Hadoop 学习笔记

Jun-10

Follow http://my.oschina.net/zhangdengpan/blog/356641 to set up development and debuggin in IntelliJ.

sbin/start-all.sh to start hadoop services.

use bin/hadoop fs -put localdir hdfsdir to upload the data to hadoop hdfs first before running.

command to run hadoop:

bin/hadoop jar /home/xiusan/IdeaProjects/Hadoop/out/artifacts/Hadoop_jar/Hadoop.jar com.xiusan.test.Hadoop.WordCount /user/xiusan/systemmonitorlog/input /user/xiusan/output

Jun 11

Prepare data for analyze: keep component and operation, get count for all 10k data files.
Decide the vector factors

Jun 12
Decide vector factors: Calibration operations, ~15 operations as a vector
Computer vacter value and normalized vector value. 1) vector by numbers; 2) vector by normalizing to 100.
Generate Hadoop archive and run a job
Try Mahout K-means algorithm

Jun 16
Set up Mahout
Follow http://hadoop.readthedocs.org/en/latest/Hadoop-Mahout.html

mahout vectordump -i /usr/mahout/output/output/data/part-m-00000 to view the result.

Jun 17

Practice KMeans

https://mahout.apache.org/users/clustering/k-means-clustering.html is the basic guidline for kmeans
https://github.com/apache/mahout/blob/master/examples/bin/cluster-reuters.sh is the script for KMeans, need set HADOOP_HOME as environment variable first.

Sequentialize directory:
mahout seqdirectory -i ./TestData/Extracted/Vectorized/ByCount/ -o ./TestData/Extracted/Vectorized/Seq to vectorized the data.

Sequence to vectors:
mahout seq2sparse -i /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/Seq -o /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/MahoutVectorized

Run K-means:

mahout kmeans -i /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/MahoutVectorized/tfidf-vectors/ -o /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/KmeansOutput --numClusters 5 --clusters /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/KmeansOutput/Clusters --maxIter 10

View Result:
mahout clusterdump -i /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/KmeansOutput/clusters-5-final --pointsDir /usr/hadoop-2.7.0/TestData/Extracted/Vectorized/KmeansOutput/clusteredPoints

Prepare the cluster, no need if --numClusters 5 is set already.

```
bin/mahout spectralkmeans \
    -i <affinity matrix directory> \
    -o <output working directory> \
    -d <number of data points> \
    -k <number of clusters AND number of top eigenvectors to use> \
    -x <maximum number of k-means iterations>
```

https://mahout.apache.org/users/clustering/k-means-clustering.html