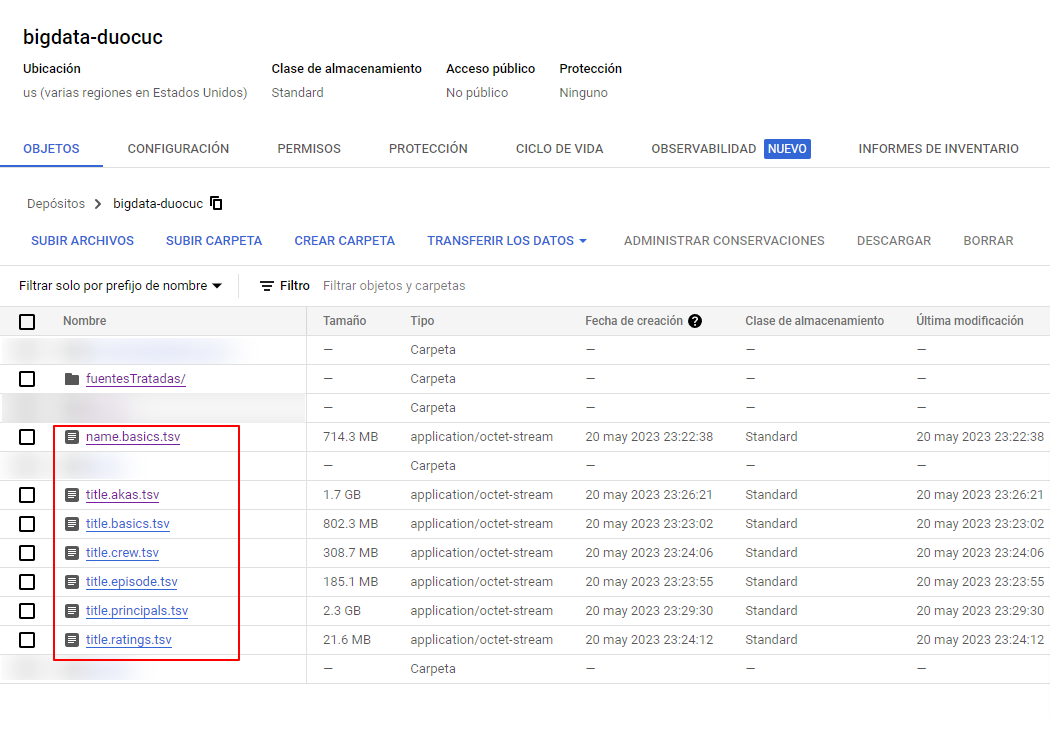
Como estamos trabajando con una gran cantidad de datos decidimos ocupar las herramientas de GCP debido a que proporciona una infraestructura altamente escalable, lo que significa que puedes manejar grandes volúmenes de datos sin problemas. Puedes escalar verticalmente o horizontalmente según las necesidades, lo que permite manejar eficientemente cualquier tamaño de conjunto de datos, además que tenemos las herramientas necesarias para almacenar los datos, limpiar datos nulos, seleccionar la información con la cual trabajaremos y con esto finalmente visualizarlos los datos, ocupando las herramientas ya anteriormente mencionadas



Como se puede observar en la imagen este es el proceso que se lleva a cabo con los datos, la primera herramienta de GCP, Cloud Storage nos permite almacenar los datos, siendo este servicio nuestro data lake, un repositorio central que contiene datos sin procesar y no estructurados, en la cual se almacenaran nuestros datos, después estos datos se llevan al servicio de dataprep que sirve para examinar, limpiar y preparar datos estructurados y no estructurados de forma visual para procesos de análisis, creación de informes y aprendizaje automático, de esta manera se puede eliminar los datos nulos y duplicados que se encuentran en los datasets, luego de transformar los datos, estos se llevan a BigQuery para poder seleccionar la información que necesitamos, esto lo realizamos por medio de consultas, como se sabe BigQuery está diseñado para analizar datos del orden de miles de millones de filas, utilizando una sintaxis similar a SQL, por lo que su uso es crucial para ayudar a tomar los datos necesarios para trabajarlos, por último utilizaremos Looker Studio la cual nos permitirá visualizar la información seleccionada en el anterior proceso, permitirá mostrar los datos de manera más visual, Looker Studio es una herramienta en línea para convertir datos en informes y paneles informativos personalizables, lo cual permite crear graficos para mostrar los datos de manera más amigable al momento de presentar la información.

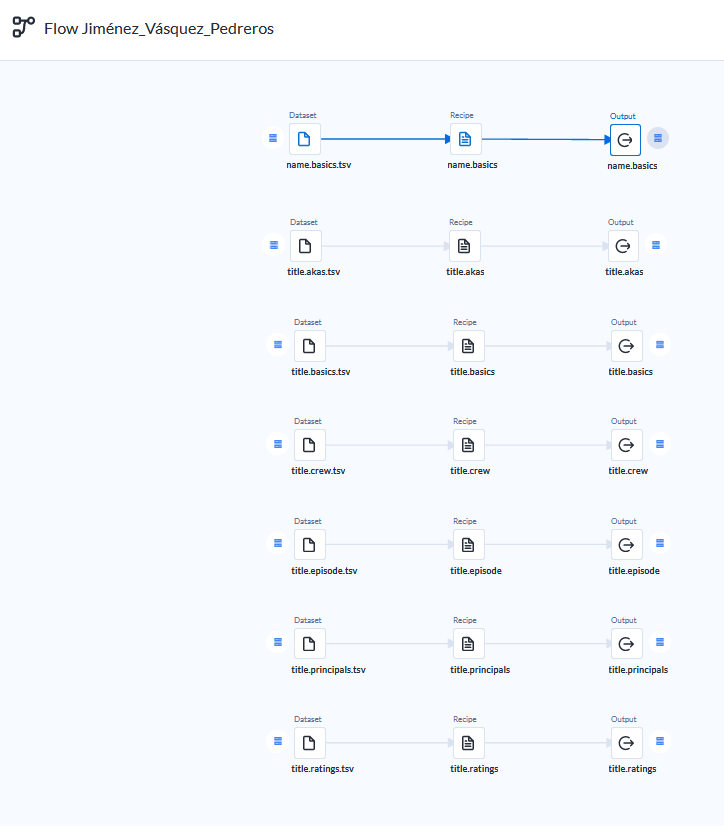
### 

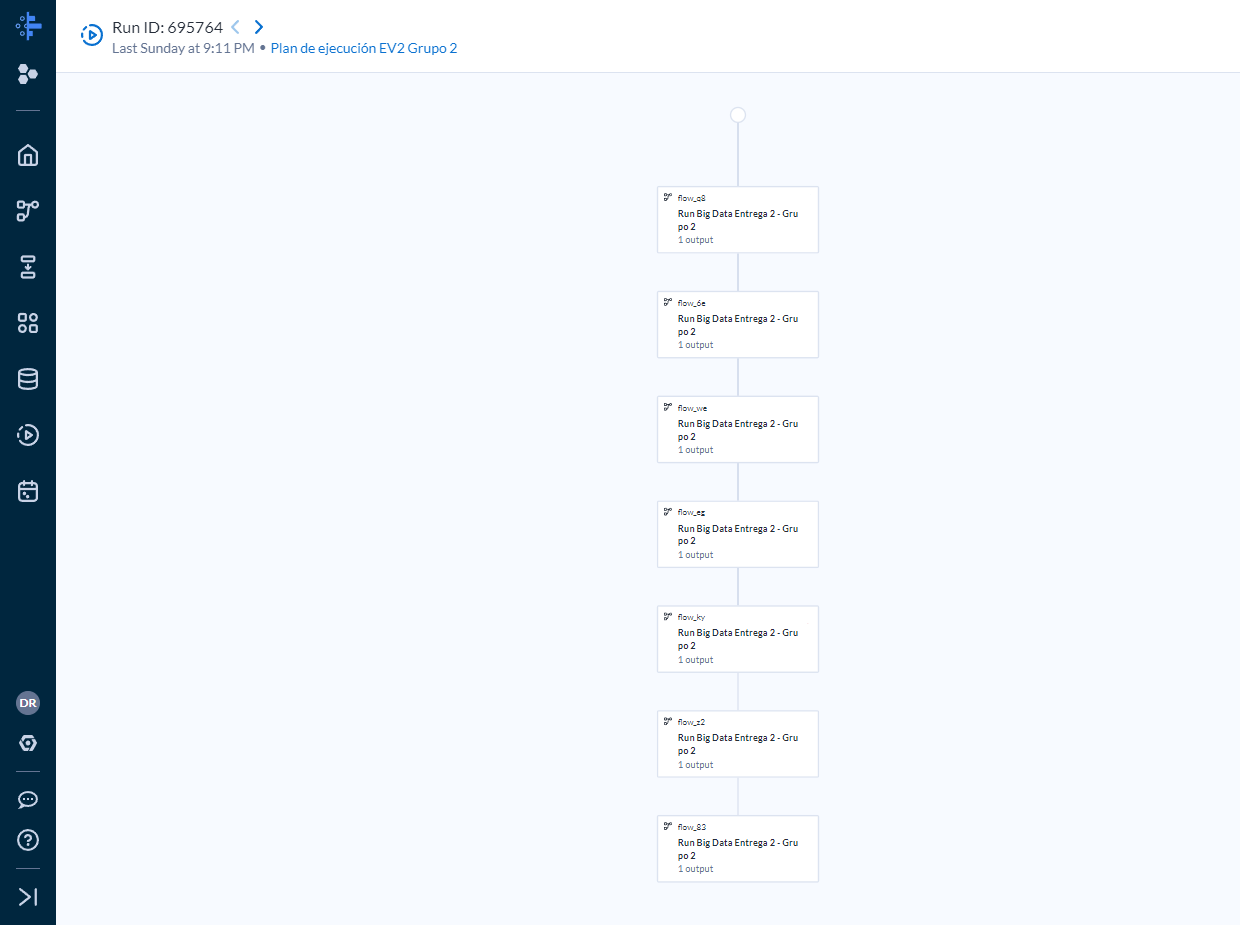
### Cloud Storage



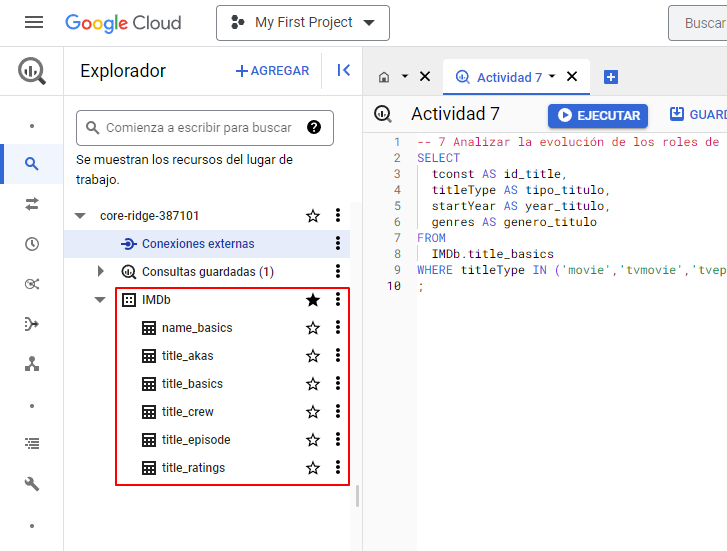
### 

### Dataprep





### BigQuery

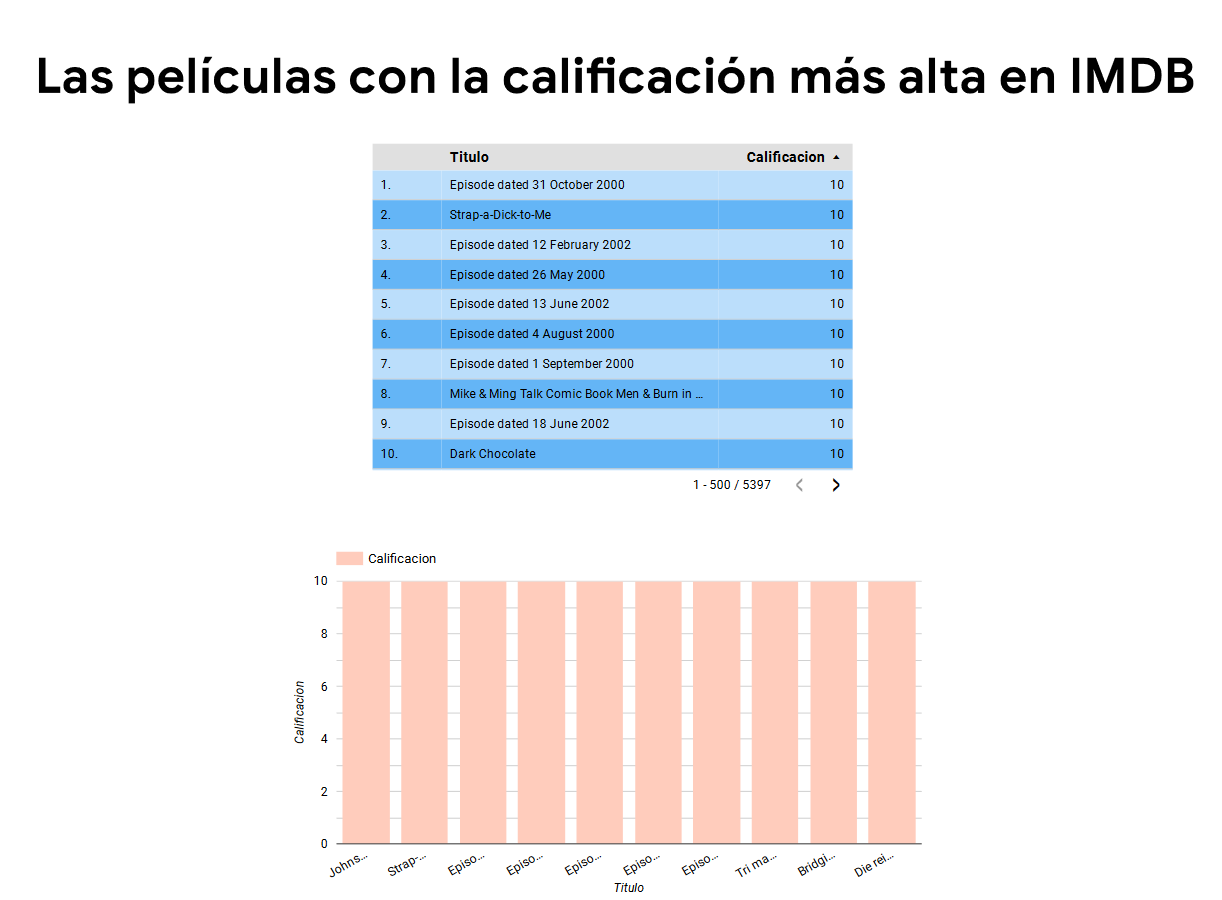


### Looker Studio

# Reportes de Actividades Desarrolladas

### Actividad 1:

*“Identificar las películas con la calificación más alta en IMDb.”*



**Código:**

SELECT titleType as Tipo, PrimaryTitle as Titulo, averageRating as Calificacion, numVotes as Votos

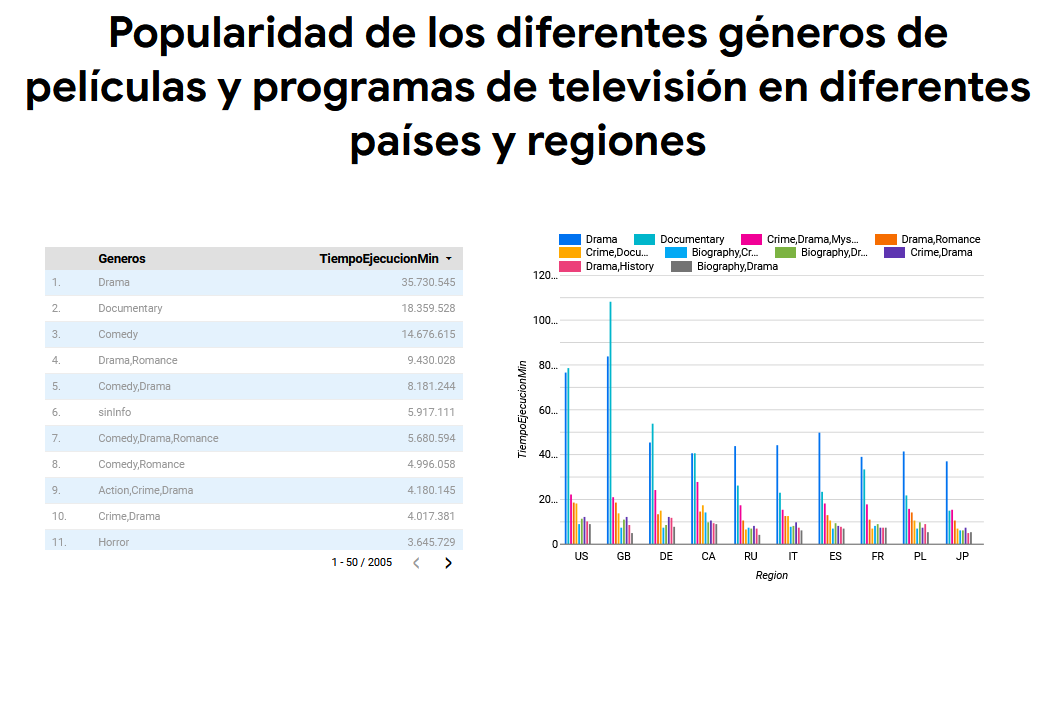
FROM d\_1.title\_basics

JOIN d\_1.title\_ratings ON title\_basics.tconst = title\_ratings.tconst

WHERE averageRating > 9.9;

### Actividad 2:

*“Analizar la popularidad de los diferentes géneros de películas y programas de televisión en diferentes países y regiones.”*



Análisis si corresponde

**Código:**

SELECT titleType as Formato,genres as Generos, region as Region, runTimeMinutes as TiempoEjecucionMin

FROM d\_1.title\_basics

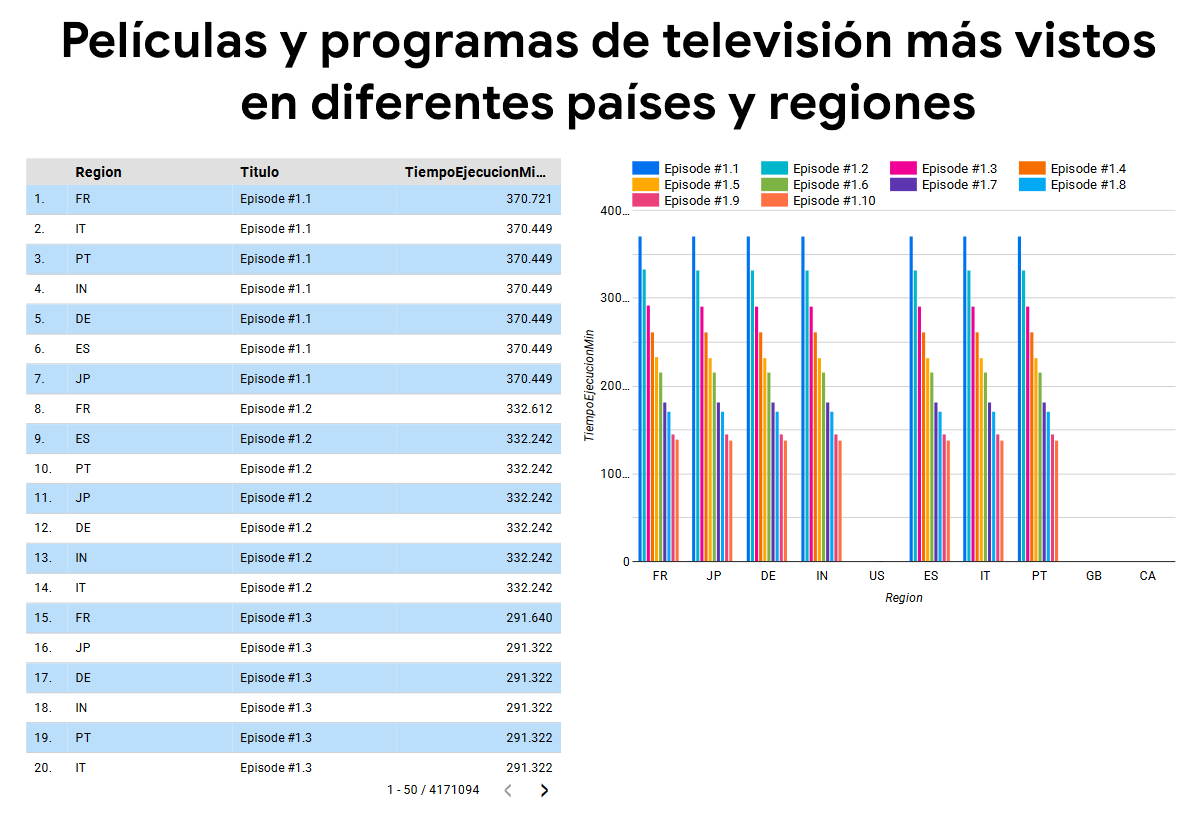
JOIN d\_1.title\_akas ON title\_basics.tconst = title\_akas.titleId

WHERE titleType = "movie" or titleType ="tvSeries" or titleType ="tvMiniSeries"

ORDER BY TiempoEjecucionMin DESC;

### Actividad 3:

*“Identificar las películas y programas de televisión más vistos en diferentes países y regiones.”*



**Código:**

SELECT region as Region, primaryTitle as Titulo, runtimeMinutes as TiempoEjecucionMin

FROM d\_1.title\_basics

JOIN d\_1.title\_akas ON title\_basics.tconst = title\_akas.titleId

WHERE titleType = "movie" or titleType ="tvSeries" or titleType ="tvMiniSeries" or titleType ="tvEpisode" or titleType ="tvMiniSeries" or titleType ="tvShort" or titleType ="tvSpecial"

ORDER BY TiempoEjecucionMin DESC;

### Actividad 4:

*“Identificar las actrices y actores que trabajan más frecuentemente en producciones.”*

(Incluir imagen del reporte en Looker)

**Código:**

SELECT

primaryName AS Nombre,

LENGTH(knownForTitles) - LENGTH(REPLACE(knownForTitles, ',', '')) + 1 AS Cantidad\_De\_Peliculas

FROM IMDb.name\_basics

WHERE primaryProfession IN ('actor','actress')

ORDER BY Cantidad\_De\_Peliculas DESC

;

### 

### 

### 

### 

### 

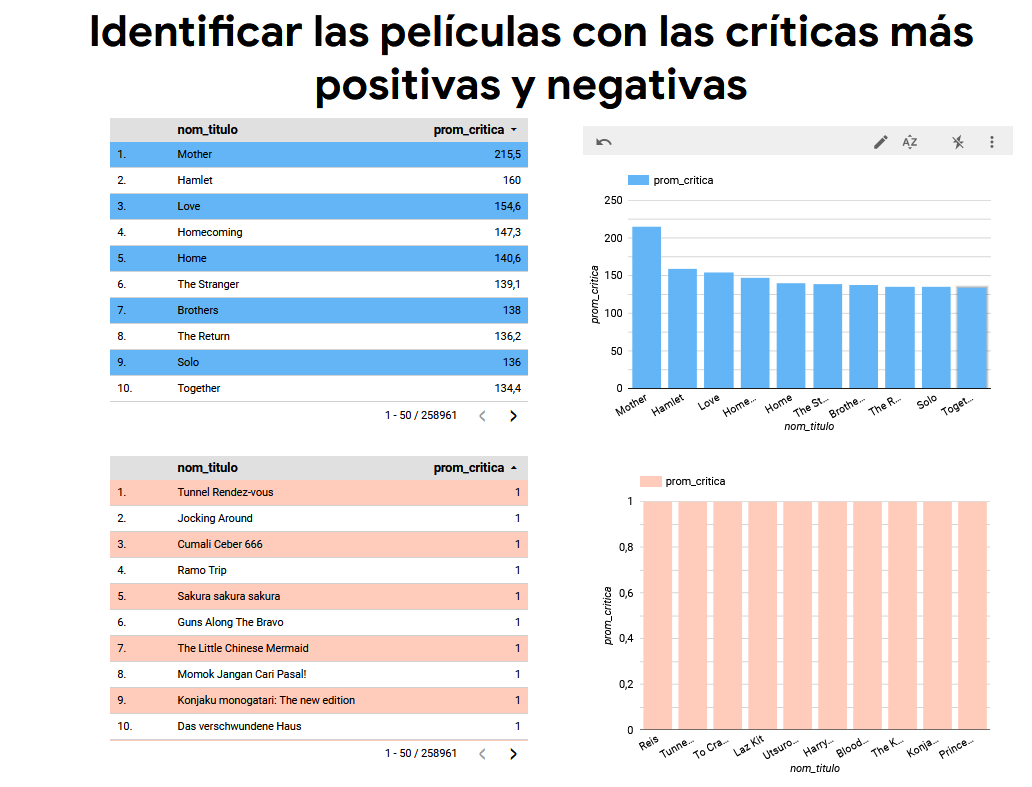
### 

### 

### 

### Actividad 5:

*“Identificar las películas con las críticas más positivas y negativas.”*



**Código:**

SELECT

TB.tconst AS id\_title,

TR.AverageRating AS prom\_critica,

TB.titleType AS tipo\_titulo,

TB.PrimaryTitle AS nom\_titulo

FROM

IMDb.title\_basics AS TB

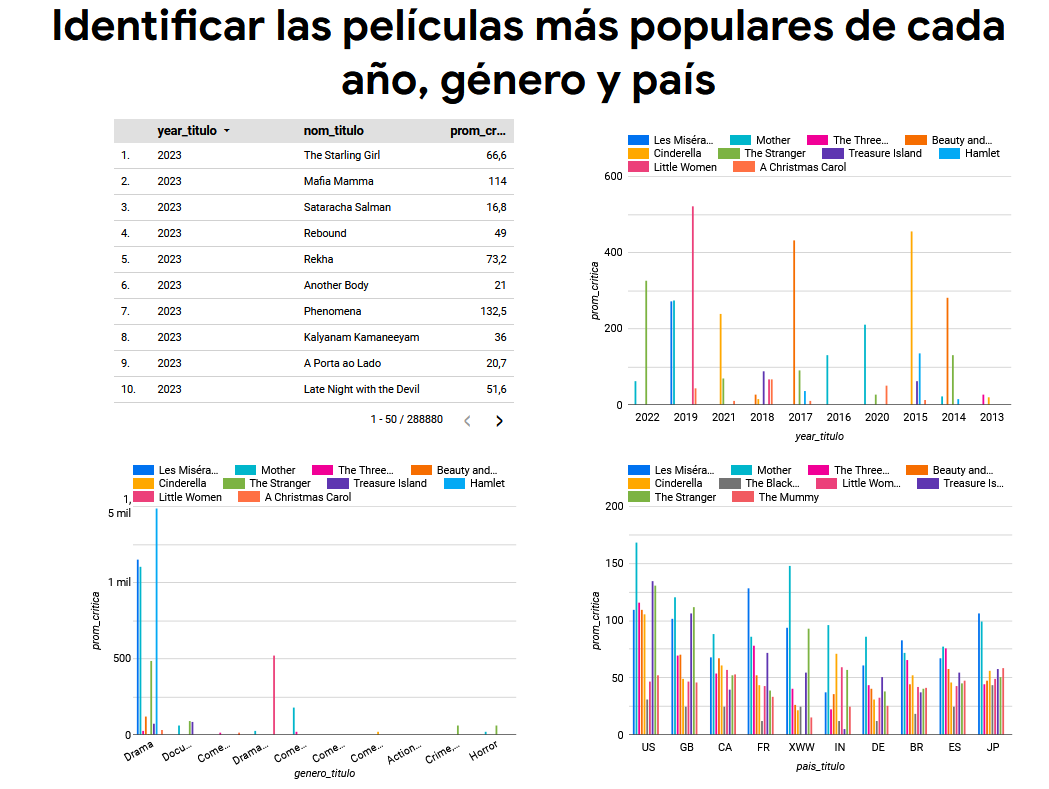
JOIN IMDb.title\_ratings AS TR ON TB.tconst = TR.tconst

WHERE titleType = "movie" OR titleType = "tvmovie"

;

### Actividad 6:

*“Identificar las películas más populares de cada año, género y país.”*



**Código:**

SELECT

TB.tconst AS id\_title,

TR.AverageRating AS prom\_critica,

TB.titleType AS tipo\_titulo,

TB.PrimaryTitle AS nom\_titulo,

TB.startYear AS year\_titulo,

TB.genres AS genero\_titulo,

TA.region AS pais\_titulo

FROM

IMDb.title\_basics AS TB

JOIN IMDb.title\_ratings AS TR ON TB.tconst = TR.tconst

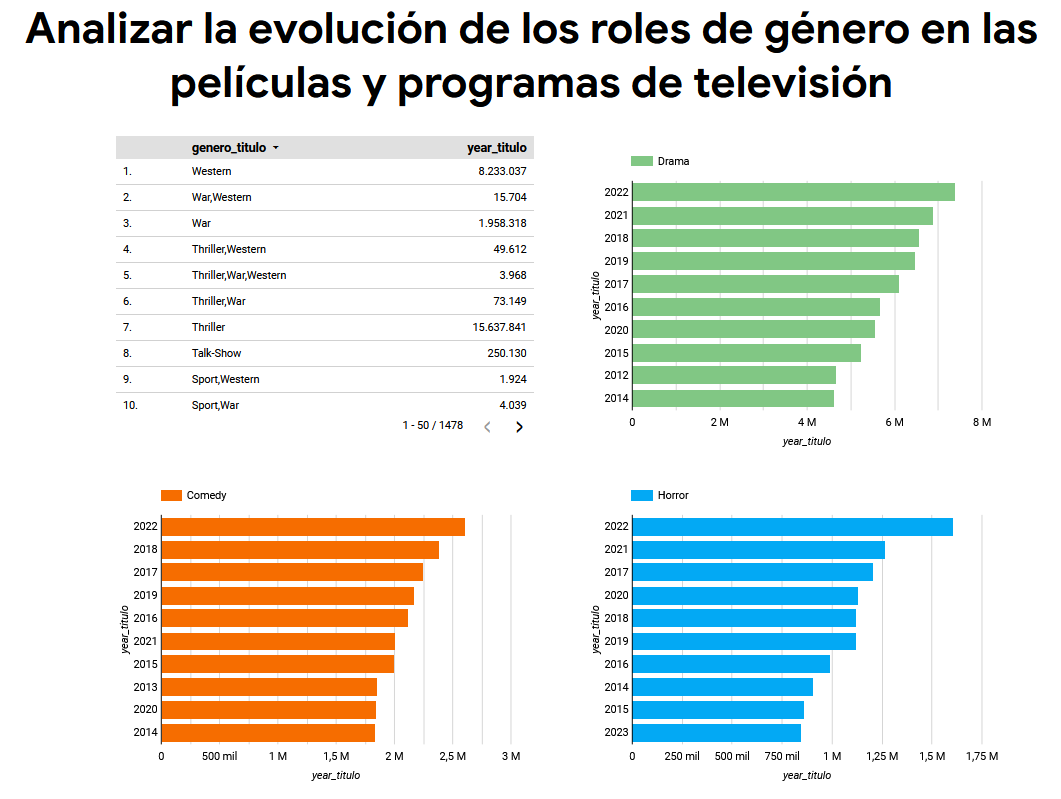
JOIN IMDb.title\_akas AS TA ON TA.titleId = TR.tconst

WHERE titleType = "movie" OR titleType = "tvmovie"

;

### Actividad 7:

*“Analizar la evolución de los roles de género en las películas y programas de televisión.”*



Análisis si corresponde

**Código:**

SELECT

tconst AS id\_title,

titleType AS tipo\_titulo,

startYear AS year\_titulo,

genres AS genero\_titulo

FROM

IMDb.title\_basics

WHERE titleType IN ('movie','tvmovie','tvepisode','tvseries','tvminiseries','tvspecial','tvshort')

;

### Actividad 8:

*“Analizar las calificaciones de IMDb para encontrar patrones de género y edad”*

(Incluir imagen del reporte en Looker)

Análisis si corresponde

Código:

SELECT

TB.tconst AS id\_title,

TR.averageRating AS prom\_calificacion,

TB.genres AS genero\_titulo,

NB.birthYear AS fecha\_nacimiento,

NB.deathYear AS fecha\_falle,

IF(NB.deathYear > 0, (NB.deathYear - NB.birthYear), (PARSE\_NUMERIC(LEFT(STRING(CURRENT\_DATE()), 4)) - NB.birthYear)) AS edad

FROM

IMDb.title\_basics AS TB

JOIN IMDb.title\_ratings AS TR ON TB.tconst = TR.tconst

JOIN IMDb.title\_principals AS TP ON TB.tconst = TP.tconst

JOIN IMDb.name\_basics AS NB ON NB.nconst = TP.nconst

;

### 

### Actividad 9:

*“Identificar las actrices y actores con mayor éxito en la industria del entretenimiento en cada época, género y país.”*

(Incluir imagen del reporte en Looker)

Código:

### Actividad 10:

*“Identificar los personajes más populares en las películas y programas de televisión.”*

(Incluir imagen del reporte en Looker)

Código:

SELECT

TB.tconst AS id\_title,

TB.titleType AS tipo\_titulo,

TP.characters AS personaje,

TR.averageRating AS prom\_calificacion

FROM

IMDb.title\_basics AS TB

JOIN IMDb.title\_principals AS TP ON TB.tconst = TP.tconst

JOIN IMDb.title\_ratings AS TR ON TR.tconst = TB.tconst

WHERE titleType IN ('movie','tvmovie','tvepisode','tvseries','tvminiseries','tvspecial','tvshort')

ORDER BY prom\_calificacion DESC

;

### 

# Conclusión

En conclusión, Las herramientas que permite GCP son variadas pero con funciones específicas lo cual permite trabajar diferentes datos con diferentes formatos, además que una de las grandes ventajas es que se puede trabajar de manera virtual, siendo accesible desde cualquier dispositivo, por lo que se puede trabajar los datos sin problemas en cualquier dispositivo, lo cual permite ser muy flexible para trabajarlo, se puede trabajar con bases de datos más grandes y de esta manera poder almacenarlos, limpiarlos, seleccionar la información que necesitamos y por último representarla de la manera más comprensible.

En resumen estas herramientas son cruciales para entender estos datos masivos o también llamados Big Data, ya que al ser muchos datos y hasta complejos es necesario tener una herramienta con el cual poder trabajarlos y comprenderlos de mejor manera, ya que esto permite facilitar el entendimiento y así poder trabajarlos, ya que es crucial para las personas que necesiten métricas o gráficos que les permitan tomar decisiones en base a muchos datos.