

Big data



Alberto Benavides

¿Qué es

Big Data?

Grandes Volúmenes de Datos

La crisis de estiércol

Londres (1900)

- 11,000 taxis (🐎)
- Miles de buses (🐎 × 12)
- 🐎 ≈ 6 kg < 💩 < 15 kg (por día)
- ~ 50,000 🐎 = ~ 500,000 kg 💩



MORTON STREET, CORNER OF BEDFORD, LOOKING TOWARD BLEECKER STREET,
MARCH 17, 1893.

Tres Cinco V del Big Data

Volumen	Grandes volúmenes de datos no estructurados de baja densidad Decenas de terabytes de datos Cientos de petabytes
Velocidad	Ritmo al que se reciben los datos (tiempo real) Evaluación y actuación en tiempo real
Variedad	Base de datos relacional Datos no estructurados y semiestructurados ([texto, audio o video] + metadatos)
Valor	Importancia (selección de características) Mejora toma de decisiones
Veracidad	Confiabilidad Certeza

Algunos casos de uso

Mantenimiento predictivo

Registrar parámetros con Internet de las cosas

Detección de fraude

Identificar patrones anómalos

Aprendizaje máquina (Machine learning)

Procesamiento Lenguaje Natural

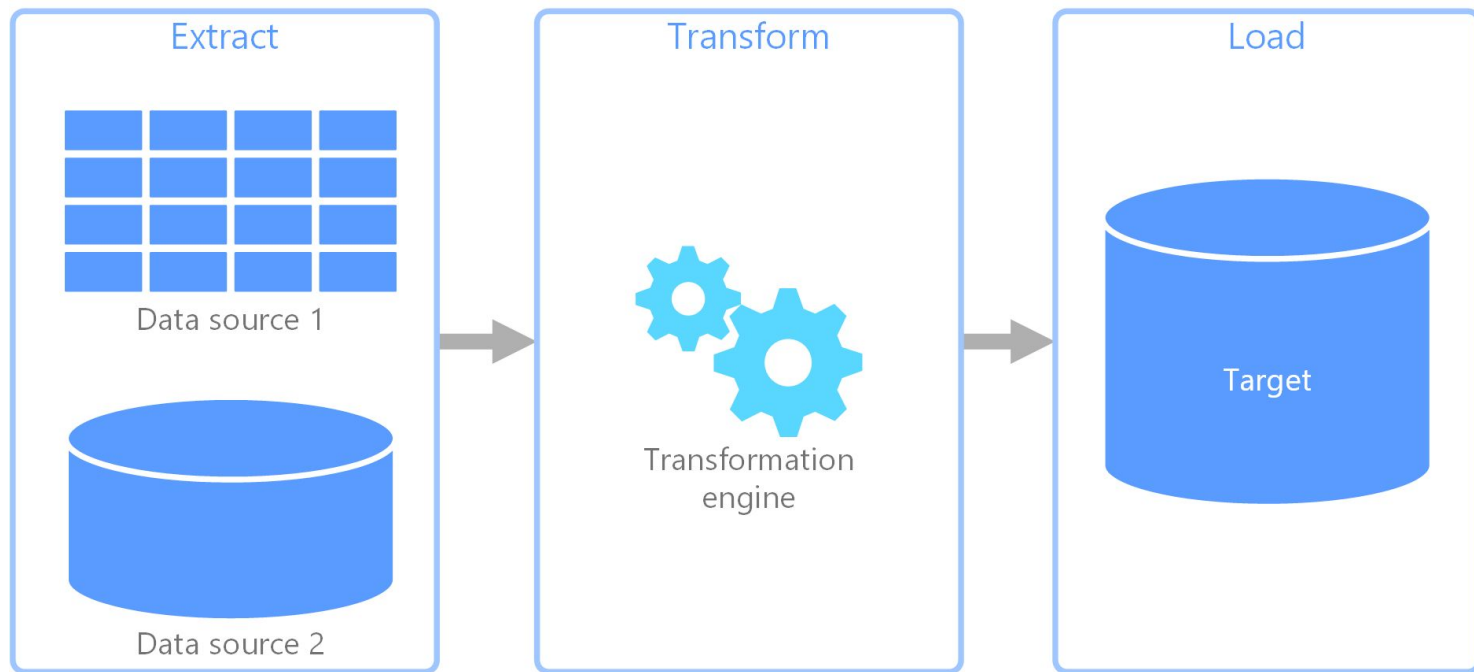
Análisis de grandes volúmenes de textos

Desarrollo de productos

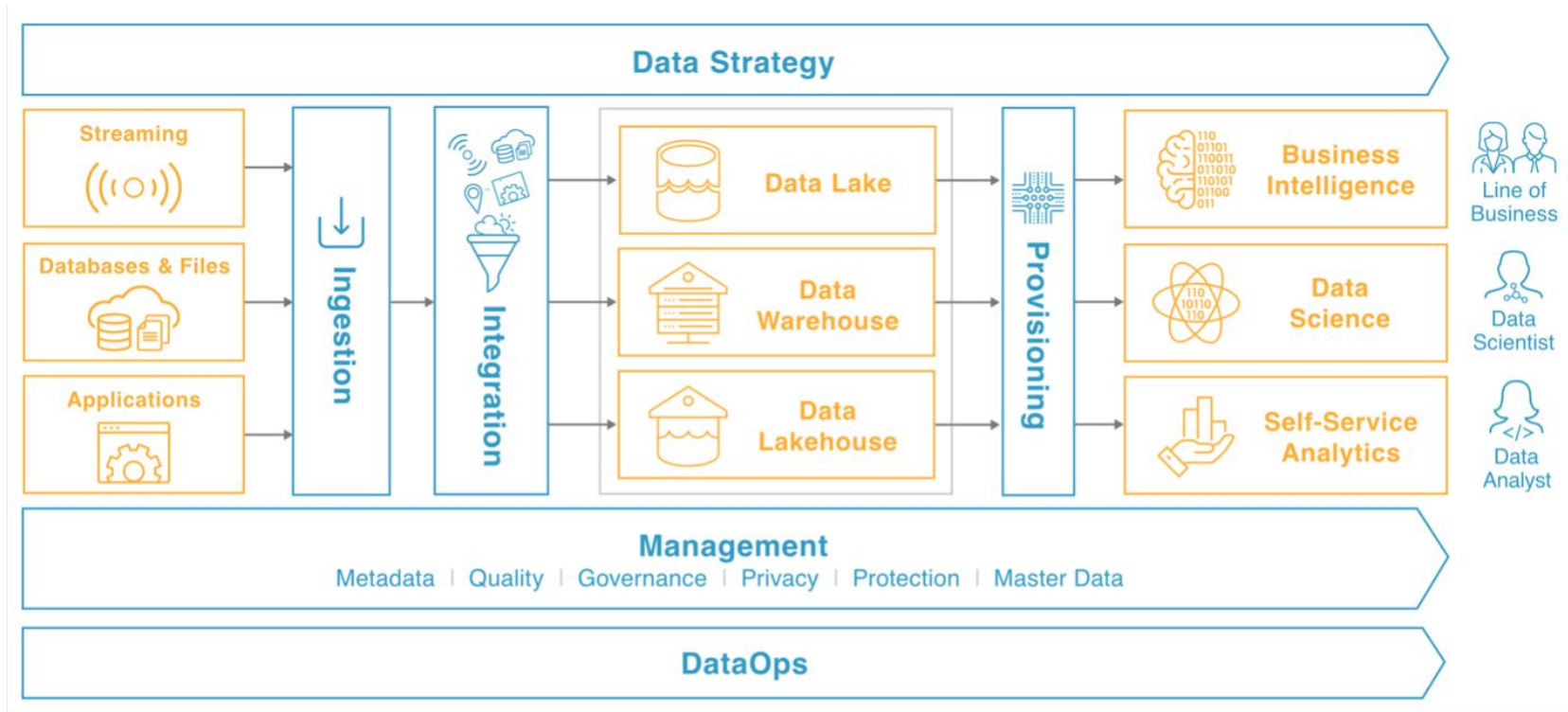
Agrupamiento y predicción de comportamiento de clientes



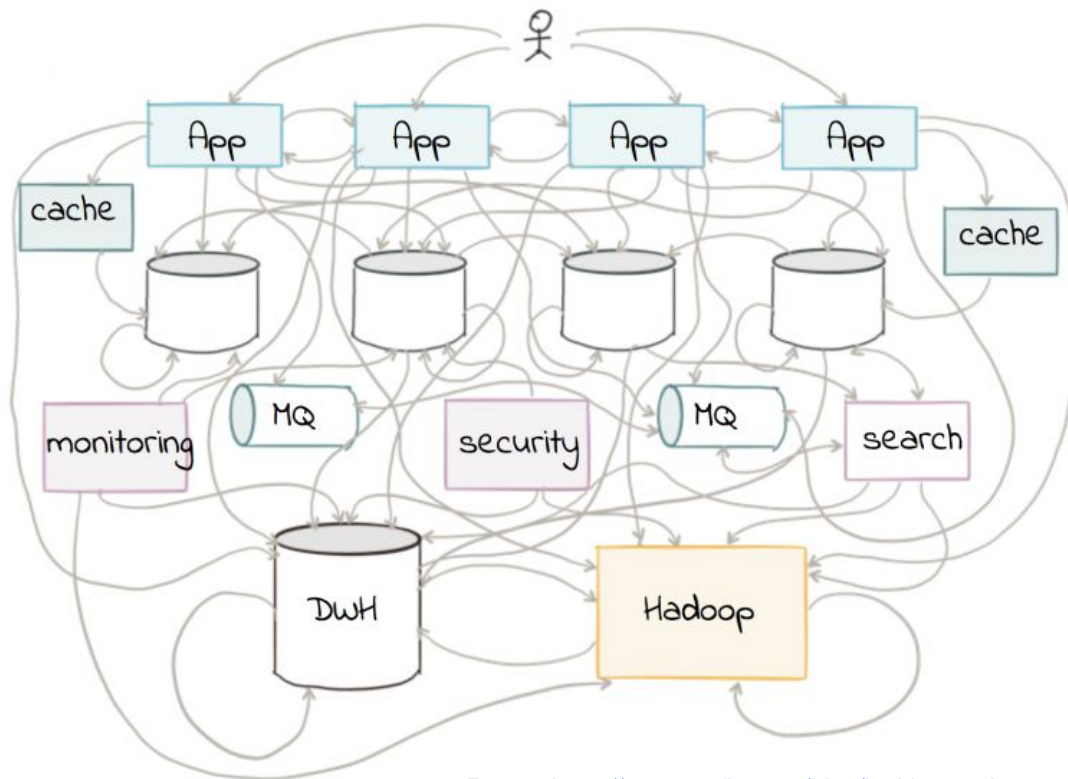
Flujo de trabajo “típico” (ETL)



Flujo de trabajo de Big Data “ideal”



Flujo de trabajo “real”



Conceptos fundamentales

- *Data warehouse* (Almacén de datos)
 - Repositorio **histórico** de datos
 - Base de datos relacional
 - Datos estructurados
 - Datos **preparados** para ser procesados
- *Data lake* (Lago de datos)
 - Datos
 - estructurados
 - semiestructurados
 - no estructurados
 - Datos no preparados (¡LO TIENE TODO!)
 - Más versátil
 - Más **preprocesamiento**
- *Data Lakehouse* = *lake* + *ware*
- *Data science* (Ciencia de datos)
 - Análisis estadístico de datos
 - Usa técnicas de inteligencia artificial
 - Pronósticos
 - Regresiones
 - Etc.
 - Perfil científico
 - Interpretación requiere especialistas
 - *Incluye* a la inteligencia de negocios
- *Business intelligence* (Inteligencia de negocios)
 - Fácil interpretación
 - Visualmente
 - breve
 - atractiva
 - cubre los puntos más importantes

Comparativas entre Datos

Característica	Data Lake	Data Warehouse	Data Lakehouse
Tipo de datos	No estructurados, semi-estructurados, estructurados	Estructurados	Ambos
Propósito	Almacenamiento masivo de datos crudos	Análisis de datos preprocesados	Almacenamiento y análisis híbrido
Costo de almacenamiento	Bajo	Alto	Moderado
Consultas	Menos optimizado	Muy optimizado	Optimizado
Flexibilidad	Alta (cualquier dato)	Baja (solo datos estructurados)	Alta

Algunos recursos de Datos

Data lakes

- [Amazon](#)
- [Google](#)
- [Azure](#) 🤪

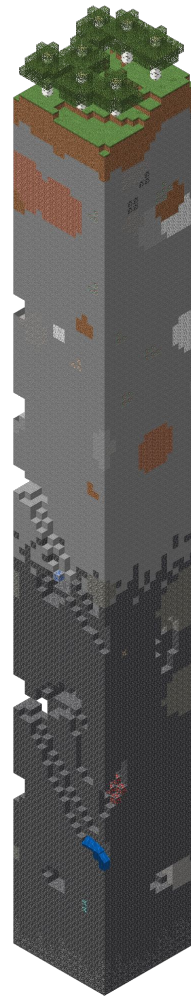
Aplicaciones para *Data lakehouses*

- [Databricks](#)
- [Gobierno de datos de Google](#)
- [Delta Lake](#)

Herramientas pequeña escala para datos grandes en Python

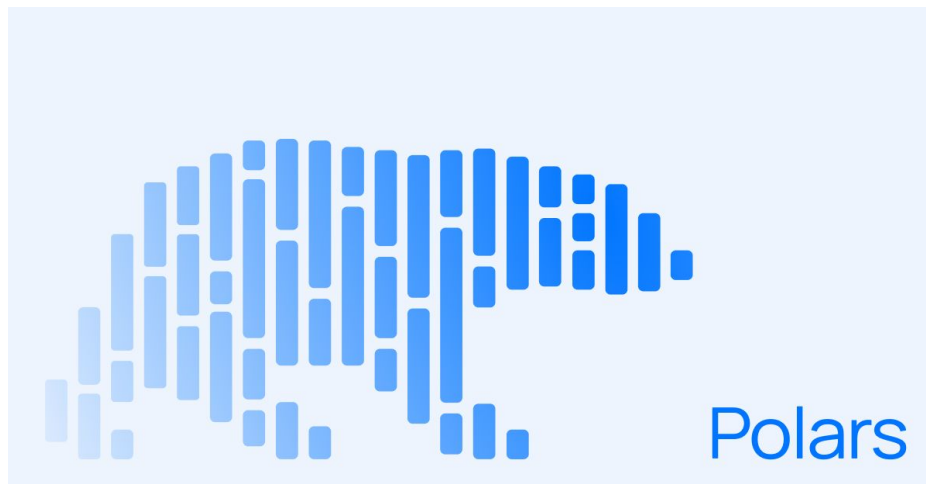
- [pandas](#): Análisis de datos tabular
- [Dask](#): Big Data pandas (trabajo por chunks)
- [Vaex](#): *Out-of-Core DataFrames* (trabajo por chunks)
- [datatable](#): Implementación del paquete homónimo de R
- [cuDF](#): Datos tabulares con CUDAs
- [Intel Extension](#): Extensión para ciencia de datos

* Recomendando aprender [Google Colab](#)



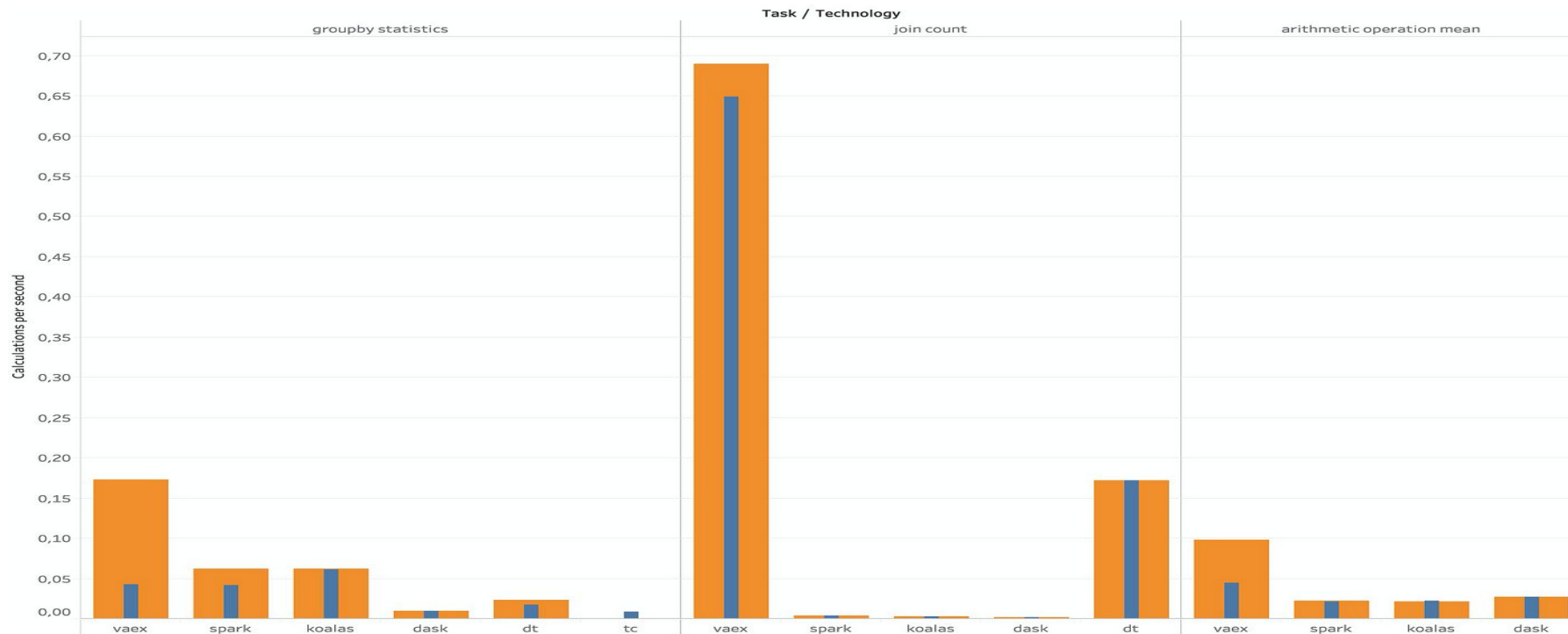
Otras herramientas

- [Polars](#)
- [DuckDB](#)



DuckDB

Comparativas de velocidad, cálculos pesados



Fuente: <https://vaex.io/blog/beyond-pandas-spark-dask-vaex-and-other-big-data-technologies-battling-head-to-head>

Cuándo (no) usar estas herramientas

Cuándo sí 👍

- Bases de datos que no caben en memoria
- Acelerar cálculos (núcleos disponibles)
- Cómputo distribuido de agrupamientos

Cuándo no 🚫

- Bases de datos que sí caben en memoria
- Datos no siguen estructura tabular

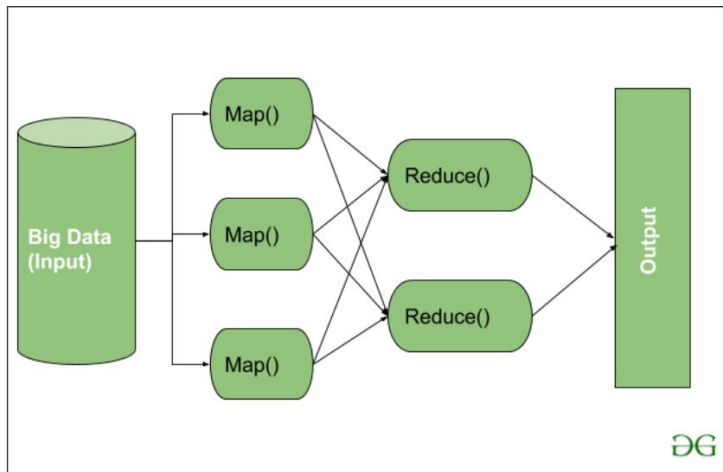
Recomendaciones de optimización

- Utilizar [tipos de datos adecuados](#)
- Eliminar [columnas no deseadas](#)
- Requiere una base de datos, usar:
 - [Postgres](#)
 - [MySQL](#)
 - [SQL Server](#)

Herramientas gran escala

Apache Hadoop

- Cómputo distribuido
- Escalable
- Tolerante a fallos



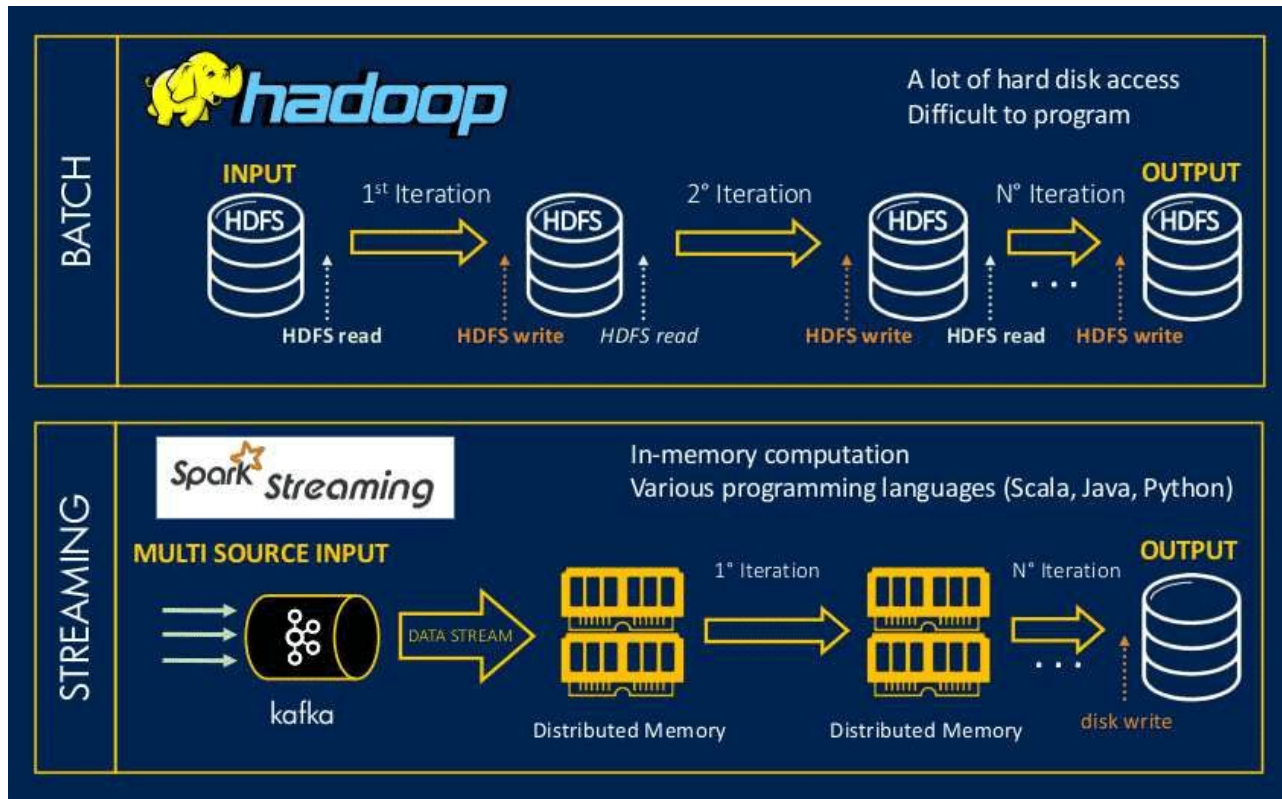
Apache Spark

- Basado en Apache Hadoop
- Optimizado para
 - Ciencia de datos
 - Aprendizaje máquina

Apache Beam

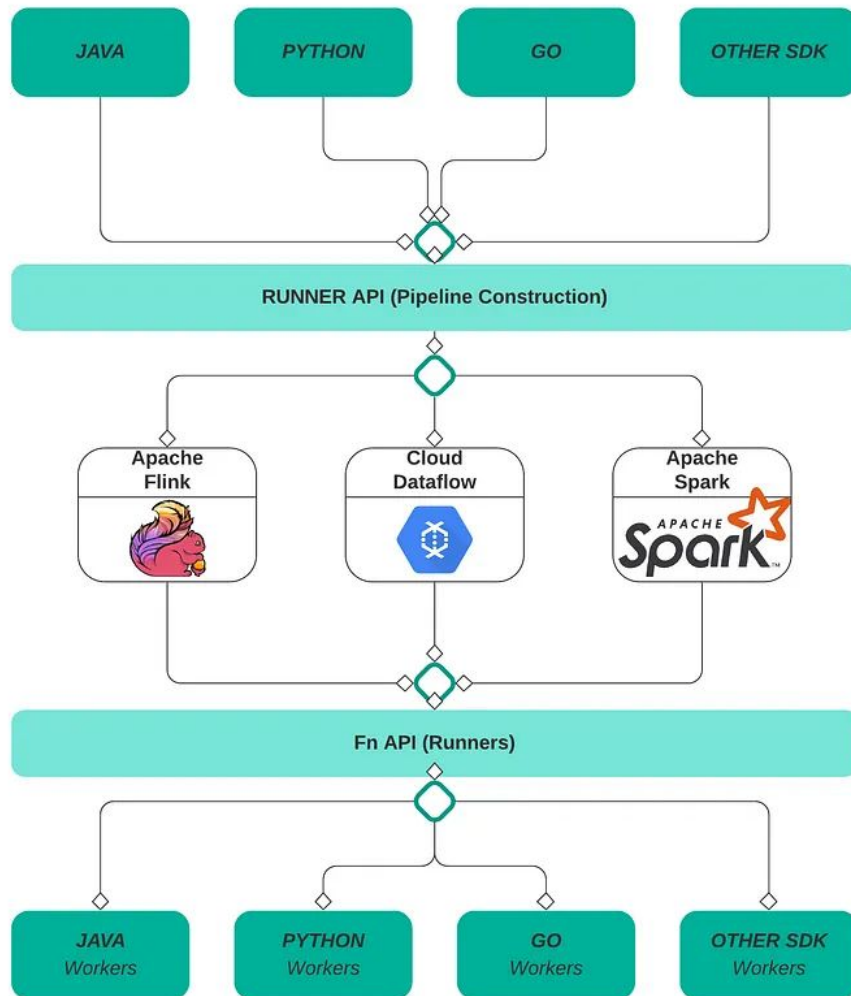
- Unificación de
 - carga
 - procesamiento
 - análisis de datos

hadoop vs Spark



Spark vs Beam








- Spark se usa para realizar operaciones de Ciencia de datos
- Beam se utiliza para crear flujos de que incluyen operaciones de Ciencia de datos



Herramientas IoT

- <https://grafana.com/>
- <https://thinger.io/>
- <https://webthings.io/gateway/>

Orígenes de datos (APIs)

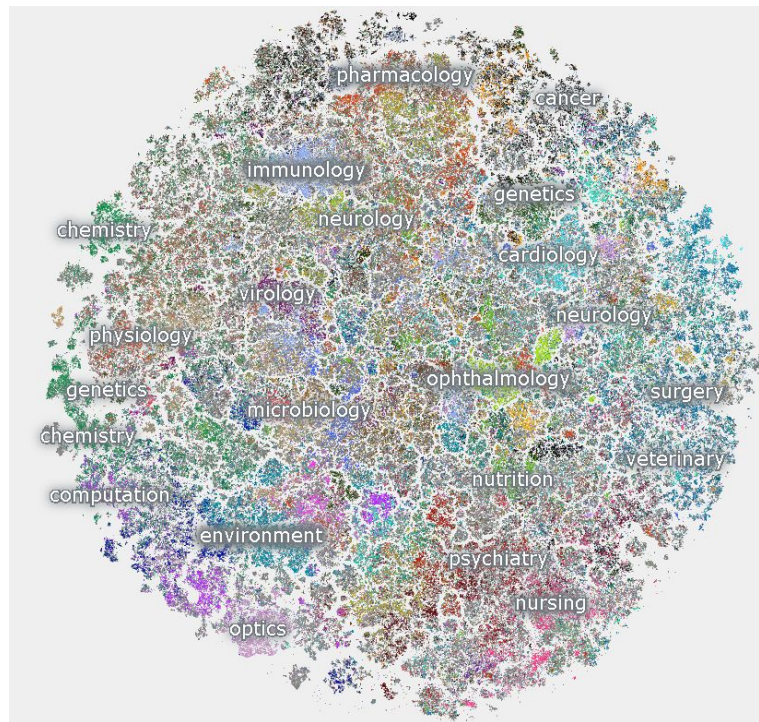
-  [New York Times](#)
 -  [Spotify](#)
 -  [Zillow](#)
 -  [Yummlly](#)
 -  [National Weather Service](#)
 -  [Sports DB](#)
 -  [Crunchbase](#)
 - [Wikipedia](#)
 - [Reddit](#)
 - [Facebook](#)
 - [Twitter](#)
 - [Project Gutenberg](#)
-
- [INEGI](#)
 - [Datos abiertos México](#)

Ejemplo: Publicaciones Reddit



Fuente: <https://anvaka.github.io/map-of-reddit/>

Ejemplo: 21 mill. artículos científicos ★



Gracias :)



Enlace a la presentación

jose.benavidesvz@uanl.edu.mx
<https://github.com/albertobenavides>

Recursos adicionales de interés

- <https://www.oracle.com/mx/big-data/what-is-big-data/>
- <https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/>