# Hadoop笔记

### Hadoop的简介

#### Hadoop与云计算

##### Hadoop的简介

Hadoop适合处理大规模数据

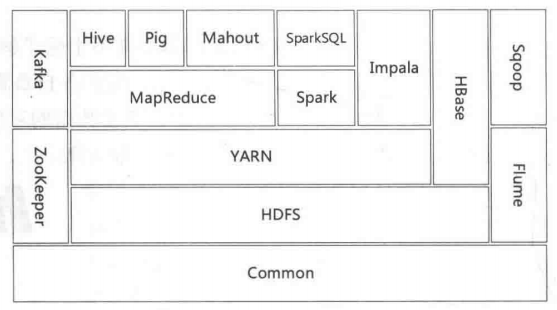
Hadoop被部署到一个集群上

Hadoop主要用于分布式存储和分布式计算

Hadoop的特点：无共享，高可用，高扩展

##### Hadoop的生态圈

Hadoop本身只有Common、HDFS、YARN和MapReduce模块，但是扩展之后，如图所示：



##### Hadoop与云计算

* + - 1. 云计算

1. 按需提供服务
2. 宽带网络访问
3. 资源池化—虚拟化技术
4. 高伸缩性，自动适用业务变化
5. 可量化，动态针对服务收费
6. 大规模，服务廉价
   * + 1. 云计算类型

基础设施即服务：相当于对用户提供裸机，常用的虚拟化产品如：

Vmware vShpere,Hyper-V,KVM,Xen,Docker等

平台即服务：相当于对用户提供系统。

软件即服务：相当于对用户提供软件使用权

1.1.3.3 Hadoop相当于云计算中的基础设置即服务层的其中一种实现。

#### 1.2 Hadoop与大数据

##### 1.2.1 大数据

数据量（volume）多，数据类型多样化（variety），速度（velocity）

数据类型：结构化（表），半结构化（json，xml），准结构化（可结构化处理），非结构化（图片，视频，文档）

##### 1.2.2 Hadoop与大数据

**HDFS**，Google的GFS的开源实现。

HDFS的设计理念：流式数据访问模式，存储超大文件，运行于廉价硬件集群上。

**MapReduce**，Google的MapReduce的开源实现。

MapReduce的分布式计算模型。

**HBase**，Google的Bigtable的开源实现。结构化存储的分布式数据库。

HBase弥补了HDFS的流式读写，其擅长实时随机读写。

##### 1.2.3 其他大数据处理平台

**Storm**，针对流式数据，现实生活中很多数据不是一个文件，而是一个源源不断的数据流，需要实时分析。Storm也可以运行在YARN之上。

**Spark**，基于内存计算，且在并行计算有很大优势。减少了HDFS的开销。提供Cache，减少了I/O开销。使用多线程池，减少了任务启动开销。

广泛的数据集操作类型：Spark SQL支持SQL的结构化查询，Spark Streaming是其流计算框架，Mllib集成了主流机器学习算法，GraphX则是Spark的图计算框架。

#### Hadoop应用领域

数据挖掘

数据仓库

商业智能

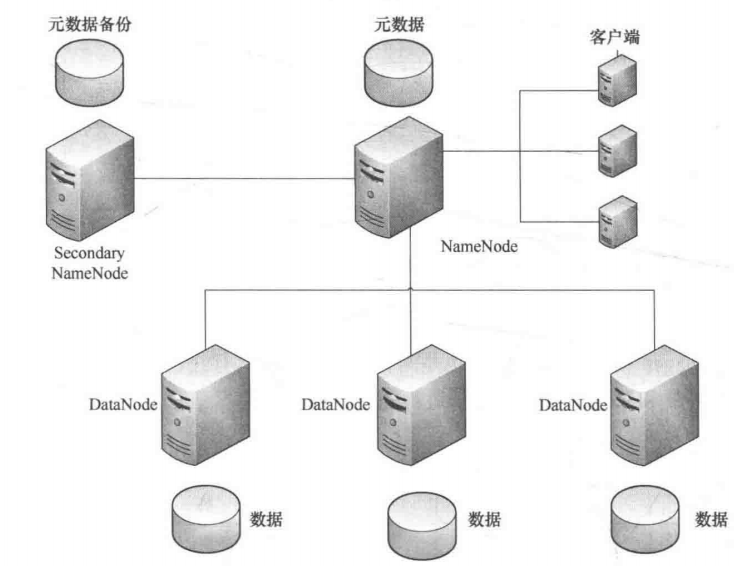
### Hadoop的安装

#### 2.1 Hadoop的架构

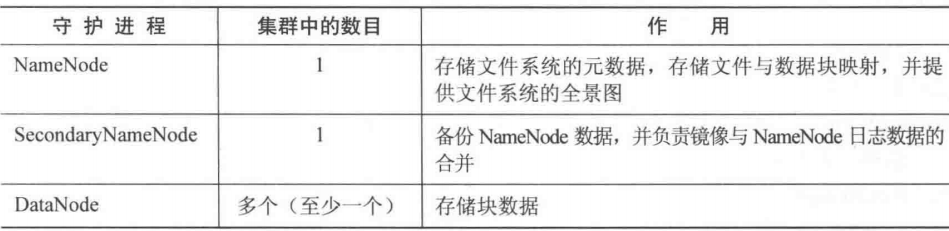
##### 2.1.1 HDFS的架构

构成HDFS集群的主要有两类节点，以主从（master/slave）模式运行，即一个NameNode和多个DataNode。而SecondaryNameNode为NameNode的备份。其实所讲的NameNode，DataNode，SecondaryNameNode是之其守护进程，当然也指运行指定进程的节点。

客户端则是用户与整个文件系统交互的客户端。

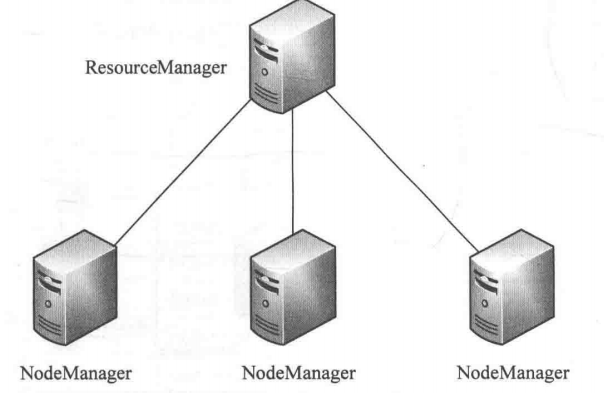


各个守护进程的作用以及在集群中的数目：

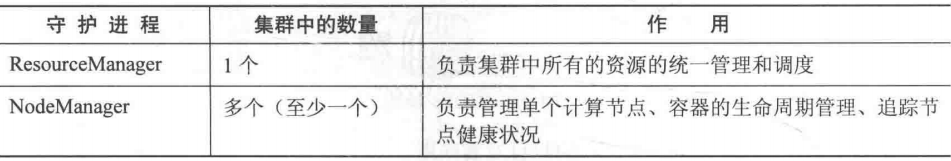


##### 2.1.2 YARN架构

构成YARN的集群也是两类节点：ResourceManager和NodeManager。也是（master/slave）架构。



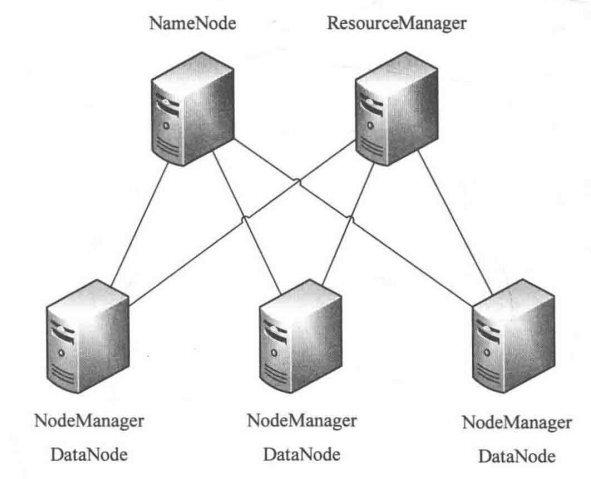
各个守护进程的作用：



##### 2.1.4 总结

综上所述，其实HDFS和YARN集群是由守护进程组成。其中NameNode进程和ResourceManager进程在一个节点上运行，同时，DataNode和NodeManager进程在同一个节点上运行。

注意：NameNode进程和ResourceManager进程不一定在同一个节点，并且在应用上建议分开。而DataNode和NodeManager进程需要配对部署到同一个节点上。



#### 2.2 Hadoop的安装

##### 2.2.1 安装Hadoop

***Hadoop的安装步骤***：

###### 2.2.1.1 安装运行环境（所有节点）

通过虚拟机安装CentOS

修改MAC地址：MACADDR=xx:xx:xx:xx

###### 2.2.1.2 修改主机名和用户名（所有节点）

a. 修改用户名和密码：useradd hadoop；passwd hadoop

b. 修改主机名：vi /etc/sysconfig/network

NETWORKING=yes

HOSTNAME=master

其他节点依次修改为HOSTNAME= slave1，slave2 ……

c. 修改所有节点hosts文件: vi /etc/hosts,可以让节点间使用主机名访问

#填写IP地址与主机名

192.168.10.3 master

#填写slave的IP地址和主机名

192.168.10.4 slave1

192.168.10.5 slave2

……

###### 2.2.1.3 配置静态IP地址（所有节点）

配置静态IP地址：

vi /etc/sysconfig/network-scripts/ifcfg-enn33

BOOTPROTO=static

NM\_CONTROLLED=yes

ONBOOT=yes

IPADDR=192.168.10.3

GATEWAY=192.168.10.1

NETMASK=255.255.255.0

###### 2.2.1.4 配置SSH无密码连接

a.关闭防火墙，service iptables stop(6.x） （所有节点）

或者永久关闭防火墙，chkconfig iptables off(6.x）

查看防火墙状态，service firewalld status

关闭防火墙，service firewalld stop

开启防火墙，service firewalld start

重启防火墙，service firewalld restart

开启开机启动，systemctl enable firewalld.service

关闭开机启动，systemctl disable firewalld.service

b.检查ssh是否安装，yum install ssh;yum install rsync（所有节点）

c.启动SSH服务，service sshd restart

d.检查是否安装成功，rpm -qa|grep openssh;rpm -qa|grep rsync

e.生成ssh公钥，ssh-keygen -t rsa(主节点)，（hadoop用户）

f.将公钥发送到从节点，（hadoop用户）

ssh-copy-id -i ~/.ssh/id\_rsa.pub hadoop@slave1

ssh-copy-id -i ~/.ssh/id\_rsa.pub hadoop@slave2

如果是root用户：

ssh-copy-id -i ~/.ssh/id\_rsa.pub root@slave1

ssh-copy-id -i ~/.ssh/id\_rsa.pub root@slave2

g.验证安装，完全分布（ssh slave1），伪分布（ssh master）

如果失败可能是权限问题，

chmod700 /home/hadoop/.ssh

chmod644 /home/hadoop/.ssh/authorized\_keys

###### 2.2.1.5 安装JDK

配置环境变量：

JAVA\_HOME=/opt/jdk1.8.0\_144

CLASSPATH=$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar

PATH=$PATH:$JAVA\_HOME/bin

Export JAVA\_HOME CLASSPATH PATH

###### 2.2.1.6 配置Hadoop

1. 下载：

<https://www.cloudera.com/documentation/enterprise/release-notes/topics/cm_vd_cdh_package_tarball_512.html#cm_vd_cdh_package_tarball_512>

2. 安装与配置

a) 配置环境变量 vi /etc/profile：

HADOOP\_HOME=/opt/hadoop

PATH=$PATH:$HADOOP\_HOME/bin

Source /etc/profile

**b**) 修改hadoop配置文件

Hadoop的配置文件：



安装hadoop只涉及前6个文件(hadoop-2.6.x/etc/hadoop/)

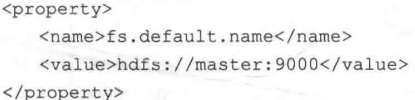
c) 修改 hadoop-env.sh

配置JAVA\_HOME：export JAVA\_HOME=/opt/jdk1.8.0\_144

配置HADOOP\_HOME：export HADOOP\_HOME=/opt/hadoop

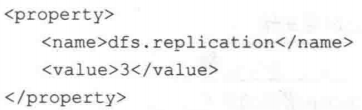
d) 修改core-site.xml

配置HDFS服务的主机名和端口号，同时也指明了NameNode节点



e) 修改hdfs-site.xml

配置冗余副本数



存放name元数据目录

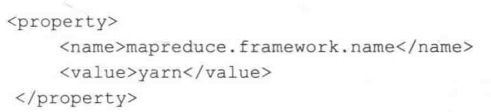


存放data的数据目录



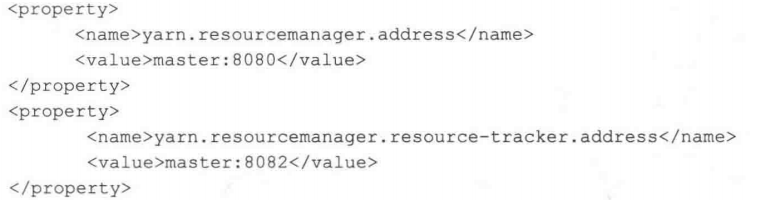
f) 修改mapred-site.xml

指明MapReduce计算框架基于YARN

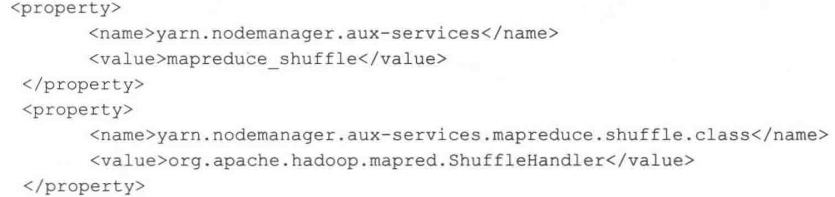


g) 修改yarn-site.xml

指明ResourceManager服务的主机名和端口号

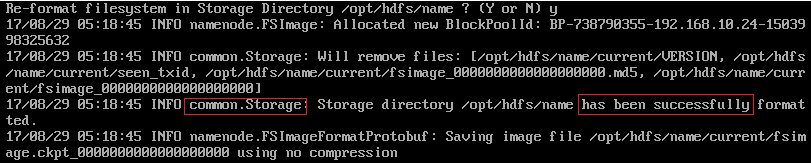


指明mapreduce\_shuffle的类



h) 服务器之间复制文件

scp -r /opt/hadoop hadoop@slave:/opt/



i) **格式化HDFS**

格式化：hadoop namenode -format

j) **启动Hadoop并验证安装**

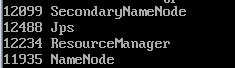
启动hadoop

/opt/hadoop/sbin/start-all.sh

使用jps验证是否成功

会显示本机节点。

Master：



Salve：



##### 2.2.2 安装Hive

###### 2.2.2.1 安装元数据库—mysql

1. 安装mysql yum install MySQL-\*.rpm

2. 启动mysq

###### 2.2.2.2 修改Hive配置文件

##### 2.2.3 安装HBase

##### 2.2.4 安装Sqoop

### 第三章 Hadoop之HDFS

#### 3.1 在HDFS上实现增删改查

##### 3.1.1 新建文件夹

hadoop fs -mkdir /user

hadoop fs -mkdir /user/root

##### 3.1.2 查看文件夹权限

查看所有文件夹

hadoop fs -ls / (根目录）

hadoop fs -ls /user (显示/user目录下的文件)

hadoop fs -ls -d /user/root （显示root文件夹的权限）

##### 3.1.3 上传文件

hadoop fs -put main.c /user/root

hadoop fs -ls -R /user/root (显示root下的文件)

##### 3.1.4 查看文件内容

hadoop fs -cat /user/root/main.c

##### 3.1.5 复制移动与删除

复制

hadoop fs -cp /user/root/main.c /user

hadoop fs -ls /user

移动

hadoop fs -mv /user/main.c /hadoop/

hadoop fs -ls /hadoop

hadoop fs -ls /user

删除

hadoop fs -rm -r /user/root/main.c

hadoop fs -ls /user/root

hadoop fs -rm -r /user

#### 3.2 MapReduce任务

### 第四章 Hadoop之YARN

### 第五章 Hadoop之MapReduce

### 第六章 Hadoop之Hive

### 第七章 Hadoop之Sqoop

### 第八章 Hadoop之HBase

### 第九章 Hadoop之应用开发