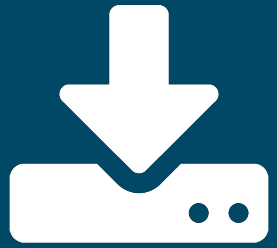


Almacenamiento y captura de datos

Claudio Aracena

Contenidos

- Captura de datos desde archivos
- Base de datos
- Captura y almacenamiento de datos en BD
- Captura de datos de la Web (Web scraping)
- Captura de datos de API (ej: Twitter)
- **Captura y almacenamiento en arquitecturas Big data**

Códigos y clase en:
https://github.com/caracena/almacenamiento-captura-datos

Clase de hoy

Captura y almacenamiento en arquitecturas Big data

- Big Data
- Ambiente Big Data

Ejemplos

- Sqoop
- Hive vs Impala
- Hbase
- Big Query

Herramientas para practicar en casa

Descargar VirtualBox

- https://www.virtualbox.org/wiki/Downloads

Descargar Cloudera 5.13 VM

- https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip

Big Data según IBM

What is Big Data?

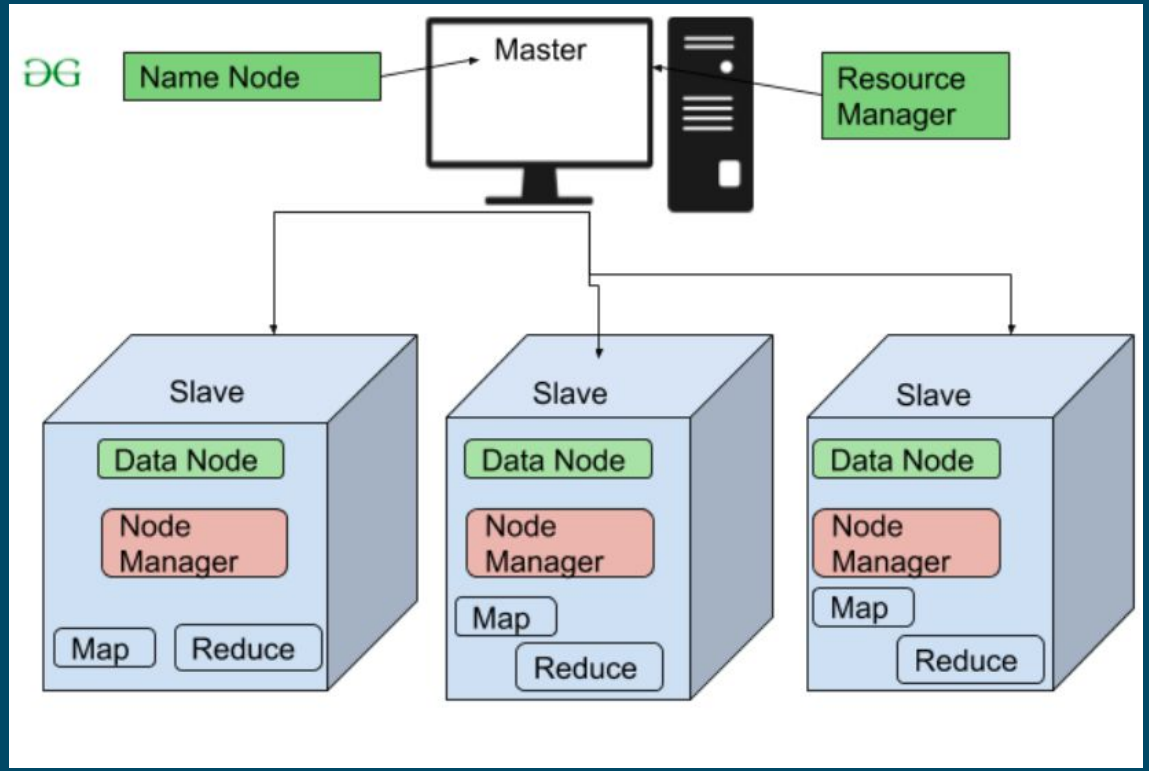
Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

Volumen - Velocidad - Variedad - Veracidad - Valor

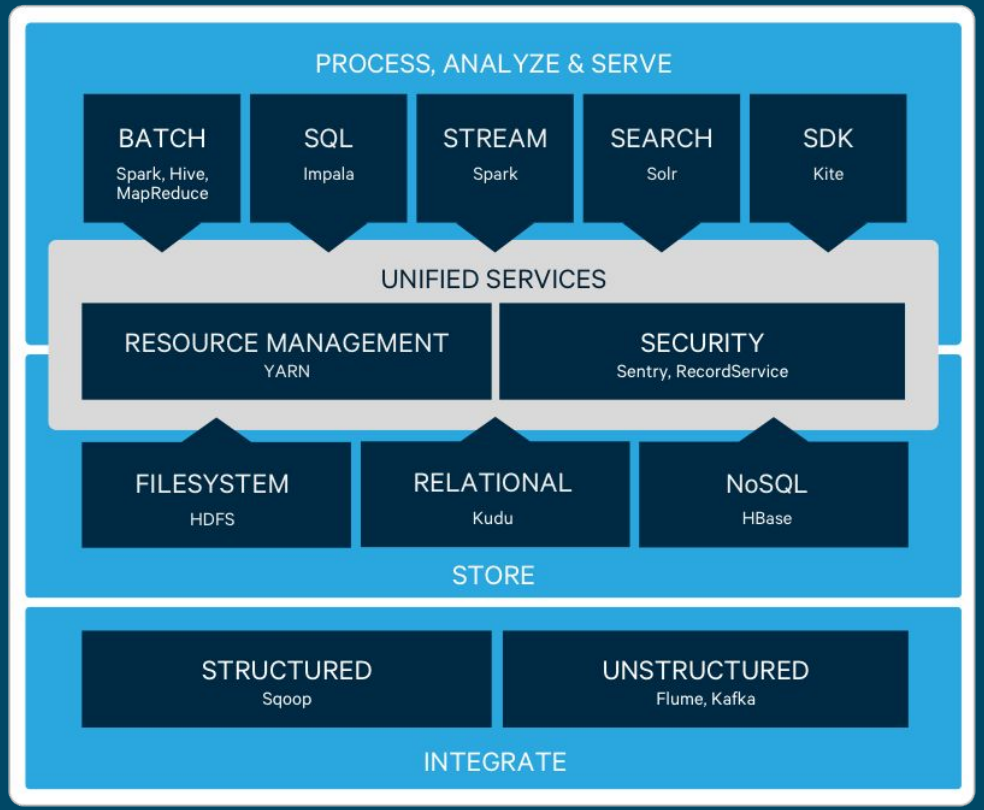
Hadoop

Framework que permite el trabajo de grandes datasets de forma distribuida. Las distribuciones más populares son:

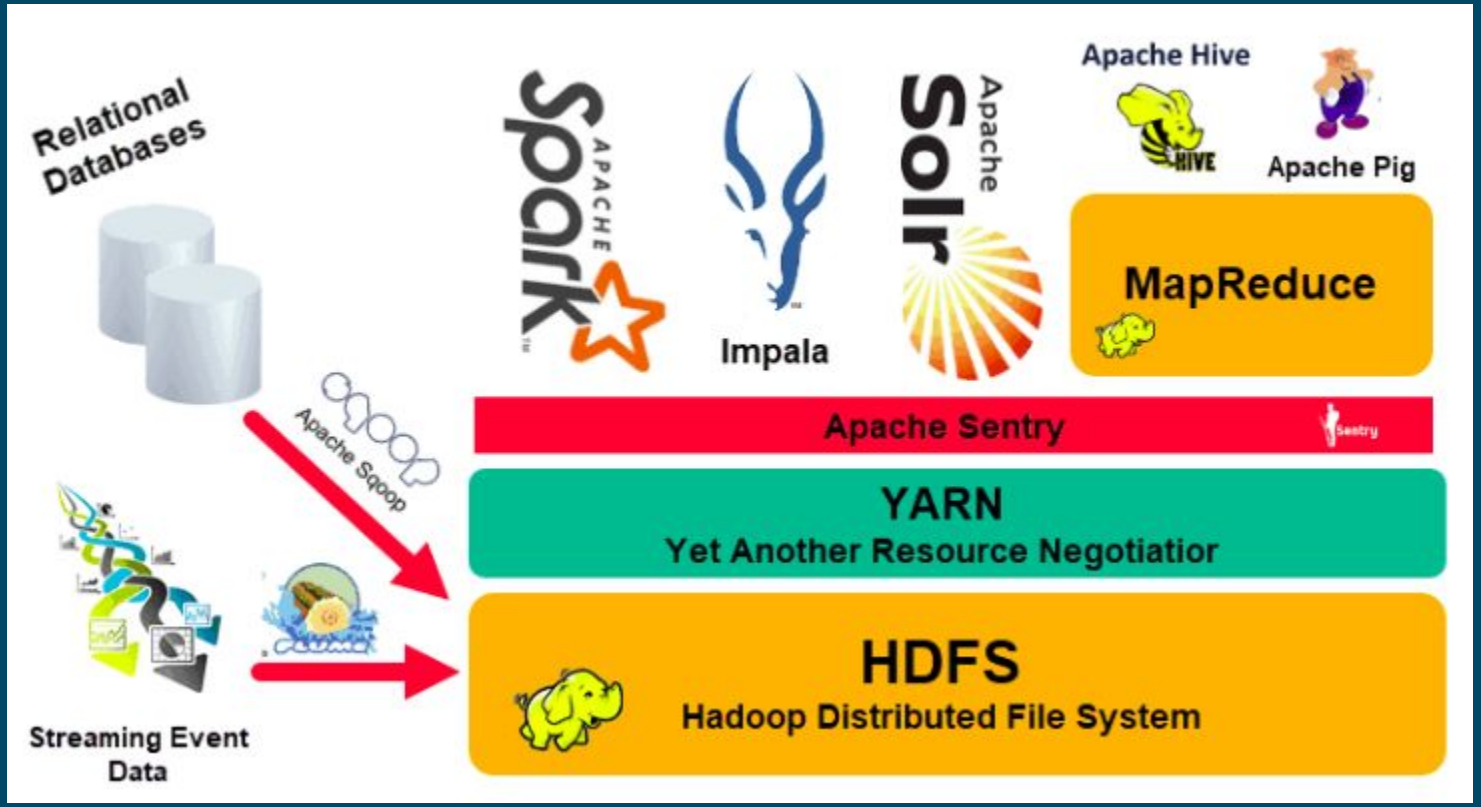
Arquitectura Hadoop



Ambiente Big Data en Cloudera



Ambiente Big Data en Cloudera



Ambiente Big Data en Cloudera

Hadoop: Framework que permite el trabajo de grandes datasets de forma distribuida

- YARN: Administrador de recursos
- MapReduce: Framework de procesamiento paralelo
- HDFS: Sistema de archivos
- Hbase: Almacenamiento orientado a columnas
- Hive: Herramienta para manejo de Data Warehouse (usando SQL)
- Impala: Herramienta analítica (usando SQL, pero más rápida que Hive)
- Spark: Sistema de análisis de datos (incluye soporte para streaming, SQL, grafos y Machine learning)

Ambiente Big Data en Cloudera

Más herramientas, sistemas y almacenamientos

- Kudu: Almacenamiento en forma similar a BD relacional
- Sqoop: Herramienta de transferencia de datos
- Flume: Herramienta para manejo de logs
- Kafka: Herramienta para manejo de pipelines (colas)
- Sentry: Sistema de seguridad (permisos)
- Solr: Motor de búsqueda
- Kite: API para simplificar manejo de datos
- Hue: Editor para consultar datos y crear dashboards

Parquet

- Es un formato de archivo open-source disponible en el ecosistema Hadoop
- Es un formato columnar de almacenamiento
- Está focalizado en la lectura eficiente de columnas y la compresión de datos

Dataset	Size on Amazon S3	Query Run Time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet Format	130 GB	6.78 seconds	2.51 GB	$0.01
Savings	87% less when using Parquet	34x faster	99% less data scanned	99.7% savings

https://github.com/dask/fastparquet

Hive vs Impala

- Hive presenta el problema de “cold start”, mientras que Impala tiene procesos corriendo desde el inicio del sistema para evitar overhead.
- Hive soporta tipo de datos complejos, mientras que Impala soporta tipo de datos simples. Por ejemplo, fechas son soportadas solo en Hive
- Hive convierte las queries en un trabajo MapReduce, Impala utiliza MPP (massively parallel processing).
- Hive tiene tolerancia a fallas, Impala debe comenzar de nuevo si un nodo falla.
- Impala es 6 a 70 veces más rápido que Hive.

Hbase

- Hbase es una base de datos columnar (basado en Big Table)

- bases relacionales: fácil de agregar y modificar, pero puede leer datos innecesarios
- bases columnares: lee solo data relevante, pero para escribir requiere múltiples accesos.