Práctica Hadoop

Procesamiento de Datos a Gran Escala

Antonio Coín Castro
Luis Antonio Ortega Andrés

18 de octubre de 2020

Puesta a punto del entorno Hadoop

Una vez que hemos instalado Hadoop correctamente en el sistema, y hemos editado los archivos de configuración necesarios para activar el modo *pseudo-distributed* con HDFS, iniciamos este último con la siguiente orden (partiendo del directorio de instalación de Hadoop):

```
sbin/start-dfs.sh
```

También creamos un directorio en el sistema de archivos HDFS donde vamos a alojar nuestros conjuntos de datos:

```
hdfs dfs -mkdir /user/bigdata/p1
```

No debemos olvidarnos de copiar nuestros archivos a HDFS con la orden `-copyFromLocal`:

```
hdfs dfs -copyFromLocal <ruta_local> <ruta_hdfs>
```

Compilación y ejecución

Utilizamos el script `compile.sh` para compilar nuestro programa y obtener un `.jar` listo para ser ejecutado. Añadimos una pequeña modificación al script original, permitiendo que el ejecutable generado tenga un nombre a nuestra elección. Para ello debemos considerar un argumento más (`$2`) y realizar el siguiente cambio en la última línea del script:

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

El fichero completo quedaría entonces como sigue:

```
#!/bin/bash
file=$1
name=$2
HADOOP_CLASSPATH=$(hadoop classpath)
rm -rf ${file}
mkdir -p ${file}
javac -classpath $HADOOP_CLASSPATH -d ${file} ${file}.java
jar -cvf ${name}.jar -C ${file} .
```

Finalmente, podemos ejecutar nuestro programa en el entorno pseudo-distribuido con la siguiente orden:

```
bin/hadoop <name>.jar <class_name> /user/bigdata/p1/<input_file> \
  <output_dir>
```

1. Solución WordCount

Nos planteamos primero el problema de construir un programa para contar el número de apariciones de cada una de las palabras de un texto, que en este caso será un fragmento del Quijote (archivo `Quijote.txt`).

La estrategia a seguir dentro del paradigma de programación MapReduce será la siguiente. En primer lugar, dividimos el texto en unidades de un cierto tamaño, y pasamos cada uno de estos trozos a los Mappers. En la fase distribuida de Map, extraemos las palabras (que en principio suponemos separadas por espacios en blanco, tabulaciones o retornos de carro) y enviamos a los Reducers parejas `(palabra, 1)`, indicando que esa palabra en concreto aparece una vez. Después, en la fase de Shuffle se juntan las parejas cuya clave es igual (en este caso, la clave es la propia palabra), y se agrupan sus segundos elementos en una lista. Finalmente, en la fase Reduce se reciben parejas `(palabra, (1, 1, 1, ...))`, y lo que hacemos es sumar todos los elementos de la lista en el segundo elemento de la pareja, para obtener el número total de apariciones. Finalmente devolvemos parejas de palabras junto a su conteo.

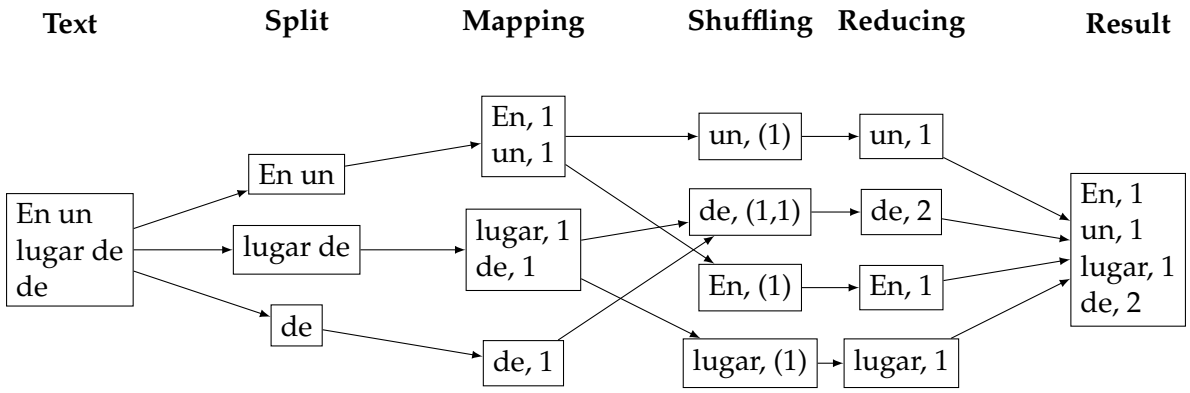

Figura 1: Esquema del programa WordCount.

En una sintaxis más funcional, podríamos escribir el programa como

```
words_by_line = text.map { line => line.split(" ") }
result = word_by_line.map { w => (w, 1) }
                     .reduceByKey { (v1, v2) => v1 + v2 }
```

Además de esto, modificamos el programa para que no tenga en cuenta mayúsculas y minúsculas, ni signos de puntuación. Veamos con detalle los principales elementos del programa en Java.

Mapper

Definimos una clase para nuestro Mapper que extiende la interfaz homónima de Hadoop, proporcionando el formato de entrada y salida de nuestras parejas. En concreto, recibimos parejas `(Object, Text)` (donde ignoramos la clave) y proporcionamos como salida parejas `(Text, IntWritable)`. Las clases `Text` e `IntWritable` son clases de Hadoop que representan a los tipos de datos de Java `String` e `Int`, respectivamente [1]. En el método `map` tokenizamos la entrada con `StringTokenizer` (dividir en palabras), y para cada palabra guardamos en la salida la propia palabra junto con la constante 1.

```
public static class TokenizerMapper
  extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context)
      throws IOException, InterruptedException {
      StringTokenizer itr =
        new StringTokenizer(value.toString(), " \n\t\r\f");
      while (itr.hasMoreTokens()) {
          word.set(itr.nextToken());
          context.write(word, one);
      }
    }
  }
```

Para conseguir que el programa 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 del constructor de la clase `StringTokenizer`.
- Para no distinguir mayúsculas y minúsculas, pasamos cada uno de los caracteres de cada palabra a minúscula utilizando la función `toLowerCase()`.

Así, la línea que modificamos queda como sigue:

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

Reducer

En el Reducer recibimos parejas de la forma `(Text, Iterable<IntWritable>)` con las palabras y una lista con '1's (uno por cada aparición contabilizada), y devolvemos parejas

[1] Estas clases optimizan y reducen el *overhead* a la hora de serializar y deserializar objetos.

`(Text, IntWritable)` con las palabras y el número total de apariciones. Para calcular este último número simplemente sumamos los elementos de la lista asociada a cada palabra.

```
public static class IntSumReducer extends
  Reducer<Text,IntWritable,Text,IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
      Context context) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}
```

En este caso no hay que realizar ninguna modificación para incluir los cambios que queríamos.

Main

Desde la función principal solo tenemos que establecer la configuración apropiada del entorno de trabajo, especificando el nombre de las clases que hemos creado y que implementan las fases de Map y Reduce. También es necesario especificar el formato final de salida de las parejas `(clave, valor)`, que como ya dijimos será `(Text, IntWritable)`. También añadimos las rutas del fichero de entrada y del directorio de salida, e invocamos el trabajo (`Job`) que hemos creado.

```
public static void main(String[] args) throws Exception {
    Job job = new Job(conf, "wordcount");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
```

1.1. Conclusiones

Los resultados de ejecución de nuestro programa sin modificar se encuentran en el archivo `wc_orig`, mientras que la salida del programa modificado está en el archivo `wc_modif`. Podemos ver como en el primero aparecen repetidas varias palabras, ya sea por capitalización de la primera letra cuando comienzan oraciones, o porque se contabilizan los signos de puntuación. Sin embargo, en el segundo archivo esto ya no ocurre.

En la solución obtenida al ejecutar directamente los ejemplos de `hadoop-map-reduce` (archivo `wc_example`) 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. Concretamente, en la versión de ejemplo el número total de palabras distintas que se contabilizan es de 10533, mientras que en la versión modificada este número desciende hasta 7492.

La versión de ejemplo y nuestra versión no modificada coinciden en que ambas detectan como palabras diferentes algunas que deberían ser consideradas la misma. Además, podemos ver que el número total de palabras contabilizadas es el mismo en ambas versiones.

1.2. 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 es `file://${hadoop.tmp.dir}/dfs/data` según la documentación de Hadoop. 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`.

Otra opción para mantener activo HDFS pero ejecutar nuestro programa con archivos de entrada locales es especificar en el código fuente que se puede recibir como entrada un archivo del sistema de archivos local y no de HDFS. Esto podemos lograrlo por ejemplo empleando la clase `FileSystem`:

```
Configuration conf = new Configuration();
Path path = new Path(args[0]);
FileSystem fs = FileSystem.get(path.toUri(), conf);
```

Y entonces al ejecutar el programa podemos especificar como argumento archivos locales con la ruta `file://path/to/file` y archivos en HDFS con `hdfs://path/to/file`.

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 del archivo de salida utilizamos código `Python` (archivo `utils.py`). En primer lugar, creamos un `dataframe` de `Pandas` con la salida facilitada por Hadoop:

```
df = pd.read_csv("wc_modif", 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 su 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` aparece 272 veces.