# 第一部分.安装hdfs集群的具体步骤：

## 一、首先需要准备N台linux服务器

先准备3台虚拟机：1个namenode节点 + 2 个datanode 节点

## 二、修改各台机器的主机名和ip地址

主机名：hdp-01 对应的ip地址：192.168.33.61

主机名：hdp-02 对应的ip地址：192.168.33.62

主机名：hdp-03 对应的ip地址：192.168.33.63

主机名：hdp-04 对应的ip地址：192.168.33.64

## 三、从windows中用MobaXterm\_Personal\_10.软件进行远程连接

在windows中将各台linux机器的主机名配置到的windows的本地域名映射文件中：

c:/windows/system32/drivers/etc/hosts

|  |
| --- |
| 192.168.33.61 hdp-01  192.168.33.62 hdp-02  192.168.33.63 hdp-03  192.168.33.64 hdp-04 |

## 四、配置linux服务器的基础软件环境

* 防火墙

关闭防火墙：service iptables stop

关闭防火墙自启： chkconfig iptables off

* 安装jdk：（hadoop体系中的各软件都是java开发的）

1. 打开sftp窗口，然后将jdk压缩包拖入sftp窗口
2. 然后在linux中将jdk压缩包解压到/root/apps 下
3. 配置环境变量：JAVA\_HOME PATH

vi /etc/profile 在文件的最后，加入：

|  |
| --- |
| export JAVA\_HOME=/root/apps/jdk1.8.0\_60  export PATH=$PATH:$JAVA\_HOME/bin |

1. 修改完成后，记得 source /etc/profile使配置生效
2. 检验：在任意目录下输入命令： java -version 看是否成功执行
3. 将安装好的jdk目录用scp命令拷贝到其他机器
4. 将/etc/profile配置文件也用scp命令拷贝到其他机器并分别执行source命令

* 集群内主机的**域名映射**配置

在hdp-01（任意一台虚拟机的名字）上，vi /etc/hosts

|  |
| --- |
| 127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4  ::1 localhost localhost.localdomain localhost6 localhost6.localdomain6  192.168.1.\* \*\*\*  192.168.1.\* \*\*\*  192.168.1.\* \*\*\*  192.168.1.\* \*\*\* |

然后，将hosts文件拷贝到集群中的所有其他机器上

scp /etc/hosts \*\*\*:/etc/

scp /etc/hosts \*\*\*:/etc/

scp /etc/hosts \*\*\*:/etc/

**如果用的是VMware虚拟机，则使用以下方法：**

|  |  |
| --- | --- |
| 补充  提示: | 如果在执行scp命令的时候，提示没有scp命令，则可以配置一个本地yum源来安装   1. 先在虚拟机中配置cdrom为一个centos的安装镜像iso文件 2. 在linux系统中将光驱挂在到文件系统中（某个目录） 3. mkdir /mnt/cdrom 4. mount -t iso9660 -o loop /dev/cdrom /mnt/cdrom 5. 检验挂载是否成功： ls /mnt/cdrom 6. 3、配置yum的仓库地址配置文件 7. yum的仓库地址配置文件目录： /etc/yum.repos.d 8. 先将自带的仓库地址配置文件批量更名：      1. 然后，拷贝一个出来进行修改        1. 修改完配置文件后，再安装scp命令： 2. yum install openssh-clients -y |

**如果是华为云服务器，则直接挂在华为云提供的centos镜像，步骤如下**

1. 以root用户登录弹性云服务器。
2. 执行以下命令，备份CentOS-Base.repo文件。

**mkdir -p /etc/yum.repos.d/repo\_bak/**

**mv /etc/yum.repos.d/\*.repo /etc/yum.repos.d/repo\_bak/**

1. 使用curl命令下载对应版本的CentOS-Base.repo文件，并将其放入/etc/yum.repos.d/目录。

针对华为云当前支持的CentOS镜像源版本，使用的curl命令分别如下：

* + CentOS 6

**curl -o /etc/yum.repos.d/CentOS-Base.repo http://mirrors.myhuaweicloud.com/repo/CentOS-Base-6.repo**

* + CentOS 7

**curl -o /etc/yum.repos.d/CentOS-Base.repo http://mirrors.myhuaweicloud.com/repo/CentOS-Base-7.repo**

1. 执行以下命令，生成缓存。

**yum makecache**

**说明：**

* 如果执行命令**yum makecache**后，回显信息为“Another app is currently holding the yum lock”，则需执行命令**rm -rf /var/run/yum.pid**删除yum.pid文件，然后再执行命令**yum makecache**生成缓存。
* 如果执行命令**yum makecache**后，回显信息为“Error: Cannot retrieve metalink for repository: epel. Please verify its path and try again”，则需执行命令**yum --disablerepo=epel -y update ca-certificates**更新证书，临时禁用epel源。

**说明**

* 如果执行命令**yum makecache**后，回显信息为“Another app is currently holding the yum lock”，则需执行命令**rm -rf /var/run/yum.pid**删除yum.pid文件，然后再执行命令**yum makecache**生成缓存。
* 如果执行命令**yum makecache**后，回显信息为“Error: Cannot retrieve metalink for repository: epel. Please verify its path and try again”，则需执行命令**yum --disablerepo=epel -y update ca-certificates**更新证书，临时禁用epel源。

## 五、安装hdfs集群

1、上传hadoop安装包到hdp-01

1. 修改配置文件

|  |  |
| --- | --- |
| 要点提示 | ***核心配置参数：***   1. *指定hadoop的默认文件系统为：hdfs* 2. *指定hdfs的namenode节点为哪台机器* 3. *指定namenode软件存储元数据的****本地目录*** 4. *指定datanode软件存放文件块的****本地目录*** |

hadoop的配置文件在：/root/apps/hadoop安装目录/etc/hadoop/

**1) 修改hadoop-env.sh**

*export JAVA\_HOME=/root/apps/jdk1.8.0\_60*

**2) 修改core-site.xml**

|  |
| --- |
| <configuration>  <property>  <name>fs.defaultFS</name>  <value>hdfs://hdp-01:9000</value>  </property>  </configuration> |

**3) 修改hdfs-site.xml**

|  |
| --- |
| <configuration>  <property>  <name>dfs.namenode.name.dir</name>  <value>/root/dfs/name</value>  </property>  <property>  <name>dfs.datanode.data.dir</name>  <value>/root/dfs/data</value>  </property>  </configuration> |

**4) 拷贝整个hadoop安装目录到其他机器**

*scp -r /root/apps/hadoop-2.8.0 hdp-02:/root/apps/*

*scp -r /root/apps/hadoop-2.8.0 hdp-03:/root/apps/*

*scp -r /root/apps/hadoop-2.8.0 hdp-04:/root/apps/*

**5) 启动HDFS**

所谓的启动HDFS，就是在对的机器上启动对的软件

|  |  |  |
| --- | --- | --- |
| 要点  提示： | *要运行hadoop的命令，需要在linux环境中配置HADOOP\_HOME和PATH环境变量*  *vi /etc/profile*   |  | | --- | | *export JAVA\_HOME=/root/apps/jdk1.8.0\_60*  *export HADOOP\_HOME=/root/apps/hadoop-2.8.0*  *export PATH=$PATH:$JAVA\_HOME/bin:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin* | |

**首先，初始化namenode的元数据目录**

要在hdp-01上执行hadoop的一个命令来初始化namenode的元数据存储目录

*hadoop namenode -format*

* 创建一个全新的元数据存储目录
* 生成记录元数据的文件fsimage
* 生成集群的相关标识：如：集群id——clusterID

**然后，启动namenode进程（在hdp-01上）**

*hadoop-daemon.sh start namenode*

启动完后，首先用jps查看一下namenode的进程是否存在

**然后，在windows中用浏览器访问namenode提供的web端口：50070**

http://hdp-01:50070

**然后，启动众datanode们（在任意地方）**

*hadoop-daemon.sh start datanode*

**6) 用自动批量启动脚本来启动HDFS**

1. 先配置hdp-01到集群中所有机器（包含自己）的免密登陆
2. 配完免密后，可以执行一次 *ssh 0.0.0.0*
3. 修改hadoop安装目录中/etc/hadoop/slaves（把需要启动datanode进程的节点列入）

|  |
| --- |
| hdp-01  hdp-02  hdp-03  hdp-04 |

1. 在hdp-01上用脚本：**start-dfs.sh** 来自动启动整个集群
2. 如果要停止，则用脚本：**stop-dfs.sh**

# 第二部分. 安装yarn集群

yarn集群中有两个角色：

主节点：Resource Manager 1台

从节点：Node Manager N台

Resource Manager一般安装在一台专门的机器上

Node Manager应该与HDFS中的data node重叠在一起

修改配置文件：

yarn-site.xml

|  |
| --- |
| <property>  <name>yarn.resourcemanager.hostname</name>  <value>hdp-04</value>  </property>  <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property> |

然后复制到每一台机器上

然后在hdp-04上，修改hadoop的slaves文件，列入要启动nodemanager的机器

然后将hdp-04到所有机器的免密登陆配置好

然后，就可以用脚本启动yarn集群：

sbin/start-yarn.sh

停止：

sbin/stop-yarn.sh

启动完成后，可以在windows上用浏览器访问resourcemanager的web端口：

<http://hdp-04:8088>

看resource mananger是否认出了所有的node manager节点

# 第三部分. zookeeper 集群搭建

上传安装包，解压

修改conf/zoo.cfg

|  |
| --- |
| # The number of milliseconds of each tick  tickTime=2000  # The number of ticks that the initial  # synchronization phase can take  initLimit=10  # The number of ticks that can pass between  # sending a request and getting an acknowledgement  syncLimit=5  # the directory where the snapshot is stored.  # do not use /tmp for storage, /tmp here is just  # example sakes.  dataDir=/root/zkdata  # the port at which the clients will connect  clientPort=2181  # Set to "0" to disable auto purge feature  #autopurge.purgeInterval=1  server.1=hdp-01:2888:3888  server.2=hdp-02:2888:3888  server.3=hdp-03:2888:3888 |

配置文件修改完后，将安装包拷贝给hdp-02 和 hdp-03

接着，到hdp-01上，新建数据目录/root/zkdata，并在目录中生成一个文件myid，内容为1

接着，到hdp-02上，新建数据目录/root/zkdata，并在目录中生成一个文件myid，内容为2

接着，到hdp-03上，新建数据目录/root/zkdata，并在目录中生成一个文件myid，内容为3

启动zookeeper集群：zookeeper集群启动时可写个脚本，运行脚本后，便不需要一台一台的启动了。

脚本请自行编写。

# 第四部分.hive安装

## 一．最简安装：用内嵌derby作为元数据库

准备工作：安装hive的机器上应该有HADOOP环境（安装目录，HADOOP\_HOME环境变量）

安装：直接解压一个hive安装包即可

此时，安装的这个hive实例使用其内嵌的derby数据库作为记录元数据的数据库

此模式不便于让团队成员之间共享协作

## 二．标准安装：将mysql作为元数据库

### mysql安装

1. 上传mysql安装包
2. 解压：

[root@mylove ~]# tar -xvf MySQL-5.6.26-1.linux\_glibc2.5.x86\_64.rpm-bundle.tar

1. 安装mysql的server包

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

依赖报错：

缺perl

yum install perl

安装完perl后 ，继续重新安装mysql-server

（可以配置一个本地yum源进行安装：

1、先在vmware中给这台虚拟机连接一个光盘镜像

2、挂在光驱到一个指定目录：mount -t iso9660 -o loop /dev/cdrom /mnt/cdrom

3、将yum的配置文件中baseURL指向/mnt/cdrom

）

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

又出错：包冲突conflict with

移除老版本的冲突包：mysql-libs-5.1.73-3.el6\_5.x86\_64

[root@mylove ~]# rpm -e mysql-libs-5.1.73-3.el6\_5.x86\_64 --nodeps

继续重新安装mysql-server

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

成功后，注意提示：里面有初始密码及如何改密码的信息

初始密码：/root/.mysql\_secret

改密码脚本：/usr/bin/mysql\_secure\_installation

1. 安装mysql的客户端包：

[root@mylove ~]# rpm -ivh MySQL-client-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

1. 启动mysql的服务端：

[root@mylove ~]# service mysql start

Starting MySQL. SUCCESS!

1. 修改root的初始密码：

[root@mylove ~]# /usr/bin/mysql\_secure\_installation 按提示

1. 测试：

用mysql命令行客户端登陆mysql服务器看能否成功

[root@mylove ~]# mysql -uroot -proot

mysql> show databases;

1. 给root用户授予从任何机器上登陆mysql服务器的权限：

mysql> grant all privileges on \*.\* to 'root'@'%' identified by '你的密码' with grant option;

Query OK, 0 rows affected (0.00 sec)

mysql> flush privileges;

Query OK, 0 rows affected (0.00 sec)

注意点：要让mysql可以远程登录访问

最直接测试方法：从windows上用Navicat去连接，能连，则可以，不能连，则要去mysql的机器上用命令行客户端进行授权：

在mysql的机器上,启动命令行客户端：

mysql -uroot -proot

mysql>grant all privileges on \*.\* to 'root'@'%' identified by 'root的密码' with grant option;  
mysql>flush privileges;

### hive的元数据库配置

vi conf/hive-site.xml

|  |
| --- |
| <configuration>  <property>  <name>javax.jdo.option.ConnectionURL</name>  <value>jdbc:mysql://localhost: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>root</value>  <description>password to use against metastore database</description>  </property>  </configuration> |

1. 上传一个mysql的驱动jar包到hive的安装目录的lib中
2. 配置HADOOP\_HOME 和HIVE\_HOME到系统环境变量中：/etc/profile
3. source /etc/profile

5、hive启动测试

然后用命令启动hive交互界面：

[root@hdp20-04 ~]# hive

## 三．hive使用方式

**最基本使用方式：**

启动一个hive交互shell

bin/hive

hive>

设置一些基本参数，让hive使用起来更便捷，比如：

1、让提示符显示当前库：

hive>set hive.cli.print.current.db=true;

1. 显示查询结果时显示字段名称：

hive>set hive.cli.print.header=true;

但是这样设置只对当前会话有效，重启hive会话后就失效，解决办法：

在linux的当前用户目录中，编辑一个.hiverc文件，将参数写入其中：

vi .hiverc

|  |
| --- |
| set hive.cli.print.header=true;  set hive.cli.print.current.db=true; |

## 四．启动hive服务使用

启动hive的服务：

[root@hdp20-04 hive-1.2.1]# bin/hiveserver2 -hiveconf hive.root.logger=DEBUG,console

上述启动，会将这个服务启动在前台，如果要启动在后台，则命令如下：

nohup bin/hiveserver2 1>/dev/null 2>&1 &

启动成功后，可以在别的节点上用beeline去连接

* 方式（1）

[root@hdp20-04 hive-1.2.1]# bin/beeline 回车，进入beeline的命令界面

输入命令连接hiveserver2

beeline> !connect jdbc:hive2//mini1:10000

（hadoop01是hiveserver2所启动的那台主机名，端口默认是10000）

* 方式（2）

启动时直接连接：

bin/beeline -u jdbc:hive2://mini1:10000 -n root

接下来就可以做正常sql查询了

## 五．脚本化运行

大量的hive查询任务，如果用交互式shell来进行输入的话，显然效率及其低下，因此，生产中更多的是使用脚本化运行机制：

该机制的核心点是：hive可以用一次性命令的方式来执行给定的hql语句

[root@hdp20-04 ~]# hive -e "insert into table t\_dest select \* from t\_src;"

然后，进一步，可以将上述命令写入shell脚本中，以便于脚本化运行hive任务，并控制、调度众多hive任务，示例如下：

vi t\_order\_etl.sh

|  |
| --- |
| #!/bin/bash  hive -e "select \* from db\_order.t\_order"  hive -e "select \* from default.t\_user"  hql="create table default.t\_bash as select \* from db\_order.t\_order"  hive -e "$hql" |

如果要执行的hql语句特别复杂，那么，可以把hql语句写入一个文件：

vi x.hql

|  |
| --- |
| select \* from db\_order.t\_order;  select count(1) from db\_order.t\_user; |

然后，用hive -f /root/x.hql 来执行

# 第五部分. 安装HBASE

HBASE是一个分布式系统

其中有一个管理角色： HMaster(一般2台，一台active，一台backup)

其他的数据节点角色： HRegionServer(很多台，看数据容量)

### 安装准备：

首先，要有一个HDFS集群，并正常运行； regionserver应该跟hdfs中的datanode在一起

其次，还需要一个zookeeper集群，并正常运行

然后，安装HBASE

角色分配如下：

Hdp01: namenode datanode regionserver hmaster zookeeper

Hdp02: datanode regionserver zookeeper

Hdp03: datanode regionserver zookeeper

### 安装步骤：

#### 安装zookeeper

自行查阅第三部分

#### 2.安装hbase

解压hbase安装包

修改hbase-env.sh

|  |
| --- |
| export JAVA\_HOME=/root/apps/jdk1.7.0\_67  export HBASE\_MANAGES\_ZK=false |

修改hbase-site.xml

|  |
| --- |
| <configuration>  <!-- 指定hbase在HDFS上存储的路径 -->  <property>  <name>hbase.rootdir</name>  <value>hdfs://hdp01:9000/hbase</value>  </property>  <!-- 指定hbase是分布式的 -->  <property>  <name>hbase.cluster.distributed</name>  <value>true</value>  </property>  <!-- 指定zk的地址，多个用“,”分割 -->  <property>  <name>hbase.zookeeper.quorum</name>  <value>hdp01:2181,hdp02:2181,hdp03:2181</value>  </property>  </configuration> |

修改 regionservers

|  |
| --- |
| hdp01  hdp02  hdp03 |

#### 启动hbase集群：

bin/start-hbase.sh

启动完后，还可以在集群中找任意一台机器启动一个备用的master

bin/hbase-daemon.sh start master

新启的这个master会处于backup状态

#### 启动hbase的命令行客户端

bin/hbase shell

Hbase> list // 查看表

Hbase> status // 查看集群状态

Hbase> version // 查看集群版本

# 第六部分.Flume的安装部署

1. Flume的安装非常简单，只需要解压即可，当然，前提是已有hadoop环境

上传安装包到数据源所在节点上

然后解压 tar -zxvf apache-flume-1.6.0-bin.tar.gz

然后进入flume的目录，修改conf下的flume-env.sh，在里面配置JAVA\_HOME

2、根据数据采集的需求**配置采集方案**，描述在配置文件中(文件名可任意自定义)

3、**指定采集方案配置文件**，在相应的节点上启动flume agent

先用一个最简单的例子来测试一下程序环境是否正常:

1. 先在flume的conf目录下新建一个配置文件（采集方案）

vi netcat-logger.properties

|  |
| --- |
| # 定义这个agent中各组件的名字  a1.sources = r1  a1.sinks = k1  a1.channels = c1  # 描述和配置source组件：r1  a1.sources.r1.type = netcat  a1.sources.r1.bind = localhost  a1.sources.r1.port = 44444  # 描述和配置sink组件：k1  a1.sinks.k1.type = logger  # 描述和配置channel组件，此处使用是内存缓存的方式  a1.channels.c1.type = memory  a1.channels.c1.capacity = 1000  a1.channels.c1.transactionCapacity = 100  # 描述和配置source channel sink之间的连接关系  a1.sources.r1.channels = c1  a1.sinks.k1.channel = c1 |

1. 启动agent去采集数据

|  |
| --- |
| bin/flume-ng agent -c conf -f conf/netcat-logger.conf -n a1 -Dflume.root.logger=INFO,console |