Principios de arquitectura e ingeniería de datos

Carlos Fernando Chicata Farfan
Data & solution engineer

30/04/2025

¿Qué es la ingeniería de datos?

Lo más simple:

#1: “mover & integrar datos”.

#2: “un superusuario y conector”.

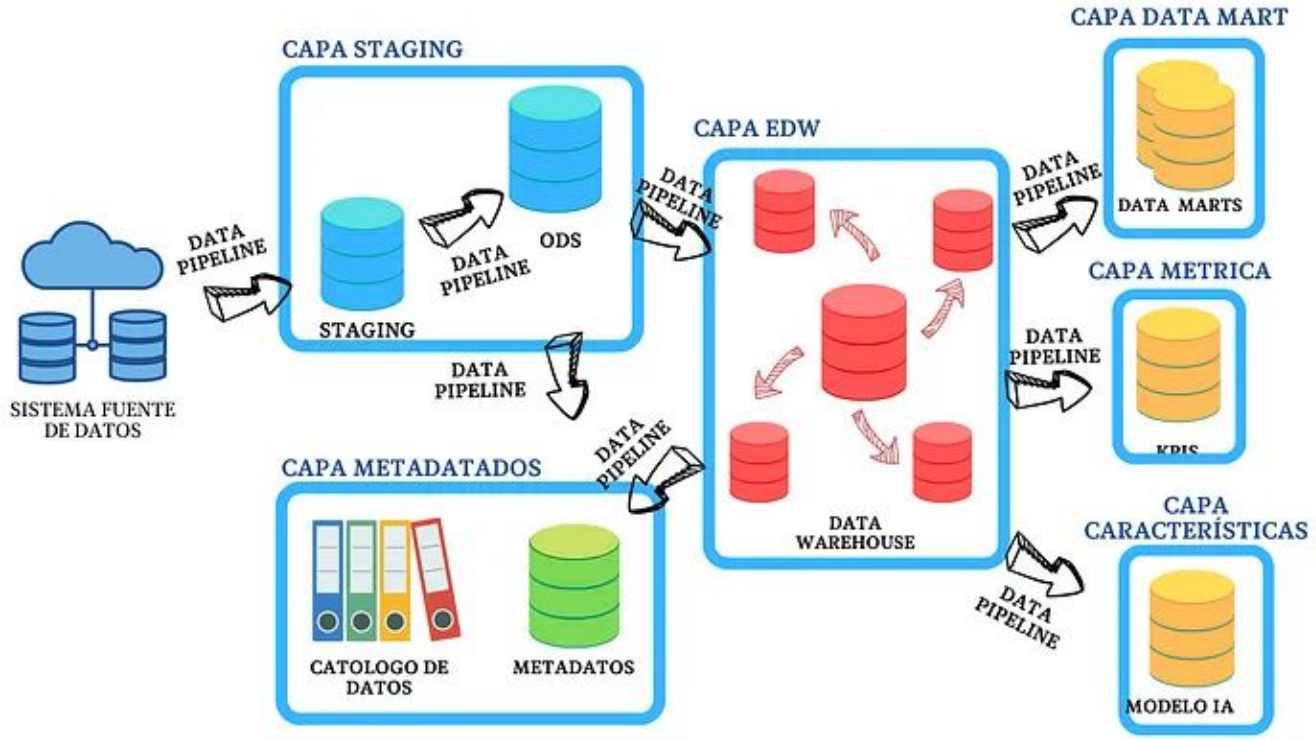
¿Qué es la arquitectura de datos?

Lo más simple:

#1: “organizar los datos”.

#2: “Entender los requerimiento”.

La relación entre ambos

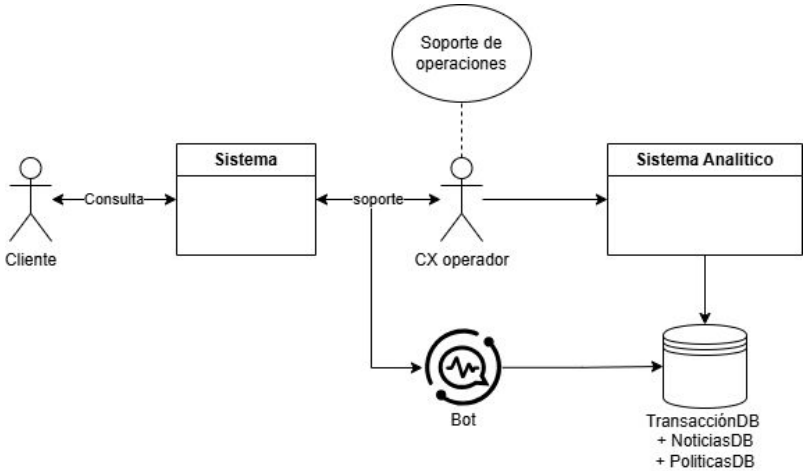


Facto #1: Naturaleza de los datos y solicitudes

La contexto de los datos analiticos

Esto se contextualiza en la realidad del equipo:

- *Proceso*: formas de cómo el negocio realiza sus actividades.
- *Personas*: la capacidad del equipo para realizar sus actividades.
- *Tecnología*: que herramientas utilizan las personas para gestionar los procesos.

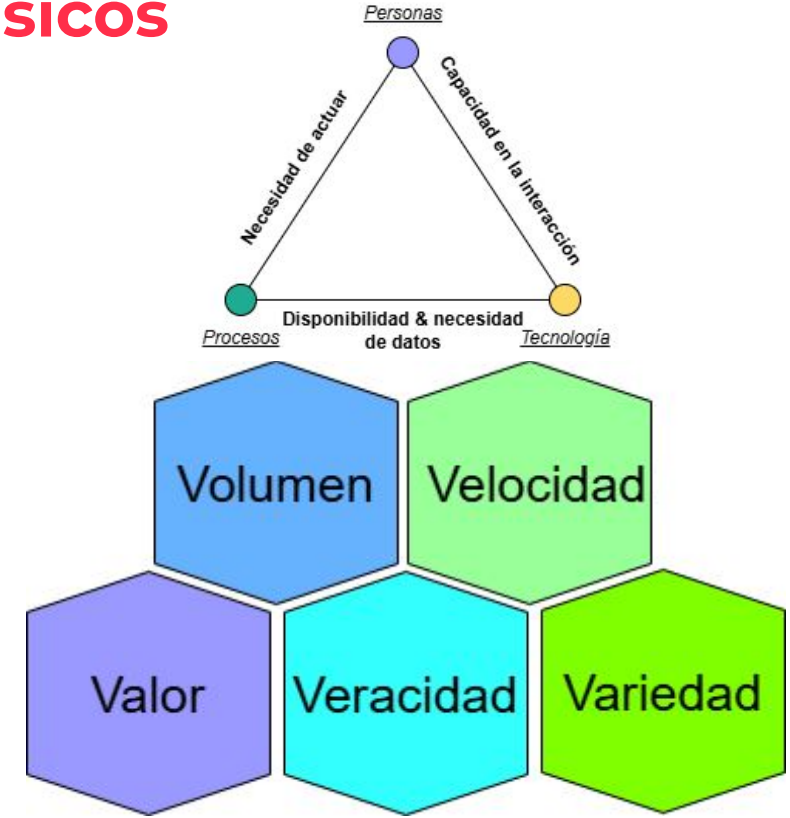


La naturaleza de los datos analiticos: los detalles fundamentales

En los detalles del alto nivel está la situación real:

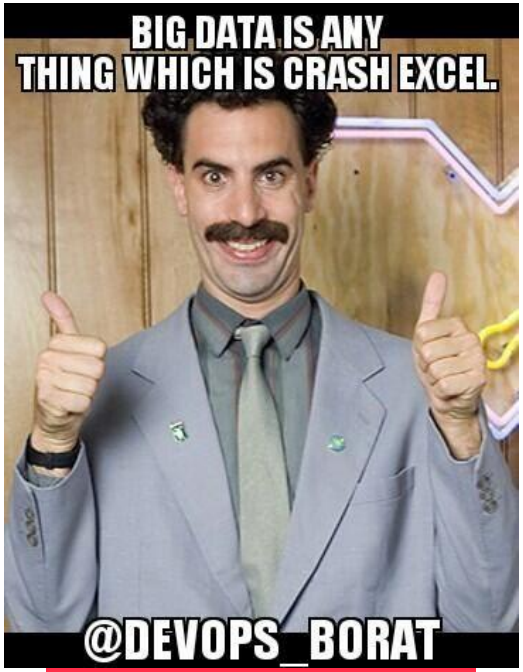
1. Necesidad del negocio para actuar.
2. Disponibilidad y necesidad de datos.
3. Capacidad en la interacción.

La naturaleza de los datos analiticos: los detalles basicos



El adjetivo big

Aquí tendremos problemas y conversaciones interesantes....



Situaciones en análisis de datos.

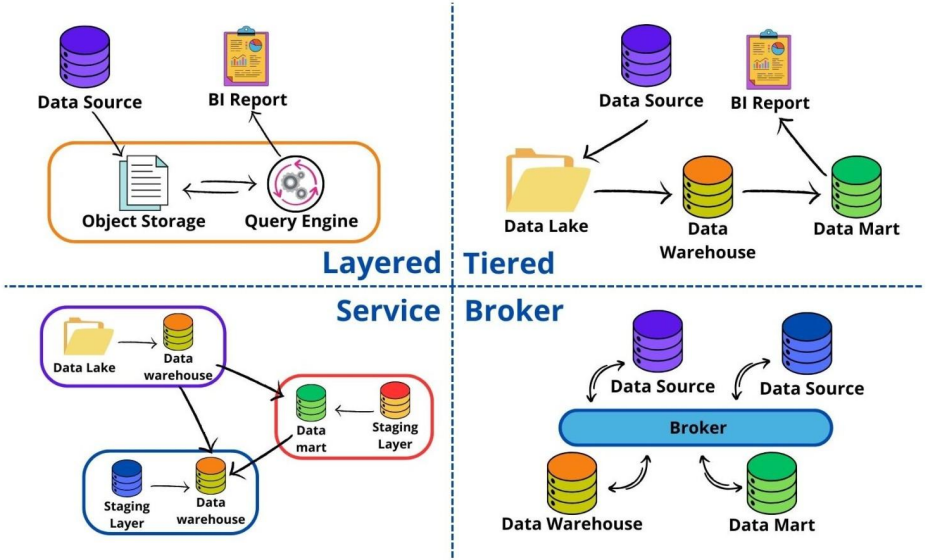
Hay diferentes situaciones que se abordan:

- Gestión de información/datos procesados 👔
- Integración de nuevas fuentes de datos 🛠️
- Seguridad en disponibilidad de datos 🔐
- Migraciones y optimizaciones 💰
- Mantenimiento de sistemas 🧰
- Archivamiento de datos 🗃️

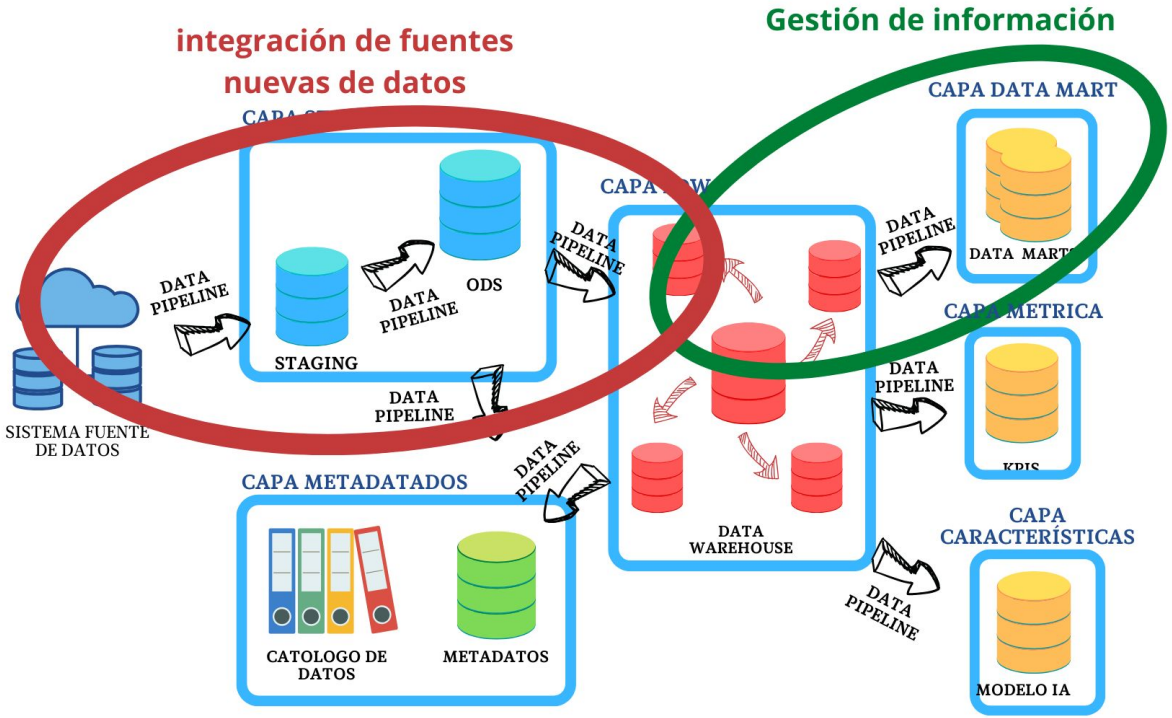
Contexto del sistema y la situación

Realidad se fundamenta en:

- *Fuentes:* API externas, bases de datos internas, Web Scraping.
- *Almacenamiento:* ¿Donde, cual y porque?
- *Organización*: interacción de los datos y sus transformaciones.



Ejemplo de una situación y realidad



El pasado sí importa en tu entorno

Como dejes el entorno en tu tiempo, puede ser la lágrima o la gloria del siguiente en tomarlo.

Facto #2: Conocer los componentes y herramientas

Lo básico para ambos roles

Como se queda cuando no sabes los componentes de la arquitectura...

Primero la aplicación y después las herramientas

Componentes de la arquitectura de datos: data lake

Consideraciones:

- *Función*: Almacenar los datos de todo tipo para uso de datos.
- Consideraciones críticas:
  - Modelo de costo.
  - Estructura de organización.
  - Mecanismo de seguridad.
  - Integración con otros servicios.

Componentes de la arquitectura de datos: data warehouse

Consideraciones:

- *Función*: Almacenar de datos históricos integrados.
- Consideraciones críticas:
  - Modelo de costo.
  - Modelo de datos y consultas.
  - Gestión de infraestructura.
  - Integración con otros servicios.

Componentes de la arquitectura de datos: data pipeline

Consideraciones:

- *Función*: Mover y transformar los datos.
- Consideraciones críticas:
  - Modelo de costo.
  - Capacidad de programación.
  - Gestión de infraestructura y despliegue.

Componentes de la arquitectura de datos: data mart

Consideraciones:

- *Función*: Almacenar los datos para BI.
- Consideraciones críticas:
  - Modelo de costo.
  - Modelo de datos y consultas.
  - Integración con otros servicios..

Componentes de la arquitectura de datos: operational data store

Consideraciones:

- *Función*: Almacenar de datos integrados.
- Consideraciones críticas:
  - Modelo de costo.
  - Modelos de datos y consultas.
  - Integración con otros servicios.

Aurora

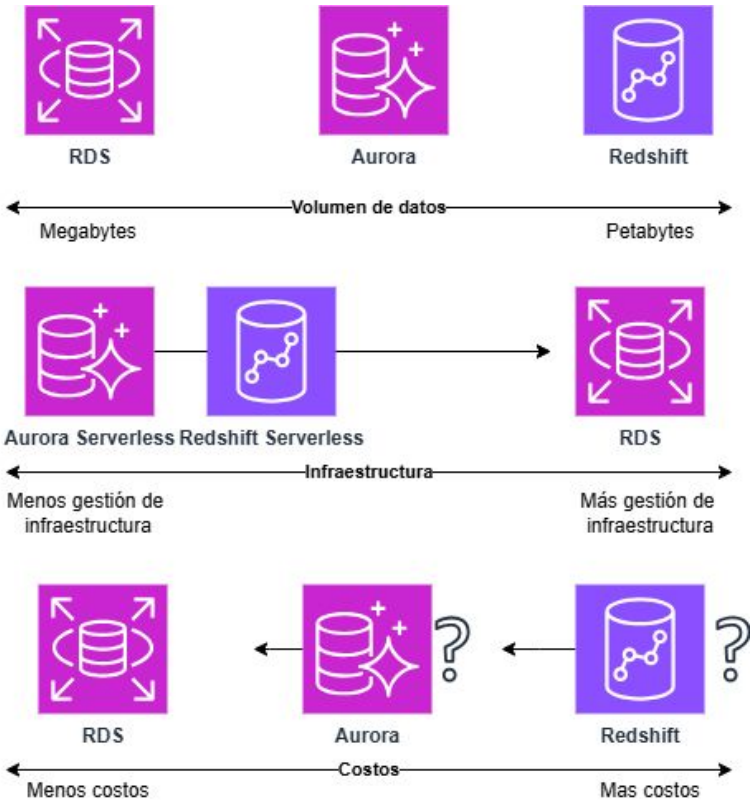
RDS

Componentes de la arquitectura de datos: staging layer

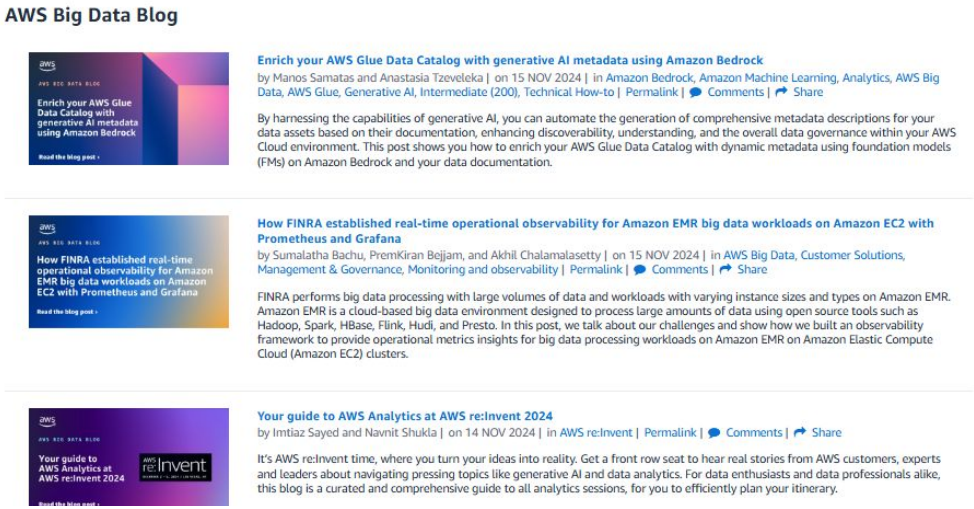
Consideraciones:

- *Función*: Almacenar los datos para soportar al data pipeline.
- Consideraciones críticas:
  - Modelo de costo.
  - Operaciones disponibles.
  - Integración con otros servicios.

Criterios de evaluación



El aprendizaje es importante en todo momento



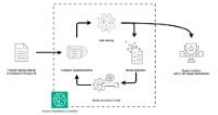
El aprendizaje es importante en todo momento

AWS Machine Learning Blog

Responsible AI in action: How Data Reply red teaming supports generative AI safety on AWS

by Cassandre Vandeputte, Davide Gallitelli, and Amine Ait el harraj | on 29 APR 2025 | in Advanced (300), Amazon Bedrock, Amazon Machine Learning, Amazon SageMaker, Amazon SageMaker Data & AI Governance, Generative AI, Responsible AI | Permalink | Comments | Share

In this post, we explore how AWS services can be seamlessly integrated with open source tools to help establish a robust red teaming mechanism within your organization. Specifically, we discuss Data Reply's red teaming solution, a comprehensive blueprint to enhance AI safety and responsible AI practices.



InterVision accelerates AI development using AWS LLM League and Amazon SageMaker AI

by Vu Le, Jaya Padma Mutta, Mohan CV, Rajesh Babu Nuvvula, and Brent Lazarenko | on 29 APR 2025 | in Amazon Machine Learning, Amazon SageMaker, Amazon SageMaker AI, Amazon SageMaker JumpStart, Amazon SageMaker Studio, Foundational (100), Partner solutions, Technical How-to, Thought Leadership | Permalink | Comments | Share

This post demonstrates how AWS LLM League's gamified enablement accelerates partners' practical AI development capabilities, while showcasing how fine-tuning smaller language models can deliver cost-effective, specialized solutions for specific industry needs.

Improve Amazon Nova migration performance with data-aware prompt optimization

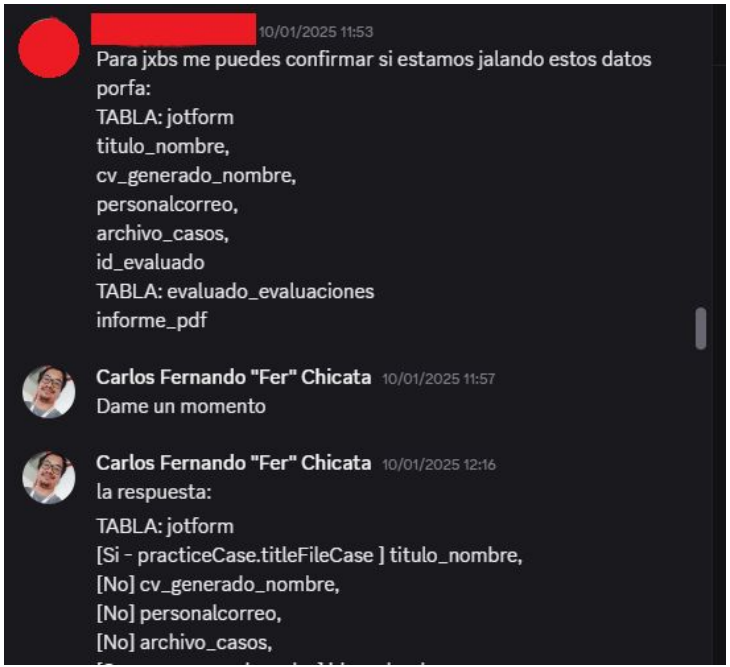
by Yunfei Bai, Anupam Dewan, Kashif Imran, and Shuai Wang | on 29 APR 2025 | in Amazon Bedrock, Amazon Nova, Generative AI, Thought Leadership | Permalink | Comments | Share

In this post, we present an LLM migration paradigm and architecture, including a continuous process of model evaluation, prompt generation using Amazon Bedrock, and data-aware optimization. The solution evaluates the model performance before migration and iteratively optimizes the Amazon Nova model prompts using user-provided dataset and objective metrics.

Facto #3: documentar todo

Una día más en la oficina...

“Parece fácil, se ve muy fácil, pero es difícil en realidad” - El tri



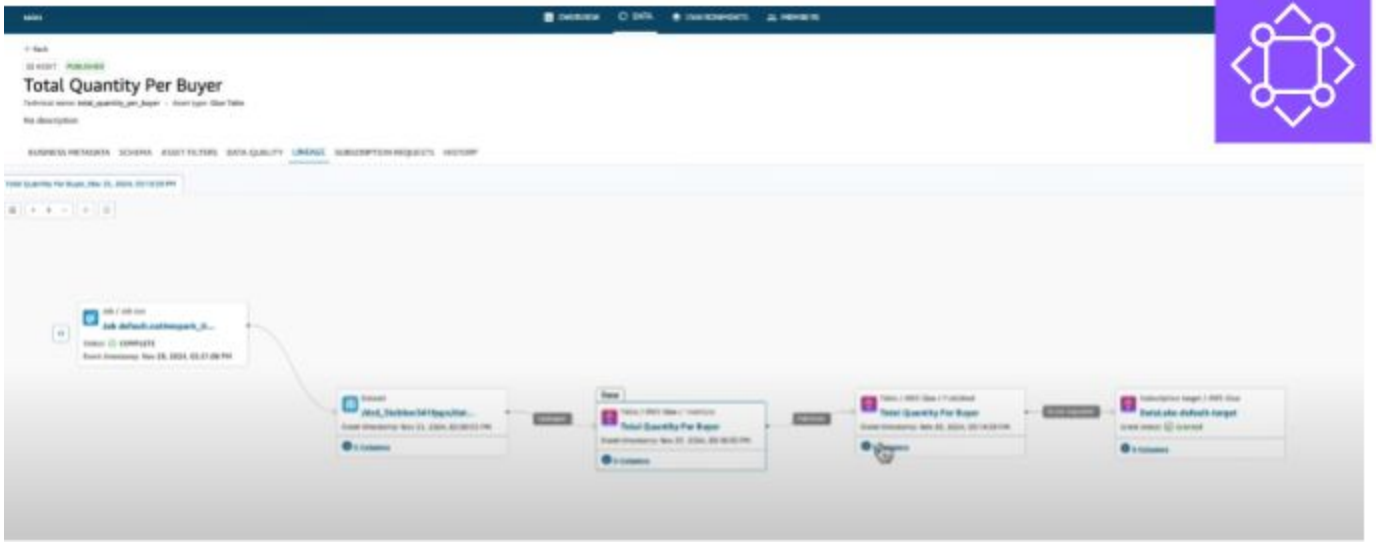
Tipos de documentación necesaria

Los más comunes más comunes y necesarios son:

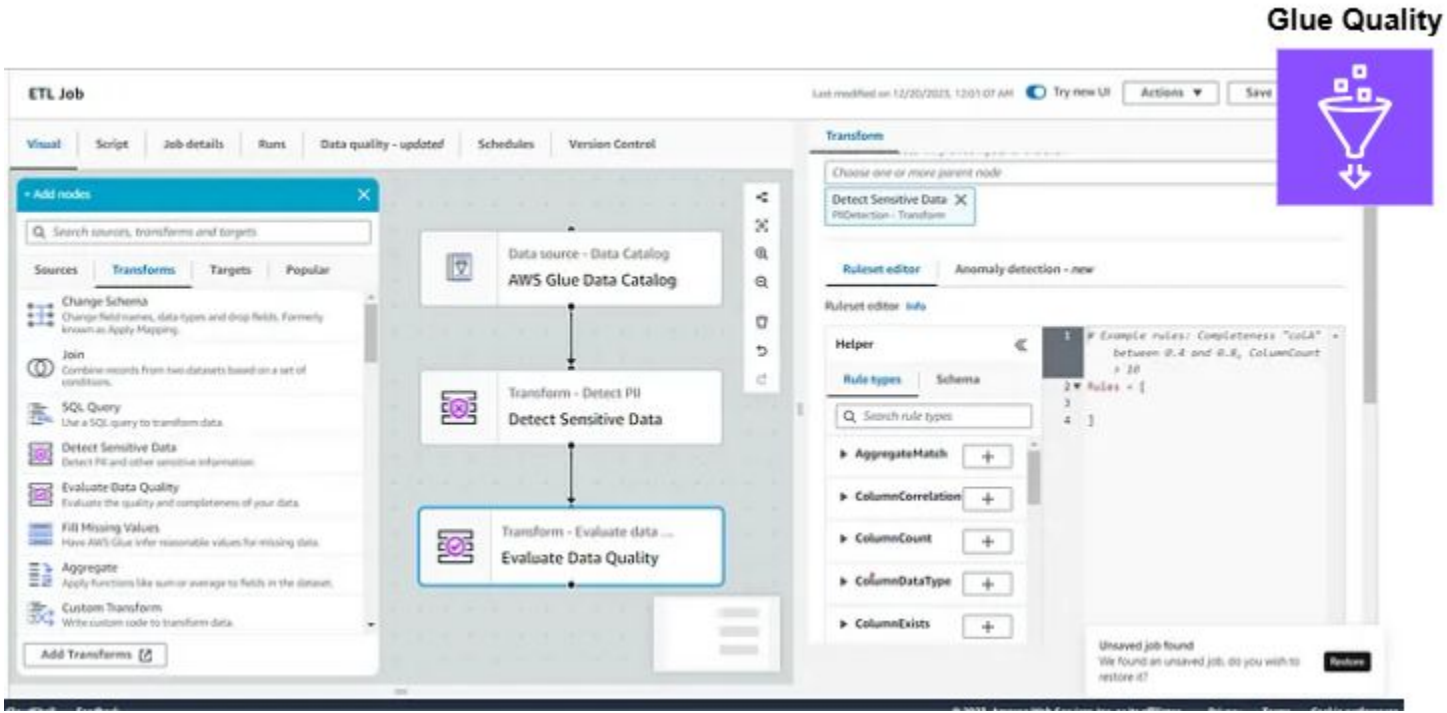
- Procesos de información.
- Linaje de datos.
- Inventario de artefactos relacionados.
- Criterios de calidad de datos.
- Mapeamiento, diccionario y modelo de datos.

Servicios que nos pueden ayudar

DataZone

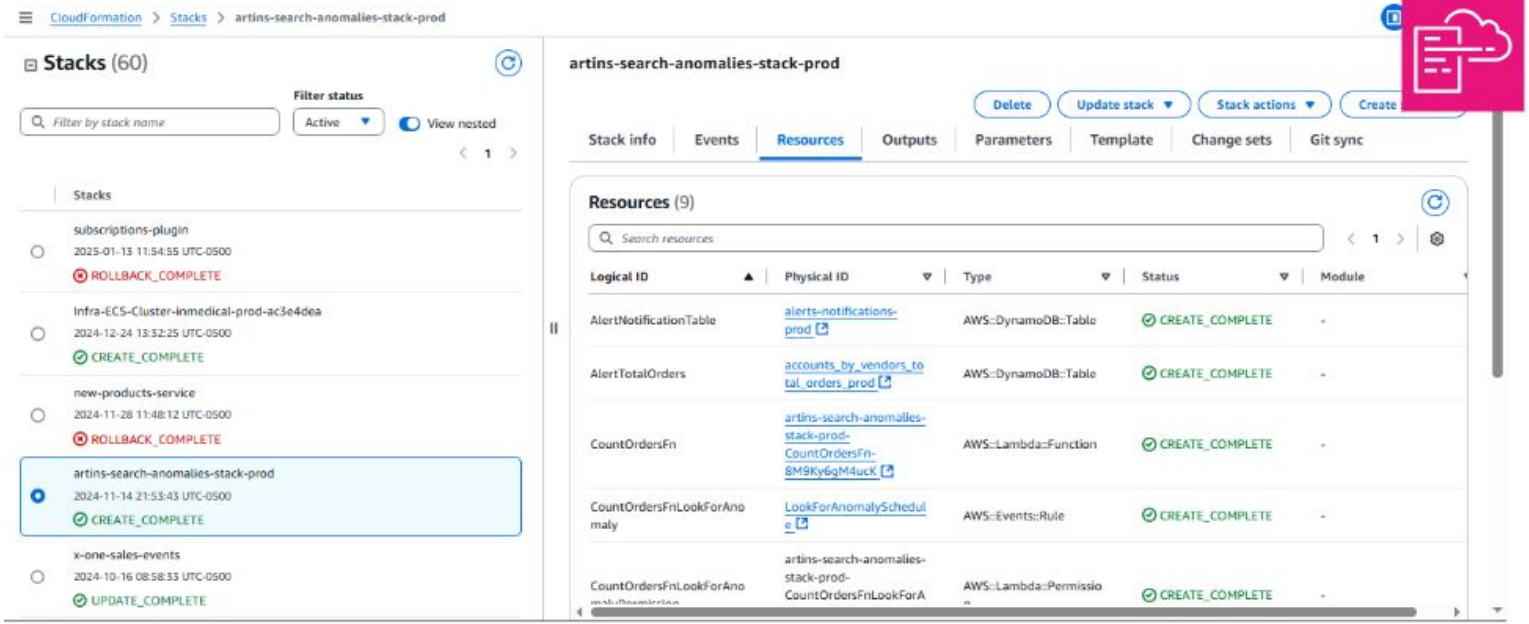


Servicios que nos pueden ayudar



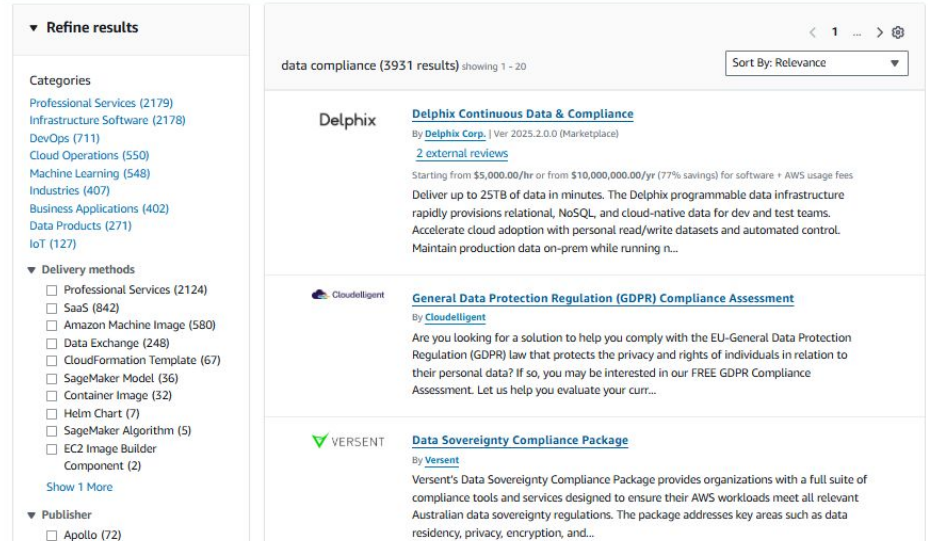
Servicios que nos pueden ayudar

Cloudformation



Siempre hay opciones disponibles

AWS Marketplace



Crealo la solución

TRD Technology Research & Data

Conclusión y moraleja

Conclusión

“Los principios son normas para hacer las cosas bien en todo momento dentro de tu vida profesional”

¡Me presento!

Community Top Voice

Carlos Fernando Chicata

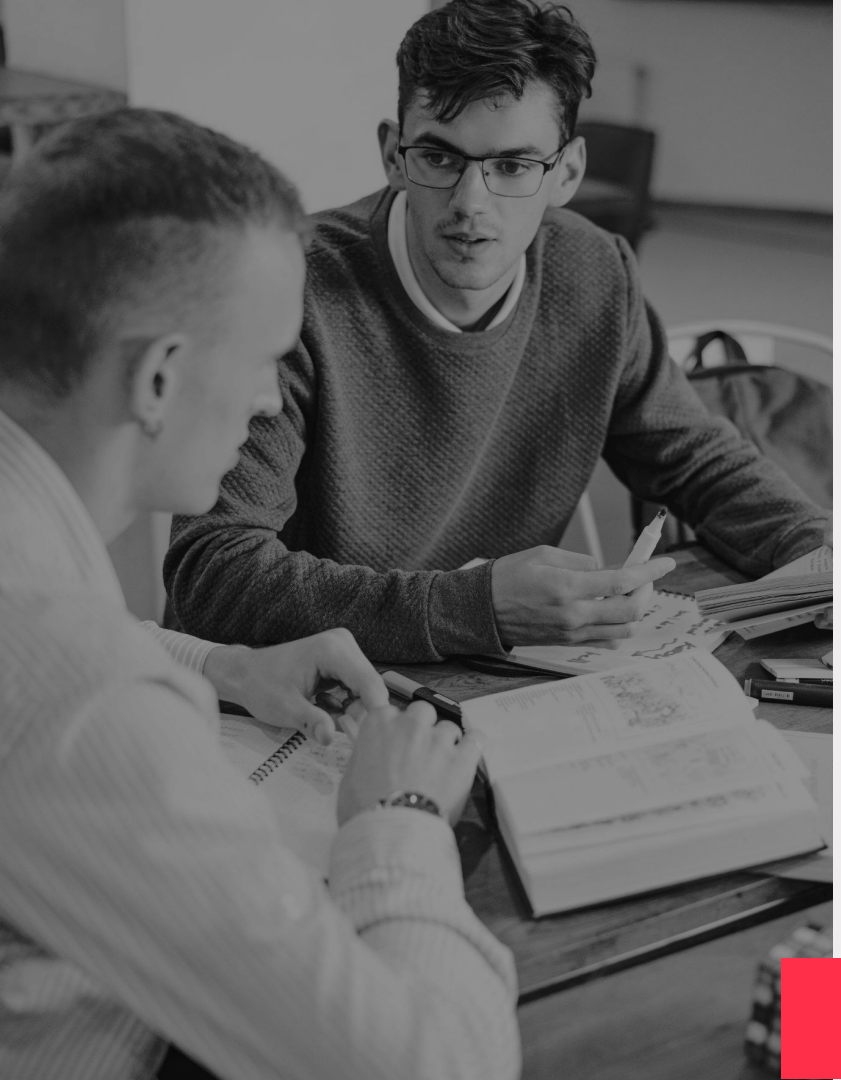
Top Data Engineering Voice

People on LinkedIn find **Carlos Fernando Chicata** an insightful contributor
8 skills.

- Data Engineering
- Data Warehousing
- Data Architecture
- Data Modeling
- Data Governance
- Data Privacy

Empecemos hoy

Contáctame por linkedin



¿Pregunta o dudas?

¡Gracias totales!

¡Referencias importantes!

1.- **S3 vs. EFS vs. FSx: Navegando por el laberinto del almacenamiento de AWS.**
Enlace:
https://medium.com/@sudhir_thakur/s3-vs-efs-vs-fsx-navigating-awss-storage-labyrinth-7d278b754dd9
2.-