Hadoop Implementation Project

Course and Certification Project

edureka!

Version 2.2

Hadoop Implementation Project

Course and Certification Project

Table of Contents

Course Project

Understand the file location on Data Nodes:

1) Setup a minimum 2 Node Hadoop Cluster, either using YARN or older version.

   Make sure all the daemons like: namenode, datanode, jobtracker, tasktracker are up.

   Note: Make sure HDFS block size is 64 MB globally and replication is N/2, where N is the number of Data Nodes.

   - Node 1 - Namenode, datanode, tasktracker
   - Node 2 - Jobtracker, datanode, tasktracker (if possible run Jobtracker on a separate node, if you have resources)

2) WordCount and find the location of the blocks:
   - Create a simple text file called "file1.txt" with some text.
   - Copy the file to HDFS.
     - Find out which node it went and why it went to that data node?
       - Run MapReduce - Find which datanode the output files are written and why?

   Note: Please document the steps used to find the blocks of file.

3) Create a file "largefile.txt" with some text.
   - Copy the file to HDFS with block size = 256 MB. All other files must have the default block size.
   - How the block size impacts performance? Give some positives and Negatives for it.
   -

4) Let say you set a spaceQuota=200MB of /projects and try to copy a file of size=70MB, with replication=2. It is not letting you copy the file, what could be the reason?
   - Assuming that this is the most important file to copy to HDFS, how will you solve this without increasing quota? Give the command.

5) Configure Rack awareness and after that copy a file to hdfs.
   - Find out the rack distribution for that file. Tell the command used for it.
   - How will you change the replication factor of an existing file?

Certification project

Problem Statement 1:

Please setup a Hadoop1.0, Single Node or 2 node Cluster depending upon the resources you have.

Please make sure the following conditions are met:

1) All daemons like namenode, datanode, jobtracker, tasktracker must run in the cluster
2) (Single node or multinode).
3) Default Block size must be 128 MB.
4) Write the cluster NamespaceID of your cluster (_____).
5) Create a directory /projects on the hadoop cluster and let the following conditions:
    - set Namespace quota of 10 on the directory.
    - set Space Quota of 100MB on the directory.
6) Use distcp command to copy /projects to /new on the same cluster.
7) Create a list of DataNodes participating in the cluster and save the list in file "list_of_datanodes" on your local machine.

Problem Statement 2:

1) Save the namespace of the Namenode, without using secondary namenode. fsimage and edits file must merge, without stopping the namenode daemon.
2) Set include file, so that no other nodes can talk to the namenode, other than what are in the include file.
3) Set cluster Re-balancer threshold to 40%.
4) Set the map and reduce slots to 4 and 2 respectively for each node.