

# Democratizando el Acceso a los Datos:

# En AWS

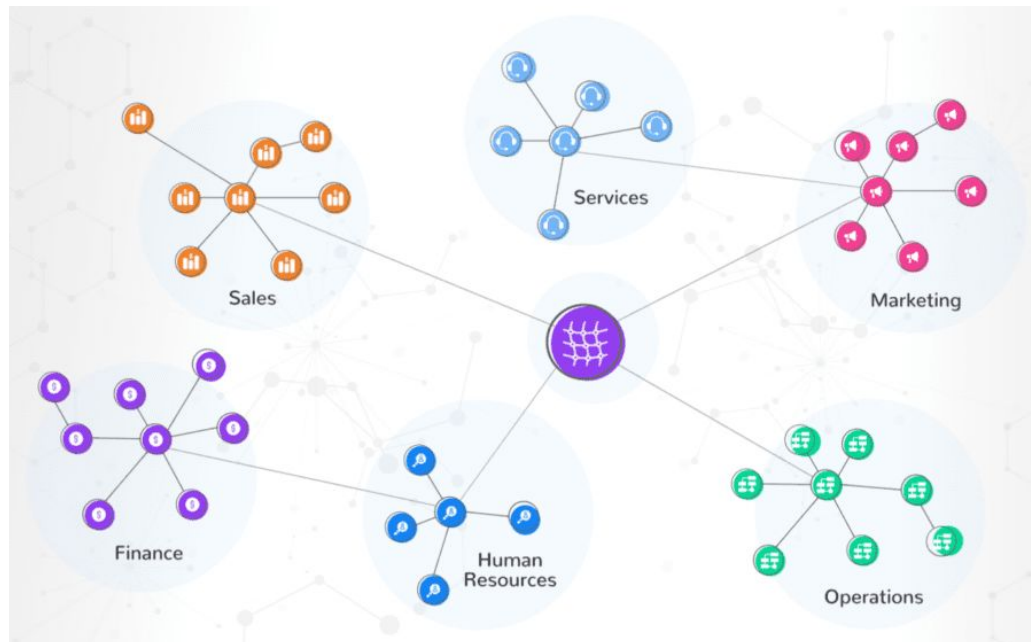
By [Lucas Trubiano](#)

Data Engineer













Experience in Banking & Finance

---

# ¿Cómo crear un ecosistema de datos distribuido y escalable?

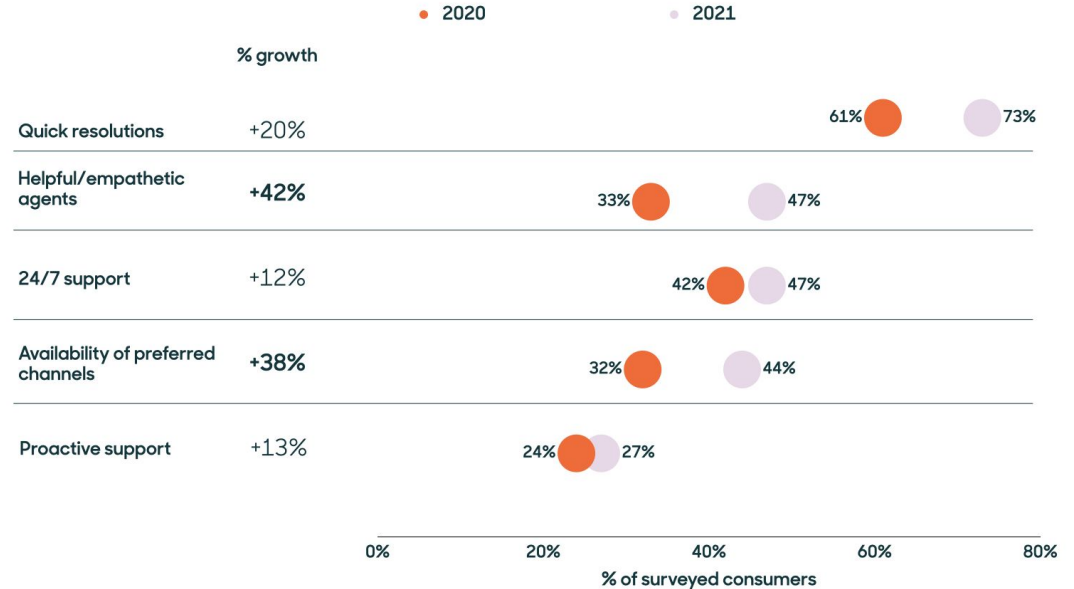


# ¿Para qué queremos un ecosistema de datos?

THE PAST	THE FUTURE
Technology Makes Experience Worse 	 Technology Makes Experience Better
Open 9-5x5 	 Open 24x7
Company Is In Control Of Where Interaction Happens 	 Customer Is In Control Of Where Interaction Happens
Customer Repeats Information On Every Channel 	 Company Knows Information From Every Channel
Call Center Volume Is High 	 Mobile Messaging Volume Is High
Customer Finds Content 	 Content Finds Customer

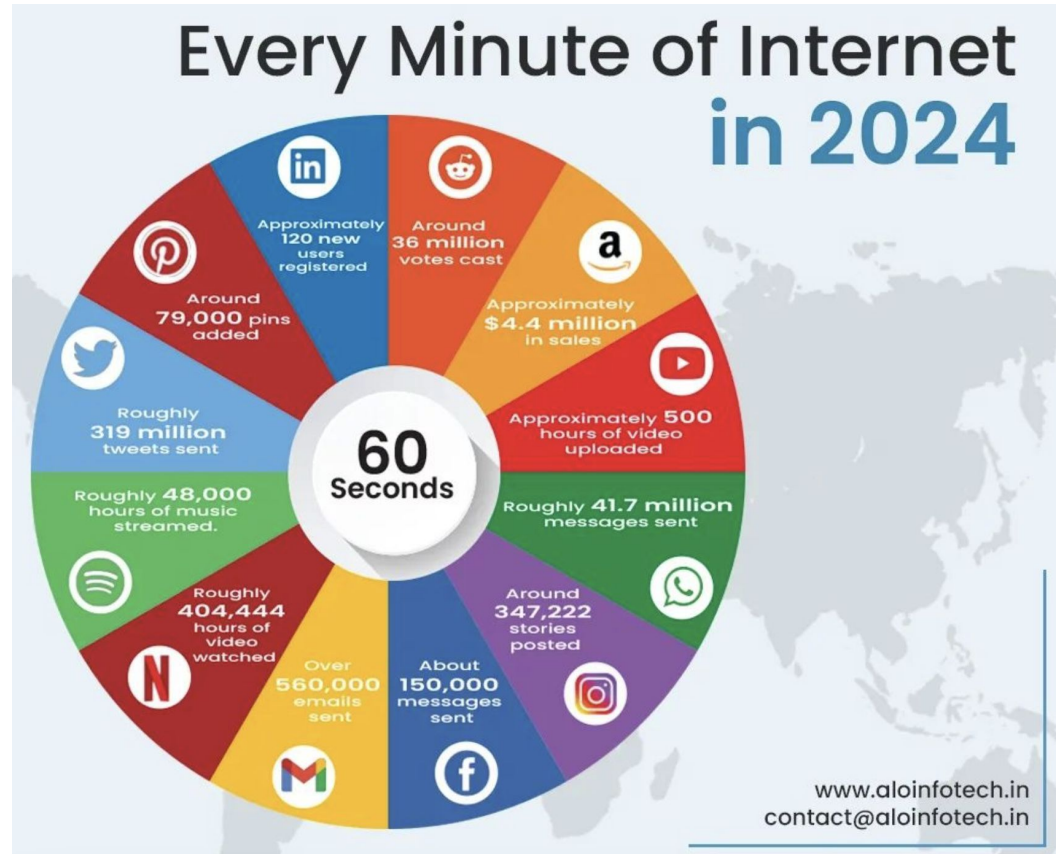
# ¿Para qué queremos un ecosistema de datos?

## Customers still value speed, but increasingly want empathy and access to preferred channels



—

¿Para qué queremos  
un ecosistema  
de datos?

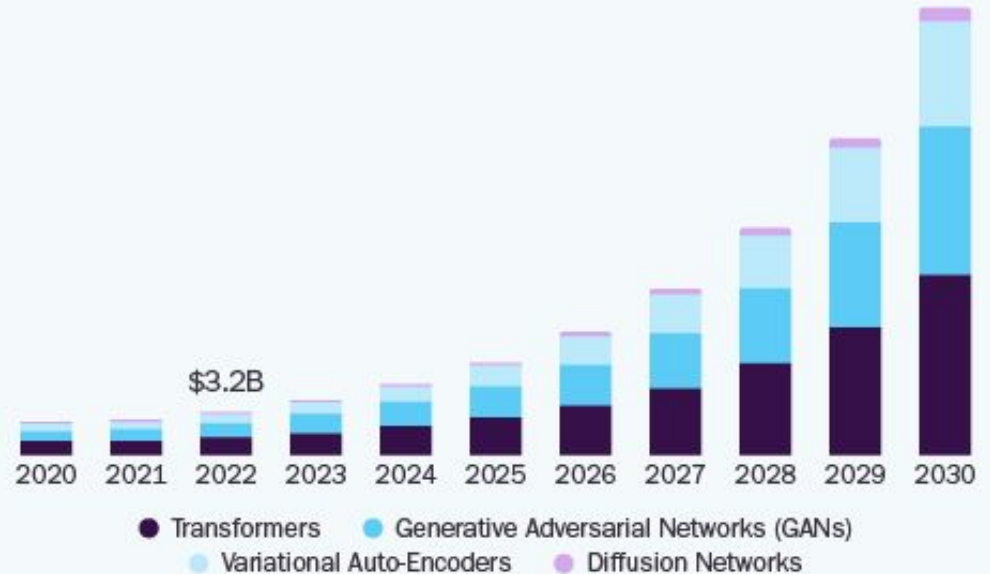


---

¿Para qué queremos  
un ecosistema  
de datos?

### U.S. Generative AI Market

Size, by Technology, 2020 - 2030 (USD Billion)



—

¿Para qué queremos  
un ecosistema  
de datos?



# ¿El Problema de los Enfoques Tradicionales?

- Dependencia de equipos centralizados.
- Retrasos en el acceso a los datos.
- Problemas de escalabilidad.



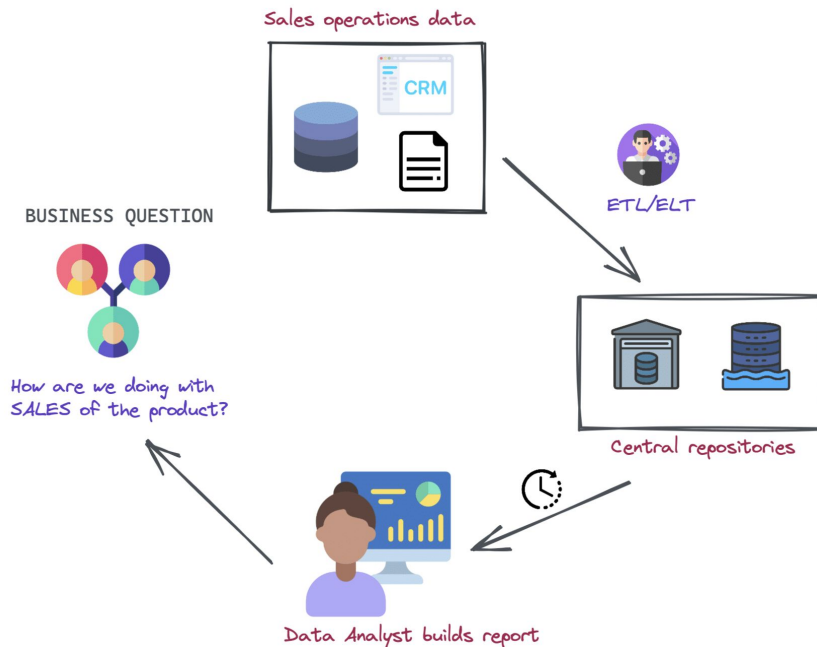
Disconnect among teams



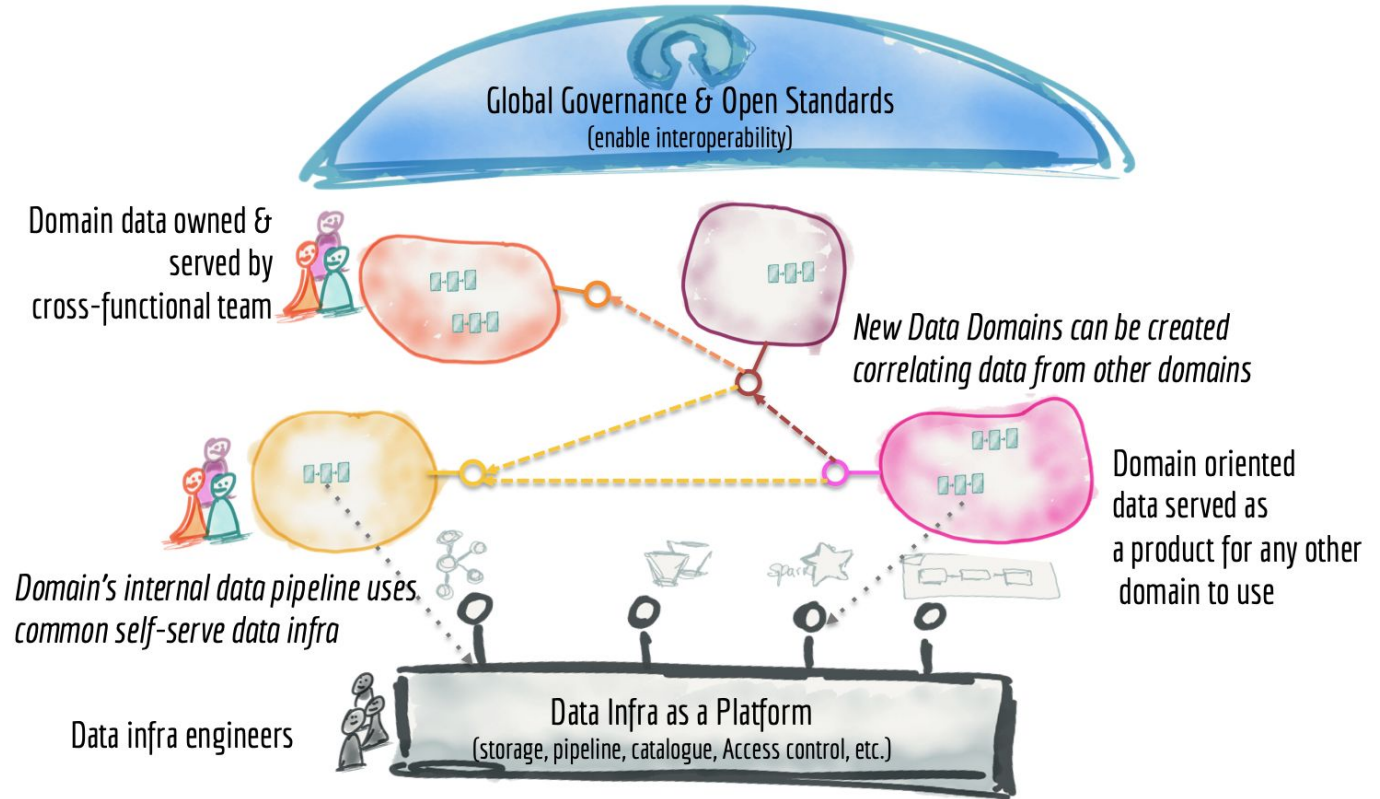
Data isn't interoperable



Overhead on central infra teams



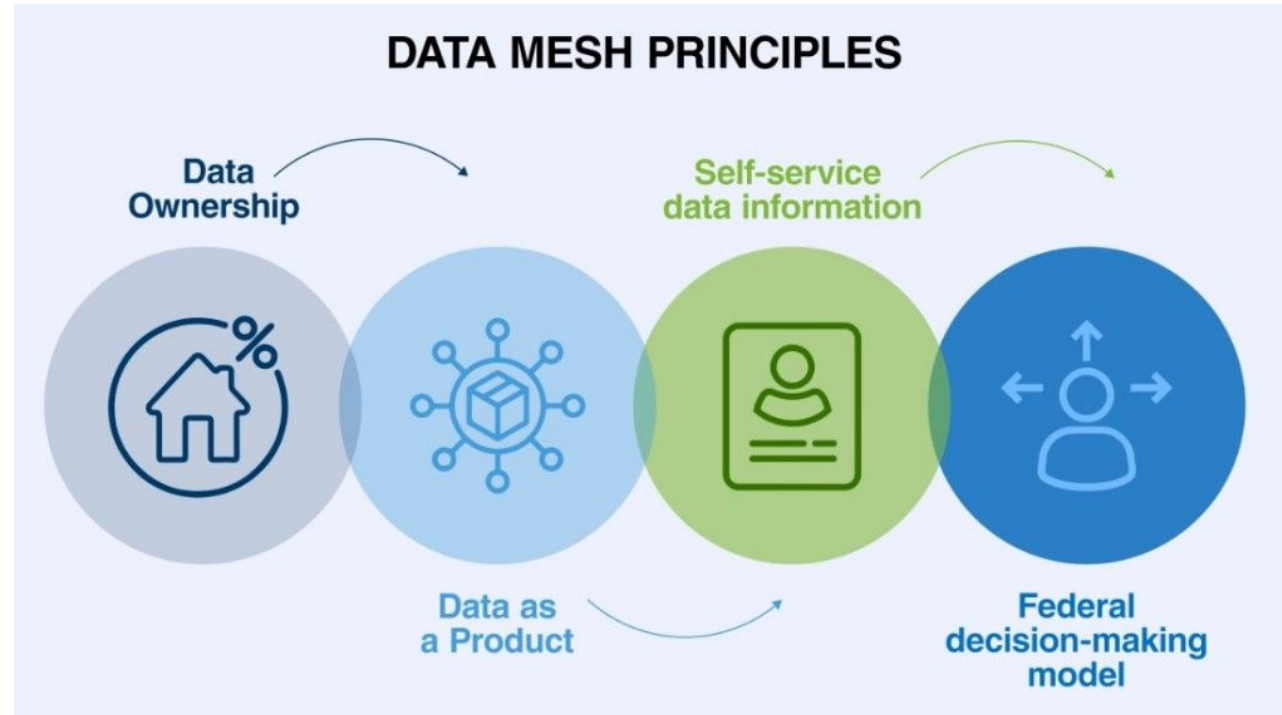
# ¿Qué es Data Mesh?



---

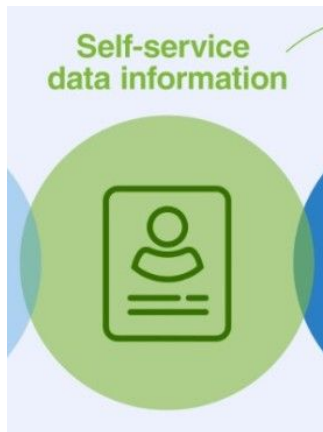
# Sus 4 pilares

- Enfoque descentralizado.
- Equipos responsables de sus propios datos.
- Datos tratados como productos.
- Gobierno y KPIs Federados.





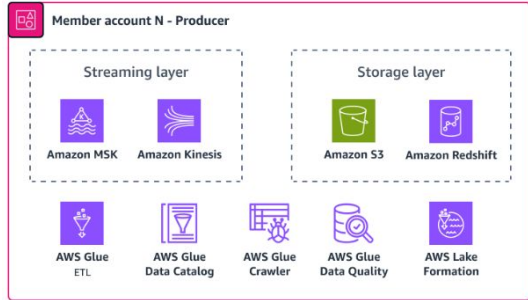
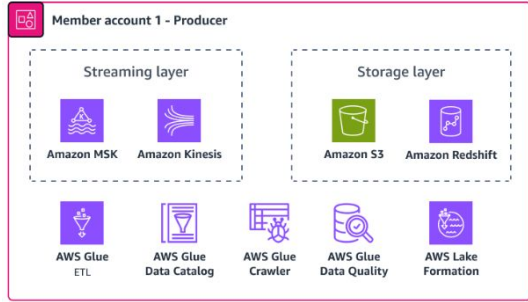
- Cada equipo es dueño de sus datos.
  - Eliminación de cuellos de botella.
  - Equipos que mejor conocen sus datos los gestionan.
  
  - Datos de alta calidad y bien documentados.
  - Disponibles para otros equipos como productos consumibles.
  - Estándares de calidad y mantenimiento.
-



- Plataforma autoservicio para todos los equipos.
  - Eliminación de la dependencia de equipos centrales.
  - Facilita la creación, gestión y consumo de datos.
  
  - Políticas comunes de seguridad y calidad.
  - Interoperabilidad entre dominios.
  - Cumplimiento de normativas de privacidad y seguridad.
-

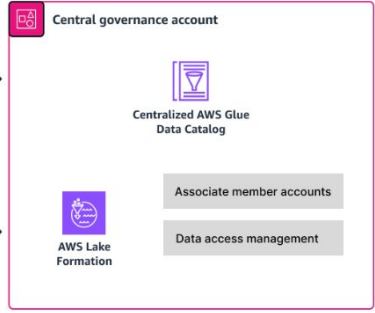


Data producers



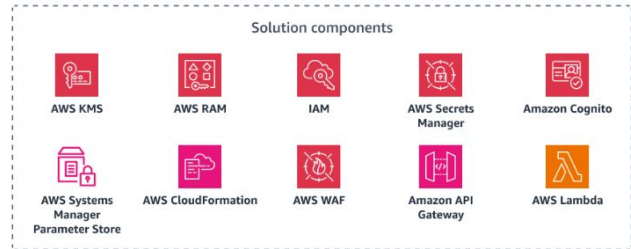
1

1



1

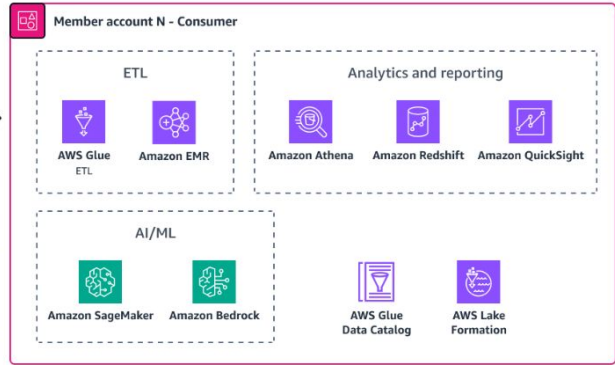
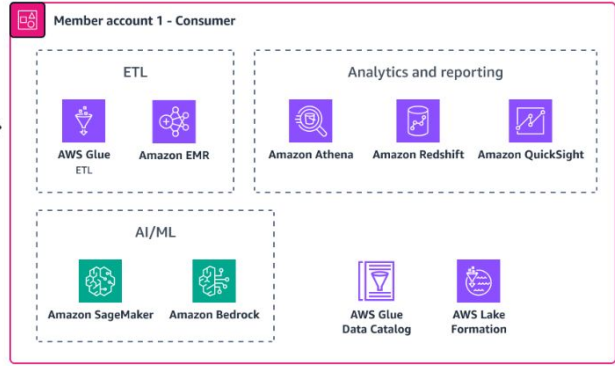
1



2

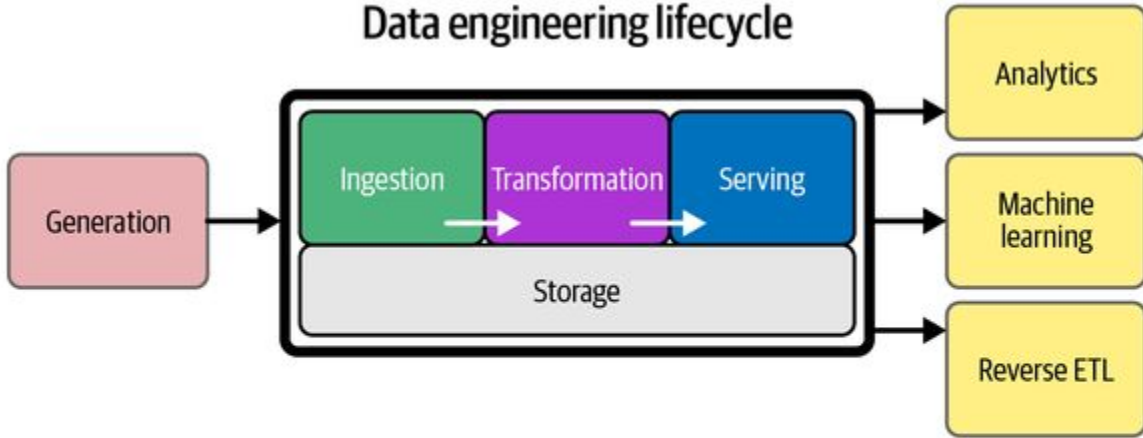
2

Data consumers



---

## Data engineering lifecycle

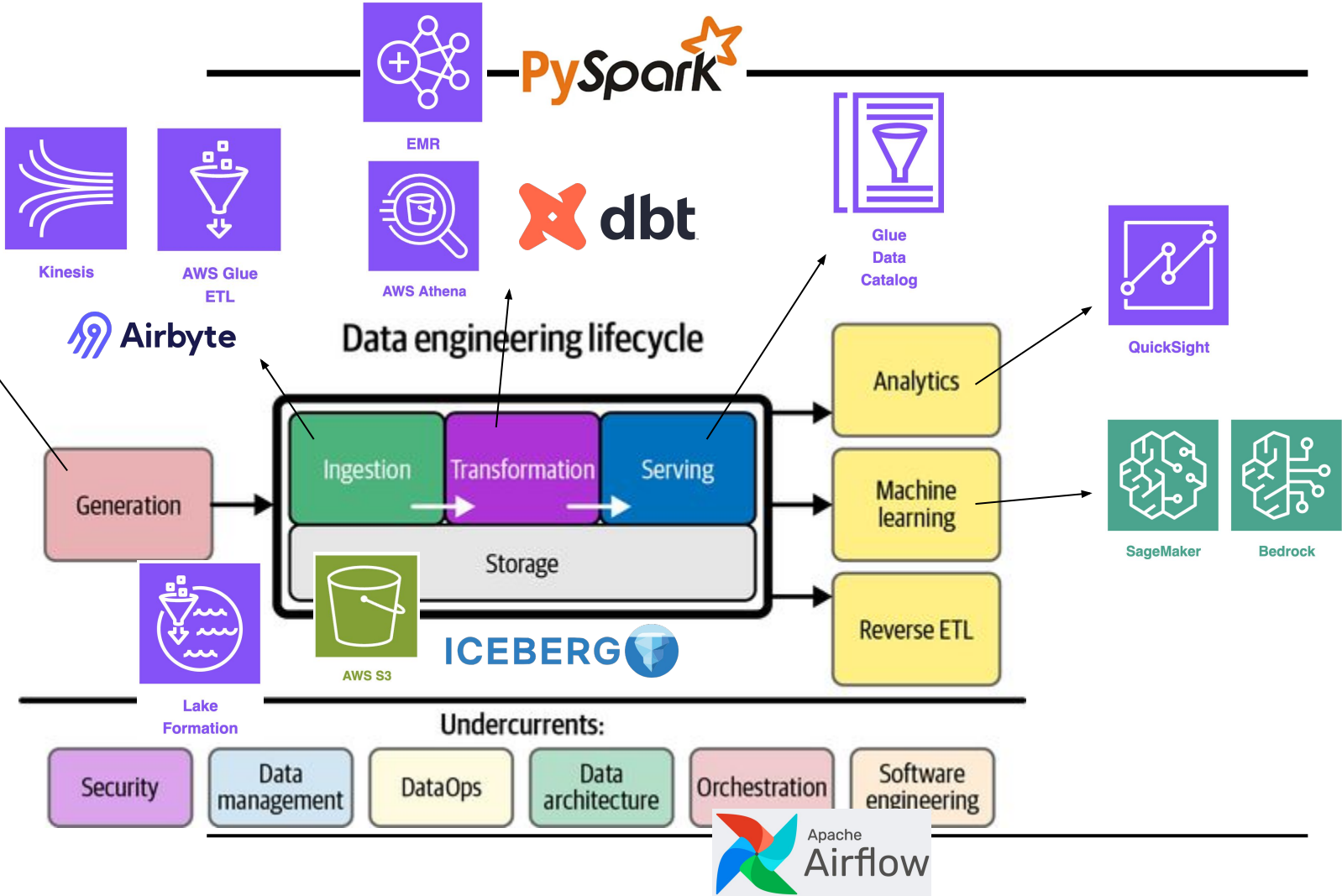


---

### Undercurrents:



Aplicación  
Backend  
Frontend  
BBDD  
APIs  
Scrapping





AWS Glue  
ETL

# 1. Ingesta

stg\_snapshot-global\_permissions-daily

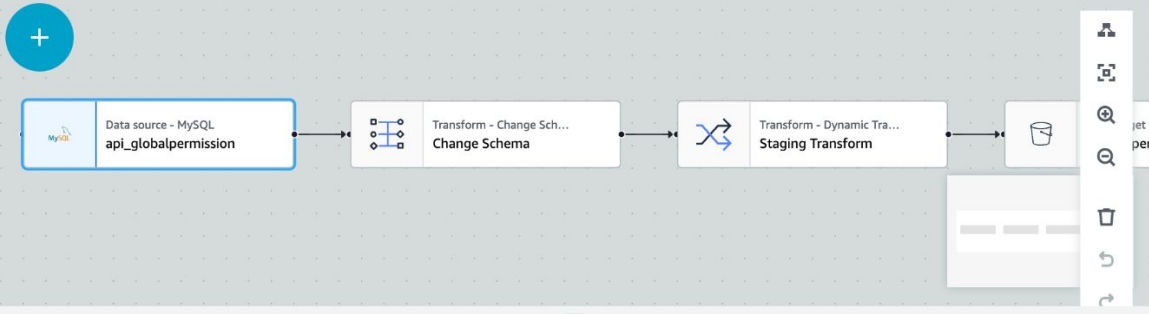
Last modified on 9/22/2024, 12:54:53 AM

Actions ▾

Save

Run

Visual | Script | Job details | Runs | Data quality | Schedules | Version Control



Data source properties - MySQL

Name

api\_globalpermission

JDBC source

JDBC connection details

Data Catalog table

Connection name [Info](#)

Choose the AWS Glue connection for your JDBC data source.

MySQL\_Prod

No description available.

Create a connection

Table name

Look up the name of the table from your JDBC source and enter it here.

chestmusic\_backend.api\_globalpermission

Data preview

Output schema

Data preview (8)

Info

READY



End session

Previewing 6 of 6 fields

Filter sample dataset

id	plays	date_added	accepted	from_user_id
15	6	2024-09-12 03:33:07.167965	true	6
4	1	2024-01-17 21:40:27.304550	true	14



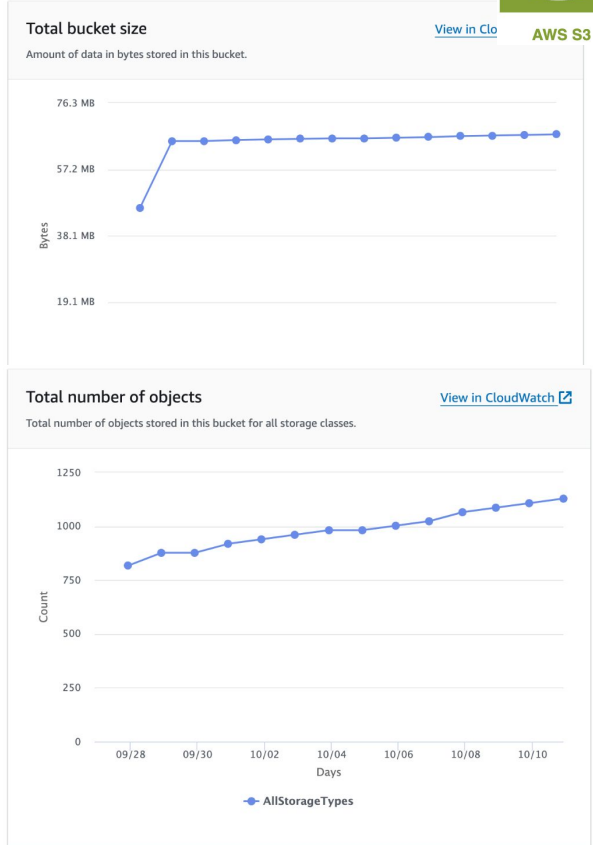
# 2. Almacenamiento

**Objects** | Properties | Permissions | Metrics | Management | Access Points

**Objects (4)** [Info](#) ↻ 📄 Copy S3 URI 📄 Copy URL 📄 Dow

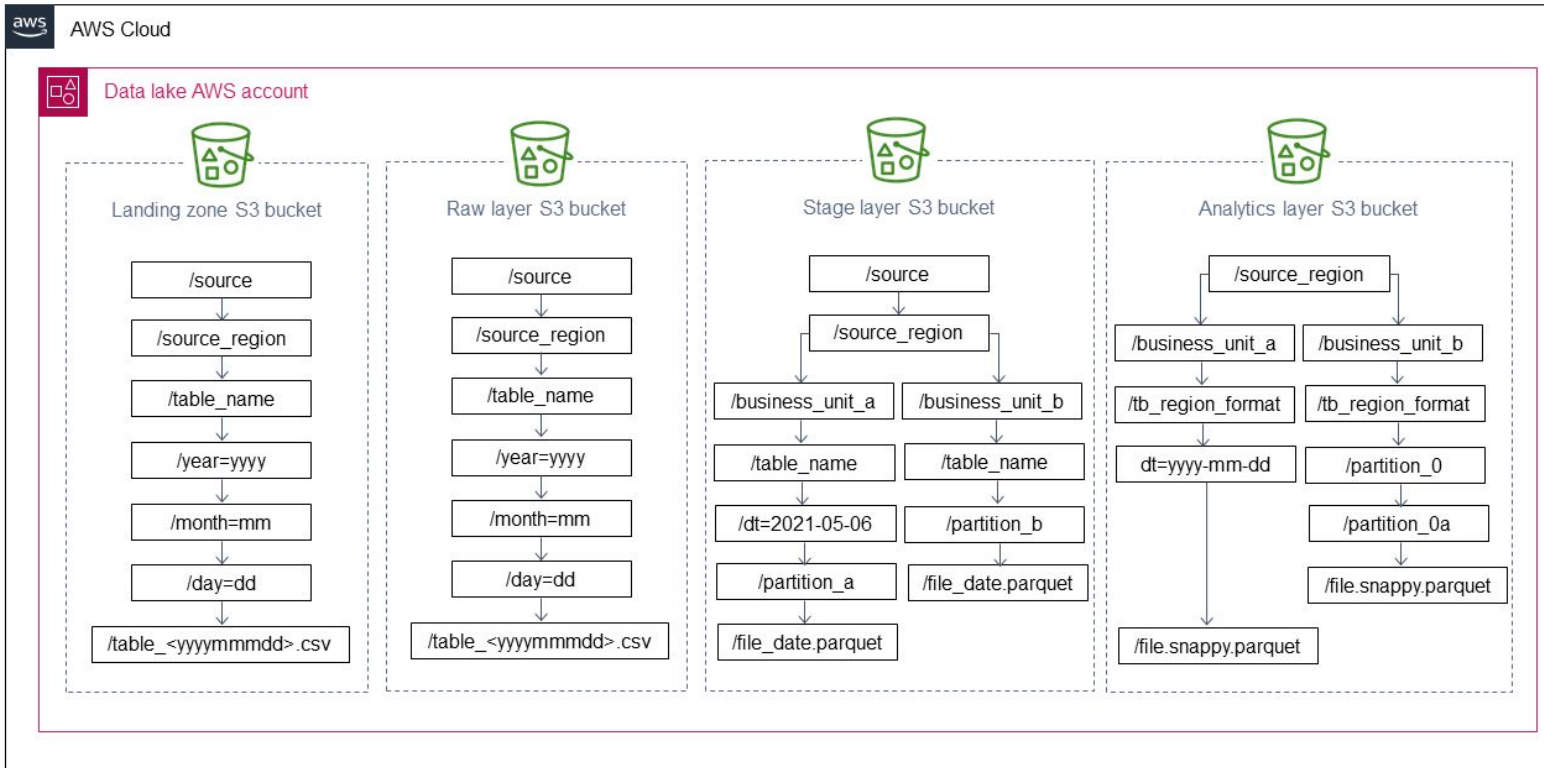
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket.

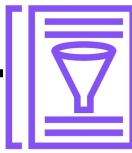
<input type="checkbox"/>	Name	Type	Last modified
<input type="checkbox"/>	<a href="#">athena/</a>	Folder	-
<input type="checkbox"/>	<a href="#">intermediate/</a>	Folder	-
<input type="checkbox"/>	<a href="#">mart/</a>	Folder	-
<input type="checkbox"/>	<a href="#">staging/</a>	Folder	-





# 2. Almacenamiento





# 3. Serving y Discovery

[AWS Glue](#) > Tables

## Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

**Tables (73)** Last updated (UTC)  
October 11, 2024 at 22:13:07 [Refresh] [Delete] [Add tables using ...]

View and manage all available tables.

<input type="checkbox"/>	Name ▲	Database ▼	Location ▼	Classification ▼	Deprecated ▼	View data	Data quality
<input type="checkbox"/>	albums_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	audio_files_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	beta_access_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	calendar	chestmusic_intermediate	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	country_payment_me	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	country_plans_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	email_invitations_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	faqs_v1	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	general_notifications_	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	global_permissions_v	chestmusic_staging	s3://datalake.ch	Parquet	-	Table data	View data quality
<input type="checkbox"/>	invitation_his	chestmusic_intermediate	-	-	-	-	View data quality





# 4. Governance

## AWS Lake Formation ×

Dashboard

### ▼ Data Catalog

Databases

Tables

Views [New](#)

Data filters

Data sharing

Crawlers [↗](#)

### ▼ Permissions

**Data lake permissions**

LF-Tags and permissions

Hybrid access mode

Data locations

### ▼ Administration

Administrative roles and tasks

Data lake locations

Application integration settings

IAM Identity Center integration

AWS Lake Formation > Data lake permissions

Too many permissions? Filter by database or table. In the navigation page, choose **Databases** or **Tables**. Then choose a database or table, and on the **Actions** menu, choose **View Permissions**.

### Data lake permissions (245 loaded, more available)



Revoke

Grant

< 1 2 3 4 5 ... >

<input type="checkbox"/>	Principal ▲	Princip... ▼	Princip... ▼	Resour... ▼	Database ▼	Table ▼	Resource ▼	Catalog ▼	LF-Tag ex...
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Database	default	-	default	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Database	chestmusi...	-	chestmusi...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Database	chestmusi...	-	chestmusi...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Database	chestmusi...	-	chestmusi...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Table	chestmusi...	vw_roles_...	vw_roles_...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Column	chestmusi...	vw_roles_...	Included: ...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Table	chestmusi...	vw_gener...	vw_gener...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Column	chestmusi...	vw_gener...	Included: ...	12976148...	-
<input type="checkbox"/>	analytica...	IAM user	arn:aws:ia...	Table	chestmusi...	vw_subscr...	vw_subscr...	12976148...	-



# 5. Transformación y Consultas

**Data** ↻ <

Data source  
AwsDataCatalog

Database  
chestmusic\_mart

Tables and views Create ⚙️

▶ **Tables** (0) < **1** >

▼ **Views** (3) < **1** >

- 🔍 proyectos
- 🔍 shares
- 🔍 usuarios

Query Control : ✕ | Show Columns : ✕ | ✔️ **emails no verificados** : ✕

```
1 SELECT email,  
2     is_email_verified  
3 FROM chestmusic_intermediate.users  
4 where is_email_verified = false
```

SQL Ln 4, Col 32

Run again Explain Cancel Clear Create

Reuse query results  
up to 60 minutes ago

**Query results** | Query stats

✔️ **Completed** | Time in queue: 70 ms | Run time: 4.612 sec | Data scanned: 10.43 KB



QuickSight

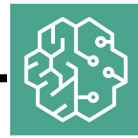
# 6. Analytics



**SPICE**

(Super-fast, Parallel, In-memory Calculation Engine)

# 7. Machine Learning



SageMaker



Bedrock

AMAZON SAGEMAKER CAPABILITY	FREE TIER USAGE PER MONTH FOR THE FIRST 2 MONTHS
Studio notebooks, and on-demand notebook instances	250 hours of ml.t3.medium instance on Studio notebooks OR 250 hours of ml.t2 medium instance or ml.t3.medium instance on on-demand notebook instances
RStudio on SageMaker	250 hours of ml.t3.medium instance on RSession app AND free ml.t3.medium instance for RStudioServerPro app
Data Wrangler	25 hours of ml.m5.4xlarge instance
Feature Store	10 million write units, 10 million read units, 25 GB storage
Training	50 hours of m4.xlarge or m5.xlarge instances
Real-Time Inference	125 hours of m4.xlarge or m5.xlarge instances
Serverless Inference	150,000 seconds of inference duration
Canvas	160 workspace instance hours/month, and up to 10 model creation requests/month, each with up to 1 million cells/model creation request

## Broad choice of models from leading AI companies

### AI21 Labs

Instruction-following FMs built for the enterprise that perform a range of tasks including text generation, question answering, summarization, and more.



### Amazon

A family of FMs for text and image generation, summarization, classification, open-ended Q&A, information extraction, and text or image search.



### Anthropic

FMs for thoughtful dialogue, content creation, complex reasoning, creative writing, and coding, trained with Constitutional AI.



### Cohere

Text-generation and representation models to generate text, summarize, search, cluster, classify, and use RAG.



### Meta

Models ideal for dialogue use cases and natural language tasks like Q&A and reading comprehension.



### Mistral AI

Models with publicly available weights supporting a variety of use cases from text summarization, text classification, and text completion to code generation and code completion.



### Stability AI

Image-generation models that produce unique, realistic, and high-quality visuals, art, logos, and designs.



# Lucas Trubiano

Consultor y Profesor en Ingeniería de Datos

Data Engineer - Fintech & Banking

Ing. en Mecatrónica UNCUYO

#Python #AWS #BigData #Finance

#Latam #Spain #USA



Santander  
Tecnología



[/in/lucastrubiano](https://www.linkedin.com/company/lucastrubiano)



[/lucastrubiano](https://www.instagram.com/lucastrubiano)



[/lucastrubiano](https://github.com/lucastrubiano)



[@lucastrubiano](https://www.youtube.com/channel/UC...)

# ¡Muchas Gracias!

## Preguntas