Source
Meridian

Temas de la charla

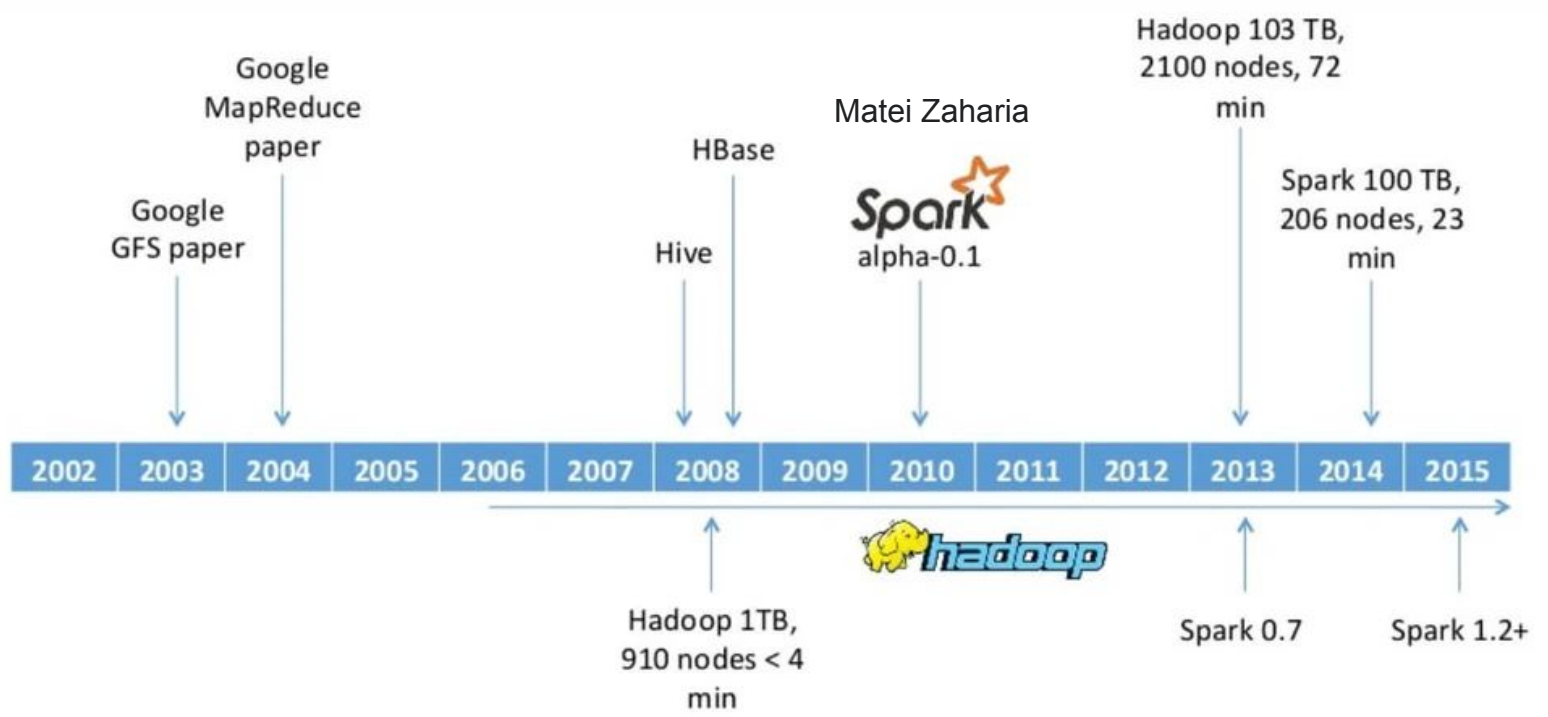
- Breve Historia de Spark
- Qué es y para qué usar Spark?
- Componentes de Spark I
  - Driver
  - Executor
  - Worker Node
- Lenguajes de la API de spark
- Lazy Evaluation
- Demo

Source Meridian
A SOFTWARE COMPANY

5 Vs De Big Data

- volumen (cantidad de datos generados y almacenados)
- variedad (tipo y naturaleza de los datos)
- velocidad (rapidez a la que se generan y procesan los datos)
- variabilidad (incoherencia del conjunto de datos)
- veracidad (la calidad de los datos puede variar mucho)

Breve Historia de Spark



Tomado de
https://www.bigdata.uma.es/apache-spark-un-poco-de-historia/

Qué es y para qué usar Spark?

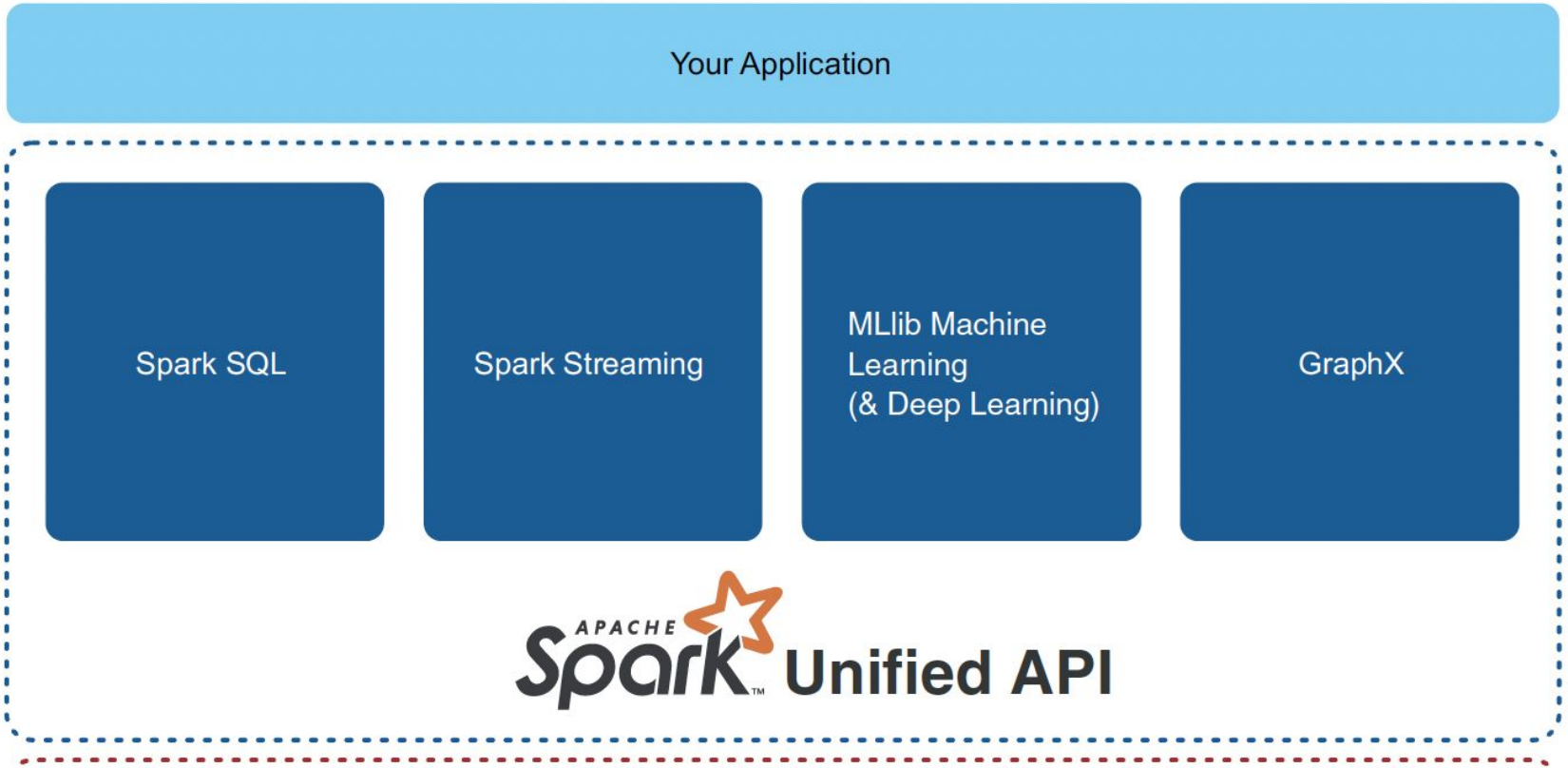
Apache Spark es un **motor de computación** unificado y un conjunto de bibliotecas para el procesamiento paralelo de datos en un clúster

Unificado: Cuenta con una plataforma completamente unificada que permite hacer diferente tipo de tareas

Motor de Computación: Spark se enfoca en realizar cálculos sobre los datos, sin importar dónde se encuentren
Spark es independiente del almacenamiento.

Librerias: Spark admite tanto bibliotecas estándar que se incluyen con el motor como una amplia gama de bibliotecas
externas

Qué es y para qué usar Spark?



Qué es y para qué usar Spark?

Spark en un escenario de procesamiento/ingeniería de datos

- Ingestión
- Mejora de la calidad de los datos (DQ)
- Transformación
- Publicación

Spark en un escenario de ciencia de datos

- Análisis exploratorios de datos (EDA) en datos a escala de petabytes sin tener que recurrir a la reducción de la muestra
- Entrene algoritmos de aprendizaje automático

Problema embarazosamente paralelo

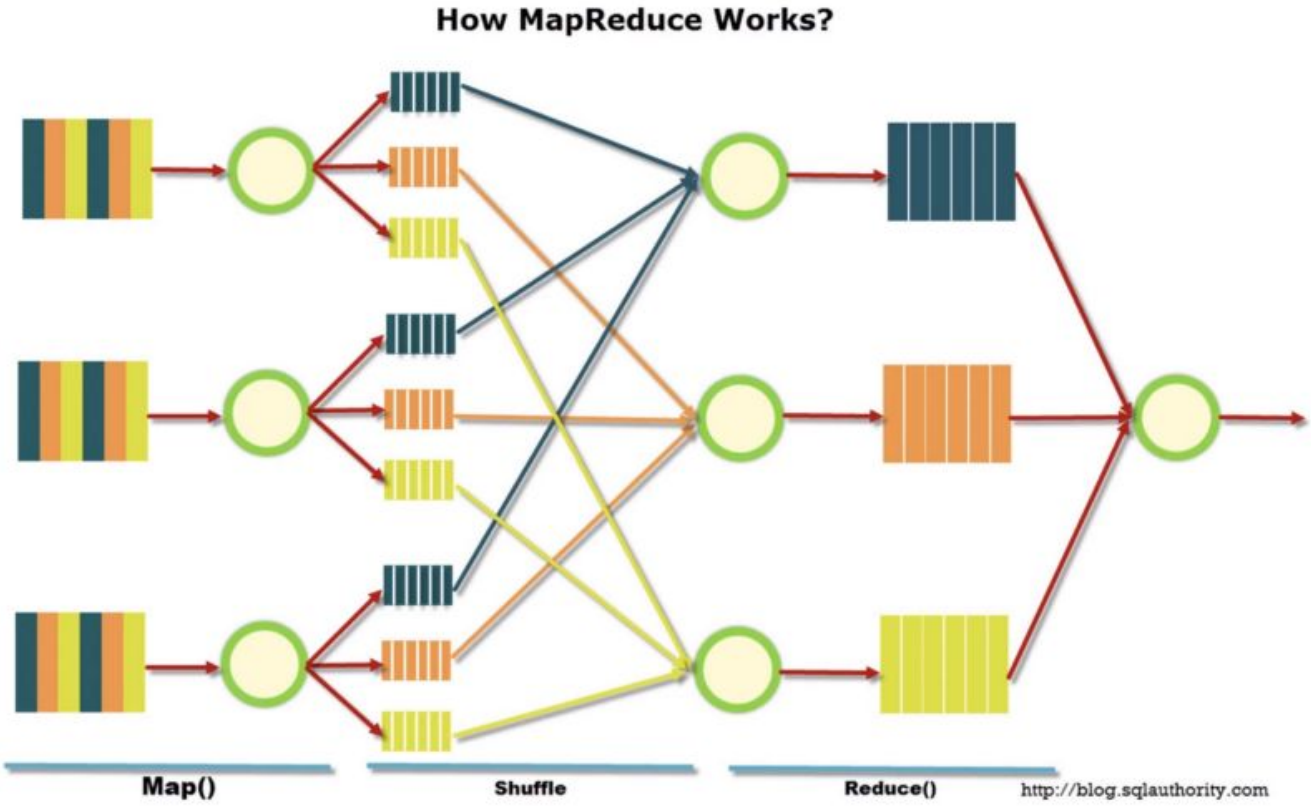
Características:

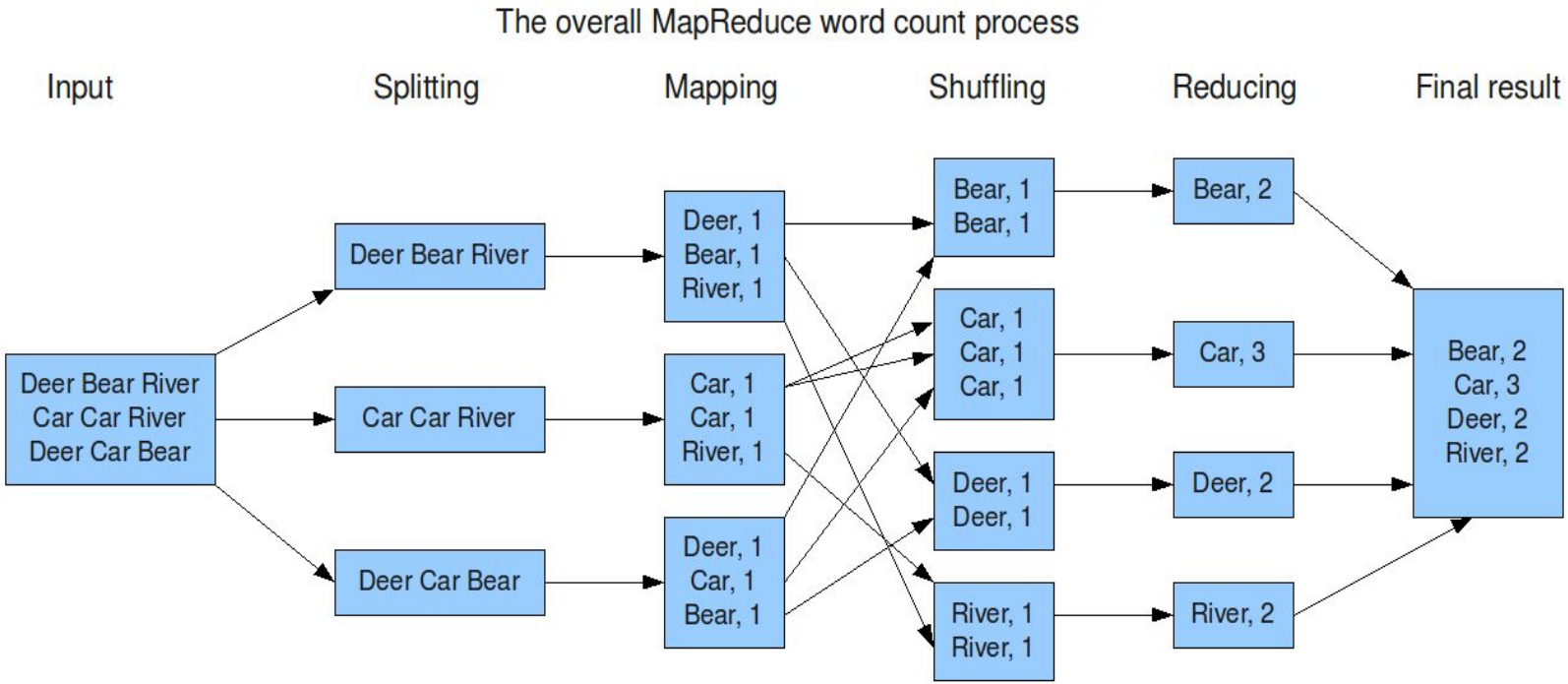
1. Independencia de Tareas
2. Escalabilidad
3. Bajo Overhead

Algunos ejemplos:

- Extracción
- Transformación
- Carga

MapReduce


How MapReduce Works?
Map()
Shuffle
Reduce()
http://blog.sqlauthority.com

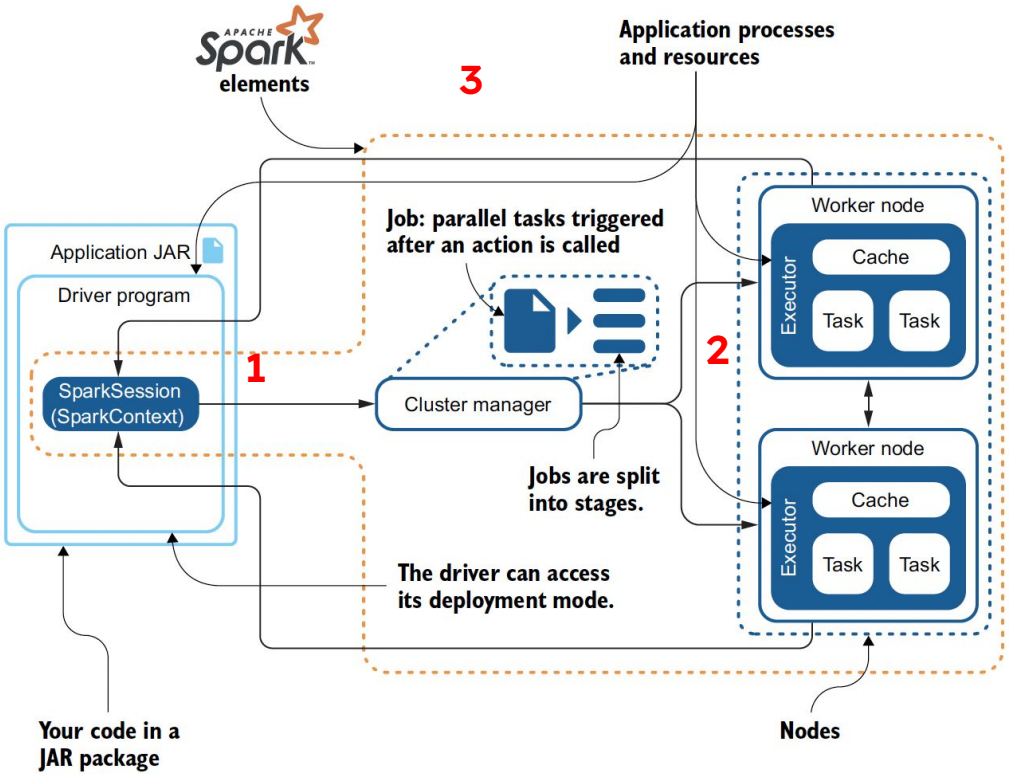


Tomado de https://www.todaysoftmag.com/article/1358/hadoop-mapreduce-deep-diving-and-tuning

Principales diferencia entre spark y hadoop

Hadoop	Spark
* MapReduce, En la fase Map, los datos se procesan en paralelo, y en la fase Reduce, los resultados intermedios se combinan *Altamente escalable, se ha probado en grandes clústeres *Proceso de re-computación basado en la re-ejecución de tareas fallidas en diferentes nodos *Más lento debido a que el procesamiento intermedio se escribe en disco	*RDD, Spark usa RDD y DAG para manejar datos distribuidos, y permite operaciones más complejas como map, reduce, filter, join *También escalable, pero con una arquitectura más eficiente en clústeres grandes debido al uso de memoria *Usa **DAGs y RDDs** para re-computar solo las particiones fallidas de datos, lo que reduce la sobrecarga de recuperación *Mucho más rápido al realizar el procesamiento intermedio en memoria *Más costoso, ya que usa RAM

Componentes de Spark I

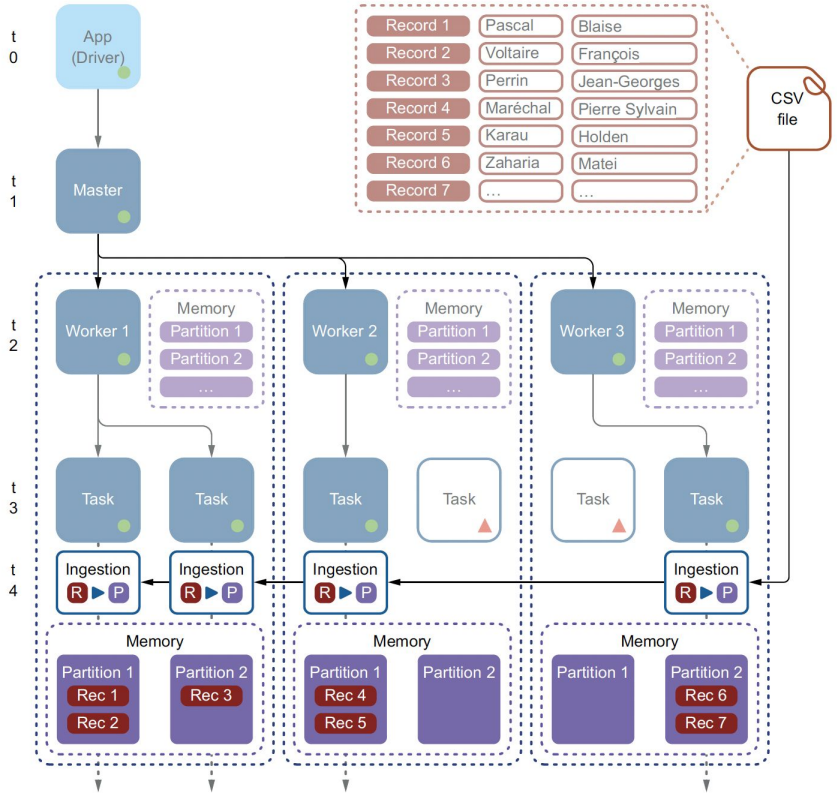


Apache Spark components

Driver: mantener información sobre la aplicación de Spark; responder al programa o entrada de un usuario; y analizar, distribuir y programar el trabajo en los ejecutores

Ejecutor: ejecutar el código asignado por el driver y reportar el estado de la computación en ese ejecutor de vuelta al nodo driver

Componentes de Spark I

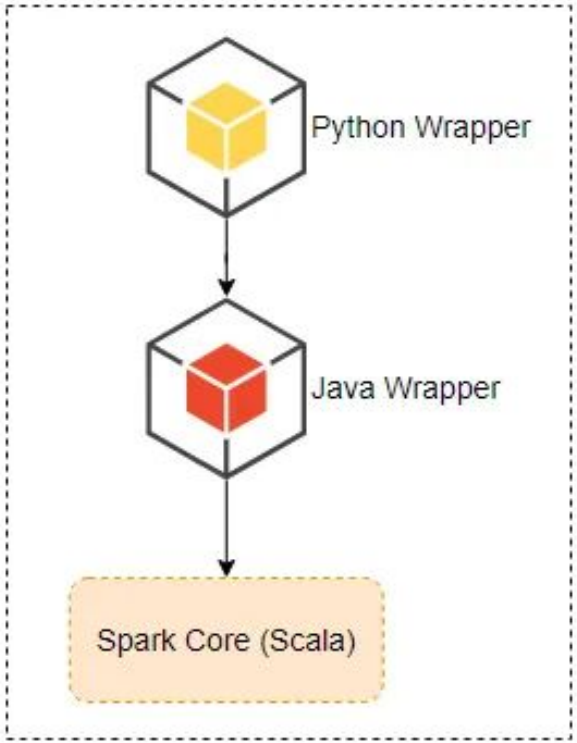


Lazy Evaluation

Lazy evaluation significa que Spark esperará hasta el último momento para ejecutar el grafo de instrucciones de computación

```
-- Default Output
EXPLAIN select k, sum(v) from values (1, 2), (1, 3) t(k, v) group by k;
+----------------------------------------------------+
|                                                plan|
+----------------------------------------------------+
| == Physical Plan ==
 *(2) HashAggregate(keys=[k#33], functions=[sum(cast(v#34 as bigint))])
 +- Exchange hashpartitioning(k#33, 200), true, [id=#59]
    +- *(1) HashAggregate(keys=[k#33], functions=[partial_sum(cast(v#34 as bigint))])
       +- *(1) LocalTableScan [k#33, v#34]
|
+----------------------------------------------------
```

Lenguajes de la API de spark



Referencias

Perrin, J.-G. (2020). *Spark in action* (2nd ed.). Manning Publications.

Chambers, B., & Zaharia, M. (2018). *Spark: The definitive guide*. O'Reilly Media.

Source Meridian

Source Meridian