Big Data: How can I add Apache Mahout to my Hadoop instance?

How can I add Apache Mahout to my Hadoop instance?

Apache Mahout is a suite of machine learning libraries. Depending on the algorithm, Mahout can work with or without Hadoop.

We will show how Mahout can be added to a Bright cluster that has a Hadoop instance already installed. In this case it is "CDH5.2.1", and uses Cloudera CDH 5.2.1. An example is given of the use of Mahout to run Mapreduce jobs.

1. Download Apache Mahout tarball and unpack it

Execute the following commands on the active head node as root user:
cd /tmp/
curl -O http://archive.cloudera.com/cdh5/cdh/5/mahout-0.9-cdh5.2.1.tar.gz
/cm/shared/apps/hadoop/Cloudera
tar xvzf /tmp/mahout-0.9-cdh5.2.1.tar.gz

2. Grant access to HDFS for user "foobar"

Granting access will create directory /user/foobar in HDFS.
cmsh
% user use user foobar
% set hadoophdfsaccess cdh5.2.1
% commit

3. Prepare execution of Mahout test

For the Naive Bayes classifier test, a sample of Wikipedia articles in xml format will be used.

su - foobar

```
$ curl -O http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles1.xml-p000000010p000010000.bz2
$ bunzip2 enwiki-latest-pages-articles1.xml-p000000010p000010000.bz2
$ module load hadoop/CDH5.2.1/Cloudera/2.5.0-cdh5.2.1
$ hdfs dfs -mkdir /user/foobar/wiki
$ hdfs dfs -copyFromLocal enwiki-latest-pages-articles1.xml-p000000010p000010000 /user/foobar/wiki
$ hdfs dfs -ls /user/foobar/wiki
```

4. Execute Mahout job (as YARN application) and check result

```
# su - foobar
$ /cm/shared/apps/hadoop/Cloudera/mahout-0.9-cdh5.2.1/bin/mahout seqwiki -i /user/foobar/wiki/enwiki-latest-pages-articles1.xml-p000000010p000010000 -o /user/foobar/wiki/seqfiles
$ hdfs dfs -ls /user/foobar/wiki/seqfiles
```

Unique solution ID: #1243
Author: Michele Lamarca
Last update: 2015-01-09 11:57