Resilient Distributed Datasets (RDD)

Conceptos importantes

- **Dataset:** una colección estructurada de datos
- **Raw data:** datos que se guardan en forma de bytes, en español serían "datos crudos" y por lo tanto, no importa su formato.
- **Unstructured data:** datos que no siguen ningún patrón.
- **Structured data:** datos con un esquema bien definido.
- **Semi-structured data:** híbrido entre datos estructurados y no estructurados (Ej. JSON)
- **POJO:** abreviación de "Plain Old Java Object"; en otras palabras, es una clase simple de Java que no sigue ningún framework.
- **Database Schema:** nombre del tipo de dato con el tipo de dato, en bases de datos nosql se le llama mapping, template, entre otros términos.
- **Provisioning:** el tiempo que uno tarda en crear la estructura necesaria para procesar datos.
- **Query-able:** permite hacer consultas de manera sencilla a los datos.
- **Fault tolerance:** para todo procesamiento, se garantiza que se esté procesando en paralelo en más de un servidor (a esto se le llama replicación), así, si hay un fallo, este está como un back up.
- **Serialization:** la manera de traducir datos estructurados a un formato para poder almacenar o transferir a través de la web (JSON/XML), esto generalmente se hace para asegurar que se envíen y reciban los datos de manera correcta sin importar el entorno (por ejemplo, se podría enviar un dato de java a C# y viceversa sin ningún problema).

Scala

Scala es un lenguaje de alto nivel inventado por Martin Odersky. Una de sus características principales es su similitud con Java, especialmente porque ambos usan una JVM (máquina virtual de java) luego de ser compilado para correr el programa. Es un lenguaje muy utilizado en el manejo de big data.

Spark

Apache Spark es un sistema de procesamiento de datos open-source de manera distribuida diseñado especialmente para big data. Este se puede usar como una interfaz en Scala (también

tiene APIs en diferentes lenguajes como Java, R y Python).

Para ejecutar, se descarga el zip, se extrae, *cd* en el folder y se ejecuta con el comando ./bin/spark-shell

```
cd <folder extraído>
./bin/spark-shell
```

- Spark tiene un spark context y un spark section para interactuar con el framework de procesamiento de datos
- SparkConf es manejada por el resource manager
- SparkSQL permite analizar big data utilizando lenguaje sql

Parquet

Parquet es un formato de datos columnar (basado en columnas) creado para optimizar el almacenamiento y obtención de datos. Este mismo provee una compresión para manejar grandes cantidades de datos. No está ligado a ningún lenguaje y acepta tipos de datos complejos.

Especulative Execution

En big data se separan los datos en partes y las transformaciones se aplican a estas en paralelo. Un sistema distribuido puede tener diferentes características desde un punto de vista de hardware y software. Esto puede causar demoras, gracias a que diferentes máquinas procesan los datos en tiempos diferentes y el tiempo de ejecución completo está dado por el mayor tiempo de ejecución. En spark, ya que este posee replicación de datos, se genera el procesamiento de las particiones en varios servidores al mismo tiempo, al mismo tiempo, se miden las velocidades de estos y se utilizan solo los más rápidos (los más lentos se eliminan).