Big Data Cluster

JieLiu (Deployment plan)

1 集群规划

1.1 总体方案

整个计算集群都是基于分布式文件系统HDFS之上，YARN用来实现集群资源的管理与调度，MapReduce用于并行计算。Hive提供了SQL标准的数据存储，管理服务，Pig是一个分析大型数据集的平台。Spark是一个内存分布式计算框架，HBase 是分布式数据库，用来管理和存储大规模的数据。

1.2 主机规划

Table 1: 主机规划

主机名	IP	用户	角色	
cu01	192.168.1.21	hadoop	master	
cu02	192.168.1.22	hadoop	slave01	
cu03	192.168.1.28	hadoop	slave02	
cu04	192.168.1.29	hadoop	slave03	
cu05	192.168.1.30	hadoop	slave04	

2 Hadoop安装配置

2.1 主机名配置

修改集群中每个主机的hosts文件

```
#master
192.168.1.21 cu01
#slaves
192.168.1.22 cu02
192.168.1.28 cu03
192.168.1.29 cu04
192.168.1.30 cu05
```

2.2 创建hadoop用户

```
root$useradd hadoop
root$passwd hadoop
```

2.3 master无密码登录slaves主机配置

生成ssh-key密钥在控制主机上

```
ssh-keygen -b 4096
```

拷贝密钥到其他节点

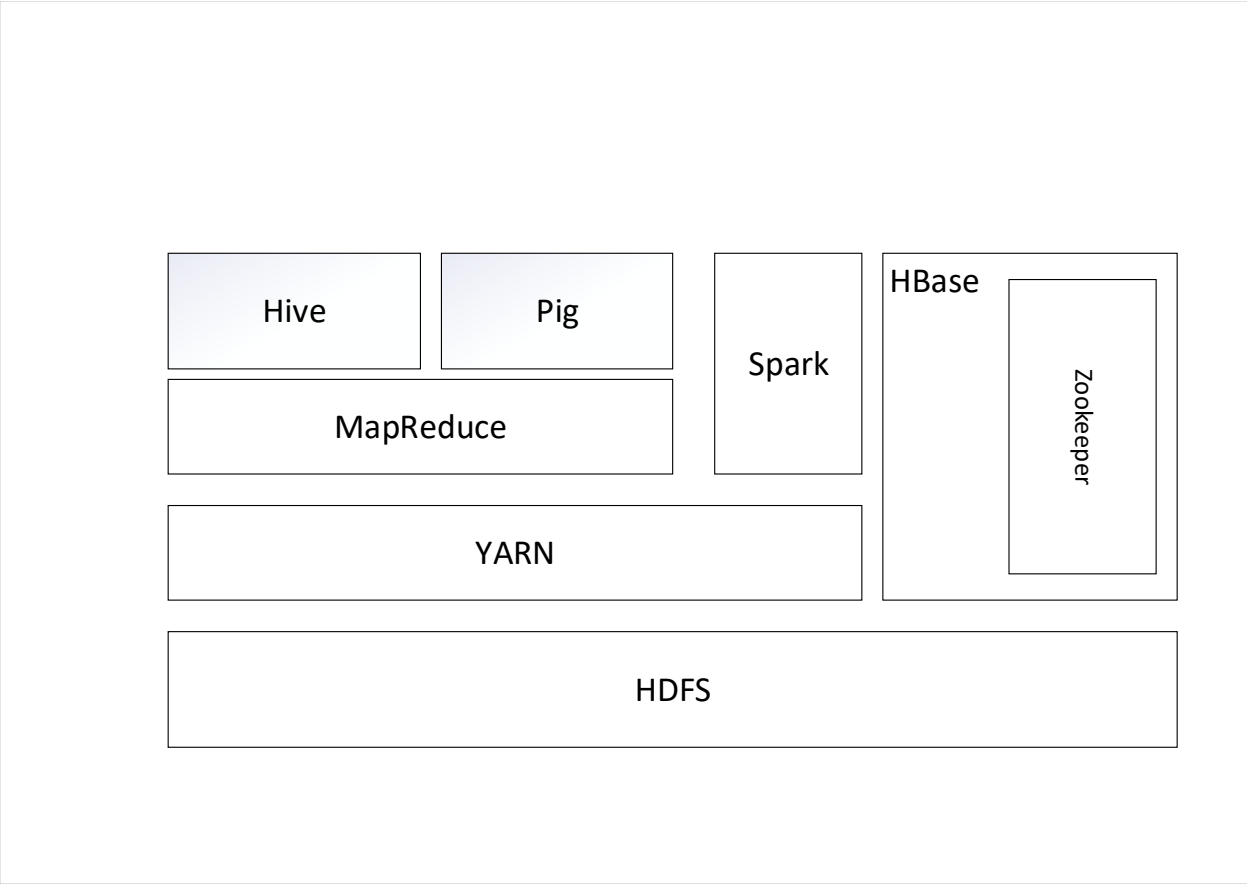

Figure 1: 软件架构

```
ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@cu01
ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@cu02
ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@cu03
ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@cu04
ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@cu05
```

2.4 下载解压安装包，设置环境变量

```
cd ~
wget http://good.ncu.edu.cn/mirrors/hadoop-2.8.1.tar.gz
tar -xzf hadoop-2.8.1.tar.gz
mv hadoop-2.8.1 hadoop
```

配置环境变量(/etc/profile文件)

```
export HADOOP_CONF_DIR=/home/hadoop/hadoop/etc/hadoop
export LD_LIBRARY_PATH=/home/hadoop/hadoop/lib/native:$LD_LIBRARY_PATH
export PATH=/home/hadoop/hadoop/bin:/home/hadoop/hadoop/sbin:$PATH
```

2.5 安装配置java环境

```
cd ~
wget http://good.ncu.edu.cn/mirrors/jdk-8u161-linux-x64.tar.gz
tar -zxvf jdk-8u161-linux-x64.tar.gz
```

配置环境变量

```
export JAVA_HOME=/home/hadoop/jdk1.8.0_161
export PATH=$PATH:$JAVA_HOME/bin
export CLASSPATH=$JAVA_HOME/jre/lib/ext:$JAVA_HOME/lib/tools.jar
```

2.6 master节点配置

修改 /hadoop/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/home/hadoop/jdk1.8.0_161
```

2.7 配置NameNode位置

修改 /hadoop/etc/hadoop/core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
            <name>fs.default.name</name>
            <value>hdfs://192.168.1.21:9000</value>
        </property>
</configuration>
```

2.8 配置HDFS路径

修改 /hadoop/etc/hadoop/hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

    <property>
            <name>dfs.namenode.name.dir</name>
            <value>/home/hadoop/data/nameNode</value>
    </property>

    <property>
            <name>dfs.datanode.data.dir</name>
            <value>/home/hadoop/data/dataNode</value>
    </property>

    <property>
            <name>dfs.replication</name>
            <value>3</value>
    </property>

</configuration>
```

2.9 配置YARN为作业调度器

```
cd ~/hadoop/etc/hadoop
mv mapred-site.xml.template mapred-site.xml
```

修改文件内容

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

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

    <property>
        <name>yarn.app.mapreduce.am.resource.mb</name>
        <value>512</value>
    </property>

    <property>
        <name>mapreduce.map.memory.mb</name>
        <value>256</value>
    </property>

    <property>
        <name>mapreduce.reduce.memory.mb</name>
```

```
        <value>256</value>
    </property>

</configuration>
```

2.10 配置YARN

修改 /hadoop/etc/hadoop/yarn-site.xml

```
<?xml version="1.0"?>

<configuration>

<!-- Site specific YARN configuration properties -->
    <property>
            <name>yarn.acl.enable</name>
            <value>0</value>
    </property>

    <property>
            <name>yarn.resourcemanager.hostname</name>
            <value>cu01</value>
    </property>

    <property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
    </property>

    <property>
        <name>yarn.nodemanager.resource.memory-mb</name>
        <value>1536</value>
    </property>

    <property>
        <name>yarn.scheduler.maximum-allocation-mb</name>
        <value>1536</value>
    </property>

    <property>
        <name>yarn.scheduler.minimum-allocation-mb</name>
        <value>128</value>
    </property>

    <property>
        <name>yarn.nodemanager.vmem-check-enabled</name>
        <value>false</value>
    </property>

</configuration>
```

2.11 配置slaves

修改 /hadoop/etc/hadoop/slaves

```
cu02
cu03
```

```
cu04
cu05
```

2.12 拷贝文件到每一个slave节点

```
cd /home/hadoop/
scp -r hadoop cu02:/home/hadoop
scp -r hadoop cu03:/home/hadoop
scp -r hadoop cu04:/home/hadoop
scp -r hadoop cu05:/home/hadoop
```

2.13 配置slave节点的环境变量

同master环境变量配置

2.14 运行测试

```
hdfs namenode -format
hadoop/sbin/start-all.sh
```

jps查看状态

3 HBase安装配置

3.1 下载解压安装包，设置环境变量

```
cd ~
wget http://good.ncu.edu.cn/mirrors/hbase-1.2.6-bin.tar.gz
tar -xzf hbase-1.2.6-bin.tar.gz
mv hbase-1.2.6 hbase
```

配置环境变量(/etc/profile文件)

```
export HBASE_HOME=/home/hadoop/hbase
export PATH=$PATH:$HBASE_HOME/bin
export CLASSPATH=$CLASSPATH:/home/hadoop/hbase/lib/*:.
```

3.2 配置hbase环境

修改hbase/conf下hbase-env.sh

```
export JAVA_HOME=/home/hadoop/jdk1.8.0_161
export HBASE_MANAGES_ZK=true
```

3.3 配置hbase-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>

<property>
```

```
<name>hbase.rootdir</name>
<value>hdfs://192.168.1.21:9000/hbase</value>
</property>
<property>
<name>hbase.cluster.distributed</name>
<value>true</value>
</property>
<property>
<name>hbase.zookeeper.property.dataDir</name>
<value>/home/hadoop/zookeeper</value>
</property>
<property>
<name>hbase.zookeeper.quorum</name>
<value>cu01,cu02,cu03,cu04,cu05</value>
</property>
<property>
<name>hbase.zookeeper.property.clientPort</name>
<value>2181</value>
</property>

</configuration>
```

3.4 配置从服务器

修改文件conf/regionservers

```
cu02
cu03
cu04
cu05
```

3.5 创建zookeeper目录

在每个节点与hbase同级目录创建zookeeper文件夹

3.6 region节点配置

拷贝hbase文件到region节点

```
scp -r hbase cu02:/home/hadoop
scp -r hbase cu03:/home/hadoop
scp -r hbase cu04:/home/hadoop
scp -r hbase cu05:/home/hadoop
```

3.7 region 节点环境变量配置

同master节点配置

3.8 启动测试hbase

```
$HBASE_HOME/bin/start-hbase.sh
hbase shell
```

4 Hive安装配置

4.1 下载解压安装包，设置环境变量

```
cd ~
wget http://good.ncu.edu.cn/mirrors/apache-hive-1.2.2-bin.tar.gz
tar -xzf apache-hive-1.2.2-bin.tar.gz
mv apache-hive-1.2.2-bin.tar.gz hive
```

配置环境变量(/etc/profile文件)

```
export HIVE_HOME=/home/hadoop/hive
export PATH=$HIVE_HOME/bin:$PATH
```

4.2 安装mysql用于存储元数据

```

wget http://repo.mysql.com/mysql-community-release-el7-5.noarch.rpm
sudo rpm -ivh mysql-community-release-el7-5.noarch.rpm
yum update

sudo yum install mysql-server
sudo systemctl start mysqld

sudo mysql_secure_installation
```

登录mysql创建hive用户,并创建hive数据库

```
mysql>CREATE USER 'hive' IDENTIFIED BY 'hive';
mysql>GRANT ALL PRIVILEGES ON *.* TO 'hive'@'cu01' WITH GRANT OPTION;
mysql>flush privileges;

mysql>create database hive;
```

4.3 配置hive-site.xml

找到hive-default.xml.template,cp一份为hive-default.xml

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
        <property>
        <name>javax.jdo.option.ConnectionURL</name>
        <value>jdbc:mysql://192.168.1.21:3306/hive?createDatabaseIfNotExist=true</value>
        <description>JDBC connect string for a JDBC metastore</description>
        </property>
        <property>
        <name>javax.jdo.option.ConnectionDriverName</name>
        <value>com.mysql.jdbc.Driver</value>
        <description>Driver class name for a JDBC metastore</description>
        </property>

        <property>
        <name>javax.jdo.option.ConnectionUserName</name>
        <value>root</value>
```

```
        <description>username to use against metastore database</description>
        </property>

        <property>
        <name>javax.jdo.option.ConnectionPassword</name>
        <value>work@good308</value>
        <description>password to use against metastore database</description>
        </property>

</configuration>
```

4.4 JDBC依赖配置

```
wget http://cdn.mysql.com/Downloads/Connector-J/mysql-connector-java-5.1.36.tar.gz
#
tar -zxvf mysql-connector-java-5.1.36.tar.gz
cp mysql-connector-java-5.1.33-bin.jar apache-hive-1.2.1-bin/lib/
```

4.5 hive客户端配置

```
<configuration>
    <property>
        <name>hive.metastore.uris</name>
    <value>thrift://192.168.1.21:9083</value>
    </property>
</configuration>
```

4.6 启动并测试hive

```
#start metastore service
hive --service metastore &
#start test hive
hive

hive> show databases;
OK
default
src
Time taken: 1.332 seconds, Fetched: 2 row(s)
hive> use src;
OK
Time taken: 0.037 seconds
hive> create table test1(id int);
OK
Time taken: 0.572 seconds
hive> show tables;
OK
abc
test
test1
Time taken: 0.057 seconds, Fetched: 3 row(s)
hive>
```

5 安装配置Spark

5.1 下载解压安装包，设置环境变量

```
cd ~
wget https://good.ncu.edu.cn/mirrors/spark-2.2.0-bin-hadoop2.7.tgz
tar -xvf spark-2.2.0-bin-hadoop2.7.tgz
mv spark-2.2.0-bin-hadoop2.7 spark
```

```
pathmunge /home/hadoop/spark/bin
```

配置环境变量(/etc/profile文件)

```
export HADOOP_CONF_DIR=/home/hadoop/hadoop/etc/hadoop
export SPARK_HOME=/home/hadoop/spark
export LD_LIBRARY_PATH=/home/hadoop/hadoop/lib/native:$LD_LIBRARY_PATH
```

5.2 配置YARN管理spark

```
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements.  See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

# Default system properties included when running spark-submit.
# This is useful for setting default environmental settings.

# Example:
spark.master                     yarn
spark.yarn.am.memory     512m
spark.executor.memory           512m

spark.eventLog.enabled           true
spark.eventLog.dir               hdfs://cu01:9000/spark-logs
# spark.serializer                 org.apache.spark.serializer.KryoSerializer
spark.driver.memory              512m
spark.history.provider            org.apache.spark.deploy.history.FsHistoryProvider
spark.history.fs.logDirectory     hdfs://cu01:9000/spark-logs
spark.history.fs.update.interval  10s
spark.history.ui.port             18080

# spark.executor.extraJavaOptions  -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"
```

6 安装配置pig

6.1 下载解压安装包，设置环境变量

```
cd ~
wget https://good.ncu.edu.cn/mirrors/pig-0.16.0.tar.gz
tar -xvf pig-0.16.0.tar.gz
mv pig-0.16.0.tar.gz pig
```

配置环境变量(/etc/profile文件)

```
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:/home/hadoop/pig/bin
export PIG_CLASSPATH=$HADOOP_HOME/conf
```