**YARN Benchmarking Project**

**Objectives and description:**

1. Identify how to configure following parameters in YARN:
   1. MapSlots/node
   2. ReduceSlots/node
   3. Block size
   4. Replication Factor
   5. Compress (Map output – intermediate data)
2. Update the script (written for Hadoop 1) given in appendix to execute micro benchmarks (randomtextwriter, randomwriter, wordcount and sort) in YARN.
3. Write a script to collate job performance results after the script execution. The script needs to take jobs output log as input and create summary report of jobs execution. The summary for each job should be in following csv format:

iteration,Date,StartTime,EndTime,freq,NoofSlvs,Compress,DataBlkSize,ReplicationFactor,JobName,InputDataSize,MapSlots,ReduceSlots,JobId,JobStartTime,JobEndTime,JobResponseTime,AvgMapTime,AvgReduceTime,No\_KilledMapTasks,No\_KilledRedTasks,No\_FailedMaps,NoFailedReduceTasks, NoofDataLocalMaps

1,12/11/2014,1323,132500,1.6,10,No,64M,10,randomtextwriter,30GB,MS2,RS2,job\_201411121323\_0001,1415778935571,1415779058404,2.04722,0.995741,-nan,14,0,0,0,10

Appendix: Hadoop 1 script

#!/bin/bash

for noofSlvs in 10 #8 6 4 2 1

do

slaves\_files=slaves"-"$noofSlvs

cat $slaves\_files > slaves

cp /home/hduser1/scripts/slaves /usr/local/hadoop/conf

for inDataGB in 5 10 15 20

do

((inDataVol=$inDataGB\*1073741824))

for Jobs in randomwriter\_sort randomtextwriter\_wordcount

do

Job1=`echo $Jobs |cut -d '\_' -f 1`

Job2=`echo $Jobs |cut -d '\_' -f 2`

JobNmLn=`expr length $Job1`

echo "JobNmLn is "$JobNmLn

let JobNmLn1=$JobNmLn-1

echo "JobNmLn1 is "$JobNmLn1

DataVar=`expr ${Job1:0:$JobNmLn1}`

echo "Job1 is "$Job1 " Job2 is "$Job2

echo "DataVar is "$DataVar

#bin/stop-mapred.sh

#bin/stop-dfs.sh

#Iteration

t=3

datetimest=`date "+%F\_%T"`

logfile=$curfreq"\_rndwtrWC\_xMS\_$datetimest"

for mapslots in 1-2-No 2-2-No 4-2-No 6-2-No

do

echo $Job1" MS"$mapslots"\_RS2\_"$curfreq > /home/hduser1/10slaves\_logs/15Apr15/$logfile

mapredfile="mapred-site-"$mapslots".xml"

cat $mapredfile > mapred-site.xml

echo "Map slots file - "$mapredfile

cp /home/hduser1/scripts/mapred-site.xml /usr/local/hadoop/conf

scp /home/hduser1/scripts/mapred-site.xml hduser1@node1:/usr/local/hadoop/conf

scp /home/hduser1/scripts/mapred-site.xml hduser1@blrkec331317d:/usr/local/hadoop/conf

scp /home/hduser1/scripts/mapred-site.xml

for hdfs in 16M-1 16M-3 16M-5 16M-7 16M-9 16M-10 64M-1 64M-3 64M-5 64M-7 64M-9 256M-1 256M-5 256M-7 256M-9 256M-10 256M-3 1GB-3 1GB-1 1GB-5 1GB-7 1GB-9 1GB-10

do

replica=`echo $hdfs | cut -d "-" -f 2`

echo "in data volume is "$inDataVol

hdfsfile="hdfs-site-"$hdfs".xml"

cat $hdfsfile > hdfs-site.xml

echo "hdfs file - "$hdfs

cp /home/hduser1/scripts/hdfs-site.xml /usr/local/hadoop/conf

scp /home/hduser1/scripts/hdfs-site.xml hduser1@node1:/usr/local/hadoop/conf

scp /home/hduser1/scripts/hdfs-site.xml hduser1@blrkec331317d:/usr/local/hadoop/conf

scp /home/hduser1/scripts/hdfs-site.xml

rm -rf /usr/local/hadoop/logs/\*

ssh hduser1@node1 rm -rf "/usr/local/hadoop/logs/\*"

logdirnm=$noofSlvs"-Slvs\_MS-"$mapslots"-compRS2\_blk-"$hdfs"-Rep-"$inDataGB"-GB"

Job1logdirpth="/home/hduser1/10slaves\_logs/15Apr15/"$curfreq"/CHNG\_MS/"$Job1"/"$logdirnm

echo "Job1logdirpth is : "$Job1logdirpth

mkdir -p $Job1logdirpth

chmod 775 $Job1logdirpth

Job2logdirpth="/home/hduser1/10slaves\_logs/15Apr15/"$curfreq"/CHNG\_MS/"$Job2"/"$logdirnm

echo "Job2logdirpth is : "$Job2logdirpth

mkdir -p $Job2logdirpth

chmod 775 $Job2logdirpth

cd /usr/local/hadoop

for ((c=1;c<=$t;c++))

do

bin/stop-mapred.sh

bin/stop-dfs.sh

rm -rf /app/hadoop/tmp/\*

ssh hduser1@node1 rm -rf "/app/hadoop/tmp/\*"

bin/hadoop namenode -format

sleep 1m

bin/start-dfs.sh

bin/start-mapred.sh

sleep 2m

bin/hadoop dfs -rmr /user/hduser1/out\*

datetimest=`date "+%F\_%T"`

echo "Start "$Job1" MS "$mapslots", RS 2, iteration = $c at $datetimest" >> /home/hduser1/10slaves\_logs/15Apr15/$logfile

bin/hadoop jar hadoop-\*examples\*.jar $Job1 -Dtest.$DataVar.total\_bytes=$inDataVol /user/hduser1/output\_RandomWriter

datetimest=`date "+%F\_%T"`

echo "Stop "$Job1" MS "$mapslots", RS 2, iteration = $c at $datetimest" >> /home/hduser1/10slaves\_logs/15Apr15/$logfile

# bin/hadoop dfs -rmr /user/hduser1/out\*

bin/stop-mapred.sh

bin/stop-dfs.sh

#cp -r /usr/local/hadoop/logs/history/\*/\*/\*/\*/\*/\*/\*/\*hduser1\* /home/hduser1/10slaves\_logs/15Apr15/onDemand/CHNG\_MS/RandomWriter/$logdirnm

cp -r /usr/local/hadoop/logs/history/\*/\*/\*/\*/\*/\*/\*/\*hduser1\* $Job1logdirpth

rm -rf /usr/local/hadoop/logs/\*

ssh hduser1@node1 rm -rf "/usr/local/hadoop/logs/\*"

bin/start-dfs.sh

bin/start-mapred.sh

sleep 2m

bin/hadoop dfs -rmr /user/hduser1/Sort\*

datetimestc=`date "+%F\_%T"`

echo "Start "$Job2" MS "$mapslots", RS 2, iteration = $c at $datetimestc" >> /home/hduser1/10slaves\_logs/15Apr15/$logfile

bin/hadoop jar hadoop-\*examples\*.jar $Job2 /user/hduser1/output\_RandomWriter /user/hduser1/Sort\_output

datetimestc=`date "+%F\_%T"`

echo "Stop "$Job2" MS "$mapslots", RS 2, iteration = $c at $datetimestc" >> /home/hduser1/10slaves\_logs/15Apr15/$logfile

bin/hadoop dfs -rmr /user/hduser1/out\*

bin/hadoop dfs -rmr /user/hduser1/Sort\*

bin/stop-mapred.sh

bin/stop-dfs.sh

#cp -r /usr/local/hadoop/logs/history/\*/\*/\*/\*/\*/\*/\*/\*hduser1\* /home/hduser1/10slaves\_logs/15Apr15/onDemand/CHNG\_MS/Sort/$logdirnm

cp -r /usr/local/hadoop/logs/history/\*/\*/\*/\*/\*/\*/\*/\*hduser1\* $Job2logdirpth

rm -rf /usr/local/hadoop/logs/\*

ssh hduser1@node1 rm -rf "/usr/local/hadoop/logs/\*"

done

cd /home/hduser1/scripts

done

done

done

done

done