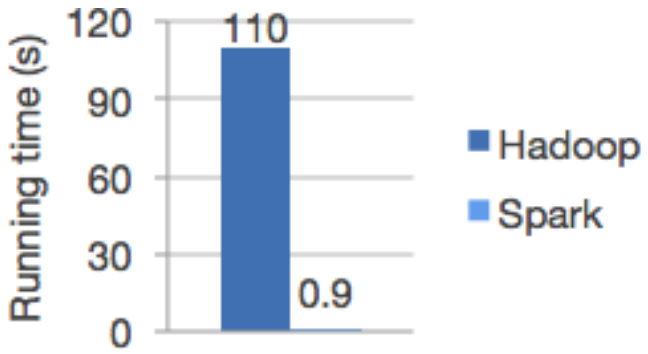
Apasoft Training

Introducción a Spark

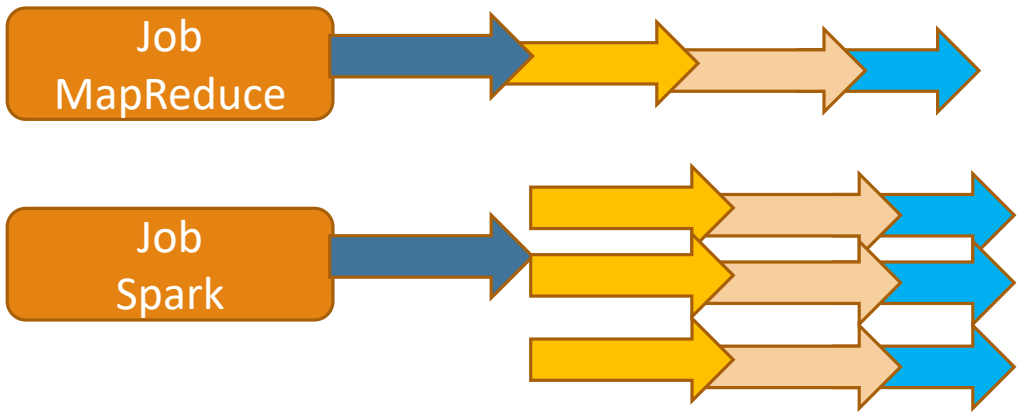
Introducción a Spark

- Apache Spark es una entorno de procesamiento distribuido y paralelo que trabaja **en memoria**
- **Permite el análisis** de grandes conjuntos de datos
- Integra diferentes entornos **como Bases de Datos NoSQL, Real Time, machine learning, o análisis de grafos, etc**
- Es mucho más rápido que MapReduce
- Compatible con Hadoop



Introducción a Spark

- Al contrario que Hadoop Map Reduce que trabaja sobre todo con procesos de tipo Batch, Spark está orientado al trabajo in-memory y el procesamiento en real
- Mientras MapReduce trabaja secuencialmente, Spark lo hace en paralelo

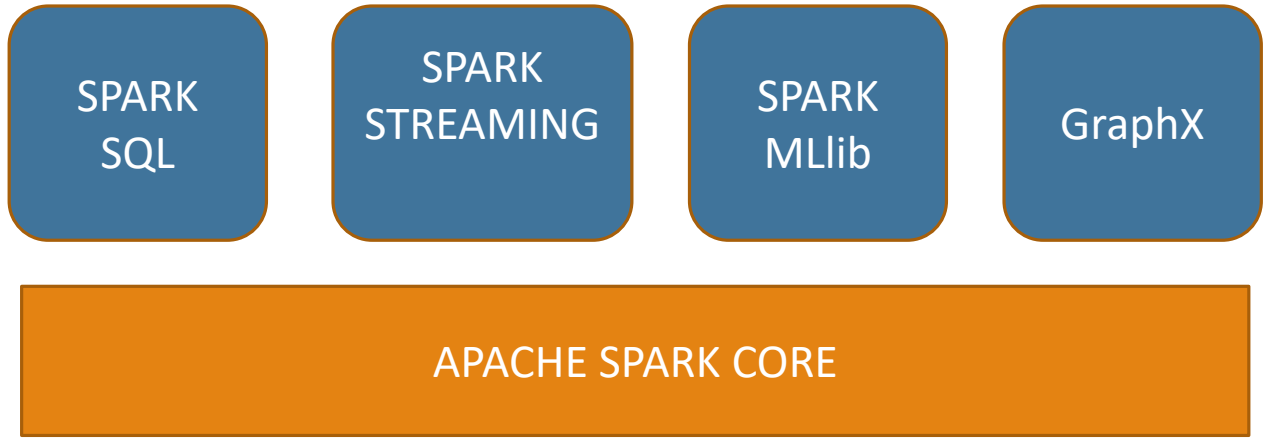


Introducción a Spark

- Compatible con Hadoop
  - Se puede ejecutar sobre HDFS
  - MapReduce. Se puede usar en el mismo cluster que MapReduce
  - YARN: una aplicación Spark se puede lanzar sobre YARN
  - Se pueden mezclar aplicaciones spark y MapReduce para batch y Real Time
- Soporta múltiples fuentes de datos
  - HIVE
  - JSON
  - CASSANDRA
  - CSV
  - RDBMS…

Introducción a Spark

- Está construido en Scala, pero se pueden escribir aplicaciones en Java, Python y R.
- Dispone de un Shell interactivo
- Consiste en un Core y en un conjunto de librerías



Introducción a Spark

- Spark Core
  - El motor base para el procesamiento en escala y distribuido
  - Aunque está construido en Scala, hay APIs para Python, Java y R.
  - Se encarga entre otras cosas de:
    - Gestión de la memoria
    - Recuperación ante fallos
    - Planificación, distribución de trabajas en el cluster
    - Monitorizar trabajo
    - Accedes a los sistemas de almacenamiento

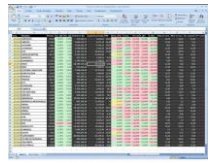
Introducción a Spark

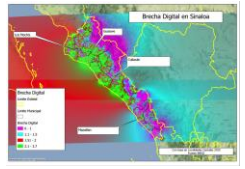
- Spark Core. RDD
  - Spark Core usa una estructura de datos especial denominada RDD (Resilient Distributed Datasets).
  - Resilient Distributed Datasets permite realizar procesos - fault tolerant ‘in-memory’.
  - Los RDD son colecciones de registros inmutables y particionadas que además pueden ser manejadas en paralelo.
  - Los RDDs pueden contener cualquier clase de objetos Python, Scala, Java opersonalizados.
  - Los RDD se crean habitualmente transformándolos de otros RDD o cargandos los datos de una Fuente externa, como por ejemplo HDFS o HBase.

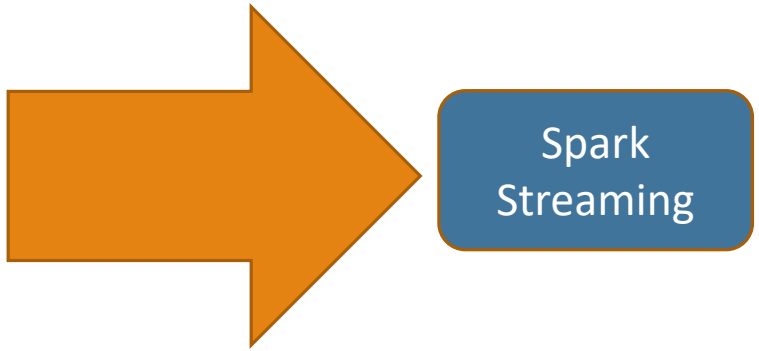
Introducción a Spark

- Spark Streaming
  - Se usa para procesar fuentes de datos en tiempo real (streaming data)
  - Permite procesar con una alta tolerancia a fallos y un gran rendimiento las fuentes “vivas” de información que le suministremos
  - Su unidad fundamental de trabajo es el Dstream (serie de RDDs, que veremos posteriormente)

- twitter
- Datos financieros
- Datos geográficos
- Etc…

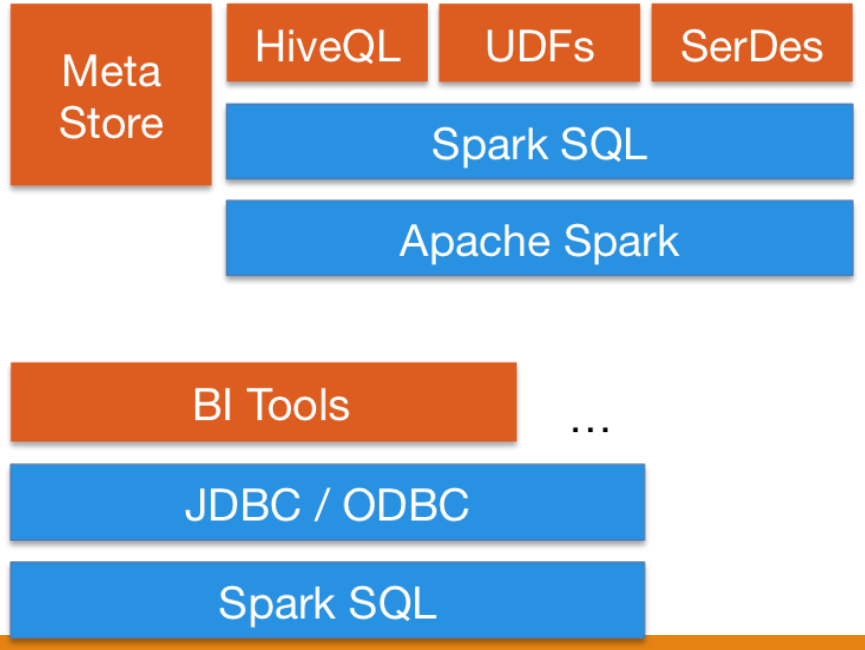






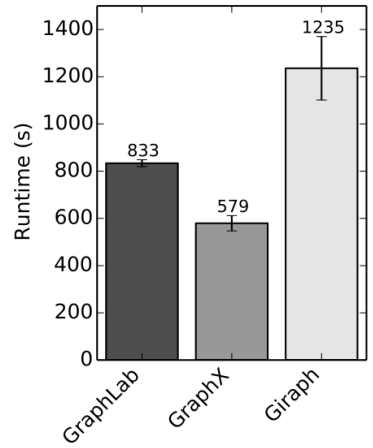
Introducción a Spark

- Spark Sql
  - Permite integrar comandos y componentes relacionales junto con la programación funcional de Spark
  - Podemos usar SQL o Hive Query Language
  - Permite el acceso a múltiples fuentes de datos
  - Dispone de 4 librería básicas
    - Data Source
    - DataFrame
    - Interpreter and Optimizer
    - Sql Service
  - Permite el acceso por JDBC o ODBC



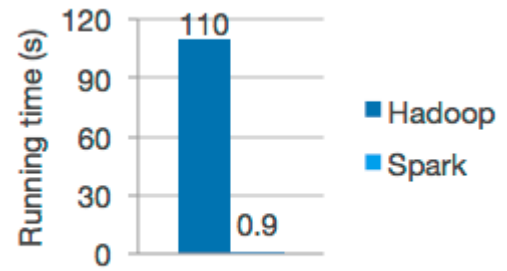
Introducción a Spark

- **GraphX** es el API para procesamientos paralelo en grafos.
  - Spark GraphX implementa Resilient Distributed Graph (RDG- una abstración de los RDD's).
  - RDG's asocia registros con los vertices y bordes de un grafo. Sin embargo, se pueden seguir viendo como colecciones tradicionales de RDD
  - Se dispone de una gran cantidad de algoritmos preparados, que permiten agilizar el proceso de construcción de aplicaciones y mejora el rendimiento y velocidad



Introducción a Spark

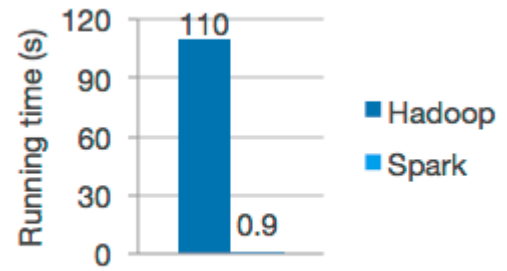
- **Mlib** se utiliza para machine learning en Spark
  - Se dispone de una variedad de algoritmos y otros procesos como “data cleaning”
  - Por ejemplo clasificación, clustering, regression, extracción etc…
  - Permite su ejecución sobre HDFS, HBAse, etc…



Logistic regression in Hadoop and Spark

Introducción a Spark

- **Mlib** se utiliza para machine learning en Spark
  - Se dispone de una variedad de algoritmos y otros procesos como “data cleaning”
  - Por ejemplo clasificación, clustering, regression, extracción etc…
  - Permite su ejecución sobre HDFS, HBAse, etc…



Logistic regression in Hadoop and Spark