DIPLOMADO

Ciencia de Datos

Módulo 12

Datos masivos

Mtro. Omar Mendoza González

DGTIC

Universidad Nacional Autónoma de México

Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Contenido

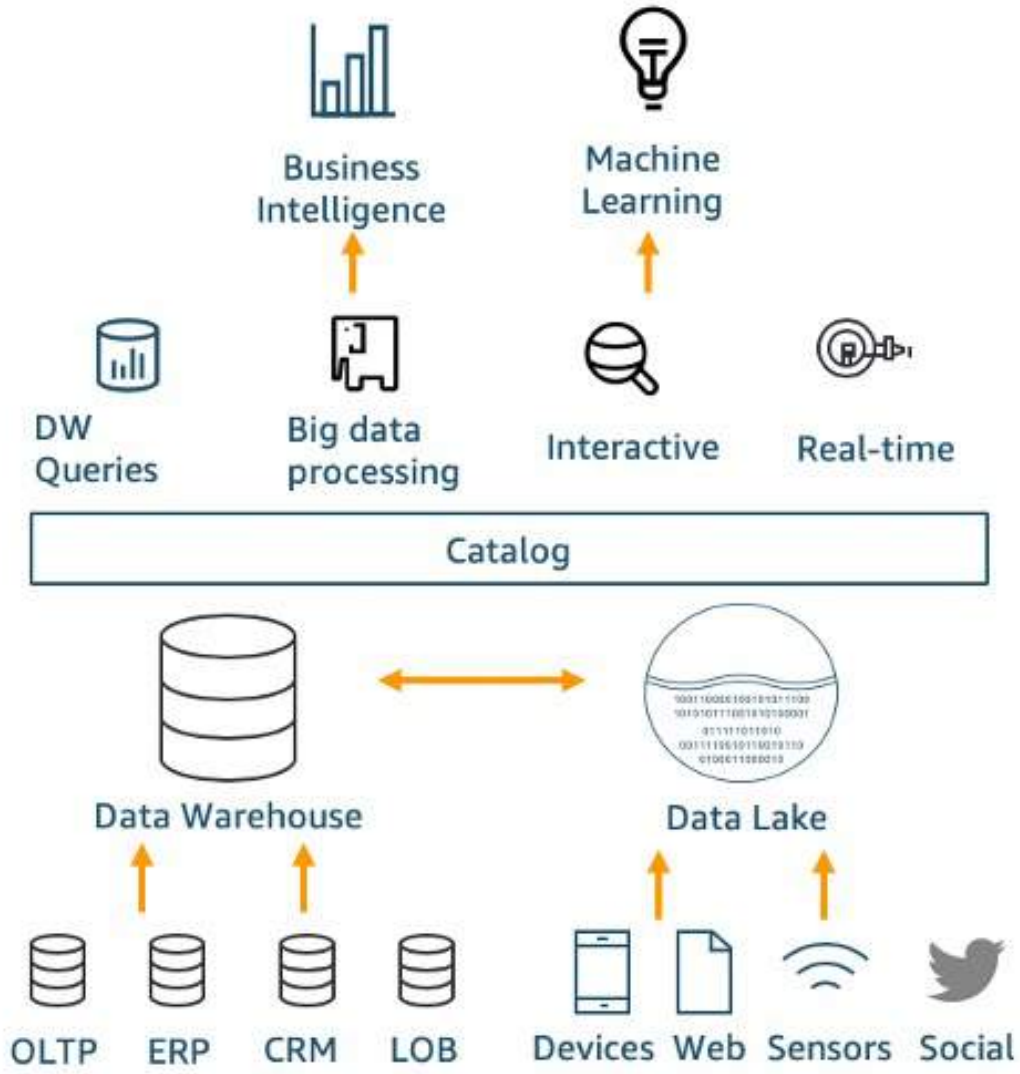
Ingesta de datos

- La **ingesta** de datos es el proceso que se usa para cargar los registros de datos de uno o varios orígenes a un data lake en Hadoop

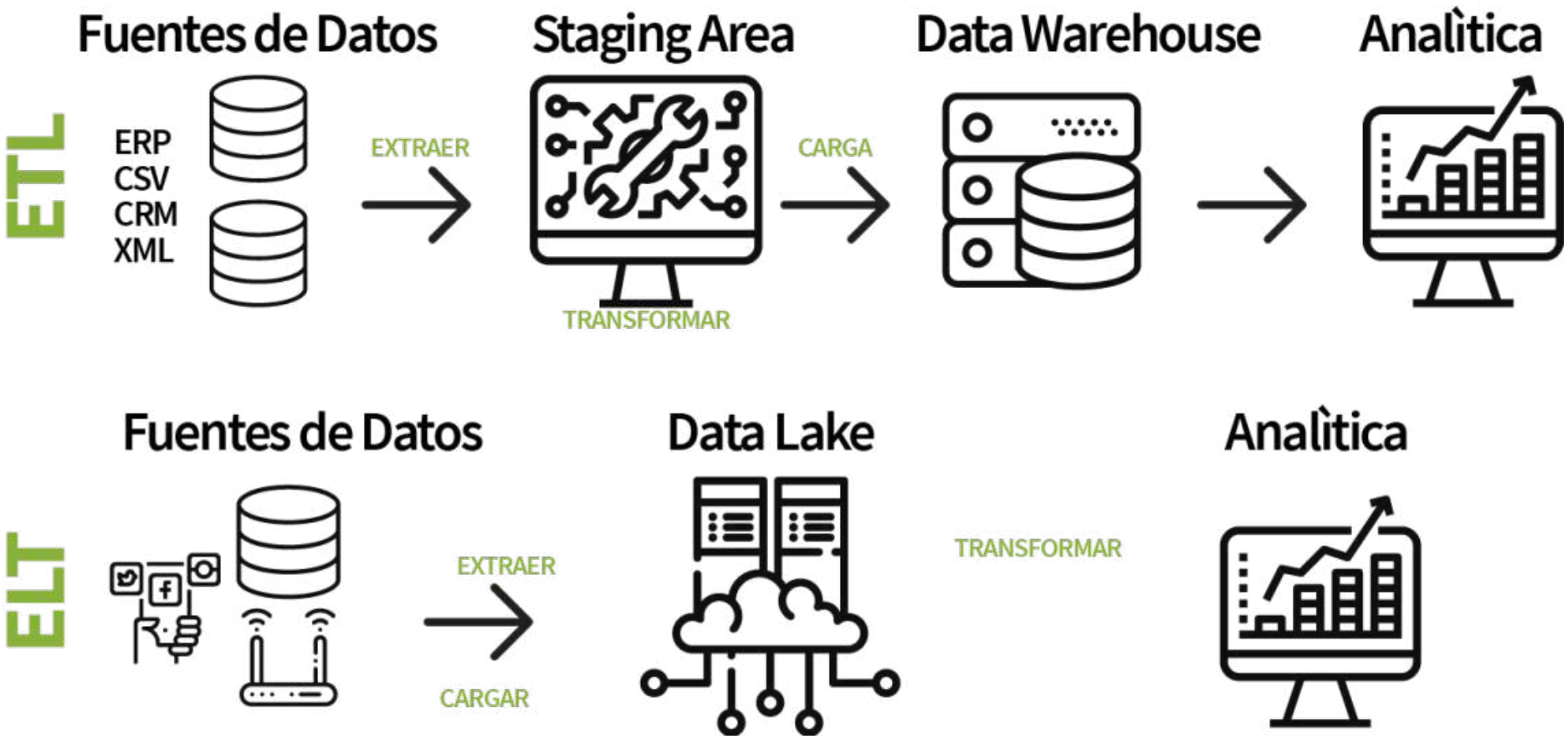
- Múltiples fuentes requieren múltiples formas de procesamiento

- Una vez ingeridos, los datos están disponibles para su consulta.

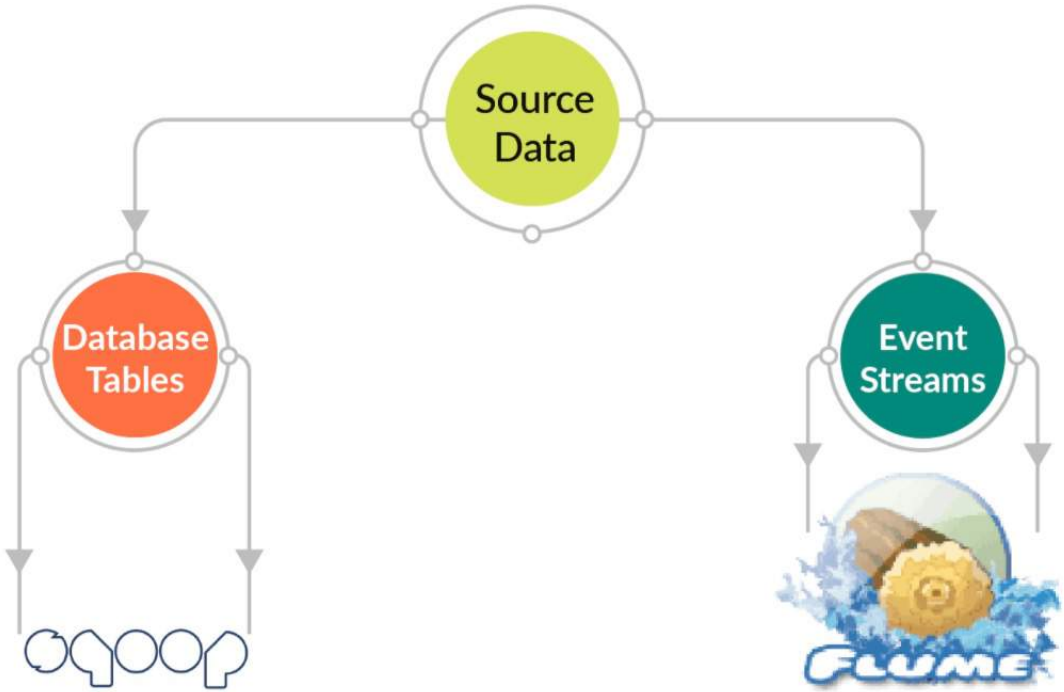
Ingesta de datos

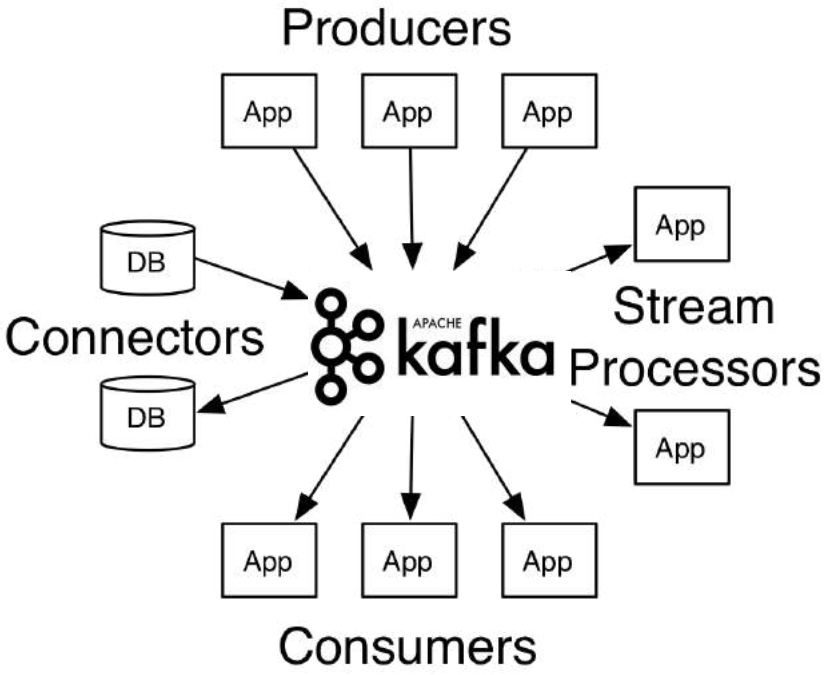


Ingesta de datos

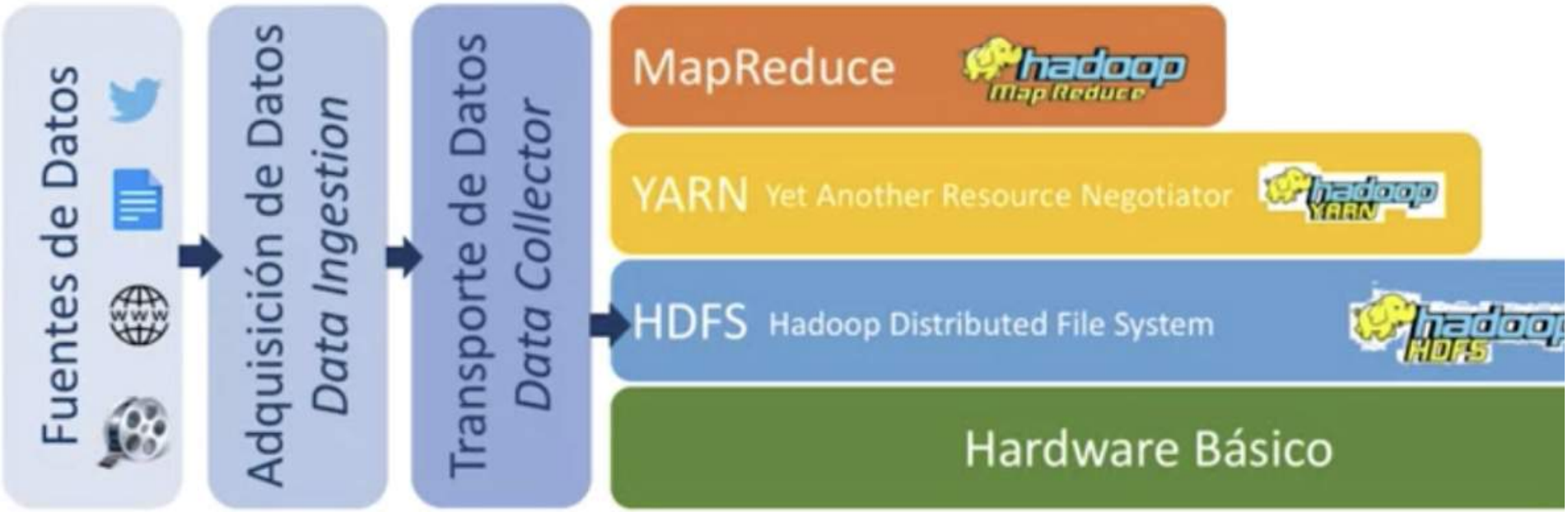


Ingesta de datos

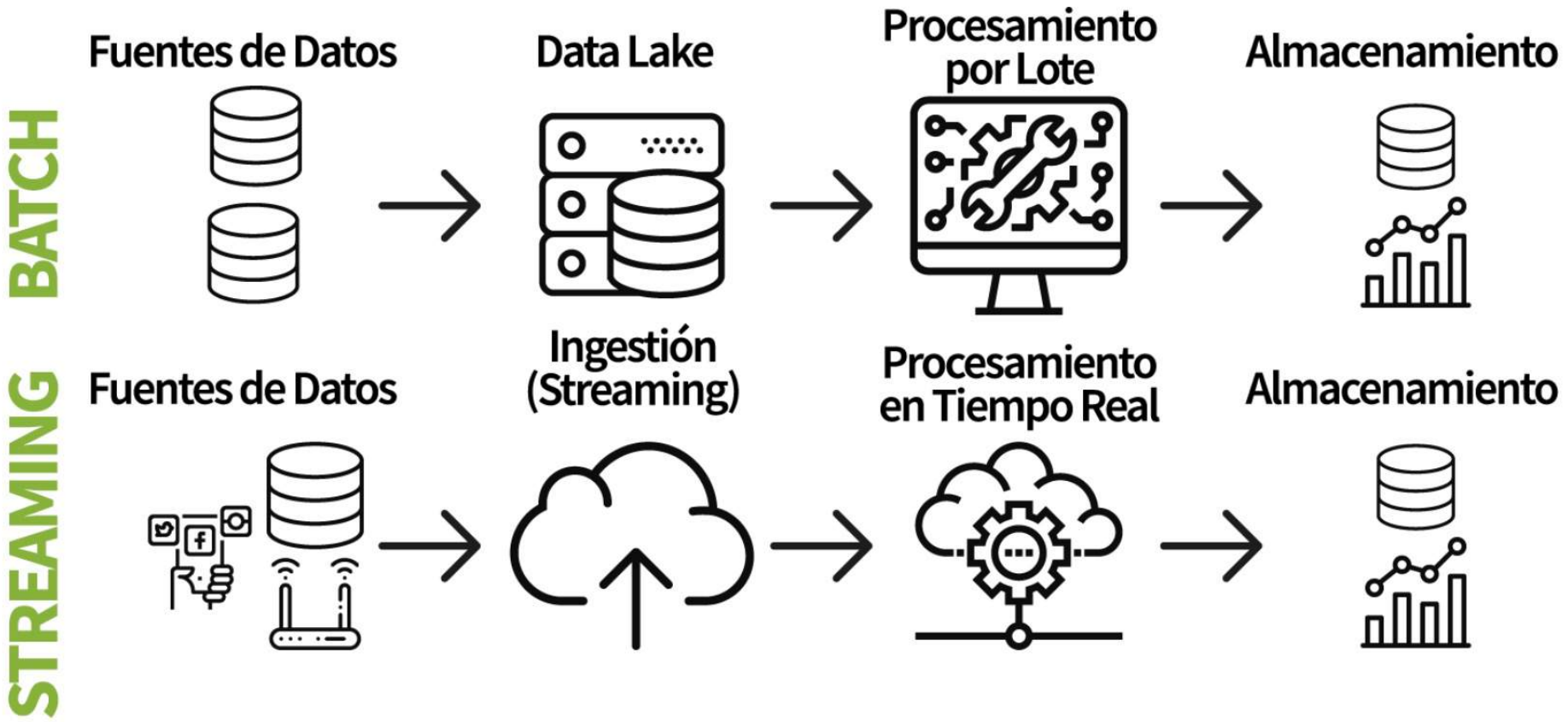




Ingesta de datos



Ingesta de datos



Ingesta de datos

Macro Batch	• > 15 min
Micro Batch	• 2 - 15 min
Near Real-Time Decision Support	• > 2 seg < 2 min
Near Real Time Event Processing	• > 100 ms < 2 seg
Real Time	• < 100 ms

Data Lake

- Respecto al Big Data se utiliza un repositorio de almacenamiento centralizado que contiene datos masivos de varias fuentes en un formato granular y sin procesar
- Generalmente se encuentra en la nube, en cluster de computadoras conformados de varios nodos
- Puede guardar datos estructurados, semiestructurados o no estructurados
- Es conocido como data lake

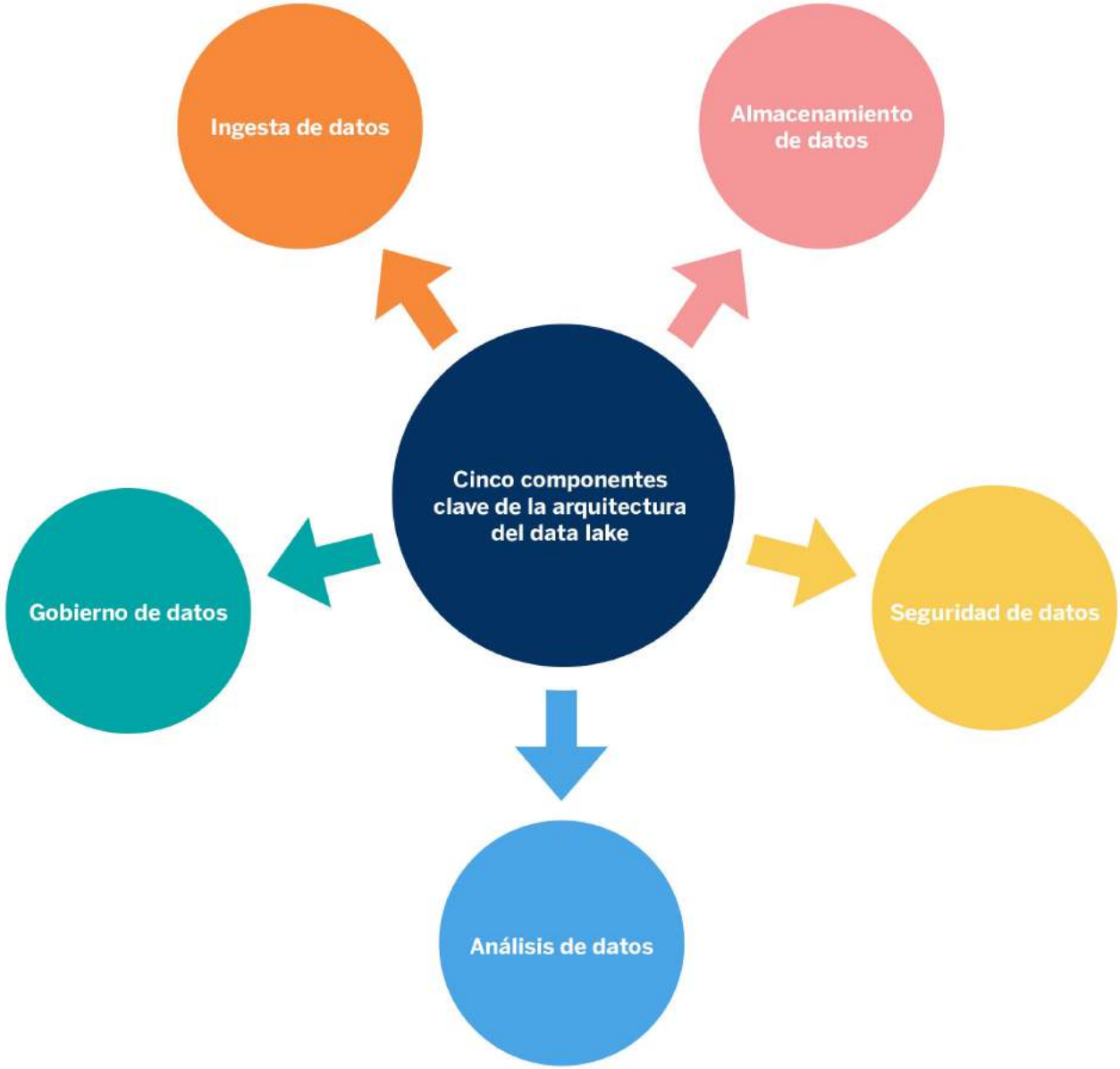
Data Lake

- Es un repositorio de almacenamiento masivo que contiene una gran cantidad de datos **sin procesar** en su formato original hasta que se les necesite.

- Capacidad para ser más flexibles.

- Es un lugar para descargar y almacenar temporalmente todos los datos hasta que el **almacén de datos** esté en funcionamiento.

Data Lake


Ingesta de datos
Almacenamiento de datos
Cinco componentes clave de la arquitectura del data lake
Gobierno de datos
Seguridad de datos
Análisis de datos

Data Lake

	Base de datos	Data Warehouse	Data Lake
Datos	Estructurados	Estructurados	Raw y Desestructurados
Procesamiento	On Write	On Write	On Read
Costo	Medio alto	Alto	Bajo
Agilidad	Variada	Mínima	Máxima
Seguridad	Madura	Madura	Inmadura
Usuarios	Todos en la organización	IT / Negocio	Data Scientists
Casos de uso	Reportes, análisis, automatización, OLTP	OLAP, Machine Learning	Ciencia de datos, investigación

Data Lake

- Precaución
  - Los lagos de datos permiten almacenar lo que sea sin cuestionar si necesita todos los datos
  - No priorizan qué datos entran en una cadena de suministro y cómo esos datos son beneficiosos
  - La latencia de los datos es mayor
  - No tienen reglas que supervisen lo que pueden incorporar
  - Fomentan el exceso de datos

Contacto

Omar Mendoza González

Profesor de carrera ICO FES Aragón

omarmendoza564@aragon.unam.mx

Referencias

- **Corea, Francesco**, **An Introduction to data : everything you need to know about AI, Big data and data science / Francesco Corea -- Cham, Switzerland :** Springer, [2019].-- xv, 131 páginas : ilustraciones (Studies in Big data, 2197-6503 ; 50)

- **Casas Roma, Jordi**, **Big data : análisis de datos en entornos masivos / Jordi Casas Roma, Jordi Nin Guerrero, Francesc Julbe López -- Barcelona :** Editorial UOC, 2019 287 páginas : ilustraciones (Tecnología ; 623).

- **Caballero, Rafael, Big data con Python recolección, almacenamiento y proceso /** Rafael Caballero Adrián Riesco Enrique Martín: Universidad Complutense de Madrid Editorial AlfaOmega, 2019 282 páginas

- **Rioux, Jonathan, Data Analysis with Python and PySpark** / Jonathan Rioux: Editorial Manning Publications, 2020 259 páginas

- **Singh, Pramod, Machine Learning with PySpark: With Natural Language Processing and Recommender Systems** / Pramod Singh: Editorial Apress, 2019 233 páginas