Hadoop 2.7 Spark 2.0伪分布式安装配置实录

Yao Ran

Nov-2016

目录

[1. 软件包列表 3](#_Toc468180974)

[2. 操作系统安装配置 3](#_Toc468180975)

[2.1. 安装操作系统 3](#_Toc468180976)

[2.2. 设置网络 3](#_Toc468180977)

[2.3. 安装&定制软件包 4](#_Toc468180978)

[2.4. 配置Java环境 4](#_Toc468180979)

[2.5. 准备Hadoop用户 4](#_Toc468180980)

[3. Hadop 2.7.3安装配置 5](#_Toc468180981)

[3.1. 解压安装包 5](#_Toc468180982)

[3.2. 配置系统环境变量 5](#_Toc468180983)

[3.3. 修改Hadoop配置文件 5](#_Toc468180984)

[3.4. 设置Hadoop用户所有权 5](#_Toc468180985)

[3.5. 生成ssh授信密钥 6](#_Toc468180986)

[3.6. HDFS操作 6](#_Toc468180987)

[3.7. 配置YARN 6](#_Toc468180988)

[3.8. 启动&测试服务 7](#_Toc468180989)

[4. Spark安装配置 7](#_Toc468180990)

[4.1. 解压软件包 7](#_Toc468180991)

[4.2. 配置Spark 7](#_Toc468180992)

[4.3. 运行Spark示例 8](#_Toc468180993)

[4.4. 启动Spark-Shell 8](#_Toc468180994)

[4.5. 测试Spark RDD 8](#_Toc468180995)

[5. Hbase安装配置 8](#_Toc468180996)

[5.1. 解压软件包 8](#_Toc468180997)

[5.2. 配置Hbase 8](#_Toc468180998)

[5.3. 启动停止Hbase 9](#_Toc468180999)

[5.4. Hbase Shell操作 10](#_Toc468181000)

[6. PostgreSQL安装配置 10](#_Toc468181001)

[6.1. 解压软件包 10](#_Toc468181002)

[6.2. 创建用户 11](#_Toc468181003)

[6.3. 编译安装 11](#_Toc468181004)

[6.4. 配置环境与数据库 11](#_Toc468181005)

[6.5. 配置系统服务 11](#_Toc468181006)

[6.6. 测试数据库使用 11](#_Toc468181007)

[6.7. 配置远程访问 12](#_Toc468181008)

[7. Hive安装配置 12](#_Toc468181009)

[7.1. 解压软件包 12](#_Toc468181010)

[7.2. 配置环境参数 12](#_Toc468181011)

[7.3. 准备Meta Store 13](#_Toc468181012)

[7.4. 配置Hive 13](#_Toc468181013)

[7.5. 配置HiveServer2 14](#_Toc468181014)

[7.6. 启动停止Hive服务 15](#_Toc468181015)

[8. Hive使用及进阶 16](#_Toc468181016)

[8.1. Hive使用 16](#_Toc468181017)

[8.2. Hive与Hbase集成 18](#_Toc468181018)

[8.3. Python 3访问Hive 18](#_Toc468181019)

[9. Sqoop安装配置 18](#_Toc468181020)

[10. Pig安装配置 18](#_Toc468181021)

[11. Presto安装配置 19](#_Toc468181022)

[12. Tableau使用Spark SQL案例 19](#_Toc468181023)

## 软件包列表

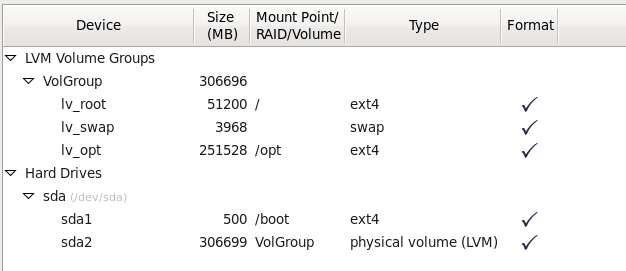
|  |  |  |
| --- | --- | --- |
| **软件包名** | **下载地址** | **说明** |
| jdk-8u65-linux-x64.tar.gz | Oracle OTN | JDK |
| hadoop-2.7.3.tar.gz | http://ftp.wayne.edu/apache/hadoop/common/hadoop-2.7.3/hadoop-2.7.3.tar.gz | Apache Hadoop |
| hbase-1.2.4-bin.tar.gz | http://www.apache.org/dyn/closer.cgi/hbase/ | Hbase |
| postgresql-9.6.1.tar.gz | https://www.postgresql.org/download/ | PostgreSQL(Hive Metastore) |
| postgresql-9.4.1212.jar | https://jdbc.postgresql.org/download.html#current | PostgreSQL JDBC |
| apache-hive-1.2.1-bin.tar.gz | http://hive.apache.org/ | Hive |
| apache-hive-1.2.1-src.tar.gz | http://hive.apache.org/ | 编译Hive-Hbase集成使用 |
| scala-2.11.8.tgz |  | Scala |
| spark-2.0.2-bin-hadoop2.7.tgz |  | Spark |
|  |  |  |
|  |  |  |

## 操作系统安装配置

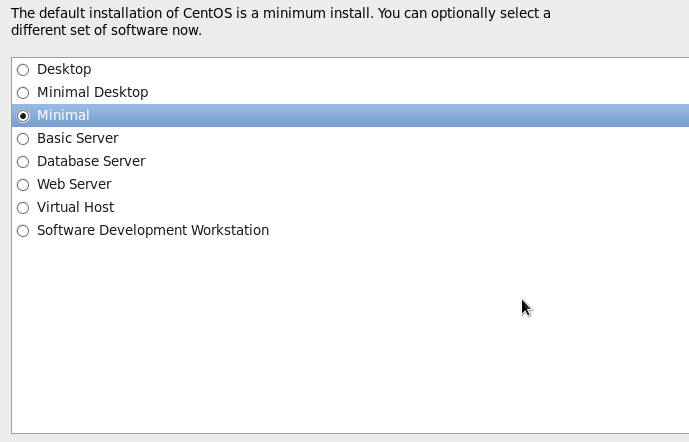
### 安装操作系统

* **安装CentOS 6.7**

将最大的空间划分到/opt卷下



最小化安装CentOS 6.7



### 设置网络

* **关闭防火墙服务**

|  |
| --- |
| [root@que01 ~]# chkconfig iptables off  [root@que01 ~]# chkconfig ip6tables off |

* **关闭SELINUX**

|  |
| --- |
| [root@que01 ~]# cd /etc/selinux/  [root@que01 selinux]# vi config  # This file controls the state of SELinux on the system.  # SELINUX= can take one of these three values:  # enforcing - SELinux security policy is enforced.  # permissive - SELinux prints warnings instead of enforcing.  # disabled - No SELinux policy is loaded.  SELINUX=disabled  # SELINUXTYPE= can take one of these two values:  # targeted - Targeted processes are protected,  # mls - Multi Level Security protection.  SELINUXTYPE=targeted |

### 安装&定制软件包

使用SSH远程工具登录刚安装好的系统

* **编辑设置本地YUM源**

|  |
| --- |
| mkdir -p /mnt/cdrom  mount -t iso9660 -o loop /dev/sr0 /mnt/cdrom  mkdir -p /etc/yum.repos.d/bak  cd /etc/yum.repos.d/  mv \*.repo bak  cd /etc/yum.repos.d  vi local.repo  [base]  name=server  baseurl=file:///mnt/cdrom  enable=1  gpgcheck=0 |

* **安装软件包**

|  |
| --- |
| [root@hdp27 ~]# yum -y install openssh-clients  [root@hdp27 ~]# yum -y install rsync  [root@hdp27 ~]# yum -y install gcc\*  [root@hdp27 ~]# yum -y install readline-devel |

### 配置Java环境

|  |
| --- |
| [root@hdp27 install]# tar -xvzf jdk-8u65-linux-x64.tar.gz  [root@hdp27 install]# mv jdk1.8.0\_65 ..  [root@hdp27 jdk1.8.0\_65]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$PATH  export JAVA\_HOME CLASSPATH PATH |

### 准备Hadoop用户

|  |
| --- |
| [root@hdp27 ~]# useradd hadoop  [root@hdp27 ~]# passwd hadoop  Changing password for user hadoop.  New password:  BAD PASSWORD: it is based on a dictionary word  Retype new password:  passwd: all authentication tokens updated successfully. |

## Hadop 2.7.3安装配置

### 解压安装包

|  |
| --- |
| [root@hdp27 install]# tar -xvzf hadoop-2.7.3.tar.gz  [root@hdp27 install]# mv hadoop-2.7.3 ..  [root@hdp27 install]# cd /opt  [root@hdp27 opt]# ln -s hadoop-2.7.3 hadoop |

### 配置系统环境变量

|  |
| --- |
| [root@hdp27 ~]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  HADOOP\_HOME=/opt/hadoop-2.7.3  HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$PATH  export JAVA\_HOME CLASSPATH PATH HADOOP\_HOME HADOOP\_COMMON\_LIB\_NATIVE\_DIR HADOOP\_OPTS |

修改完成后，建议重启系统，使环境变量生效

### 修改Hadoop配置文件

|  |
| --- |
| [root@hdp27 hadoop]# cd /opt/hadoop  [root@hdp27 hadoop]# vi etc/hadoop/core-site.xml  <configuration>  <property>  <name>fs.defaultFS</name>  <value>hdfs://hdp27:9000</value>  </property>  </configuration>  [root@hdp27 hadoop]# vi etc/hadoop/hdfs-site.xml  <configuration>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  </configuration>  修改Hadoop-env.sh，否则会报JAVA\_HOME找不到错误  [hadoop@hdp27 hadoop]# vi etc/hadoop/hadoop-env.sh  #export JAVA\_HOME=${JAVA\_HOME} 注释这行  export JAVA\_HOME=/opt/jdk1.8.0\_65 |

### 设置Hadoop用户所有权

|  |
| --- |
| [root@hdp27 ~]# cd /opt  [root@hdp27 opt]# chown -R hadoop:hadoop hadoop-2.7.3    [root@hdp27 opt]# chown -R hadoop:hadoop jdk1.8.0\_65 |

### 生成ssh授信密钥

|  |
| --- |
| [root@hdp27 opt]# su – hadoop  [hadoop@hdp27 ~]$ ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa  [hadoop@hdp27 ~]$ cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys  [hadoop@hdp27 ~]$ chmod 0600 ~/.ssh/authorized\_keys |

### HDFS操作

|  |
| --- |
| 格式化Namenode  [hadoop@hdp27 ~]$ hdfs namenode -format  启动HDFS  [hadoop@hdp27 ~]$ start-dfs.sh  [hadoop@hdp27 ~]$ hdfs dfs -ls /  [hadoop@hdp27 ~]$ hdfs dfs -mkdir /user  [hadoop@hdp27 ~]$ hdfs dfs -mkdir /user/hadoop  [hadoop@hdp27 ~]$ hdfs dfs -ls /  Found 1 items  drwxr-xr-x - hadoop supergroup 0 2016-11-22 14:08 /user  停止HDFS服务  [hadoop@hdp27 ~]$ stop-dfs.sh |

### 配置YARN

|  |
| --- |
| [hadoop@hdp27 ~]$ cd /opt/hadoop-2.7.3/etc/hadoop  [hadoop@hdp27 hadoop]$ cp mapred-site.xml.template mapred-site.xml  [hadoop@hdp27 hadoop]$ vi mapred-site.xml  <configuration>  <property>  <name>mapreduce.framework.name</name>  <value>yarn</value>  </property>  </configuration>  [hadoop@hdp27 hadoop]$ vi yarn-site.xml  <configuration>  <!-- Site specific YARN configuration properties -->  <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property>  </configuration> |

### 启动&测试服务

|  |
| --- |
| [hadoop@hdp27 hadoop]$ start-dfs.sh  [hadoop@hdp27 hadoop]$ start-yarn.sh  浏览器访问  HDFS服务  <http://10.0.0.27:50070>  YARN服务  <http://10.0.0.27:8088> |

## Spark安装配置

### 解压软件包

|  |
| --- |
| [hadoop@hdp27 ~]#cd /opt/install  [root@hdp27 install]# tar -xvzf spark-2.0.2-bin-hadoop2.7.tgz  [root@hdp27 install]# mv spark-2.0.2-bin-hadoop2.7 ..  [root@hdp27 install]# cd ..  [root@hdp27 opt]# chown -R hadoop:hadoop spark-2.0.2-bin-hadoop2.7/ |

### 配置Spark

|  |
| --- |
| [root@hdp27 opt]# su - hadoop  [hadoop@hdp27 opt]$ cd spark-2.0.2-bin-hadoop2.7/  [hadoop@hdp27 spark-2.0.2-bin-hadoop2.7]$ cd conf  [hadoop@hdp27 conf]$ cp spark-env.sh.template spark-env.sh  [hadoop@hdp27 conf]$ cp spark-defaults.conf.template spark-defaults.conf  [hadoop@hdp27 conf]$ vi spark-env.sh  export HADOOP\_HOME=/opt/hadoop-2.7.3  export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop  export SPARK\_DIST\_CLASSPATH=$(hadoop classpath)  切换回root帐户  [root@hdp27 opt]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  HADOOP\_HOME=/opt/hadoop-2.7.3  HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  SPARK\_HOME=/opt/spark-2.0.2-bin-hadoop2.7  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$SPARK\_HOME/bin: $PATH  export JAVA\_HOME CLASSPATH PATH HADOOP\_HOME HADOOP\_COMMON\_LIB\_NATIVE\_DIR HADOOP\_OPTS SPARK\_HOME |

### 运行Spark示例

|  |
| --- |
| 返回解压目录  [hadoop@hdp27 spark-2.0.2-bin-hadoop2.7]$ ./bin/run-example SparkPi 10 |

### 启动Spark-Shell

|  |
| --- |
| [hadoop@hdp27 ~]$ spark-shell  浏览器可以访问：  <http://10.0.0.27:4040/> |

### 测试Spark RDD

|  |
| --- |
| HDFS上传一个文件  [hadoop@hdp27 jdk1.8.0\_65]$ pwd  /opt/jdk1.8.0\_65  [hadoop@hdp27 jdk1.8.0\_65]$ hdfs dfs -ls /user/hadoop  [hadoop@hdp27 jdk1.8.0\_65]$ hdfs dfs -put COPYRIGHT /user/hadoop  [hadoop@hdp27 jdk1.8.0\_65]$ hdfs dfs -ls /user/hadoop    运行Spark-Shell  [hadoop@hdp27 ~]$ spark-shell  scala> val textFile = sc.textFile("/user/hadoop/COPYRIGHT")  scala> textFile.count() |

## Hbase安装配置

### 解压软件包

|  |
| --- |
| [root@hdp27 install]# tar -xvzf hbase-1.2.4-bin.tar.gz  [root@hdp27 install]# mv hbase-1.2.4 ..  [root@hdp27 install]# cd ..  [root@hdp27 opt]# chown -R hadoop:hadoop hbase-1.2.4  [root@hdp27 ~]# su - hadoop |

### 配置Hbase

Root帐户操作

|  |
| --- |
| [root@hdp27 hbase-1.2.4]# pwd  /opt/hbase-1.2.4  [root@hdp27 hbase-1.2.4]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  HADOOP\_HOME=/opt/hadoop-2.7.3  HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  SPARK\_HOME=/opt/spark-2.0.2-bin-hadoop2.7  HBASE\_HOME=/opt/hbase-1.2.4  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$SPARK\_HOME/bin:$HBASE\_HOME/bin:$PATH  export JAVA\_HOME CLASSPATH PATH HADOOP\_HOME HADOOP\_COMMON\_LIB\_NATIVE\_DIR HADOOP\_OPTS SPARK\_HOME HBASE\_HOME |

Hadoop帐户操作

|  |
| --- |
| [hadoop@hdp27 ~]$ source /etc/profile  [hadoop@hdp27 ~]$ hdfs dfs -mkdir /hbase  [hadoop@hdp27 ~]$ cd $HBASE\_HOME  [hadoop@hdp27 hbase-1.2.4]$ vi conf/hbase-site.xml  <configuration>  <property>  <name>hbase.cluster.distributed</name>  <value>true</value>  </property>  <property>  <name>hbase.rootdir</name>  <value>hdfs://hdp27:9000/hbase</value>  </property>  <property>  <name>hbase.master</name>  <value>hdp27:60000</value>  </property>  <property>  <name>hbase.master.info.port</name>  <value>60010</value>  </property>  <property>  <name>hbase.master.maxlockskew</name>  <value>180000</value>  </property>  <property>  <name>hbase.zookeeper.quorum</name>  <value>hdp27</value>  </property>  <property>  <name>hbase.zookeeper.property.clientPort</name>  <value>2181</value>  </property>  <property>  <name>hbase.zookeeper.property.dataDir</name>  <value>/home/hadoop/tmp/zookeeper</value>  </property>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  </configuration>  [hadoop@hdp27 hbase-1.2.4]$ vi conf/hbase-env.sh  export JAVA\_HOME=/opt/jdk1.8.0\_65  export HBASE\_MANAGES\_ZK=true  [hadoop@hdp27 hbase-1.2.4]$ vi conf/regionservers  hdp27 |

### 启动停止Hbase

|  |
| --- |
| [hadoop@hdp27 hbase-1.2.4]$ start-hbase.sh  [hadoop@hdp27 hbase-1.2.4]$ stop-hbase.sh  查看HBASE状态  <http://10.0.0.27:60010> |

### Hbase Shell操作

|  |
| --- |
| [hadoop@hdp27 ~]$ hbase shell  hbase(main):001:0> create 'test','cf'  0 row(s) in 1.7800 seconds  => Hbase::Table - test  hbase(main):002:0> list 'test'  TABLE  test  1 row(s) in 0.0320 seconds  => ["test"]  hbase(main):003:0> put 'test', 'row1', 'cf:a', 'value1'  0 row(s) in 0.2160 seconds  hbase(main):004:0> put 'test', 'row2', 'cf:b', 'value2'  0 row(s) in 0.0860 seconds  hbase(main):005:0> put 'test', 'row3', 'cf:c', 'value3'  0 row(s) in 0.0290 seconds  hbase(main):006:0>  hbase(main):007:0\* scan 'test'  ROW COLUMN+CELL  row1 column=cf:a, timestamp=1479884649433, value=value1  row2 column=cf:b, timestamp=1479884672804, value=value2  row3 column=cf:c, timestamp=1479884713581, value=value3  3 row(s) in 0.0810 seconds  hbase(main):008:0> get 'test', 'row1'  COLUMN CELL  cf:a timestamp=1479884649433, value=value1  1 row(s) in 0.0300 seconds  hbase(main):009:0> disable 'test'  0 row(s) in 2.3450 seconds  hbase(main):010:0> enable 'test'  0 row(s) in 1.3250 seconds  hbase(main):011:0> disable 'test'  0 row(s) in 2.2710 seconds  hbase(main):012:0> drop 'test'  0 row(s) in 1.3210 seconds  hbase(main):013:0> exit |

## PostgreSQL安装配置

PostgreSQL是安装Hadoop Hive等服务的必须项，为Hive等提供元数据管理服务 。

### 解压软件包

|  |
| --- |
| [root@hdp27 install]# tar -xvzf postgresql-9.6.1.tar.gz |

### 创建用户

|  |
| --- |
| [root@hdp27 install]# adduser postgres  [root@hdp27 install]# passwd postgres  Changing password for user postgres.  New password:  BAD PASSWORD: it is based on a dictionary word  Retype new password:  passwd: all authentication tokens updated successfully. |

### 编译安装

|  |
| --- |
| [root@hdp27 install]# mkdir -p /opt/postgres/pgsql  [root@hdp27 install]# cd postgresql-9.6.1  [root@hdp27 postgresql-9.6.1]# ./configure --prefix=/opt/postgres/pgsql  [root@hdp27 postgresql-9.6.1]# gmake  [root@hdp27 postgresql-9.6.1]# gmake install |

### 配置环境与数据库

|  |
| --- |
| **配置环境变量**  [root@hdp27 ~]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  HADOOP\_HOME=/opt/hadoop-2.7.3  HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  SPARK\_HOME=/opt/spark-2.0.2-bin-hadoop2.7  HBASE\_HOME=/opt/hbase-1.2.4  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$SPARK\_HOME/bin:$HBASE\_HOME/bin:/opt/postgres/pgsql/bin:$PATH  export JAVA\_HOME CLASSPATH PATH HADOOP\_HOME HADOOP\_COMMON\_LIB\_NATIVE\_DIR HADOOP\_OPTS SPARK\_HOME HBASE\_HOME  **初始化数据库**  [root@hdp27 ~]# mkdir /opt/postgres/pgsql/data  [root@hdp27 ~]# chown postgres:postgres /opt/postgres/pgsql/data  [root@hdp27 ~]# su - postgres  [postgres@hdp27 ~]$ initdb -D /opt/postgres/pgsql/data |

### 配置系统服务

|  |
| --- |
| [root@hdp27 ~]# cd /opt/install/postgresql-9.6.1  [root@hdp27 postgresql-9.6.1]# cp contrib/start-scripts/linux /etc/init.d/postgresql  [root@hdp27 postgresql-9.6.1]# vi /etc/init.d/postgresql  prefix=/opt/postgres/pgsql  PGDATA="/opt/postgres/pgsql/data"  [root@hdp27 postgresql-9.6.1]# chmod +x /etc/init.d/postgresql  [root@hdp27 postgresql-9.6.1]# chkconfig --add postgresql  [root@hdp27 postgresql-9.6.1]# chkconfig --levels 345 postgresql on  [root@hdp27 postgresql-9.6.1]# service postgresql start  Starting PostgreSQL: ok  [root@hdp27 postgresql-9.6.1]# mkdir -p /home/postgres/pgsql  [root@hdp27 postgresql-9.6.1]# touch /home/postgres/pgsql/.pgsql\_history  [root@hdp27 postgresql-9.6.1]# chown postgres:postgres /home/postgres/pgsql/.pgsql\_history |

### 测试数据库使用

|  |
| --- |
| [root@hdp27 postgresql-9.6.1]# su - postgres  [postgres@hdp27 ~]$ createdb test  [postgres@hdp27 ~]$ psql test |

### 配置远程访问

|  |
| --- |
| [postgres@hdp27 ~]$ cd /opt/postgres/pgsql/data  [postgres@hdp27 data]$ vi postgresql.conf  listen\_addresses = '\*'  [postgres@hdp27 data]$ vi pg\_hba.conf  host all all 0.0.0.0/0 trust    重启数据库服务  [root@hdp27 ~]# service postgresql restart |

## Hive安装配置

### 解压软件包

|  |
| --- |
| [root@hdp27 install]# tar -xvzf apache-hive-1.2.1-bin.tar.gz  [root@hdp27 install]# mv apache-hive-1.2.1-bin ..  [root@hdp27 install]# cd ..  [root@hdp27 opt]# chown -R hadoop:hadoop apache-hive-1.2.1-bin |

### 配置环境参数

|  |
| --- |
| [root@hdp27 ~]# vi /etc/profile  JAVA\_HOME=/opt/jdk1.8.0\_65  HADOOP\_HOME=/opt/hadoop-2.7.3  HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native  HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  SPARK\_HOME=/opt/spark-2.0.2-bin-hadoop2.7  HBASE\_HOME=/opt/hbase-1.2.4  HIVE\_HOME=/opt/apache-hive-1.2.1-bin  CLASSPATH=$CLASSPATH:.:$JAVA\_HOME/lib/tools.jar  PATH=$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$SPARK\_HOME/bin:$HBASE\_HOME/bin:/opt/postgres/pgsql/bin:$HIVE\_HOME/bin: $PATH  export JAVA\_HOME CLASSPATH PATH HADOOP\_HOME HADOOP\_COMMON\_LIB\_NATIVE\_DIR HADOOP\_OPTS SPARK\_HOME HBASE\_HOME HIVE\_HOME |

### 准备Meta Store

**创建用户和元数据库**

|  |
| --- |
| [root@hdp27 ~]# su – postgres  [postgres@hdp27 ~]$ cd /opt/apache-hive-1.2.1-bin/scripts/metastore/upgrade/postgres  [postgres@hdp27 postgres]$ psql  psql (9.6.1)  Type "help" for help.  postgres=# create user hiveuser with password 'admin1234';  CREATE ROLE  postgres=# create database metastore;  CREATE DATABASE  postgres=# \c metastore  You are now connected to database "metastore" as user "postgres".  metastore=# \i hive-schema-1.2.0.postgres.sql |

**赋权给hiveuser**

|  |
| --- |
| metastore=# \c metastore  You are now connected to database "metastore" as user "postgres".  metastore=# \pset tuples\_only on  Tuples only is on.  metastore=# \o /tmp/grant-privs  metastore=# select 'GRANT SELECT,INSERT,UPDATE,DELETE ON "' || schemaname || '"."' || tablename ||'" TO hiveuser ;'  metastore-# FROM pg\_tables  metastore-# where tableowner = CURRENT\_USER and schemaname = 'public';  metastore=# \o  metastore=# \pset tuples\_only off  Tuples only is off.  metastore=# \i /tmp/grant-privs |

**测试hiveuser用户**

|  |
| --- |
| [postgres@hdp27 ~]$ psql -h hdp27 -U hiveuser -d metastore  psql (9.6.1)  Type "help" for help.  metastore=> |

### 配置Hive

**复制数据库驱动**

|  |
| --- |
| [root@hdp27 ~]# su - hadoop  [hadoop@hdp27 ~]$ cp /opt/install/postgresql-9.4.1212.jar /opt/apache-hive-1.2.1-bin/lib/postgresql-jdbc.jar |

**准备HDFS目录**

|  |
| --- |
| [hadoop@hdp27 ~]$ hdfs dfs -mkdir /tmp  mkdir: `/tmp': File exists  [hadoop@hdp27 ~]$ hdfs dfs -mkdir -p /user/hive/warehouse  [hadoop@hdp27 ~]$ hdfs dfs -chmod g+w /tmp  [hadoop@hdp27 ~]$ hdfs dfs -chmod g+w /user/hive/warehouse  [hadoop@hdp27 ~]$ hdfs dfs -chmod 1777 /user/hive/warehouse |

**修改hive配置文件**

|  |
| --- |
| [hadoop@hdp27 ~]$ cd /opt/apache-hive-1.2.1-bin/bin  [hadoop@hdp27 bin]$ vi hive-config.sh  export JAVA\_HOME=/opt/jdk1.8.0\_65  export HADOOP\_HOME=/opt/hadoop-2.7.3  export HIVE\_HOME=/opt/apache-hive-1.2.1-bin  [hadoop@hdp27 bin]$ cd ..  [hadoop@hdp27 apache-hive-1.2.1-bin]$ cd conf  [hadoop@hdp27 conf]$ cp hive-default.xml.template hive-site.xml  [hadoop@hdp27 conf]$ vi hive-site.xml  <property>  <name>hive.exec.stagingdir</name>  <value>/tmp/hive</value>  </property>  <property>  <name>hive.server2.logging.operation.log.location</name>  <value>/tmp/hive/operation\_logs</value>  </property>  <property>  <name>hive.exec.local.scratchdir</name>  <value>/tmp/hive</value>  </property>  <property>  <name>hive.downloaded.resources.dir</name>  <value>/tmp/hive/resources</value>  <description>Temporary local directory for added resources in the remote file system.</description>  </property>  <property>  <name>javax.jdo.option.ConnectionURL</name>  <value>jdbc:postgresql://hdp27/metastore</value>  </property>  <property>  <name>javax.jdo.option.ConnectionDriverName</name>  <value>org.postgresql.Driver</value>  </property>  <property>  <name>javax.jdo.option.ConnectionUserName</name>  <value>hiveuser</value>  </property>  <property>  <name>javax.jdo.option.ConnectionPassword</name>  <value>admin1234</value>  </property>  <property>  <name>datanucleus.autoCreateSchema</name>  <value>false</value>  </property>  <property>  <name>hive.metastore.uris</name>  <value>thrift://hdp27:9083</value>  </property>  <property>  <name>hive.metastore.schema.verification</name>  <value>true</value>  </property>  修改log4j配置  [hadoop@hdp27 conf]$ cp hive-log4j.properties.template hive-log4j.properties  [hadoop@hdp27 conf]$ vi hive-log4j.properties  #log4j.appender.EventCounter=org.apache.hadoop.hive.shims.HiveEventCounter  log4j.appender.EventCounter=org.apache.hadoop.log.metrics.EventCounter |

### 配置HiveServer2

|  |
| --- |
| [root@hdp27 ~]# su - hadoop  [hadoop@hdp27 ~]$ cd /opt/apache-hive-1.2.1-bin/  [hadoop@hdp27 apache-hive-1.2.1-bin]$ vi conf/hive-site.xml  <property>  <name>hive.support.concurrency</name>  <value>true</value>  </property>  <property>  <name>hive.zookeeper.quorum</name>  <value>hdp27</value>  </property>  <property>  <name>hive.server2.thrift.bind.host</name>  <value>hdp27</value>  </property> |

### 启动停止Hive服务

|  |
| --- |
| 启动metastore服务  [hadoop@hdp27 ~]$ hive --service metastore  [hadoop@hdp27 ~]$ hive  ls: cannot access /opt/spark-2.0.2-bin-hadoop2.7/lib/spark-assembly-\*.jar: No such file or directory  Logging initialized using configuration in file:/opt/apache-hive-1.2.1-bin/conf/hive-log4j.properties  hive> show databases;  OK  default  Time taken: 1.006 seconds, Fetched: 1 row(s)  hive>  **启动Hiveserver2**  ***启动Hiveserver2服务，必须先启动Zookeeper服务，本安装配置使用hbase自带的Zookeeper，故先启动hbase***  [hadoop@hdp27 ~]$ start-hbase.sh  [hadoop@hdp27 ~]$ jps    [hadoop@hdp27 ~]$ hive --service hiveserver2  连接Hiveserver2测试  [root@hdp27 ~]# su - hadoop  [hadoop@hdp27 ~]$ beeline  !connect jdbc:hive2://hdp27:10000/default  用户名密码可不填 |

## Hive使用及进阶

### Hive使用

#### Hive基本使用

|  |
| --- |
| **#创建数据(文本以tab分隔)**  [hadoop@hdp27 ~]$ cd /home/hadoop  [hadoop@hdp27 ~]$ mkdir demo  [hadoop@hdp27 ~]$ cd demo  [hadoop@hdp27 demo]$ vi t\_hive.txt  1 12 11  12 131 188  119 134 180  32 56 799  137 150 228  11 345 187  121 67 87  456 873 190  **#创建新表**  [hadoop@hdp27 ~]$ beeline  beeline> !connect jdbc:hive2://hdp27:10000/default  0: jdbc:hive2://hdp27:10000/default> create table t\_hive(a int,b int,c int) row format delimited fields terminated by '\t';  No rows affected (0.617 seconds)  **#导入数据t\_hive.txt到t\_hive表**  0: jdbc:hive2://hdp27:10000/default> load data local inpath '/home/hadoop/demo/t\_hive.txt' overwrite into table t\_hive;  INFO : Loading data to table default.t\_hive from file:/home/hadoop/demo/t\_hive.txt  INFO : Table default.t\_hive stats: [numFiles=1, numRows=0, totalSize=131, rawDataSize=0]  No rows affected (0.776 seconds)  **#查看表**  0: jdbc:hive2://hdp27:10000/default> show tables;  +-----------+--+  | tab\_name |  +-----------+--+  | t\_hive |  +-----------+--+  1 row selected (0.164 seconds)  **#正则匹配表名**  0: jdbc:hive2://hdp27:10000/default> show tables '\*t\*';  +-----------+--+  | tab\_name |  +-----------+--+  | t\_hive |  +-----------+--+  1 row selected (0.084 seconds)  **#查看表数据**  0: jdbc:hive2://hdp27:10000/default> select \* from t\_hive;  +-----------+-----------+-----------+--+  | t\_hive.a | t\_hive.b | t\_hive.c |  +-----------+-----------+-----------+--+  | 1 | 12 | 11 |  | 12 | 131 | 188 |  | 119 | 134 | 180 |  | 32 | 56 | 799 |  | 137 | 150 | 228 |  | 11 | 345 | 187 |  | 121 | 67 | 87 |  | 456 | 873 | 190 |  +-----------+-----------+-----------+--+  8 rows selected (0.133 seconds)  **#查看表结构**  0: jdbc:hive2://hdp27:10000/default> desc t\_hive;  +-----------+------------+----------+--+  | col\_name | data\_type | comment |  +-----------+------------+----------+--+  | a | int | |  | b | int | |  | c | int | |  +-----------+------------+----------+--+  3 rows selected (0.145 seconds)  **#增加一个字段**  0: jdbc:hive2://hdp27:10000/default> alter table t\_hive add columns (new\_col string);  No rows affected (0.173 seconds)  0: jdbc:hive2://hdp27:10000/default> desc t\_hive;  +-----------+------------+----------+--+  | col\_name | data\_type | comment |  +-----------+------------+----------+--+  | a | int | |  | b | int | |  | c | int | |  | new\_col | string | |  +-----------+------------+----------+--+  4 rows selected (0.126 seconds)  **#重命令表名**  0: jdbc:hive2://hdp27:10000/default> alter table t\_hive rename to t\_hadoop;  No rows affected (0.143 seconds)  0: jdbc:hive2://hdp27:10000/default> show tables;  +-----------+--+  | tab\_name |  +-----------+--+  | t\_hadoop |  +-----------+--+  1 row selected (0.061 seconds)  **#删除表**  0: jdbc:hive2://hdp27:10000/default> drop table t\_hadoop;  No rows affected (0.753 seconds)  0: jdbc:hive2://hdp27:10000/default> show tables;  +-----------+--+  | tab\_name |  +-----------+--+  +-----------+--+  No rows selected (0.045 seconds) |

#### Hive交互式模式

#### 数据导入

#### 数据导出

#### Hive查询HiveQL

#### Hive视图

#### Hive分区表

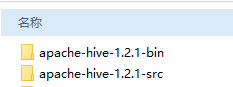
### Hive与Hbase集成

Hive默认的Hbase版本较早，故必须源码编译，否则建表时会报错

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. org.apache.hadoop.hbase.HTableDescriptor.addFamily(Lorg/apache/hadoop/hbase/HColumnDescriptor;)V

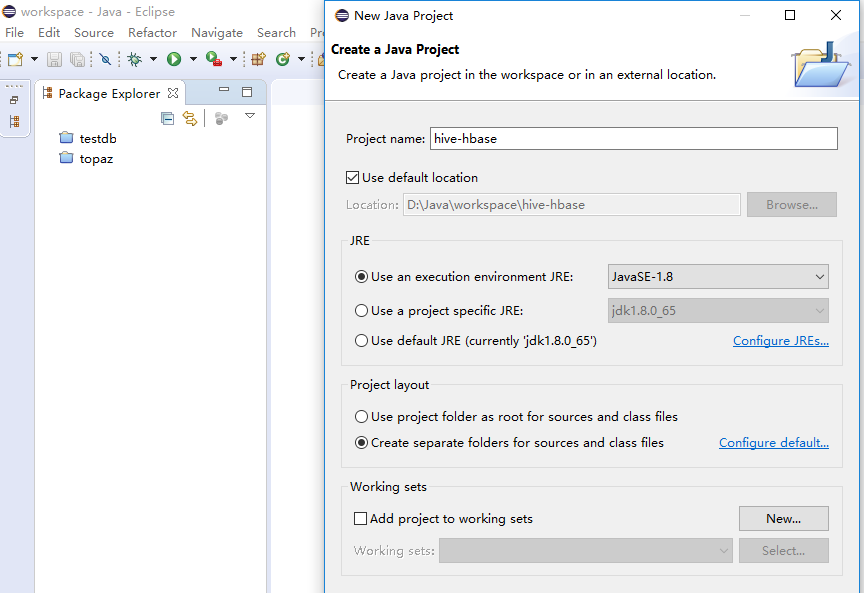
#### 下载解压Hive源码

下载Hive相同版本的源码文件apache-hive-1.2.1-src.tar.gz到本地目录，并解压缩

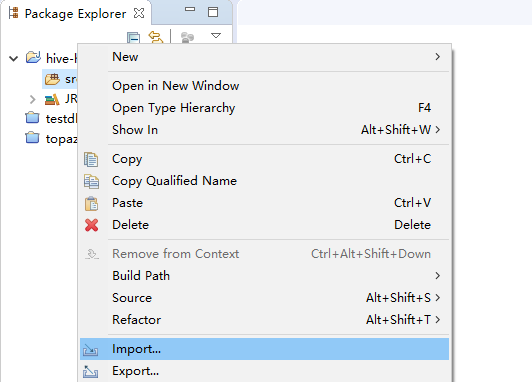


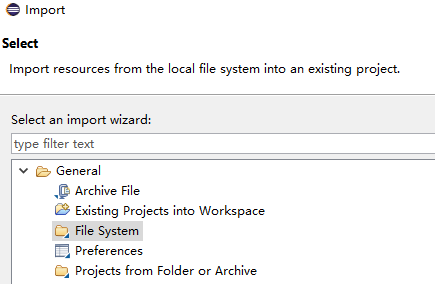
#### 创建Eclipse工程

新建Eclipse Java工程，命名为hive-hbase

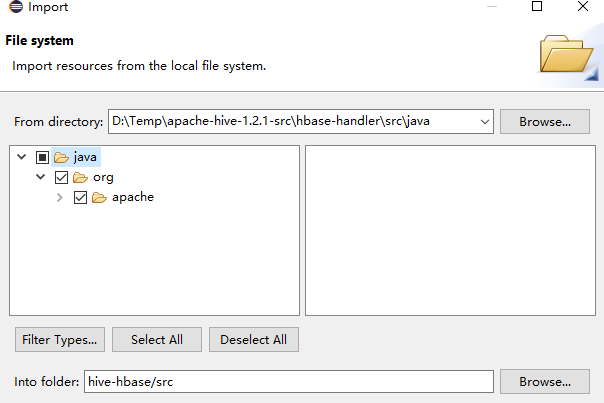


导入Hive Hbase Handler部分的源码

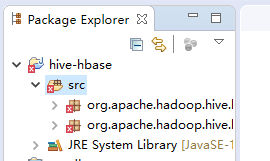




指定导入源码的目录

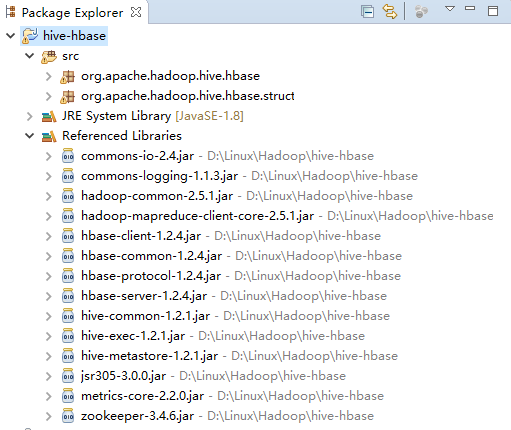


源码导入之后的工程项目结构



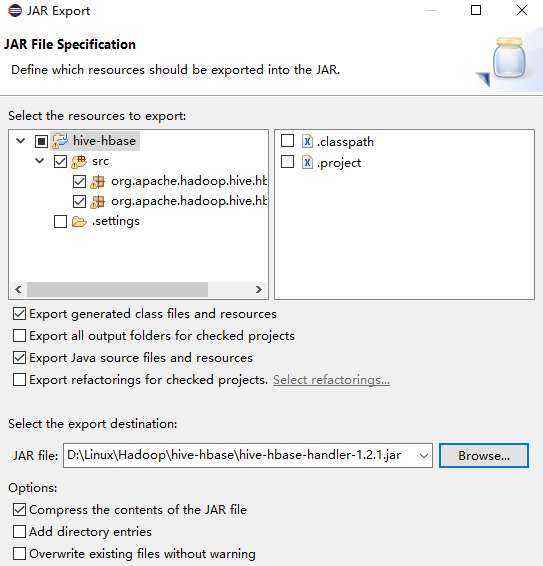
#### 引入相关Jar包与编译

引入如下14个Jar包到Eclipse工程中，Jar文件从Hadoop/Hive/Hbase安装目录中收集



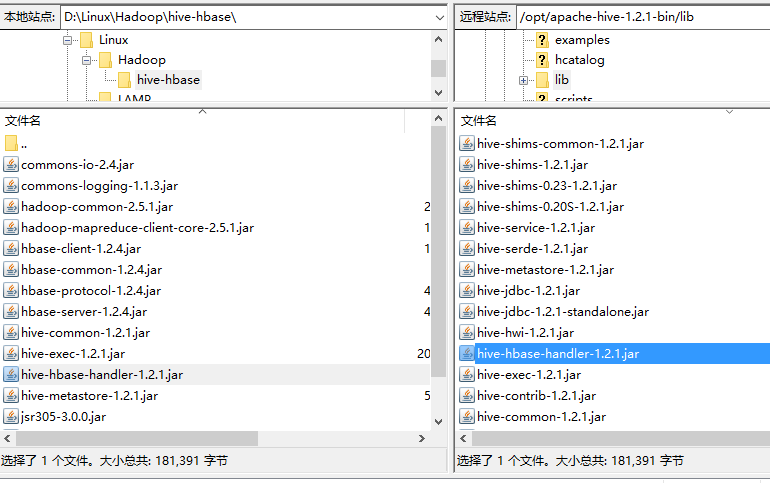
选中源码src目录，导出->Jar File，将目标文件命名为*hive-hbase-handler-1.2.1.jar*

**必须要用此文件名，与Hive的Lib目录下文件名一致**



#### 更新Jar文件

将刚编译好的jar文件及前面依赖的14个文件上传到hive的lib目录下，覆盖现有文件。



#### 修改Hive配置文件

|  |
| --- |
| [hadoop@hdp27 apache-hive-1.2.1-bin]$ vi conf/hive-site.xml  <property>  <name>hive.aux.jars.path</name>  <value>file:///opt/apache-hive-1.2.1-bin/lib/hive-hbase-handler-1.2.1.jar,file:///opt/apache-hive-1.2.1-bin/lib/guava-14.0.1.jar,file:///opt/apache-hive-1.2.1-bin/lib/hbase-common-1.2.4.jar,file:///opt/apache-hive-1.2.1-bin/lib/hbase-client-1.2.4.jar,file:///opt/apache-hive-1.2.1-bin/lib/zookeeper-3.4.6.jar</value>  </property>  <property>  <name>hbase.zookeeper.quorum</name>  <value>hdp27</value>  </property> |

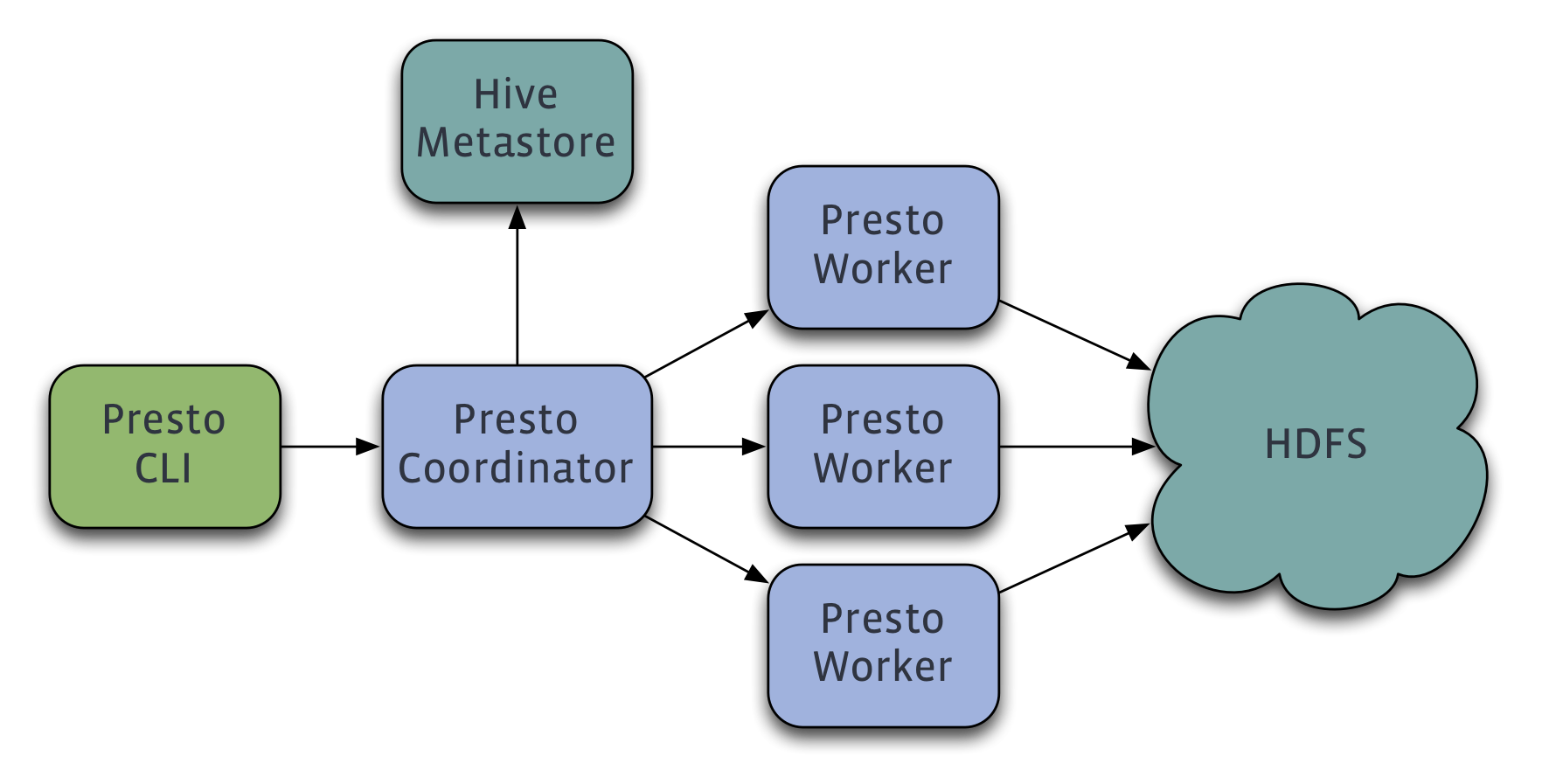
#### 测试Hive与Hbase集成

|  |
| --- |
| [hadoop@hdp27 bin]$ hive  ls: cannot access /opt/spark-2.0.2-bin-hadoop2.7/lib/spark-assembly-\*.jar: No such file or directory  Logging initialized using configuration in file:/opt/apache-hive-1.2.1-bin/conf/hive-log4j.properties  hive> CREATE TABLE hbase\_table\_1(key int, value string)  > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'  > WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf1:val")  > TBLPROPERTIES ("hbase.table.name" = "xyz", "hbase.mapred.output.outputtable" = "xyz");  OK  Time taken: 2.365 seconds  hive> drop table hbase\_table\_1;  OK  Time taken: 3.795 seconds |

### Python 3访问Hive

## Presto安装配置

### Presto体系架构



### Presto安装配置

#### 上传&解压Presto

|  |
| --- |
| [root@hdp27 install]# tar -xvzf presto-server-0.157.tar.gz  [root@hdp27 install]# mv presto-server-0.157 ..  [root@hdp27 opt]# chown -R hadoop:hadoop presto-server-0.157  [root@hdp27 ~]# mkdir -p /var/presto  [root@hdp27 ~]# chown -R hadoop:hadoop /var/presto |

#### 编写配置文件

|  |
| --- |
| [root@hdp27 ~]# su - hadoop  [hadoop@hdp27 ~]$ cd /opt/presto-server-0.157/  [hadoop@hdp27 presto-server-0.157]$ mkdir etc  [hadoop@hdp27 presto-server-0.157]$ vi etc/node.properties  node.environment=production  node.id=ffffffff-ffff-ffff-ffff-ffffffffffff  node.data-dir=/var/presto/data  [hadoop@hdp27 presto-server-0.157]$ vi etc/jvm.config  -server  -Xmx16G  -XX:+UseG1GC  -XX:G1HeapRegionSize=32M  -XX:+UseGCOverheadLimit  -XX:+ExplicitGCInvokesConcurrent  -XX:+HeapDumpOnOutOfMemoryError  -XX:OnOutOfMemoryError=kill -9 %p  [hadoop@hdp27 presto-server-0.157]$ vi etc/config.properties  coordinator=true  node-scheduler.include-coordinator=true  http-server.http.port=8080  query.max-memory=5GB  query.max-memory-per-node=1GB  discovery-server.enabled=true  discovery.uri=http://hdp27:8080  [hadoop@hdp27 presto-server-0.157]$ vi etc/log.properties  com.facebook.presto=INFO |

#### 配置Hive Connector

|  |
| --- |
| [hadoop@hdp27 presto-server-0.157]$ mkdir -p etc/catalog  [hadoop@hdp27 presto-server-0.157]$ vi etc/catalog/hive.properties  connector.name=hive  hive.metastore.uri=thrift://hdp27:9083 #修改为 hive-metastore 服务所在的主机名称，这里我是安装在Hadoop节点  hive.config.resources=/opt/hadoop-2.7.3/etc/hadoop/core-site.xml,/opt/hadoop-2.7.3/etc/hadoop/hdfs-site.xml |

#### 启动Presto

### Presto连接访问

## Sqoop安装配置

## Pig安装配置

## Tableau使用Spark SQL案例