Spark部署文档

集群信息

|  |  |  |
| --- | --- | --- |
| 主机名 | IP地址 | 角色 |
| master | 133.133.10.1 | ResourceManager & MR JobHistory Server & NameNode |
| slave1 | 133.133.10.11 | SecondaryNameNode &  DataNode & NodeManager |
| slave2 | 133.133.10.12 | DataNode & NodeManager |
| slave3 | 133.133.10.13 | DataNode & NodeManager |
| slave4 | 133.133.10.14 | DataNode & NodeManager |
| slave5 | 133.133.10.15 | DataNode & NodeManager |
| slave6 | 133.133.10.16 | DataNode & NodeManager |
| slave7 | 133.133.10.17 | DataNode & NodeManager |
| slave8 | 133.133.10.18 | DataNode & NodeManager |
| slave9 | 133.133.10.19 | DataNode & NodeManager |

无密码访问

Master节点上

ssh-keygen -t rsa

cp .ssh/id\_rsa.pub .ssh/authorized\_keys

chmod 600 .ssh/authorized\_keys

scp .ssh/authorized\_keys username@hostname:/home/xxx/

Slave节点上

ssh-keygen -t rsa

mv authorized\_keys .ssh

chmod 600 .ssh/authorized\_keys

集群问题：

如果要测试性能，需要统一集群的时间

master访问slave6和slave7的速度远远慢于访问其他slave的速度。

参考文档

<http://blog.csdn.net/zhu_xun/article/details/42077311>

部署

1. 修改spark-1.6.2-bin-hadoop2.7.1-dynamic/conf/spark-defaults.conf

|  |
| --- |
| spark.master spark://master:7077  spark.eventLog.enabled true  spark.eventLog.dir hdfs://master:9000/sparklogs  spark.serializer org.apache.spark.serializer.KryoSerializer  spark.driver.memory 2g  spark.yarn.historyServer.address master:18080  spark.executor.extraClassPath /home/lijie/spark-1.6.2-bin-hadoop2.7.1-dynamic/lib/hbase/\*  # spark.executor.extraJavaOptions -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"  spark.yarn.jar hdfs://master:9000/spark\_lib/spark-assembly-1.6.3-SNAPSHOT-hadoop2.7.1.jar  spark.dynamicAllocation.enabled true  spark.dynamicAllocation.executorIdleTimeout 60s  spark.dynamicAllocation.initialExecutors 1  spark.dynamicAllocation.minExecutors 0  spark.dynamicAllocation.maxExecutors 4  spark.dynamicAllocation.schedulerBacklogTimeout 1s  spark.shuffle.service.enabled true  spark.executor.cores 2  spark.executor.memory 2g |

1. 修改spark-1.6.2-bin-hadoop2.7.1-dynamic/conf/spark-env.sh

|  |
| --- |
| HADOOP\_CONF\_DIR=/home/lijie/hadoop-2.7.1/etc/Hadoop  SPARK\_YARN\_QUEUE=default  SPARK\_HISTORY\_OPTS="-Dspark.history.fs.logDirectory=hdfs://master:9000/sparklogs"  JAVA\_HOME=/home/lijie/apps/jdk1.7.0\_79  SCALA\_HOME=/home/lijie/apps/scala-2.10.5 |

1. 需要修改hadoop-2.7.1/etc/hadoop/yarn-site.xml，添加

|  |
| --- |
| <property>  <name>yarn.nodemanager.aux-services</name>  <value>spark\_shuffle</value>  </property>  <property>  <name>yarn.nodemanager.aux-services.spark\_shuffle.class</name>  <value>org.apache.spark.network.yarn.YarnShuffleService</value>  </property> |

1. 需要将spark-1.6.2-bin-hadoop2.7.1-dynamic/lib/spark-1.6.2-yarn-shuffle.jar拷贝到所有节点的hadoop-2.7.1/share/hadoop/yarn/lib/
2. 复制HBase/lib下的包到Spark

|  |
| --- |
| 复制HBase目录下的lib文件 到 spark目录/lib/hbase。spark 依赖此lib清单如下：  guava-12.0.1.jar  htrace-core-3.1.0-incubating.jar  protobuf-java-2.5.0.jar  以hbase开头所有jar |

1. 在hdfs上新建一个文件夹为hdfs://master:9000/sparklogs
2. 将 spark-1.6.2-bin-hadoop2.7.1-dynamic/lib/ spark-assembly-1.6.3-SNAPSHOT-hadoop2.7.1.jar存放到hdfs的spark\_lib目录下

|  |
| --- |
| hdfs dfs -mkdir /spark\_lib  hdfs dfs -put spark-assembly-1.6.3-SNAPSHOT-hadoop2.7.1.jar /spark\_lib |

测试

1. 上传测试数据

|  |
| --- |
| lijie@master:~$ hdfs dfs -mkdir -p /user/lijie/test  lijie@master:~$ hdfs dfs -mkdir /user/lijie/test/input  lijie@master:~$ hdfs dfs -put hadoop-2.7.1/etc/hadoop/\*.xml /user/lijie/test/input |

1. 测试MapReduce程序

|  |
| --- |
| hdfs dfs -rm -r /user/lijie/test/output  hadoop jar ~/hadoop-2.7.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-\*.jar \  grep -Dmapreduce.job.queuename=default \  /user/lijie/test/input \  /user/lijie/test/output \  'dfs[a-z.]+' |

1. 测试Spark程序

(1)建立一个表scores，有两个列族grad和courese

*hbase(main):001:0> create 'scores', 'grade', 'course'*

可以使用list命令来查看当前HBase里有哪些表。使用describe命令来查看表结构。（记得所有的表明、列名都需要加上引号）

(2)按设计的表结构插入值：

*put 'scores','Tom','grade:','5'  
put 'scores','Tom','course:math','97'  
put 'scores','Tom','course:art','87'  
put 'scores','Jim','grade','4'  
put 'scores','Jim','course:','89'  
put 'scores','Jim','course:','80'*

Hive on Spark

1. 配置$HIVE\_HOME

2. 生成和修改hive-site.xml

|  |
| --- |
| lijie@master:~/apache-hive-1.2.1-bin/conf$ cp hive-default.xml.template hive-site.xml |

3. 拷贝hive-site.xml到Spark/conf

4. 测试HiveFromSpark