> Hadoop 是一个分布式系统基础架构，主要就是解决数据存储和数据分析计算的问题（通过HDFS和MapReduce实现）。分布式就是多个服务器做同样的一件事。

JDK安装

卸载原有JDK

在安装JDK之前，要确保已经卸载CentOS自带的JDK。

查看当前Java：

```
java -version
```

卸载Java：

```
rpm -qa | grep -i java | xargs -n1 rpm -e --nodeps
```

- rpm -qa：查询已安装的所有rpm软件包
- grep -i：查找。参数-i：忽略大小写
- xargs：表示每次只传递一个参数
- rpm -e --nodeps：强制卸载软件

安装JDK

将jdk上传至新建的目录 `/opt/software/` 目录下，并将jar解压至 `/usr/local/src/` 目录下。

配置环境变量

全局生效： `/etc/profile.d/java.sh`

> 在 `/etc/profile` 文件中自带了一些命令。这些命令可以使系统启动的时候自动执行 `/etc/profile.d` 文件夹中的所有sh脚本，因此只需要在 /etc/profile.d 这个文件夹中新增java.sh脚本即可。

只针对root账户生效： `/root/.bash_profile`

使用vim打开以上文件，进行编辑，在下面添加以下代码。

```
export JAVA_HOME=/usr/local/src/jdk  (jdk安装目录)
export PATH=$PATH:$JAVA_HOME/bin
```

- `$`：使用变量
- `:`：拼接

配置完成，重启系统或使用source命令重新加载文件使其生效：

```
source /etc/profile
或
source /root/.bash_profile
```

查看版本

输入 `java -version` 查看jdk版本。

Hadoop安装

安装Hadoop

将Hadoop上传至新建的目录 `/opt/software/` 目录下，并解压至 `/usr/local/src/` 目录下。

配置环境变量

全局生效： `/etc/profile.d/hadoop.sh`

只针对root账户生效： `/root/.bash_profile`

```
export HADOOP_HOME=/usr/local/src/hadoop  (hadoop安装目录)
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

配置完成，重启系统或使用source命令重新加载文件使其生效：

```
source /etc/profile
或
source /root/.bash_profile
```

查看版本

输入 `hadoop version` 查看hadoop版本。

Hadoop配置

修改配置文件

目录：

```
$HADOOP_HOME/etc/hadoop/
```

0. 配置 `hadoop-env.sh` & `yarn-env.sh`

（env就是环境变量的意思）

这两个一般可以不改，如果要改只需要改Jdk的路径即可。

1. 配置 `core-site.xml`

```
<!-- 配置hadoop中Namenode的地址，master为hostname主机名 -->
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
</property>

<!-- 该路径自己指定，一般设置为hadoop下的data/tmp，该目录会自动创建，标识hadoop运行时产生的缓存文件
-->
<property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/data/hadoop</value>
</property>
```

2. 配置 `hdfs-site.xml`

```
<!-- 配置副本数量 -->
<property>
    <name>dfs.replication</name>
    <value>3</value>
</property>
```

3. 配置 `mapred-site.xml` （先cp复制一个template修改）

```
cp mapred-site.xml.template mapred-site.xml
```

```
<!-- 指定MapReduce运行时框架，这里指定在yarn上，默认是local -->
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
```

4. 配置 `yarn-site.xml`

```
<!-- 配置yarn资源调度时的机器，设置为本机 -->
<property>
    <name>yarn.resouremanager.hostname</name>
    <value>master</value>
</property>
<!-- 配置辅助服务 -->
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
```

关闭yarn的内存检测

```
<!-- 关闭物理内存检测 -->
<property>
    <name>yarn.nodemanager.pmem-check-enabled</name>
    <value>false</value>
</property>
<!-- 关闭虚拟内存检测 -->
<property>
    <name>yarn.nodemanager.vmem-check-enabled</name>
    <value>false</value>
</property>
```

5. 配置 `slaves`

```
<!-- 将集群主机都填入 -->
master
slave1
slave2
```

拷贝至其他集群

将安装好的Hadoop与环境变量拷贝至其他的集群。

格式化Hadoop

初次启动HDFS集群时，必须对主节点(namenode)进行格式化处理，格式化文件系统指令如下：（格式化之前确保namenode和datanode进程结束）

```
hdfs namenode -format
```

格式化完毕能够看到 has been successful XXXX。

启动Hadoop

`start-all.sh`

关闭Hadoop

`stop-all.sh`

查看进程

输入 `jps` 用于显示当前系统的java进程情况及进程id。

master机器上后会有5个进程：

- Namenode
- Datanode
- Resourcemanager
- Nodemanager
- Secondarynamenode

slave机器上一般只能看到2个：

- Datanode
- Nodemanager

进入WebUI

安装配置完毕后，在实体机的浏览器地址栏输入 `虚拟机IP:50070` 即可进入Haddop的WebUI。

以及 `虚拟机IP:8088`

`IP:18080`

Hadoop使用

HDFS文件系统

> - **HDFS** 是 Hadoop 下的分布式文件系统，具有高容错、高吞吐量等特性，可以部署在低成本的硬件上。是 Hadoop 核心组件之一，作为最底层的分布式存储服务而存在。
> - 分布式文件系统解决的问题就是大数据存储。它们是横跨在多台计算机上的存储系统。
> - HDFS使用Master和Slave结构对集群进行管理。一般一个 HDFS 集群只有一个 Namenode 和一定数目的Datanode 组成。Namenode 是 HDFS 集群主节点，Datanode 是 HDFS 集群从节点，两种角色各司其职，共同协调完成分布式的文件存储服务。

将文件上传至hdfs文件系统

```
hdfs dfs -put 本地文件路径 目标HDFS路径/文件名
```

将hdfs中文件取出到本地

```
hdfs dfs -get hdfs文件目录 目标本地路径/文件名
```

删除hdfs中的文件/目录

```
hdfs dfs -rm -rf hdfs目录文件
```

查看hdfs中的目录内容

```
hdfs dfs -ls hdfs目录
```

查看hdfs中文件内容

```
hdfs dfs -cat hdfs目标文件目录
```

案例-WordCount

1. 在Centos下新建文本文件，放入一系列单词，使用空格分开
2. 将该文件上传至HDFS文件系统。
3. 使用Hadoop运行maperduce程序的命令格式：

```
hadoop jar 目标程序jar所在目录 jar下的程序名 目标文件目录 输出目录
```

例：

```
hadoop jar /usr/local/src/hadoop-2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar wordcount /myword.txt /out
```

报错案例

退出安全模式

如果在执行 Hadoop 时出现带 **safe mode** 关键词的错误，则说明 Hadoop 进入了安全模式，输入以下命令退出安全模式。

```
hdfs dfsadmin -safemode leave
```

或者在 `hdfs-site.xml` 中永久关闭安全模式：

找到 **dfs.safemode** 相关的配置，将value修改为1，即可永久关闭安全模式。

```
<property>
  <name>dfs.safemode.threshold.pct</name>
  <value>0.999f</value>
  <description>
    Specifies the percentage of blocks that should satisfy
    the minimal replication requirement defined by dfs.replication.min.
    Values less than or equal to 0 mean not to wait for any particular
    percentage of blocks before exiting safemode.
    Values greater than 1 will make safe mode permanent.
  </description>
</property>
```

查看安全模式状态：

```
hdfs dfsadmin -safemode get
```