
datapath

Presentación

datapath

Kremlin Stwart Huaman Santos.
Lic. en Ingeniería de software de la UNMSM.
Senior Data Engineer
Big Data Engineer - Interbank

- Se desempeña como Big Data Software Engineer. Neo4J Certified Professional. Cuenta con certificaciones y especializaciones en Azure, Spark, Python. Cuento con +5 años de experiencia en TI y +3 años de experiencia en Data Engineering.

- Soy un apasionado estratega en tecnologías de la información, con una trayectoria de más de una década dedicada a la excelencia y la innovación en el sector TI.

https://www.linkedin.com/in/khuamans/

Agenda

01
Reglas de la clase

02
Modo de evaluación

03
—---

04
—---

05
—------

06
—--------

Reglas de clases

- Mantener el micrófono apagado en caso no vayan a hablar.
- Preguntar en caso que tengan dudas
- Mantenerse atento a la clase.

Modo de evaluación y Producto de la sesión

Evaluación continua

Ejercicios, challenges y/o test.

Producto de la sesión

Al cierre de la sesión todos los estudiantes deben tener instalado MYSQL

datapath

Módulo 1

Data Analytics Fundamentals

www.datapath.ai

datapath

Sesión 2

El Papel Fundamental del Ingeniero de Datos en Business Intelligence

www.datapath.ai

datapath

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- **Modelado de datos**
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

Ingesta de datos

ETL/ELT

Monitoreo y Mantenimiento

Almacenamiento de datos

Automatización y Orquestación

Modelado de datos

Colaboración interdiciplinaria

Calidad de datos

Optimización del rendimiento

Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

Ingesta de datos

ETL/ELT

Monitoreo y Mantenimiento

Almacenamiento de datos

Automatización y Orquestación

Modelado de datos

Colaboración interdiciplinaria

Calidad de datos

Optimización del rendimiento

Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- **Colaboración interdiciplinaria**
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

- Ingesta de datos
- ETL/ELT
- Monitoreo y Mantenimiento
- Almacenamiento de datos
- Automatización y Orquestación
- Modelado de datos
- Colaboración interdiciplinaria
- Calidad de datos
- Optimización del rendimiento
- Seguridad y Gobernanza

www.datapath.ai

Ingeniero de datos

datapath

Ingesta de datos

ETL/ELT

Monitoreo y Mantenimiento

Almacenamiento de datos

Automatización y Orquestación

Modelado de datos

Colaboración interdiciplinaria

Calidad de datos

Optimización del rendimiento

Seguridad y Gobernanza

www.datapath.ai

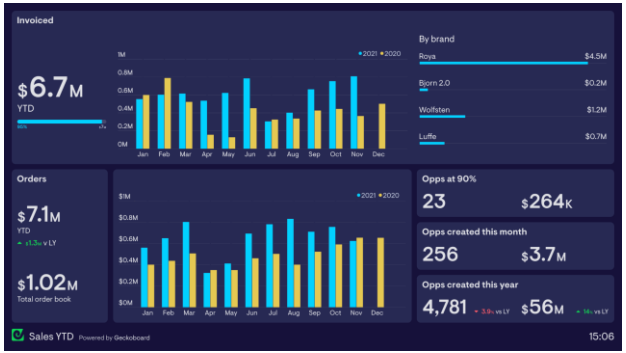
Herramientas - Tecnologias

datapath

- Lenguajes de Programación: Python, SQL, Scala, Java.
- Sistemas de Almacenamiento: SQL Server, PostgreSQL, MongoDB, Hadoop, Spark.
- Herramientas ETL/ELT: Talend, Informatica, Apache Nifi, Airflow.
- Cloud Platforms: AWS, Google Cloud, Azure.
- Herramientas de Data Warehousing: Redshift, BigQuery, Snowflake.

Contribución BI

- Generación de Informes
- Análisis Avanzados
- Toma de Decisiones



www.datapath.ai

datapath

¿Qué es Big Data?

Interfaz de usuario

datapath

"Big data son datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto se conoce como "las tres V".

Gartner's - Doug Laney

Las V´s de Big Data

datapath

Es un marco de trabajo **(conceptos + tecnologías)** que permite procesar grandes volúmenes de datos, de diferentes estructuras o con carencia de estas, que pueden variar en el tiempo, a grandes velocidades y que generen valor al negocio.

Principios de un DataLake

datapath

Arquitectura desacoplada

Arquitectura de capa con componentes modulares que aíslan el almacenamiento y el cómputo, minimizando los riesgos de pérdida de datos

Persistencia Polygot

Es el principio de una variedad de tecnologías de procesamiento y almacenamiento especialmente diseñadas para resolver distintos casos de uso del negocio **(streaming, nosql)**

Serveless

Aplicación de distintas tecnologías sin servidor (IaaS o PaaS) permitiendo reducir el costo y la carga operativa de la plataforma.

Autoservicio

Un datalake debe ser fácil de usar y de autoservicio para todos los interesados. Almacenamiento de datos con SQL vistas comerciales para BI operativo, así como datos RAW catalogados para exploración y descubrimiento.

Organización de un DataLake

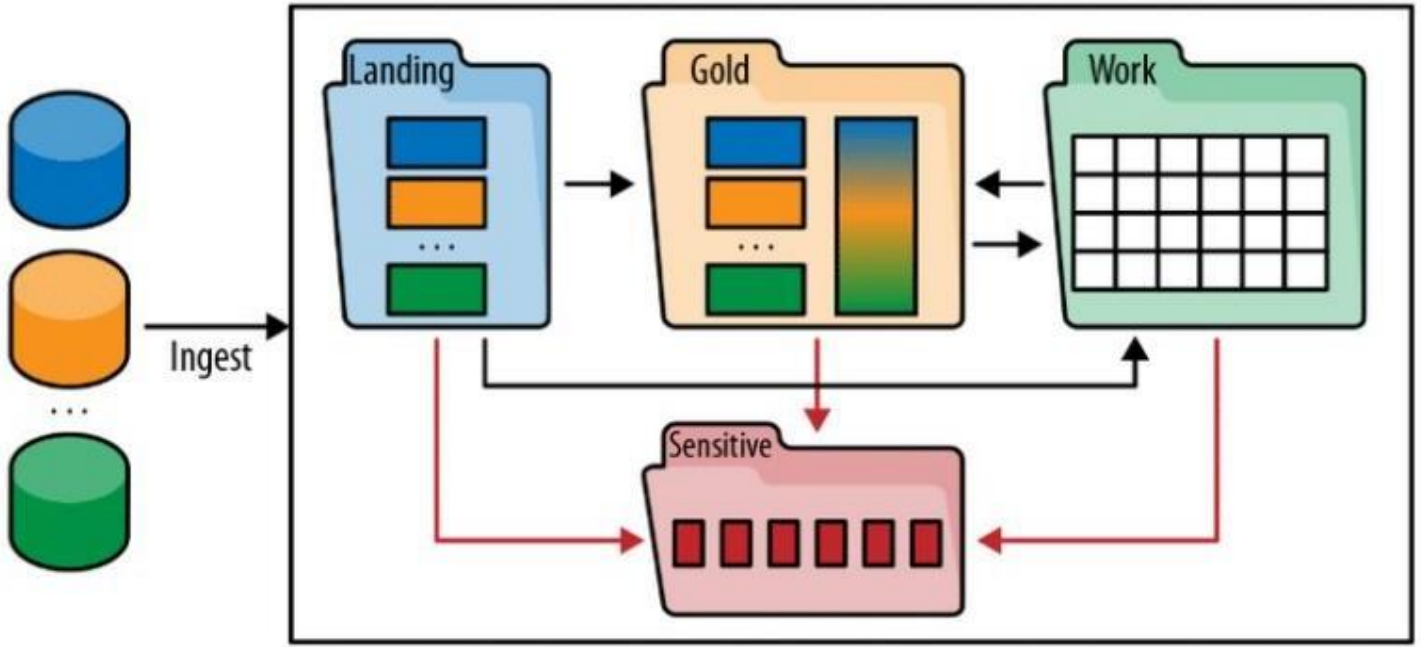


Figure 7-1. Sample breakdown of data lake into workspaces or zones

www.datapath.ai

Zona Landing - Raw

La zona **landing**, algunas veces también llamada raw o zona staging, es usada para almacenar los datos crudos. Por lo general, solo los desarrolladores altamente técnicos, ingenieros de datos y científicos de datos tienen acceso a la zona landing. En general, los usuarios de la zona landing deben tener una razón convincente para realizar su propio tratamiento y procesamiento de datos.

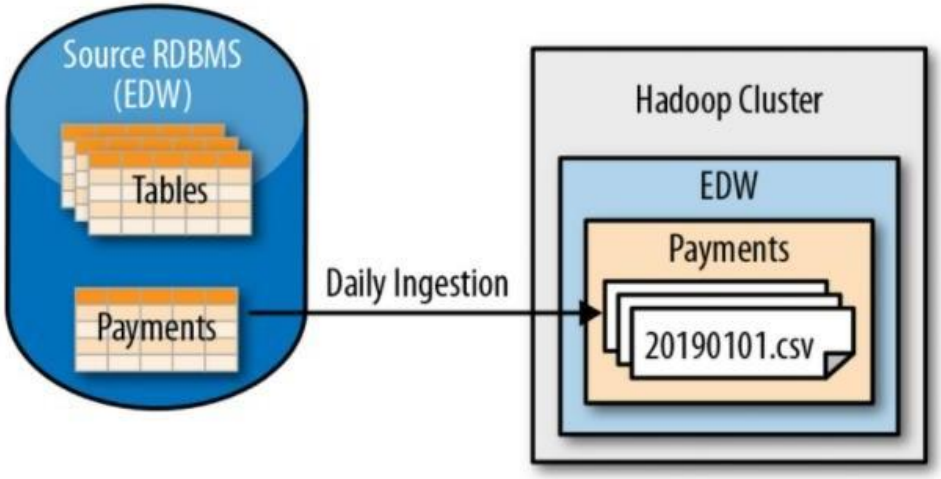
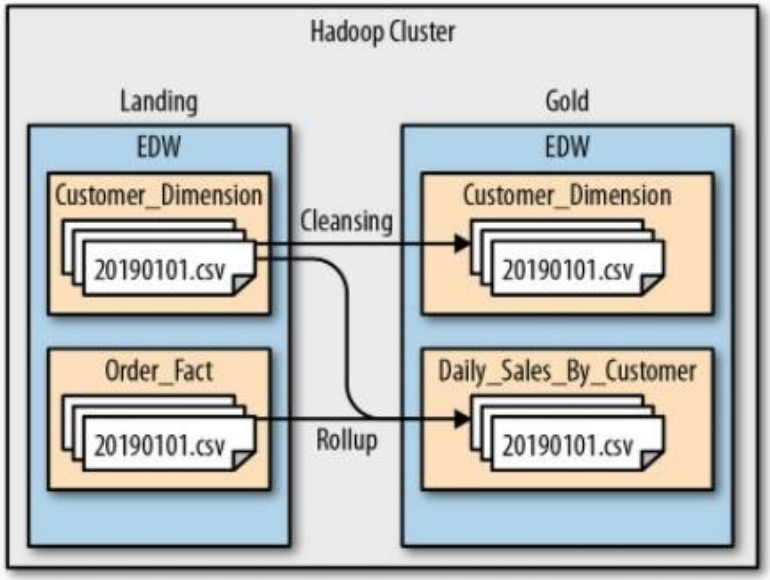


Figure 7-3. Ingesting raw or landing data into folders and files

Zona Gold

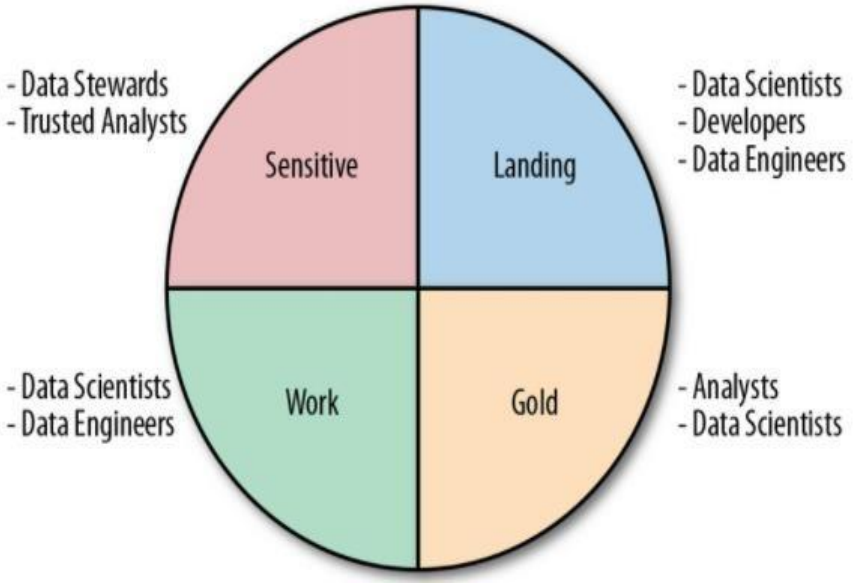
La zona **Gold** frecuentemente es espejo de la zona landing, pero contiene datos limpios y enriquecidos con alguna otra fuente del landing. Esta zona es ocasiones es llamada prod que indica que los datos que contienen son productivos y no tienen problemas de calidad. Usualmente este es la zona más popular. Muchos desarrolladores y data scientist tienen acceso a esta zona ya que tienen los datos mucho más limpios y estandarizados.



Organización de un DataLake

La mayor parte del análisis ocurre en la zona de **Work**, también conocida como zona de desarrollo o proyectos. Esta zona generalmente está estructurada para reflejar la estructura organizativa de la empresa. Normalmente es el dominio de desarrolladores, científicos de datos e ingenieros de datos, aunque los analistas a menudo lo utilizan para realizar la preparación de datos de autoservicio para sus proyectos.

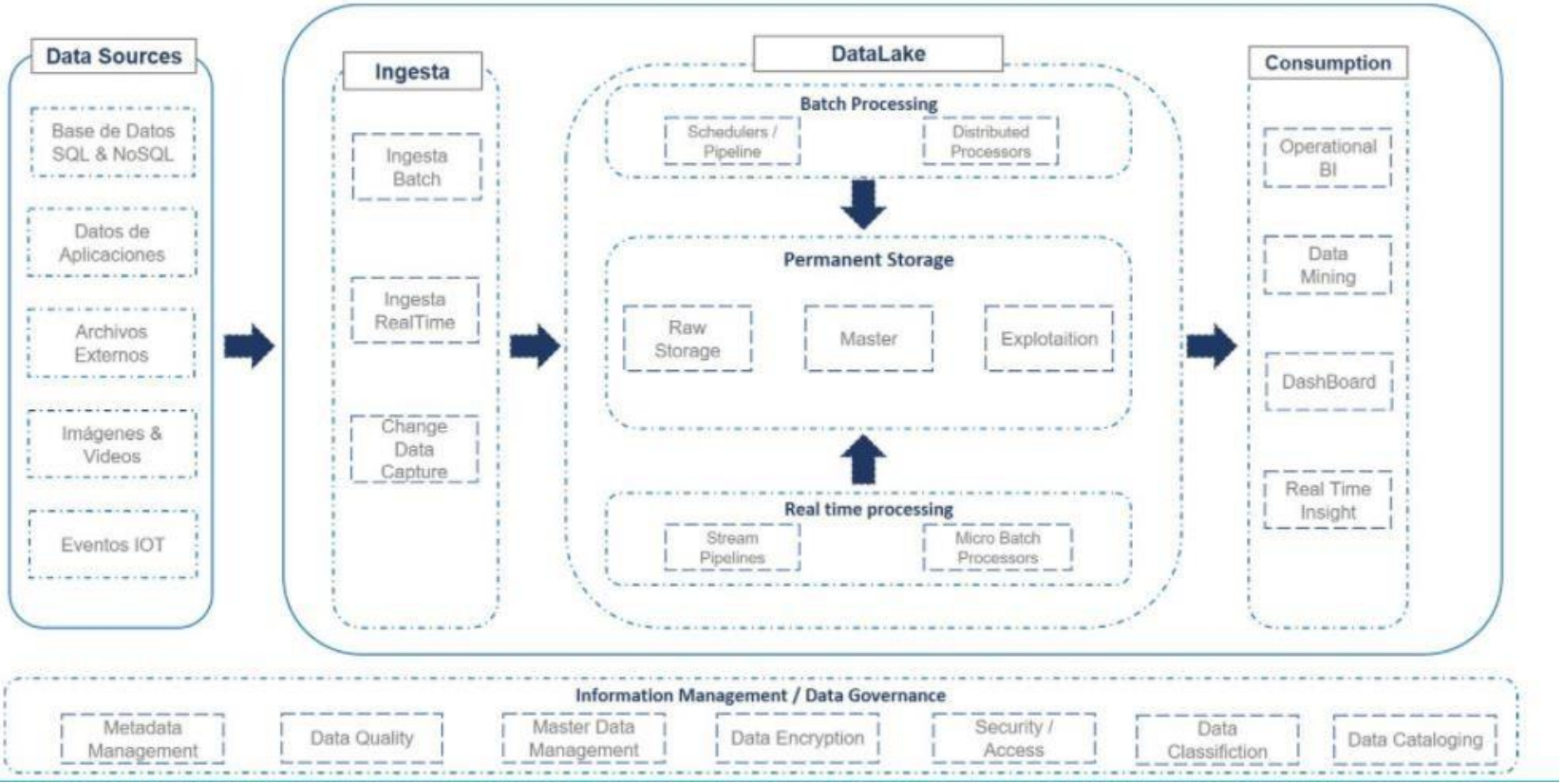
La mayor parte del análisis ocurre en la zona de **Work**, también conocida como zona de desarrollo o proyectos. Esta zona generalmente está estructurada para reflejar la estructura organizativa de la empresa. Normalmente es el dominio de desarrolladores, científicos de datos e ingenieros de datos, aunque los analistas a menudo lo utilizan para realizar la preparación de datos de autoservicio para sus proyectos.



www.datapath.ai

Plataforma de Datos

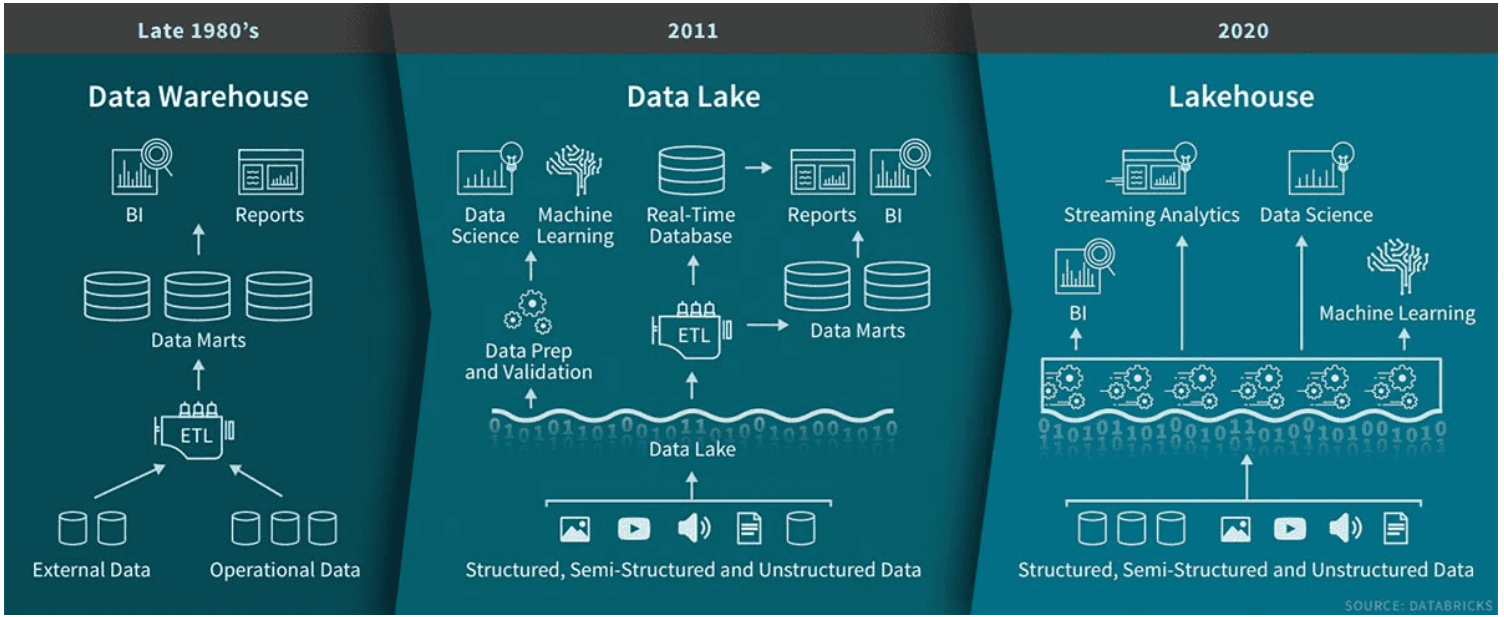
datapath



www.datapath.ai

Lakehouse

datapath



www.datapath.ai

Lakehouse

datapath

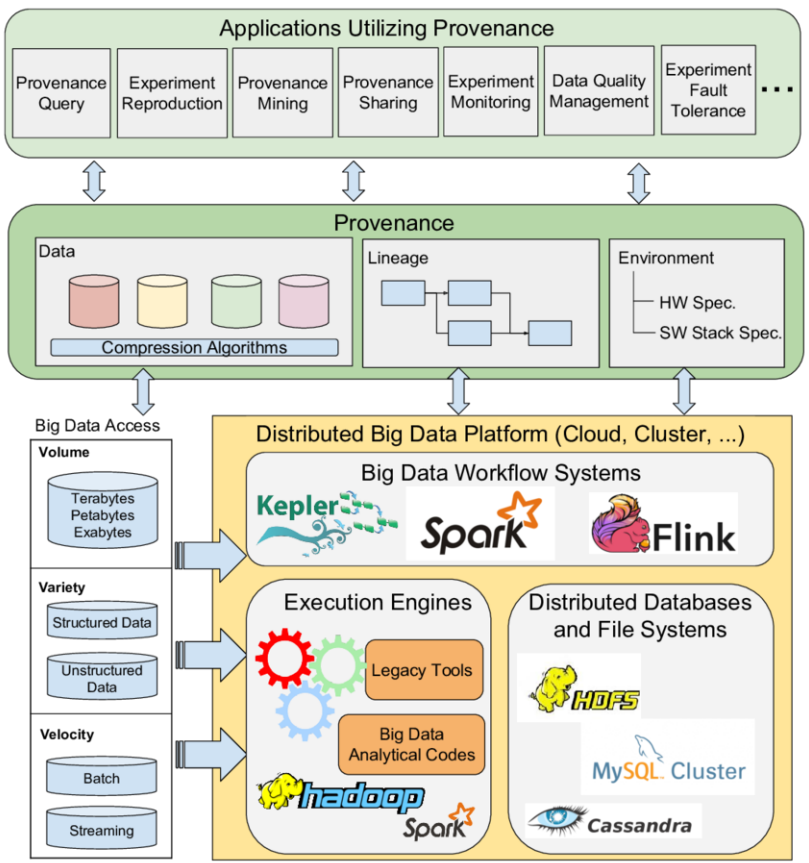
Característica	Data Warehouse	Data Lake	Lakehouse
Estructura	Altamente estructurado	Sin estructura definida	Estructurado y no estructurado
Transformación	ETL (previa al almacenamiento)	ELT (transformación post-almacenamiento)	ETL y ELT flexibles
Rendimiento	Optimizado para consultas SQL	Alta capacidad de almacenamiento	Consultas SQL eficientes y big data
Transacciones	ACID	No garantizadas	ACID
Casos de Uso	BI, análisis histórico	Análisis de big data, ML, IoT	BI, análisis en tiempo real, ML
Ejemplos	Amazon Redshift, Snowflake	Amazon S3, Azure Data Lake	Databricks Lakehouse Platform, Delta Lake

datapath

Arquitecturas de referencia

Open

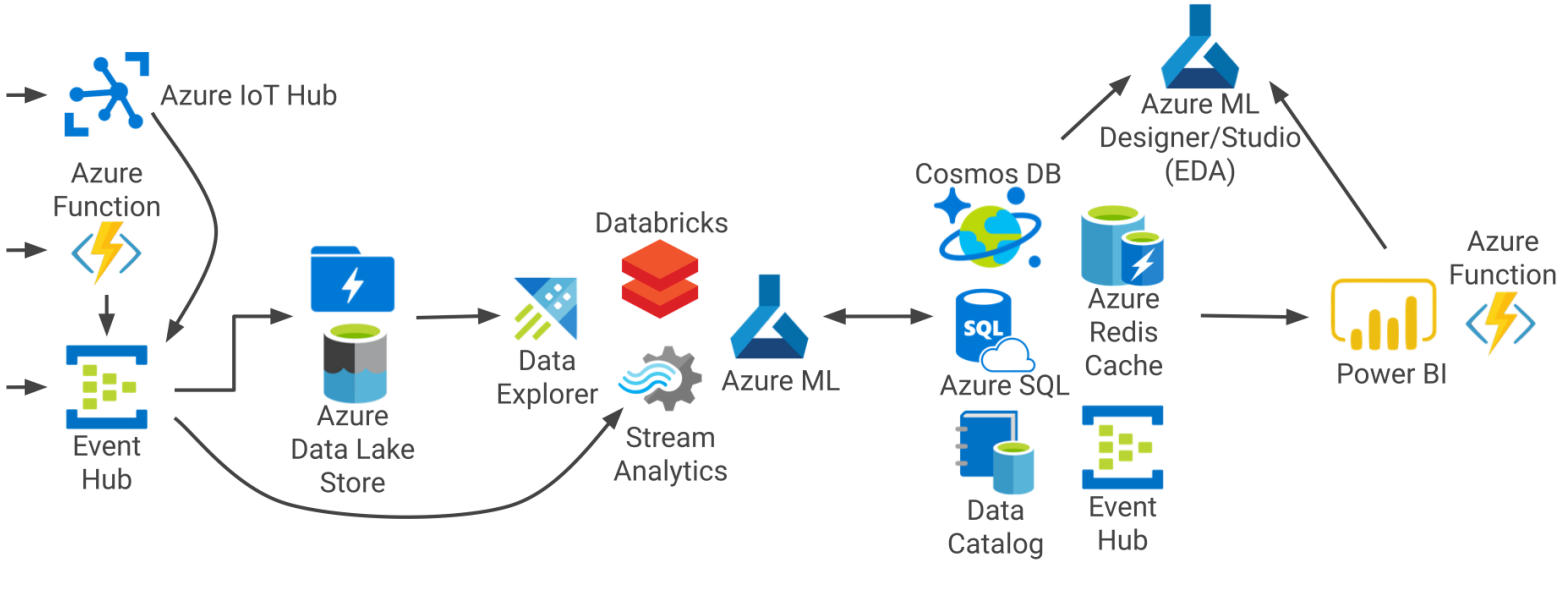
datapath



www.datapath.ai

Azure

datapath

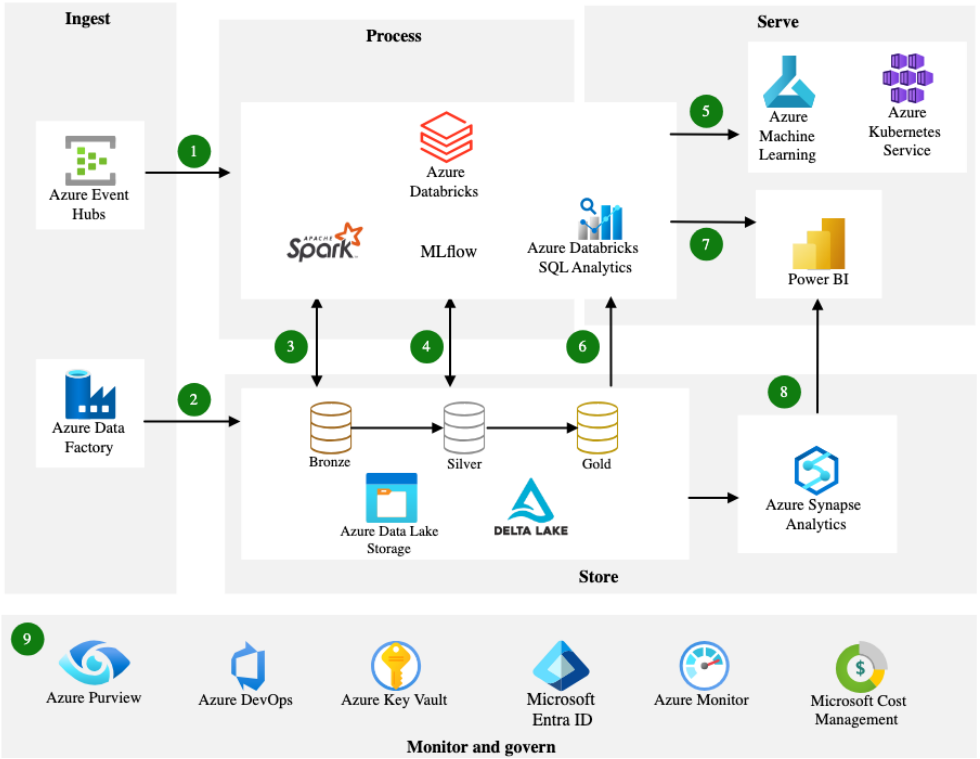


@scgupta

www.datapath.ai

Azure

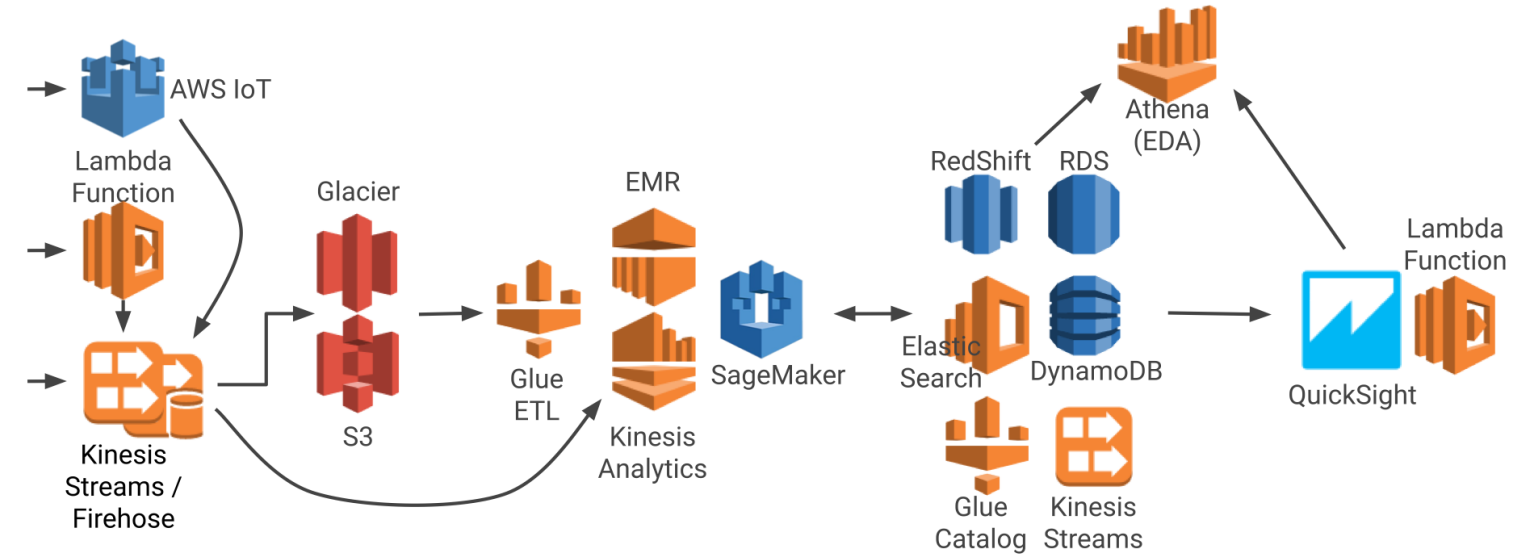
datapath



www.datapath.ai

AWS

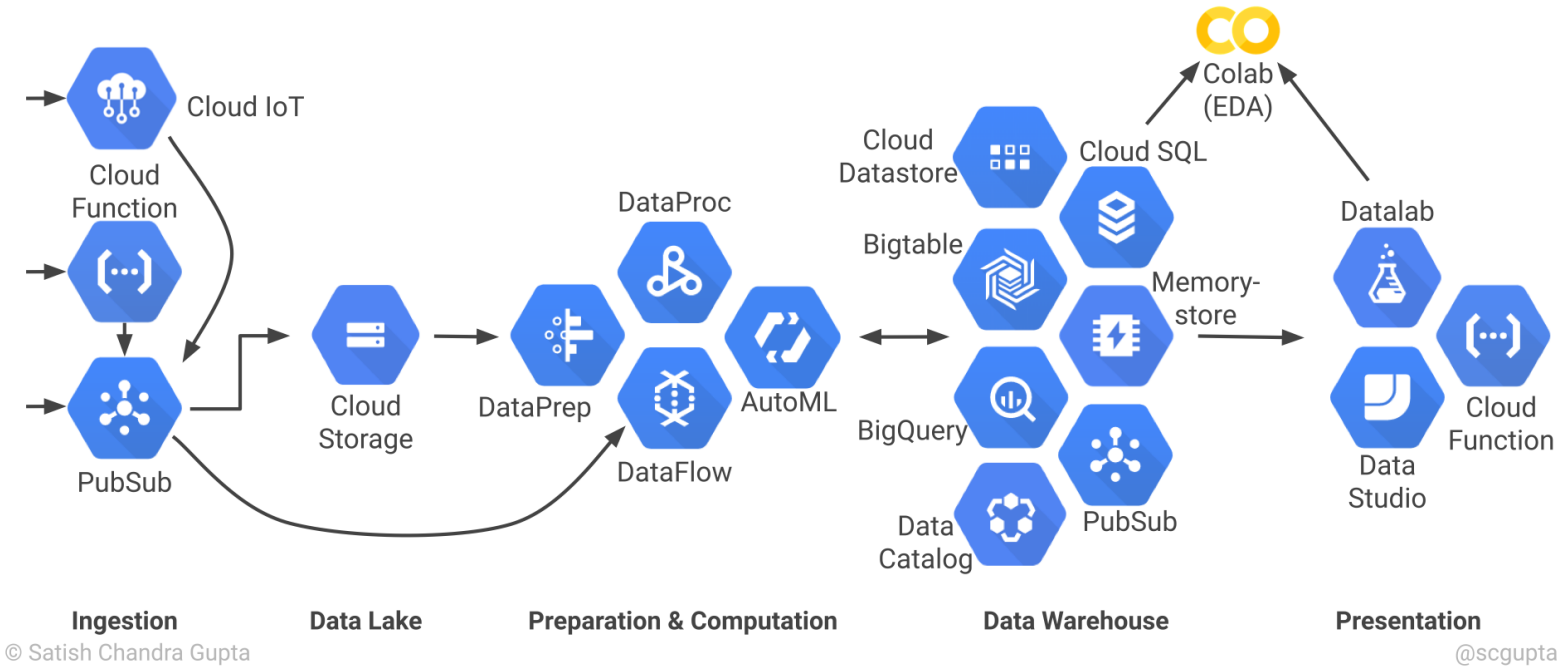
datapath



www.datapath.ai

GCP

datapath



www.datapath.ai

datapath

Instalación de MYSQL

www.datapath.ai

datapath

Manual de instalación : https://drive.google.com/file/d/1BnnrI7YAHGlh3s-Xv0M8XxUStkNWMX3f/view?usp=drive_link

www.datapath.a
i

datapath

¡Gracias!

Aprende, aplica y crece

www.datapath.ai