Práctica Hadoop

Procesamiento de Datos a Gran Escala

Antonio Coín Castro
Luis Antonio Ortega Andrés

13 de octubre de 2020

Solución WordCount

Para hacer que el fichero `.jar` generado por el script tenga otro nombre debemos considerar un argumento mas (`$2`) y realizar el siguiente cambio en la última línea de ejecución del script:

```
jar -cvf $2.jar -C ${file} .
```

Para conseguir que el programa WordCount que hemos tomado no tenga en cuenta los signos de puntuación ni las mayúsculas y minúsculas, debemos realizar dos cambios distintos:

- Para evitar los signos de puntuación, los añadimos al conjunto de delimitadores de palabras. Para ello configuramos el parámetro de delimitadores de la clase `StringTokenizer` de `Java`.
- Para no distinguir mayúsculas y minúsculas, pasamos cada uno de los caracteres de cada palabra a minúscula utilizando la función `toLowerCase()`.

```
StringTokenizer itr = new StringTokenizer(value.toString()
            .toLowerCase(), " \n\t\r\f.,;:-¡¿?!()\"\'");
```

En la solución obtenida al ejecutar directamente los ejemplos hadoop map-reduce (archivo `Resultados_ejemplo`) podemos ver como no se utilizan separadores correctos, diferenciando por ejemplo las palabras `que` y `(que`. Ante esto, los resultados obtenidos mediante la versión modificada resultan ser más precisos.

Cuestiones planteadas

Pregunta 1. *¿Dónde se crea hdfs? ¿Cómo se puede elegir su localización?*

La localización del sistema de archivos HDFS la determina la variable `dfs.datanode.data.dir`, cuyo valor por defecto según la documentación de hadoop es `file://$hadoop.tmp.dir/dfs/data` Este valor se puede cambiar en el archivo `hdfs-site.xml`:

```
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/hadoop/hdfs/datanode</value>
</property>
```

Pregunta 2. *Si estás utilizando hdfs, ¿cómo puedes volver a ejecutar WordCount como si fuese single.node?*

Para volver a configurar hadoop para funcionar como single.node debemos revertir los cambios que hicimos en los archivos de configuración para habilitar HDFS. Es decir, eliminar las secciones `fs.defaultFS` del archivo `core-site.xml` y `dfs.replication` de `hdfs-site.xml`.

Pregunta 3. *En el fragmento del Quijote, ¿cuales son las 10 palabras más utilizadas? ¿Cuántas veces aparecen el artículo "el" y la palabra "dijo"?*

Para resumir la información utilizamos codigo `Python`. Creamos un dataframe de `Pandas` con la salida facilitada por hadoop:

```
df = pd.read_csv("salida", delimiter="\t", header=None)
```

Para obtener las 10 palabras más frecuentes ordenamos el dataframe:

```
df.sort_values(by=1, ascending=False, inplace=True)
print(df.head(10))
```

Las 10 palabras mas utilizadas son: `que` (3055), `de` (2816), `y` (2585), `a` (1428), `la` (1423), `el` (1232), `en` (1155), `no` (916), `se` (753) y `los` (696).
Para buscar el número de ocurrencias de una palabra, buscamos si fila correspondiente:

```
print(df.loc[df[0] == "el"])
print(df.loc[df[0] == "dijo"])
```

La palabra `el` aparece un total de 1232 veces y la palabra `dijo` 272