Práctica 1. Hadoop y Spark

Blanca Cano Camarero y Iker Villegas Labairu

8-10-22

Índice

Práctica 1 Hadoop y Spark

Instalación

Para la instalación se ha seguido estrictamente el tutorial de clase, además para agilizar el proceso y puesto que en ambiente actual es interesante el uso de contenedores, hemos realizado un script de bash, para automatizar el proceso de instalación hasta la parte de ssh en un contenedor de docker con `CentOS:7`:

```
echo "Task 1: Updating system"
yum -y update && yum clean all
echo "Task 2: Installing packages"
yum -y install wget
yum -y install which
yum -y install rsync
yum -y install java-1.7.0-openjdk
echo "Task 3: Hadoop installation (may take a while)"
wget https://archive.apache.org/dist/hadoop/common/hadoop-2.8.1/hadoop-2.8.1.tar.gz
tar xvzf hadoop-2.8.1.tar.gz
mv hadoop-2.8.1 opt/
cd opt/ && ln -s hadoop-2.8.1 hadoop
export JAVA_HOME=/usr/lib/jvm/jre-1.7.0-openjdk
echo 'export JAVA_HOME=/usr/lib/jvm/jre-1.7.0-openjdk' >> ./hadoop/etc/hadoop/hadoop-env.sh
# Comando de prueba
#./opt/hadoop/bin/hadoop

yum -y install openssh-server openssh-clients
```

El respectivo Dockerfile sería el siguiente:

```
FROM centos:7
COPY ./initial.sh /
COPY ./material /
RUN bash ./initial.sh
EXPOSE 50070
```

Donde la carpeta material contiene:

```
(main)> tree material/
material/
|--WordCount_final.java
|-- quijote.txt
```

Preguntas expuestas en las diapositivas

Las diapositivas con la guía de instalación) estaba salpicada de preguntas, vamos a proceder a reponderlas:

Ejecicio 1.1

¿Qué ficheros ha modificado para activar la configuración del HDFS?

Bajo supuesto de que la instalación anterior es correcta y las variables de entorno (la versión de Java) ya se hayan exportado.

Los ficheros que debemos de instalar son

`/hadoop/hdfs/hds-site.xml`

Al cual se le ha añadido la configuración (copiar de la pag 38)

¿Qué líneas ha sido necesario modificar?

Ha sido necesario añadir el valor de la la variable tal escrita como.

Es neceario además ejecutar el comando `bin/dfs namenode -format` que afecta a ciertos ficheros así que estos también se ven modificados.

Ejercicio 1.2

Para pasar a la ejecución de Hadoop sin HDFS ¿es suficiente con parar el servicio con `stop-dfs.sh`? ¿Cómo se consigue? Sí, hay que ejecutar `stop-dfs.sh` esto es con el comando `sbin/stop-dfs.sh` ya que para todos los servicios, en particular la ejecución de Hadoop con HDFS.

Ejerciccio 3.1

Modificar el ejemplo de WordCount que hemos tomado como partida, para que no tenga en cuenta signos de puntuación ni las mayúsculas ni las minúsculas volver a ejecutar la aplicación

Para conseguir esto hemos modificado el método map de la siguiente manera

```java
public void map(Object key, Text value, Context context) throws
    IOException, InterruptedException {
       String minimized_value = value.toString().toLowerCase();
       String processed_value = minimized_value.replaceAll("[^áéíóúüña-
           z0-9]+","");
       StringTokenizer itr = new StringTokenizer(processed_value));
       while (itr.hasMoreTokens()) {
         word.set(itr.nextToken());
         context.write(word, one);
       }
     }
```