**Apuntes**

* Hadoop es un framework opensource para almacenar datos y ejecutar aplicaciones en clusters de hardware básicos.
* Los datos se almacenan en disco, al contrario de Spark (por eso Spark es más rápido ya que trabaja en memoria)
* Hadoop es escalable horizontalmente y con coste reducido, con tolerancia a fallos.
* HDFS almacena datos en un clúster y MapReduce procesa los datos en un clúster
* En un clúster únicamente debe de contener un Job. Cada nodo esclavo realzar paralelamente un Task.
* En HDFS los datos se almacenan en diferentes nodos, por lo que es bastante costoso actualizar datos, por lo que es más recomendable directamente borrar los datos y volver a insertarlos
* MapReduce se compone de 3 fases:
  + Map
  + Sort & Shuffle
  + Reduce

**Ejercicios**

* Para entrar al sistema de ficheros de Hadoop: **hadoop fs**
* En ella, se puede realizar la mayoría de comandos bash del sistema de ficheros de linux: hadoop fs –ls
* Por defecto, la ruta absoluta principal (al menos para los ejercicios) es /user/cloudera
* Para poner un fichero dentro de hdfs: hadoop fs –put “/home/cloudera/ejercicios/shakespeare” /user/cloudera/shakespeare
* Comando para leer los 50 últimas líneas: hadoop fs -cat /user/cloudera/shakespeare/histories | tail -n 50
* Copiar del sharedFolders el archivo wordcount: cp /media/sf\_sharedFolders/wordcount.txt /home/cloudera/ejercicios
* Para ejecutar nuestro primer MapReduce: hadoop jar wordcount/wc.jar solution.WordCount /user/cloudera/shakespeare /user/cloudera/wordcounts
* Para ver el contenido de lo ejecutado: hadoop fs -ls /user/cloudera/wordcounts
* Para terminar un proceso MapReduce mientras se está ejecutando: mapred job –list. mapred job -kill jobid