1.软件

HDP/CDH 🡪 hdfs/zookeeper

scala-2.10.4.zip

spark-1.6.3-bin-hadoop2.6.tgz

master:hdp1

smater:hdp2

worker : hdp1,hdp2,hdp3

前提：配置ssh免密码

2.基于HDP的搭建

2.1 安装软件准备

# cd /usr/local/greenplum-db

# source greenplum\_path.sh

软件位置：/usr/local

# gpscp -f slaves /usr/local/scala-2.10.4.zip =:/usr/local/

# gpssh -f all\_hosts -e "unzip /usr/local/scala-2.10.4.zip -d /usr/local/"

在hdp1上：

# vi /etc/profile

export SCALA\_HOME=/usr/local/scala-2.10.4

export PATH=.:$JAVA\_HOME/bin:$SCALA\_HOME/bin:$PATH

# gpscp -f slaves /etc/profile =:/etc/profile

# gpssh -f all\_hosts -e "source /etc/profile"

# gpssh -f all\_hosts -e "chmod 755 /usr/local/scala-2.10.4 -R"

# gpssh -f all\_hosts -e "scala -version"

# gpscp -f slaves /usr/local/spark-1.6.3-bin-hadoop2.6.tgz =:/usr/local/

# gpssh -f all\_hosts -e "tar -zxvf /usr/local/spark-1.6.3-bin-hadoop2.6.tgz -C /usr/local/"

2.2 spark配置文件

# cd /usr/local/spark-1.6.3-bin-hadoop2.6/conf

# cp slaves.template slaves

# vi slaves

hdp1

hdp2

hdp3

# cp spark-env.sh.template spark-env.sh

# vi spark-env.sh

export JAVA\_HOME=/usr/local/jdk1.8.0\_101

export SCALA\_HOME=/usr/local/scala-2.10.4

export HADOOP\_HOME=/usr/hdp/2.4.0.0-169/hadoop

export HADOOP\_CONF\_DIR=/etc/hadoop/conf

export SPARK\_HOME=/usr/local/spark-1.6.3-bin-hadoop2.6

export SPARK\_LOG\_DIR=/usr/local/spark-1.6.3-bin-hadoop2.6/logs

export SPARK\_PID\_DIR=/usr/local/spark-1.6.3-bin-hadoop2.6/run

export SPARK\_MASTER\_IP=hdp1

export SPARK\_MASTER\_PORT=7077

export SPARK\_DAEMON\_JAVA\_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:/usr/local/spark-1.6.3-bin-hadoop2.6/logs -XX:+UseParallelGC -XX:+UseParallelOldGC -XX:+DisableExplicitGC -Xms1024m -Xmx2048m -XX:PermSize=128m -XX:MaxPermSize=256m"

# cp spark-defaults.conf.template spark-defaults.conf

# vi spark-defaults.conf

spark.master=spark://hdp1:7077

spark.eventLog.enabled=true

spark.eventLog.dir=hdfs://hdp1:8020/user/spark/applicationHistory

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/slaves =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/spark-defaults.conf =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/spark-env.sh =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

启动集群

# cd sbin/

# ./start-all.sh

FAILED SelectChannelConnector@0.0.0.0:8080: java.net.BindException: Address already in use

因为ambari占用8080端口：

1是更改ambari端口，如8090

2.更改spark端口为其他，如8090

# vi start-master.sh

SPARK\_MASTER\_WEBUI\_PORT=8090

# vi start-slave.sh 个人习惯更改，可选

SPARK\_WORKER\_WEBUI\_PORT=8091

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/sbin/start-master.sh =:/usr/local/spark-1.6.3-bin-hadoop2.6/sbin/

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/sbin/start-slave.sh =:/usr/local/spark-1.6.3-bin-hadoop2.6/sbin/

Hive整合和HA配置

# gpssh -f all\_hosts -e

"cp /usr/hdp/current/hive-client/conf/hive-site.xml /usr/local/spark-1.6.3-bin-hadoop2.6/conf"

# vi spark-env.sh

export SPARK\_DAEMON\_JAVA\_OPTS="-Dspark.deploy.recoveryMode=ZOOKEEPER -Dspark.deploy.zookeeper.url=hdp2:2181,hdp3:2181,hdp1:2181 -Dspark.deploy.zookeeper.dir=/spark -verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:/usr/local/spark-1.6.3-bin-hadoop2.6/logs -XX:+UseParallelGC -XX:+UseParallelOldGC -XX:+DisableExplicitGC -Xms1024m -Xmx2048m -XX:PermSize=128m -XX:MaxPermSize=256m"

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/spark-env.sh =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

在hdp1

# ./start-all.sh

在 hdp2

# ./start-master.sh

3.基于CDH的搭建

3.1准备工作和hdp一样

3.2spark配置

# cp slaves.template slaves

# vi slaves

cdh1

cdh2

cdh3

# cp spark-defaults.conf.template spark-defaults.conf

spark.master=spark://cdh1:7077

spark.eventLog.enabled=true

spark.eventLog.dir=hdfs://cdh1:8020/user/spark/applicationHistory

# cp spark-env.sh.template spark-env.sh

# vi spark-env.sh

export JAVA\_HOME=/usr/local/jdk1.8.0\_101

export SCALA\_HOME=/usr/local/scala-2.10.4

export HADOOP\_HOME=/opt/cloudera/parcels/CDH-5.8.0-1.cdh5.8.0.p0.42/lib/hadoop

export HADOOP\_CONF\_DIR=/etc/hadoop/conf

export SPARK\_HOME=/usr/local/spark-1.6.3-bin-hadoop2.6

export SPARK\_LOG\_DIR=/usr/local/spark-1.6.3-bin-hadoop2.6/logs

export SPARK\_PID\_DIR=/usr/local/spark-1.6.3-bin-hadoop2.6/run

export SPARK\_MASTER\_IP=cdh1 //在cdh2上配置的是cdh2主机hostname

export SPARK\_MASTER\_PORT=7077

export SPARK\_DAEMON\_JAVA\_OPTS="-Dspark.deploy.recoveryMode=ZOOKEEPER -Dspark.deploy.zookeeper.url=cdh1:2181,cdh2:2181,cdh3:2181 -Dspark.deploy.zookeeper.dir=/spark -verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:/usr/local/spark-1.6.3-bin-hadoop2.6/logs -XX:+UseParallelGC -XX:+UseParallelOldGC -XX:+DisableExplicitGC -Xms1024m -Xmx2048m -XX:PermSize=128m -XX:MaxPermSize=256m"

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/slaves =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/spark-defaults.conf =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

# gpscp -f slaves

/usr/local/spark-1.6.3-bin-hadoop2.6/conf/spark-env.sh =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

3.3hive和spark整合

# gpscp -f all\_host

/etc/hive/conf.cloudera.hive/hive-site.xml =:/usr/local/spark-1.6.3-bin-hadoop2.6/conf

3.4在cdh1上 start-all.sh

在cdh2启动start-master.sh