



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?

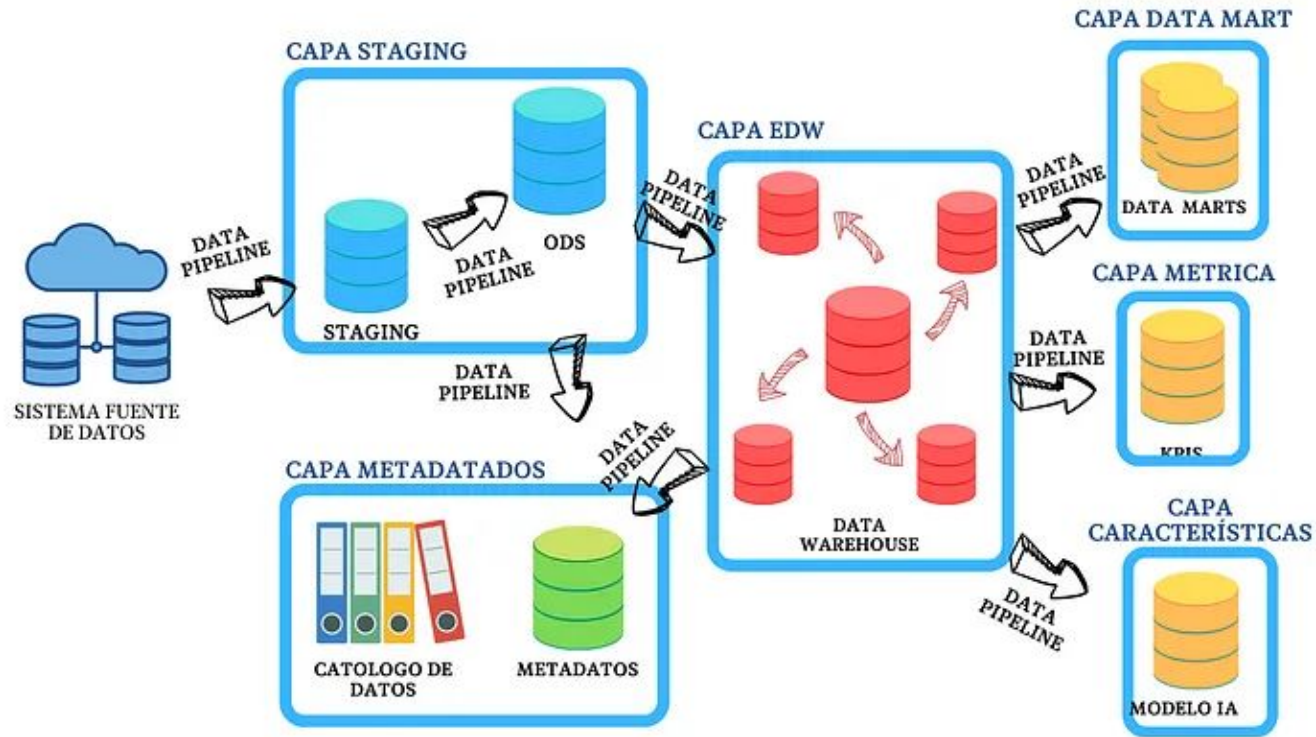
Lo más simple:

#1: “organizar los datos”.

#2: “Entender los requerimiento”.



La relación entre ambos



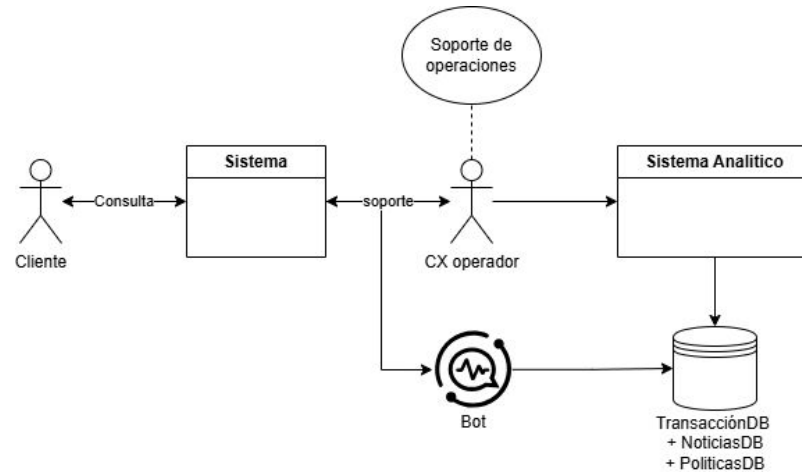
Facto #1: Naturaleza de los datos y solicitudes



La contexto de los datos analíticos

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.



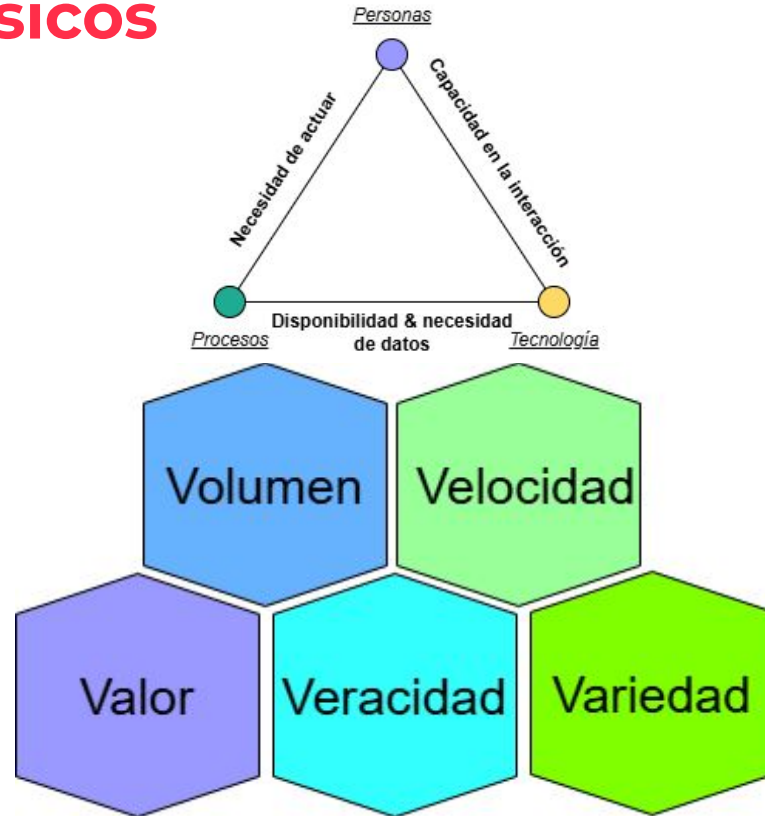
La naturaleza de los datos analiticos: los detalles fundamentales

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.

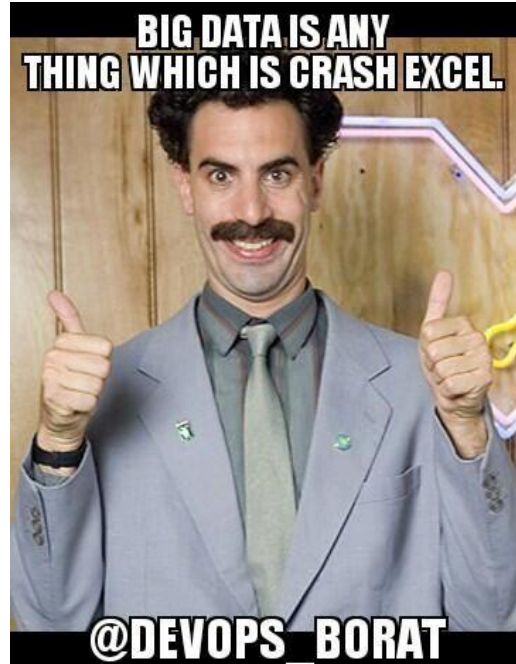


La naturaleza de los datos analiticos: los detalles basicos









El adjetivo big

Aquí tendremos problemas y conversaciones interesantes....



Situaciones en análisis de datos.

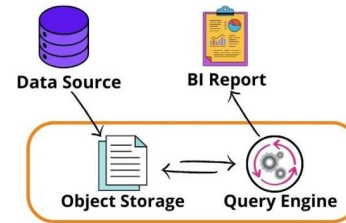
Hay diferentes situaciones que se abordan:

- 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

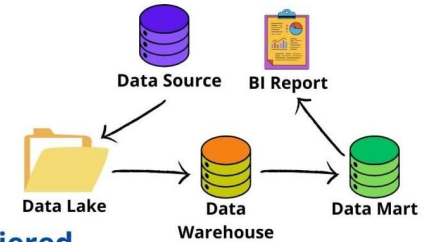
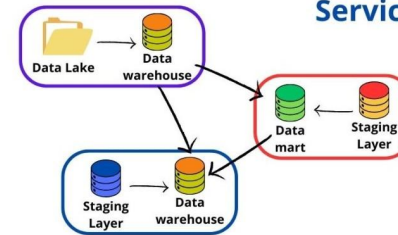
Realidad se fundamenta en:

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



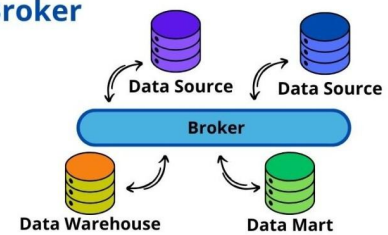
Layered

Service

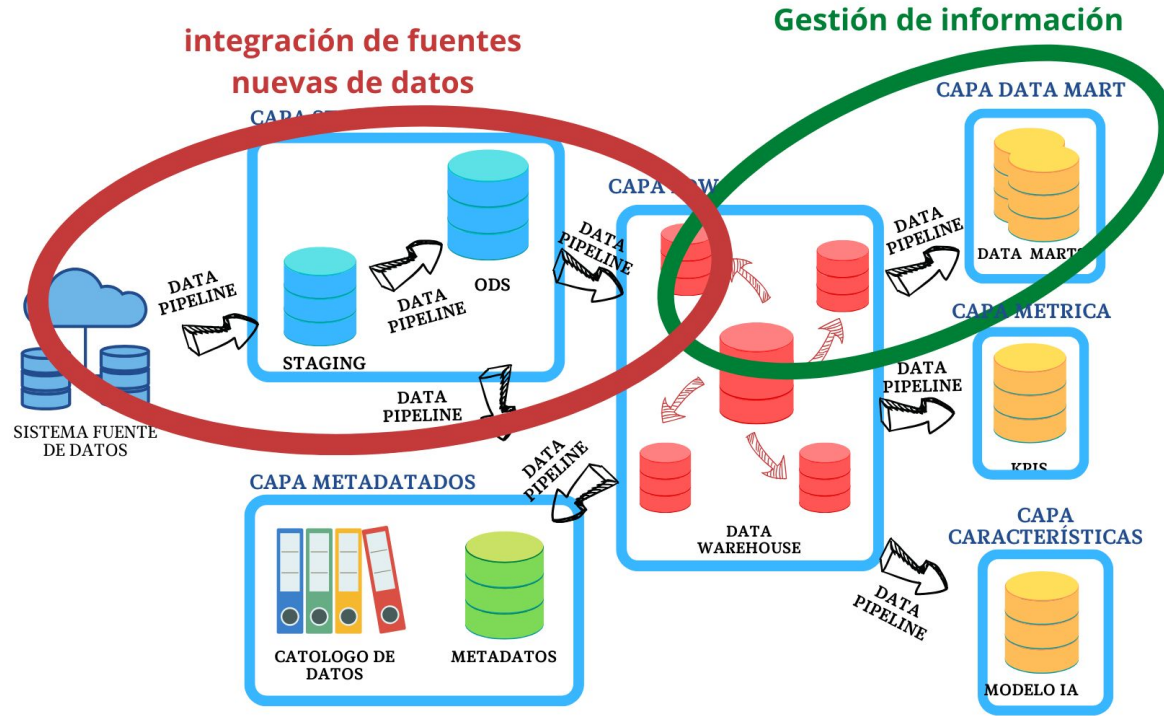


Tiered

Broker



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.



DO WE TRUST THIS DATA?

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.



S3



EFS



FSx

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.



Aurora



Redshift



RDS

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.



Lambda



ECS



Glue



EMR

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..



Aurora



Redshift



RDS

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.



DynamoDB

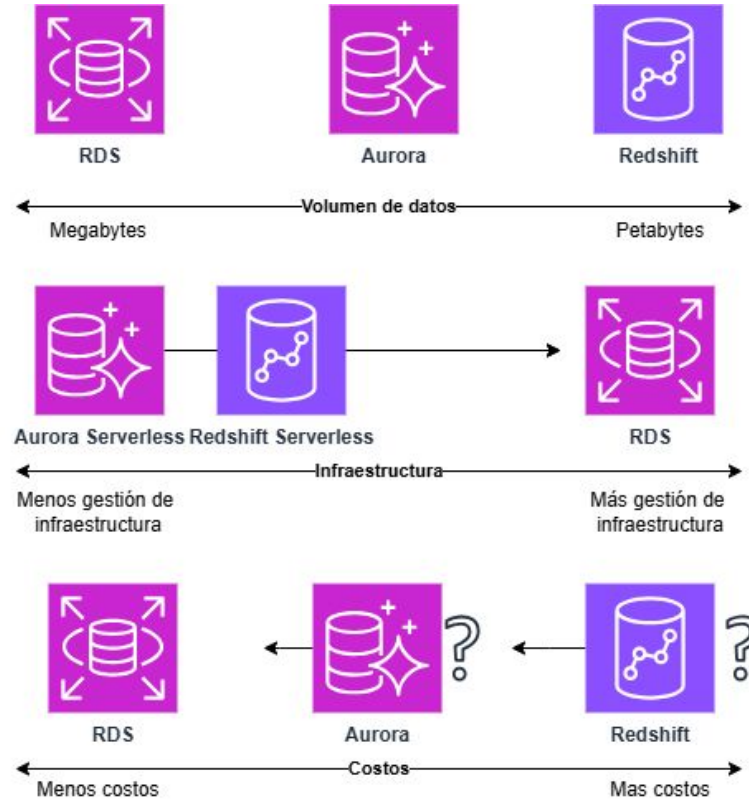


S3



MemoryDB

Criterios de evaluación



El aprendizaje es importante en todo momento



AWS Big Data Blog



Enrich your AWS Glue Data Catalog with generative AI metadata using Amazon Bedrock

by Manos Samantas and Anastasia Tzeveleka | on 15 NOV 2024 | in Amazon Bedrock, Amazon Machine Learning, Analytics, AWS Big Data, AWS Glue, Generative AI, Intermediate (200), Technical How-to | [Permalink](#) | [Comments](#) | [Share](#)

By harnessing the capabilities of generative AI, you can automate the generation of comprehensive metadata descriptions for your data assets based on their documentation, enhancing discoverability, understanding, and the overall data governance within your AWS Cloud environment. This post shows you how to enrich your AWS Glue Data Catalog with dynamic metadata using foundation models (FMs) on Amazon Bedrock and your data documentation.



How FINRA established real-time operational observability for Amazon EMR big data workloads on Amazon EC2 with Prometheus and Grafana

by Sumalatha Bachu, PremKiran Bejjam, and Akhil Chalamalasetty | on 15 NOV 2024 | in AWS Big Data, Customer Solutions, Management & Governance, Monitoring and observability | [Permalink](#) | [Comments](#) | [Share](#)

FINRA performs big data processing with large volumes of data and workloads with varying instance sizes and types on Amazon EMR. Amazon EMR is a cloud-based big data environment designed to process large amounts of data using open source tools such as Hadoop, Spark, HBase, Flink, Hudi, and Presto. In this post, we talk about our challenges and show how we built an observability framework to provide operational metrics insights for big data processing workloads on Amazon EMR on Amazon Elastic Compute Cloud (Amazon EC2) clusters.



Your guide to AWS Analytics at AWS re:Invent 2024

by Imtiaz Sayed and Navnit Shukla | on 14 NOV 2024 | in AWS re:Invent | [Permalink](#) | [Comments](#) | [Share](#)

It's AWS re:Invent time, where you turn your ideas into reality. Get a front row seat to hear real stories from AWS customers, experts and leaders about navigating pressing topics like generative AI and data analytics. For data enthusiasts and data professionals alike, this blog is a curated and comprehensive guide to all analytics sessions, for you to efficiently plan your itinerary.

El aprendizaje es importante en todo momento



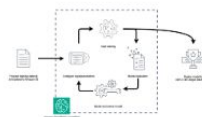
AWS Machine Learning Blog



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

by Cassandre Vandoputte, 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 Yu Le, Jaya Padma Murta, 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](#)

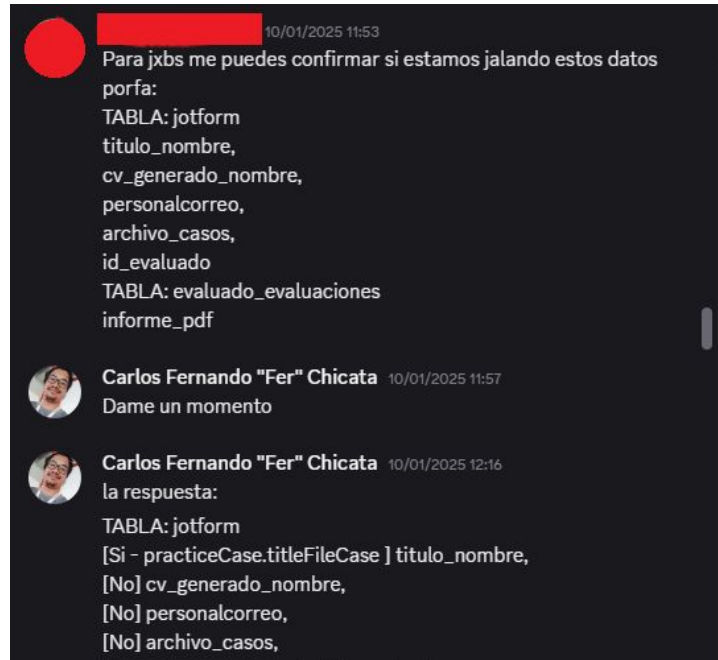
In this post, we present an LLM migration paradigm and architecture, including a continuous process of model evaluation, prompt generation using Amazon Bedrock, and data-aware optimization. The solution evaluates the model performance before migration and iteratively optimizes the Amazon Nova model prompts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



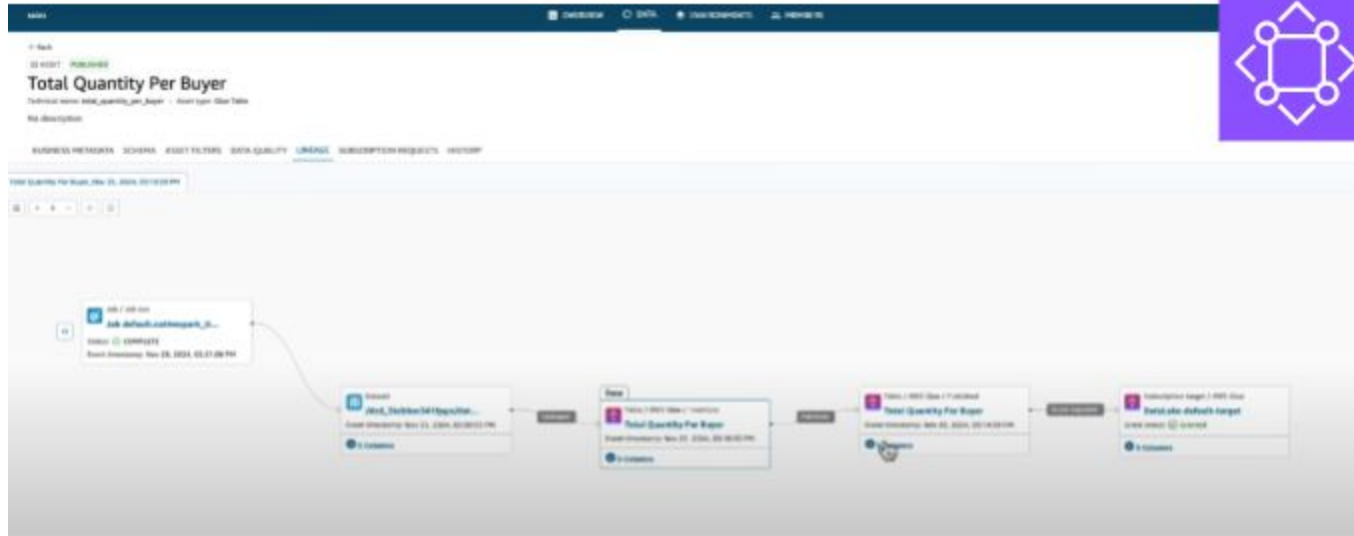
Tipos de documentación necesaria

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

Glue Quality



ETL Job

Last modified on 12/20/2023, 12:01:07 AM Try new UI Actions Save

Visual Script Job details Run Data quality - updated Schedules Version Control

+ Add nodes

Search sources, transforms and targets

Sources Transforms Targets Popular

- Change Schema
Change field names, data types and drop fields. Formerly known as Apply Mapping.
- Join
Combine records from two datasets based on a set of conditions.
- SQL Query
Use a SQL query to transform data.
- Detect Sensitive Data
Detect PII and other sensitive information.
- Evaluate Data Quality
Evaluate the quality and completeness of your data.
- Fill Missing Values
Have AWS Glue infer reasonable values for missing data.
- Aggregate
Apply functions like sum or average to fields in the dataset.
- Custom Transform
Write custom code to transform data.

Add Transforms

Data source - Data Catalog
AWS Glue Data Catalog

Transform - Detect PII
Detect Sensitive Data

Transform - Evaluate data ...
Evaluate Data Quality

Transform

Choose one or more parent node

Detect Sensitive Data
PIIDetection - Transform

Ruleset editor Anomaly detection - new

Ruleset editor info

Helper

Rule types Schema

Search rule types

- AggregateMatch
- ColumnCorrelation
- ColumnCount
- ColumnDataType
- ColumnExists

1 # Example rules: Completeness "colA" =
2 between 0.4 and 0.8, ColumnCount
3 > 10
4]

Unsaved job found
We found an unsaved job, do you wish to restore it?

Servicios que nos pueden ayudar

Cloudformation



CloudFormation > Stacks > artins-search-anomalies-stack-prod

Stacks (60)

Filter by stack name

Filter status: Active View nested

Stacks

- subscriptions-plugin
2025-01-13 11:54:55 UTC-0500
ROLLBACK_COMPLETE
- Infra-ECS-Cluster-inmedical-prod-ac3e4dea
2024-12-24 13:52:25 UTC-0500
CREATE_COMPLETE
- new-products-service
2024-11-28 11:48:12 UTC-0500
ROLLBACK_COMPLETE
- artins-search-anomalies-stack-prod**
2024-11-14 21:53:43 UTC-0500
CREATE_COMPLETE
- x-one-sales-events
2024-10-16 08:58:33 UTC-0500
UPDATE_COMPLETE

artins-search-anomalies-stack-prod

Delete Update stack Stack actions Create

Stack info Events **Resources** Outputs Parameters Template Change sets Git sync

Resources (9)

Search resources

Logical ID	Physical ID	Type	Status	Module
AlertNotificationTable	alerts-notifications-prod	AWS::DynamoDB::Table	CREATE_COMPLETE	-
AlertTotalOrders	accounts_by_vendors_to tal_orders_prod	AWS::DynamoDB::Table	CREATE_COMPLETE	-
CountOrdersFn	artins-search-anomalies-stack-prod-CountOrdersFn-8M9Ky6gM4ucl	AWS::Lambda::Function	CREATE_COMPLETE	-
CountOrdersFnLookForAnomaly	LookForAnomalySchedule	AWS::Events::Rule	CREATE_COMPLETE	-
CountOrdersFnLookForAnomaly	artins-search-anomalies-stack-prod-CountOrdersFnLookForAnomaly	AWS::Lambda::Permission	CREATE_COMPLETE	-

Siempre hay opciones disponibles

AWS Marketplace

▼ **Refine results**

Categories

- Professional Services (2179)
- Infrastructure Software (2178)
- DevOps (711)
- Cloud Operations (550)
- Machine Learning (548)
- Industries (407)
- Business Applications (402)
- Data Products (271)
- IoT (127)

▼ **Delivery methods**

- ☐ Professional Services (2124)
- ☐ SaaS (842)
- ☐ Amazon Machine Image (580)
- ☐ Data Exchange (248)
- ☐ CloudFormation Template (67)
- ☐ SageMaker Model (36)
- ☐ Container Image (32)
- ☐ Helm Chart (7)
- ☐ SageMaker Algorithm (5)
- ☐ EC2 Image Builder Component (2)

Show 1 More

▼ **Publisher**

- ☐ Apollo (72)

data compliance (3931 results) showing 1 - 20

Sort By: Relevance

Delphix


[Delphix Continuous Data & Compliance](#)

By [Delphix Corp.](#) | Ver 2025.2.0.0 (Marketplace)

[2 external reviews](#)


Starting from \$5,000.00/hr or from \$10,000,000.00/yr (77% savings) for software + AWS usage fees

Deliver up to 25TB of data in minutes. The Delphix programmable data infrastructure rapidly provisions relational, NoSQL, and cloud-native data for dev and test teams. Accelerate cloud adoption with personal read/write datasets and automated control. Maintain production data on-prem while running n...

 **General Data Protection Regulation (GDPR) Compliance Assessment**

By [Cloudelligent](#)

Are you looking for a solution to help you comply with the EU-General Data Protection Regulation (GDPR) law that protects the privacy and rights of individuals in relation to their personal data? If so, you may be interested in our FREE GDPR Compliance Assessment. Let us help you evaluate your curr...

 **Data Sovereignty Compliance Package**

By [Versent](#)

Versent's Data Sovereignty Compliance Package provides organizations with a full suite of compliance tools and services designed to ensure their AWS workloads meet all relevant Australian data sovereignty regulations. The package addresses key areas such as data residency, privacy, encryption, and...

Crealo la solución



Glue



Bedrock



Step Function



Lambda



NeptuneDB



Lake Formation



S3



Event Bridge

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 & 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.-