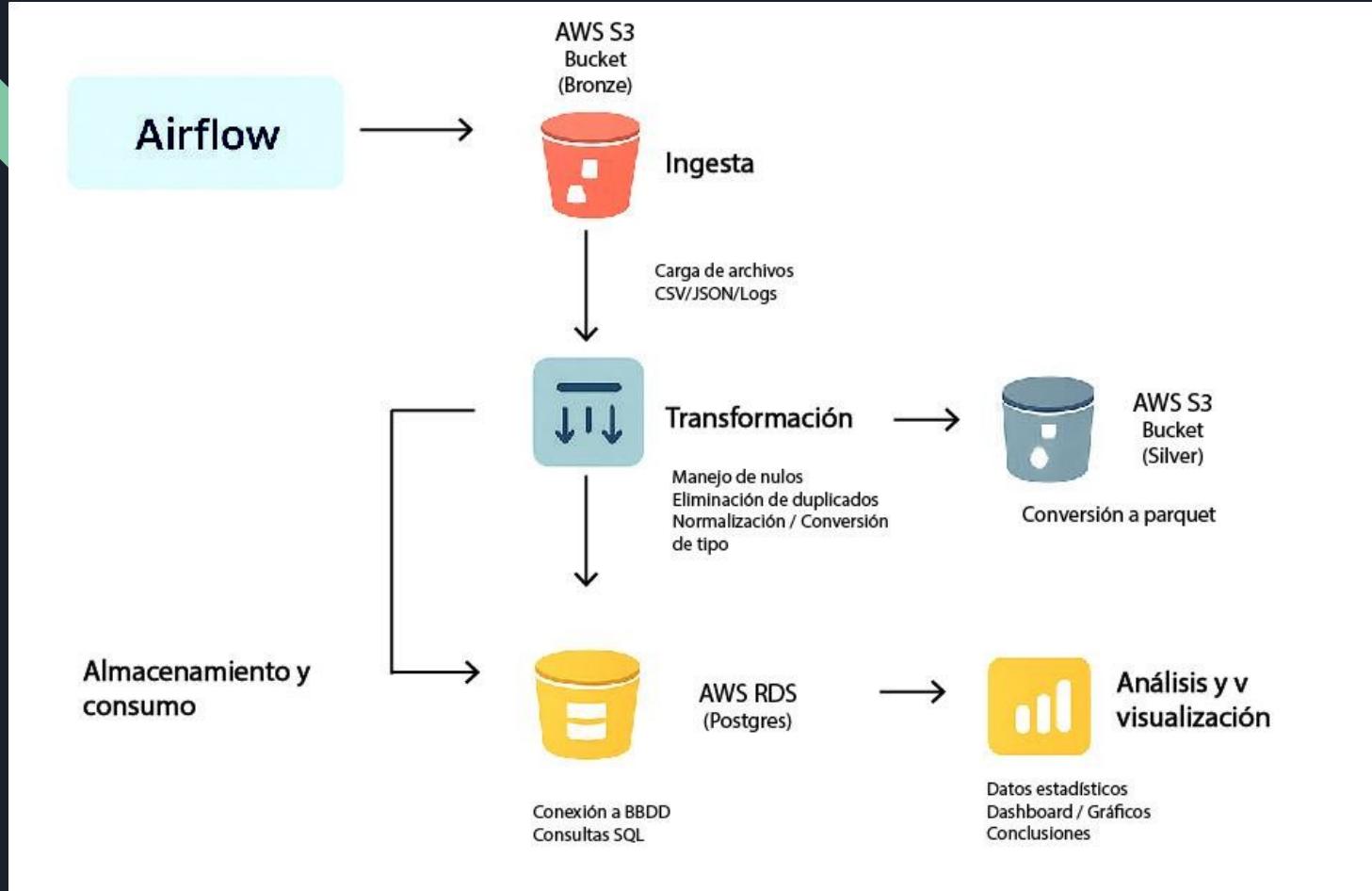




Extracción, Transformación y Carga de Datos para el Análisis de Contenido en una Plataforma de Streaming

Arquitectura Medallón



Amazon S3

General purpose buckets

Directory buckets

Table buckets

Vector buckets

Access Grants

Access Points (General Purpose Buckets, FSx file systems)

Access Points (Directory Buckets)

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

▼ Storage Lens

Dashboards

Storage Lens groups

bronze/

Copy S3 URI

Objects Properties

Objects (17)



Copy S3 URI

Copy URL

Download

Open

Delete

Actions ▾

Create folder

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

< 1 > |

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	content_genres.csv	csv	August 26, 2025, 12:02:01 (UTC-03:00)	1.2 MB	Standard
<input type="checkbox"/>	content_people.csv	csv	August 26, 2025, 12:02:03 (UTC-03:00)	6.8 MB	Standard
<input type="checkbox"/>	content.csv	csv	August 26, 2025, 12:02:01 (UTC-03:00)	4.9 MB	Standard
<input type="checkbox"/>	devices.csv	csv	August 26, 2025, 12:02:04 (UTC-03:00)	1.8 MB	Standard

Apache Airflow

- Plataforma de orquestación de flujos de trabajo
- DAG: Gráfico Acíclico Dirigido (Directed Acyclic Graph)
- Las tareas tienen una dirección y dependencias definidas.
- El flujo no puede tener bucles o ciclos infinito

The screenshot shows the Apache Airflow web interface with a dark theme. On the left is a vertical sidebar with icons for Home, Dags, Assets, Browse, Admin, and Security. The main area is titled "Dags" and contains a search bar, filter buttons for Failed, Running, and Success tasks, and dropdowns for sorting and filtering. It displays one DAG named "etl_pipeline" with tags airflow, postgres, etl. Below the DAG are sections for Schedule, Latest Run, and Next Run.

Dags Runs Task Instances

Q Search Dags Advanced Search `Ctrl+K`

All Failed Running Success All Filter by tag

1 Dag

Sort by Display Name (A-Z)

etl_pipeline airflow, postgres, etl

Schedule Latest Run Next Run

Dag
etl_pipeline

Home Dags Assets Browse Admin Security Docs User

React Flow

Search Dags **ctrl+K**

Trigger

etl_pipeline **Toggle**

Schedule Latest Run Next Run Owner Tags Latest Dag Version

2025-09-02, 00:15:03 **✓** airflow airflwo, etl, postgres v15

Overview Runs Tasks Events Code Details

Last 24 hours 2025-09-01, 00:15:08 - 2025-09-02, 00:15:08

0 Failed Tasks

0 Failed Runs

Last 1 Day Run

Duration (seconds)

385.46

0.81

2025-09-02, 12:15:03 Run After

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Vector buckets

Access Grants

Access Points (General Purpose Buckets, FSx file systems)

Access Points (Directory Buckets)

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

▼ Storage Lens

Dashboards

Storage Lens groups

silver/

 Copy S3 URI

Objects

Properties

Objects (17)



 Copy S3 URI

 Copy URL

 Download

 Open

 Delete

 Actions

 Create folder

 Upload

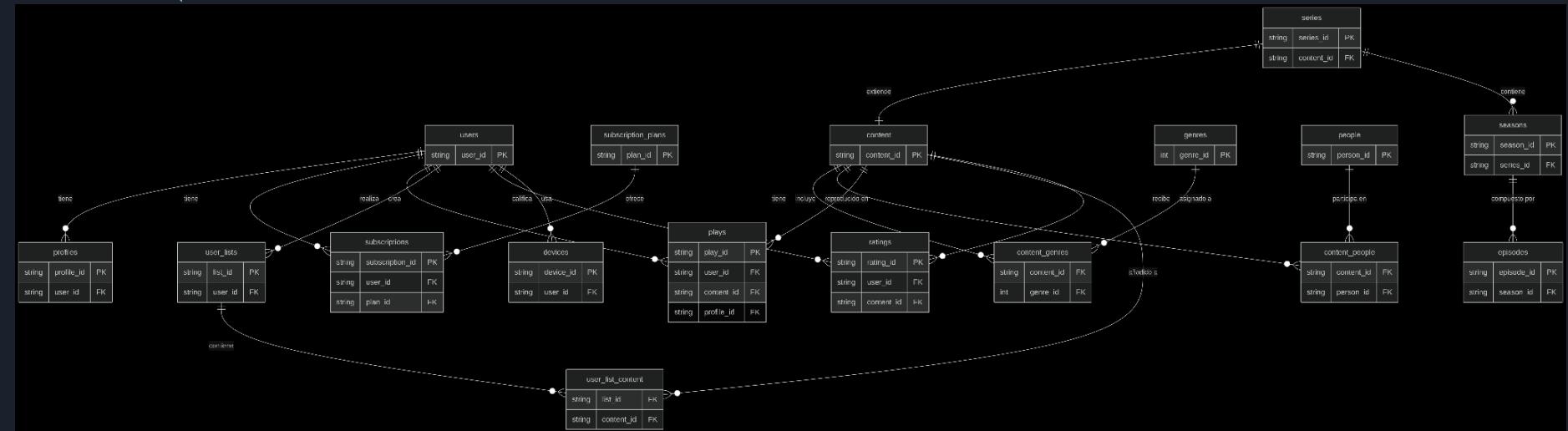
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

 Find objects by prefix

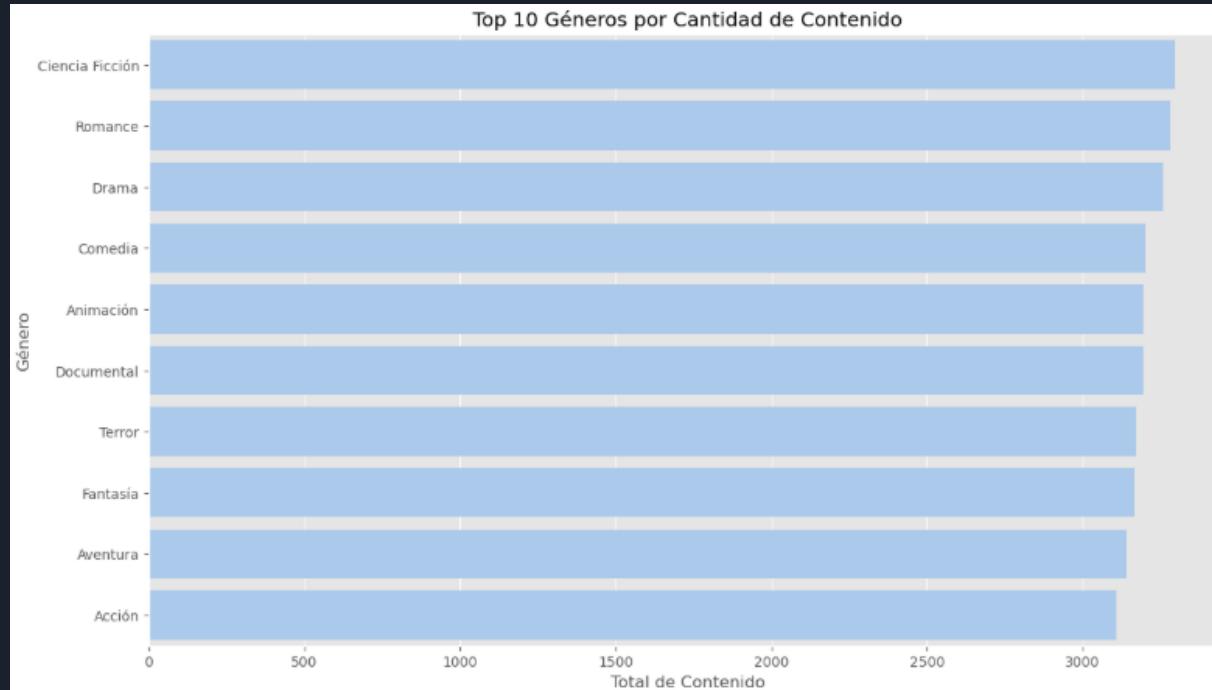
< 1 > 

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	 content_genres.parquet	parquet	August 31, 2025, 17:22:08 (UTC-03:00)	637.9 KB	Standard
<input type="checkbox"/>	 content_people.parquet	parquet	August 31, 2025, 17:22:09 (UTC-03:00)	856.8 KB	Standard
<input type="checkbox"/>	 content.parquet	parquet	August 31, 2025, 17:22:06 (UTC-03:00)	2.7 MB	Standard
<input type="checkbox"/>	 devices.parquet	parquet	August 31, 2025, 17:22:09 (UTC-03:00)	1.3 MB	Standard

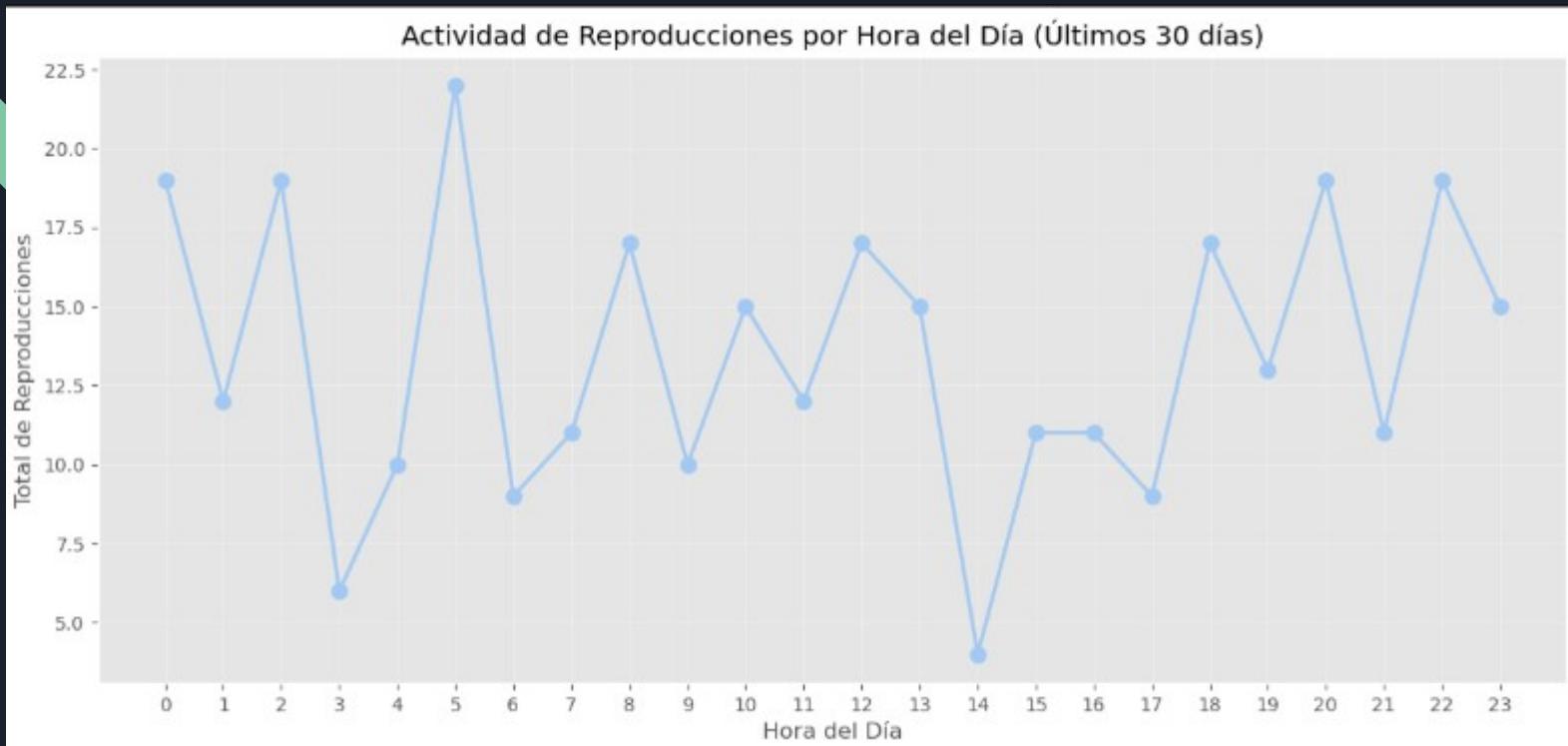
Esquema de BBDD



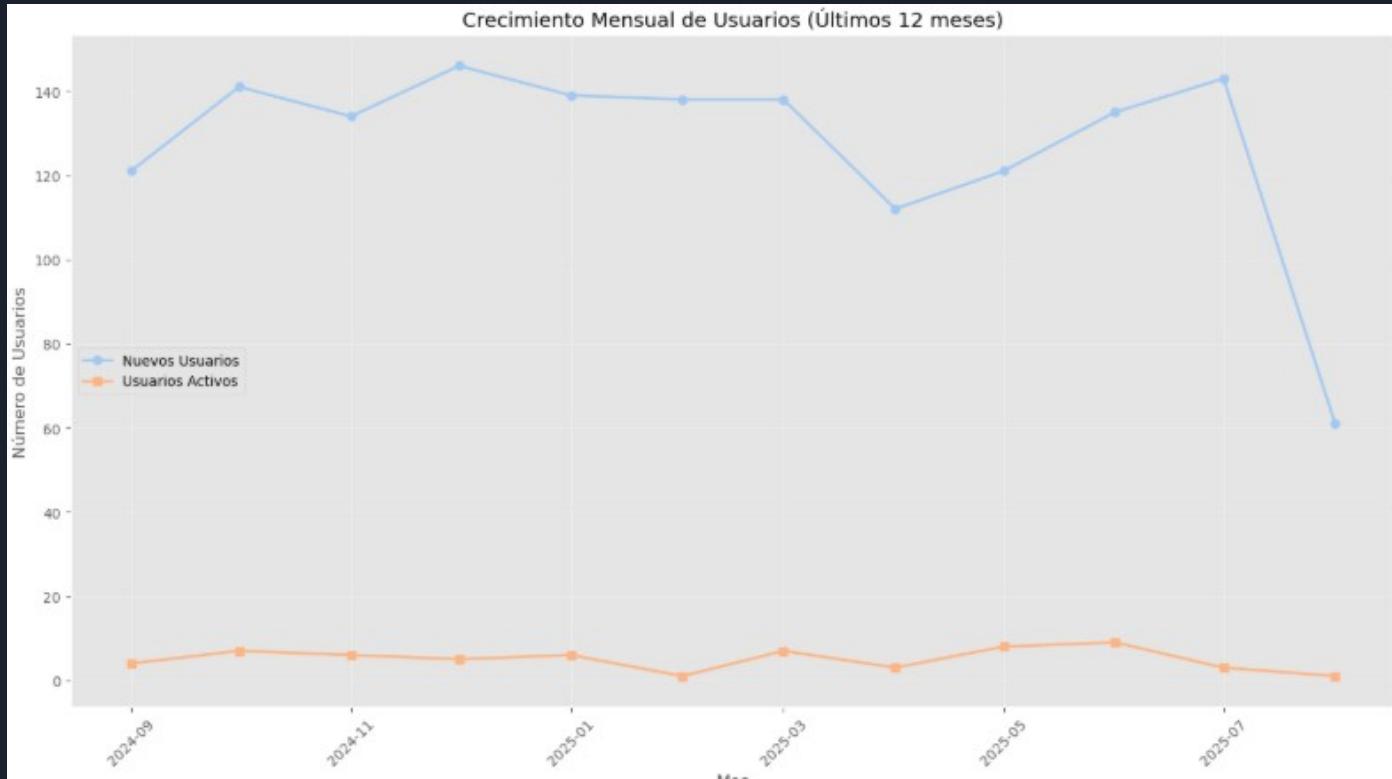
Análisis de datos



Los géneros incluidos (Ciencia Ficción, Romance, Drama, Comedia, Animación, Documental, Terror, Fantasía, Aventura y Acción) son justamente los más comunes en catálogos de cine y series, lo cual confirma su relevancia en la industria. La diferencia entre el primero y el último es mínima. Esto puede deberse a que la clasificación de géneros es amplia y muchos contenidos se catalogan en múltiples géneros, generando conteos muy similares.



El horario con menor cantidad de reproducciones es las 14 h, con menos de 5 reproducciones. También se observa baja actividad alrededor de las 3 h, 6 h, 9 h y 17 h. No parece haber una tendencia clara de “mañana” vs “tarde”, ya que la actividad es bastante dispersa a lo largo del día. Sin embargo, se notan más picos de actividad en la madrugada y noche (0–5 h y 20–23 h), lo que podría indicar que los usuarios consumen más contenido fuera de horarios laborales/estudio.



El número de nuevos usuarios muestra una gran fluctuación a lo largo del año. Hay picos significativos, como el de octubre de 2024 y el de junio de 2025, donde se supera la marca de los 140 usuarios. Estos picos pueden estar relacionados con campañas de marketing, promociones o lanzamientos de nuevos productos. Por otro lado, también hay caídas abruptas, como la del último mes, julio de 2025, donde la cifra cae a poco más de 60 nuevos usuarios. La empresa es buena atrayendo a gente nueva, pero tiene un grave problema para retenerla. Es fundamental que investiguen por qué los usuarios no se quedan y trabajen en mejorar la experiencia del producto para que se vuelvan usuarios activos a largo plazo.



Conclusión

Se puede concluir que la plataforma de streaming tiene una base de usuarios muy sólida y estable, con los usuarios existentes siendo el motor principal.

Sin embargo, hay un desafío claro: la retención de los usuarios nuevos. Aunque estos usuarios pueden ser ligeramente más activos por sesión al principio, no se quedan en la plataforma, lo que se refleja en una baja cantidad de días de actividad.

Para que la plataforma crezca, un área clave a mejorar es la retención de los usuarios nuevos. Es importante investigar por qué no regresan y optimizar el proceso de registro o las primeras impresiones para mejorar su experiencia inicial.