# Preguntas a responder

Las rutas que no empiezan por “/” son relativas al directorio de hadoop “/opt/hadoop/”.

## ¿Dónde se crea hdfs? ¿Cómo se puede elegir su localización?

HDFS está constituido por *namenodes* y *datanodes*. El lugar donde un *datanode* almacena sus metadatos se define en la propiedad “dfs.name.dir” del fichero “etc/hadoop/hdfs-site.xml”. Por otra parte, el *datanode* guarda sus datos en el directorio indicado en la propiedad “dfs.datanode.data.dir” del mismo fichero. Si no se indica alguna de estas propiedades (teniendo en cuenta que hemos ejecutado todos los comandos como *root*), por defecto se ubican en “/tmp/hadoop-root/dfs/data/”.

## Si estás utilizando hdfs ¿cómo puedes volver a ejecutar WordCount como si fuese *single node*?

Para ejecutar *WordCount* tomando los datos del sistema de ficheros local, podemos eliminar la propiedad “fs.defaultFS” del fichero “etc/hadoop/core-site.xml”.

## En el fragmento de El Quijote

### ¿Cuáles son las 10 palabras más utilizadas?

|  |  |
| --- | --- |
| Palabra | Apariciones |
| que | 3055 |
| de | 2816 |
| y | 2585 |
| a | 1428 |
| la | 1423 |
| el | 1232 |
| en | 1155 |
| no | 916 |
| se | 753 |
| los | 696 |

### ¿Cuántas veces aparecen las palabras “el” y “dijo”?

La palabra “el” aparece 1232 veces y “dijo”, 272.

# Memoria: Modificaciones a *WordCount*

En un primer momento, ejecutamos el programa *WordCount* original. Para hacerlo seguimos los siguientes pasos:

* Creamos un directorio en el que guardar los datos de entrada del programa.

> hdfs dfs -mkdir /user/bigdata/entrada\_cabana\_gomez

* Copiamos los datos del sistema de ficheros local al directorio recién creado en hdfs.

> hdfs dfs -copyFromLocal quijote.txt /user/bigdata/entrada\_cabana\_gomez/datos.txt

* Ejecutamos el *WordCount* de ejemplo.

> hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.1.jar wordcount /user/bigdata/entrada\_cabana\_gomez/datos.txt /user/bigdata/salida\_cabana\_gomez

* Recuperamos la salida.

> hdfs dfs -get /user/bigdata/salida\_cabana\_gomez/

* La salida de este programa se encuentra en la carpeta de entrega con el nombre “salida\_cabana\_gomez\_repaso/”.

El problema de esta implementación es que considera como dos palabras distintas a la misma palabra comenzando con mayúscula o con minúscula. Además, si hay un signo de puntuación pegado a la palabra, lo considera parte de ella y distingue entre las distintas combinaciones.