Analyzing Apache Access Logs in Apache Spark

The following statistics are analyzed:

1. The count of response code's returned from the server.
2. The content size of responses returned from the server to host.
3. The top ten most popular URL's in the Apache log.
4. The average, min, and max content size of responses returned from the server.

The steps to set up and execute Log Analyzer Program by using Spark and Java in Cloudera

Before proceed the below steps, we have to install the Cloudera Quickstart vm 5.5 and VMwareplayer. The Hadoop 2.6, Java 1.7, Eclipse Luna, Hive, Hbase, Spark, and all required libraries have been included in cloudera.

1. In order to use Lambda expression, we have to upgrade the java from 1.7 to 1.8. The bellow steps illustrate the installation of java 8.
   1. Download the .tar.gz file for one of the supported versions of the Oracle JDK from Java SE 8 Downloads
   2. Put the package to the installation path, say, /usr/java/
   3. Unzip the package

      sudo tar zxf jdk-8u5-linux-x64.tar.gz
   4. Set JAVA_HOME and PATH for current user, in ~/.bashrc or ~/.bash_profile

      export JAVA_HOME=/usr/java/jdk1.8.0_60

      export PATH=$PATH:$JAVA_HOME/bin
   5. Verify Java 8 is properly installed by typing

      cloudera@quickstart ~]$ java -version

      java version "1.8.0_60"

      Java(TM) SE Runtime Environment (build 1.8.0_60-b27)
2. Download the apache log file from http://www.monitorware.com/en/logsamples/apache.php and unzip it.
3. Create a new maven project in eclipse.
4. Configure pom.xml to download the dependency of the Spark Core library.

   <dependency>

   <groupId>org.apache.spark</groupId>

   <artifactId>spark-core_2.10</artifactId>

   <version>1.2.0</version>

   </dependency>
5. Create and implement the LogAccess and LogAnalyzer classes by using Java and Spark API.
6. Build a maven project(Right click on project in Eclipse and select Run As -> Maven build).

7. Copy the jar file in the target directory in Eclipse project(LogAnalyzer-0.0.1-SNAPSHOT.jar) and downloaded apache log file(access_log) into cloudera's home directory.
8. Create a loganalyzer/input directory named path in HDFS.

   hadoop fs -mkdir -p /user/cloudera/loganalyzer/input
9. Copy the log file from the local file system to directory within the HDFS.

   hadoop fs -put access_log /user/cloudera/loganalyzer/input/
10. Execute the jar file with created HDFS input by using Linux command in Terminal

    spark-submit --class edu.mum.cs522.logs.LogAnalyzer --master local[1] LogAnalyzer-0.0.1-SNAPSHOT.jar /user/cloudera/loganalyzer/input/ access_log /user/cloudera/loganalyzer/rescoderesult /user/cloudera/loganalyzer/topurlresult /user/cloudera/loganalyzer/ressizeresult
11. The analyzed results can be viewed in Hue
    1. Go to the browser and click on Hue in browser(http://quickstart.cloudera:8888)
    2. Give username as cloudera and password as cloudera
    3. Click on File Browser and then go to the directory and then open the file