## Hadoop部署经验总结

Hadoop版本：2.7.2

系统：ubuntu16.04

1. 非安全模式的hadoop集群搭建
2. 软件准备

~~在集群所有机器上安装hadoop2.7.2和java，建议每台机器都安装在相同路径，方便后续设置。~~

在集群master（namenode所在的）机器上安装hadoop2.7.2和java。

1. 修改配置

添加环境变量$HADOOP\_PREFIX=你的hadoop文件夹所在路径。

Hadoop的设置都是通过修改各种配置文件，这些文件可以大致分为三类：

* 只读的默认配置：core-default.xml, hdfs-default.xml, yarn-default.xml and mapred-default.xml
* 根据需要修改的配置文件，都在$HADOOP\_PREFIX/etc/hadoop文件夹中：etc/hadoop/core-site.xml，etc/hadoop/hdfs-site.xml,

etc/hadoop/yarn-site.xml，etc/hadoop/mapred-site.xml

* 环境变量配置文件,也在$HADOOP\_PREFIX/etc/hadoop中：hadoop-env.sh， yarn-env.sh

在hadoop-env.sh中修改$JAVA\_HOME为你的jdk所在路径;

* core-site.xml:

<configuration>  
 <property>  
 <name>fs.defaultFS</name>  
 <value>hdfs://singlenode.ustc.edu:9000</value>  
 </property>  
  
 <property>  
 <name>hadoop.tmp.dir</name>  
 <value>/opt/hadoop-2.7.2/tmp</value>  
 </property>

</configuration>

上面是示例，将singlenode.ustc.edu换为master的全域名（通过hostname -f查看，如果想修改可以修改hosts文件），/opt/hadoop-2.7.2/tmp可以自己设置，默认情况下这是集群数据存放的位置。

* hdfs-site.xml：

<configuration>  
 <property>  
 <name>dfs.replication</name>  
 <value>1</value>  
 </property>  
  
 <property>  
 <name>dfs.webhdfs.enabled</name>  
 <value>true</value>  
 </property>  
  
</configuration>

Dfs.replication是文件的副本数量，因为我搭建的是试验性质，只设置了1,即不备份，一般设置为3,重要数据的话可以更高，但是会增加消耗的存储空间和数据传输时间，dfs.webhdfs.enabled设为true，允许使用webhdfs方式操作hdfs。

* yarn-site.xml：

<configuration>  
  
 <!-- Site specific YARN configuration properties -->  
 <property>  
 <name>yarn.nodemanager.aux-services</name>  
 <value>mapreduce\_shuffle</value>  
 </property>  
  
 </configuration>

* mapred-site.xml:

<configuration>  
 <property>  
 <name>mapreduce.framework.name</name>  
 <value>yarn</value>  
 </property>  
</configuration>

1. 配置ssh无密码登录，hosts和slaves文件

/etc/hosts文件中要包含集群所有机器的host和ip的映射；

$HADOOP\_PREFIX/etc/hadoop/slaves文件中要包含datanode的host或者ip，每个datanode一行，示例：

localhost

datanode1

datanode2

datanode3

ssh的无密码登录指令如下：

ssh-keygen -t dsa -P '' -f ~/.ssh/id\_dsa

cat ~/.ssh/id\_dsa.pub >> ~/.ssh/authorized\_keys

chmod 0600 ~/.ssh/authorized\_keys

这三条指令生成了master主机的公钥id\_dsa.pub和认证文件authorized\_keys，

并将公钥内容加入认证文件，使得master可以ssh无密码登录本机，要让master能无密码登录其他机器，需要将公钥文件内容添加到其他机器的认证文件中。

1. 启动集群
2. 格式化hdfs

$HADOOP\_PREFIX/bin/hdfs namenode -format <cluster\_name>

cluster\_name自己设定

1. 开启hdfs  
    $HADOOP\_PREFIX/sbin/start-dfs.sh
2. 开启yarn  
    $HADOOP\_PREFIX/sbin/start-yarn.sh
3. 开启MapReduce Jobhistory服务  
    $HADOOP\_PREFIX/sbin/mr-jobhistory-daemon.sh \

--config $HADOOP\_CONF\_DIR start historyserver

其中$HADOOP\_CONF\_DIR是hadopp配置文件路径，一般是

$HADOOP\_PREFIX/etc/hadoop

2.安全模式的hadoop集群搭建