

一组分享

Hadoop集群安装和搭建

Hadoop 是一个开源的分布式计算和存储框架，主要用于处理海量数据。

它的核心包含两大组件：

- HDFS（分布式文件系统）：将大文件分割成小块，分布式存储在多个节点上，提供高容错性和高吞吐量的数据访问
- YARN（资源管理器）：负责集群资源（CPU、内存等）的管理和任务调度，协调计算任务在集群中高效运行

Hadoop 适合处理 PB 级甚至更大规模的数据，支持分布式并行计算，广泛应用于大数据存储、数据分析、日志处理等场景，是大数据技术生态的基础框架之一。

1.修改主机名：

```
1 hostnamectl set-hostname Hadoop01
2 hostnamectl set-hostname Hadoop02
3 hostnamectl set-hostname Hadoop03
```

2.配置hosts文件和ssh免密登录

```
1 vim /etc/hosts
2 10.0.0.139 hadoop01
3 10.0.0.140 hadoop02
4 10.0.0.142 hadoop03
5
6 每台机器都执行确保都能免密登录
7 ssh-keygen
8 ssh-copy-id hadoop01
9 ssh-copy-id hadoop02
10 ssh-copy-id hadoop03
```

3.配置目录

```
1 [root@Hadoop01 ~]#mkdir -p /export/data
2 [root@Hadoop01 ~]#mkdir -p /export/servers
3 [root@Hadoop01 ~]#mkdir -p /export/software
```

4.准备安装包

hadoop-2.7.4.tar.gz

jdk-8u161-linux-x64.tar.gz

上传至/export/software

5.安装jdk

```
1 [root@Hadoop01 ~]#cd /export/software
2 [root@Hadoop01 /export/software]#tar -zxvf jdk-8u161-linux-x64.tar.gz -C
  /export/servers/
3 [root@Hadoop01 /export/software]#cd /export/servers
4 [root@Hadoop01 /export/servers]#mv jdk1.8.0_161 jdk
5
6 配置环境变量
7 [root@Hadoop01 /export/servers]#vim /etc/profile
8 export JAVA_HOME=/export/servers/jdk
9 export PATH=$PATH:$JAVA_HOME/bin
10 export CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar
11
12 使环境变量生效
13 [root@Hadoop01 /export/servers]#source /etc/profile
14 检查是否生效
15 [root@Hadoop01 /export/servers]#java -version
16 java version "1.8.0_161"
17 Java(TM) SE Runtime Environment (build 1.8.0_161-b12)
18 Java HotSpot(TM) 64-Bit Server VM (build 25.161-b12, mixed mode)
```

6.安装hadoop

```
1 [root@Hadoop01 /export/servers]#cd /export/software
2 [root@Hadoop01 /export/software]#tar -zxvf hadoop-2.7.4.tar.gz -C
  /export/servers/
3
4 [root@Hadoop01 /export/software]#vim /etc/profile
5 export HADOOP_HOME=/export/servers/hadoop-2.7.4
6 export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
7
8 [root@Hadoop01 /export/software]#source /etc/profile
9 [root@Hadoop01 /export/software]#hadoop version
10 Hadoop 2.7.4
11 Subversion https://shv@git-wip-us.apache.org/repos/asf/hadoop.git -r
  cd915e1e8d9d0131462a0b7301586c175728a282
12 Compiled by kshvachk on 2017-08-01T00:29Z
13 Compiled with protoc 2.5.0
14 From source with checksum 50b0468318b4ce9bd24dc467b7ce1148
15 This command was run using /export/servers/hadoop-
  2.7.4/share/hadoop/common/hadoop-common-2.7.4.jar
```

7.Hadoop集群配置

7.1 进入主节点配置目录

```
1 [root@Hadoop01 /export/software]#cd /export/servers/hadoop-2.7.4/etc/hadoop/
```

7.2 修改hadoop-env.sh文件

```
1 指定 Hadoop 运行时依赖的 Java 环境路径
2 [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#vim hadoop-env.sh
   #找到相应位置，添加这段话
3 export JAVA_HOME=/export/servers/jdk
```

7.3 修改core-site.xml文件

```
1 此配置文件主要用于定义 HDFS 集群的全局属性和通用配置，这里我定义集群的命名服务入口，以及临时文件的存储目录
2 [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#vim core-site.xml
3 <configuration>
4     <!--用于设置Hadoop的文件系统，由URI指定-->
5     <property>
6         <name>fs.defaultFS</name>
7         <!--用于指定namenode地址在hadoop01机器上-->
8         <value>hdfs://hadoop01:9000</value>
9     </property>
10    <!--配置Hadoop的临时目录，默认/tmp/hadoop-`${user.name}`-->
11    <property>
12        <name>hadoop.tmp.dir</name>
13        <value>/export/servers/hadoop-2.7.4/tmp</value>
14    </property>
15 </configuration>
```

7.4 修改hdfs-site.xml文件

```
1 此配置文件主要用于定义 HDFS 集群的具体运行参数，我这里定义了副本数为3，还定义了二级名称节点，为主节点分压
2 [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#vim hdfs-site.xml
3 <configuration>
4     <!--指定HDFS的数量-->
5     <property>
6         <name>dfs.replication</name>
7         <value>3</value>
8     </property>
9     <!--secondary namenode 所在主机的IP和端口-->
10    <property>
11        <name>dfs.namenode.secondary.http-address</name>
12        <value>hadoop02:50090</value>
13    </property>
14 </configuration>
```

7.5 修改mapred-site.xml文件

```
1  此配置文件主要作用是定义 MapReduce 作业的运行方式和相关参数，我这里将 MapReduce 作业的运
   行模式切换到 “YARN 模式”，实现计算资源的集中管理和高效调度
2  [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#cp mapred-
   site.xml.template mapred-site.xml
3  [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#vim mapred-site.xml
4  <configuration>
5      <!--指定MapReduce运行时的框架，这里指定在YARN上，默认在local-->
6      <property>
7          <name>mapreduce.framework.name</name>
8          <value>yarn</value>
9      </property>
10 </configuration>
```

7.7 修改slaves文件

```
1  这个文件的核心作用是指定集群中所有从节点（DataNode 和 NodeManager）的主机名或 IP 地址
2  [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#vim slaves
3  hadoop01
4  hadoop02
5  hadoop03
```

7.8将主节点中配置好的文件和hadoop目录copy给子节点

```
1  [root@Hadoop01 ~]#scp /etc/profile hadoop02:/etc/profile
2  [root@Hadoop01 ~]#scp /etc/profile hadoop03:/etc/profile
3  [root@Hadoop01 ~]#scp -r /export/ hadoop02:/
4  [root@Hadoop01 ~]#scp -r /export/ hadoop03:/
```

7.9 使子节点中的配置环境变量文件生效

```
1  [root@Hadoop02 ~]#source /etc/profile
2  [root@Hadoop03 ~]#source /etc/profile
```

7.10 在主节点格式化文件系统（successfully formatted 格式化成功）

```
1  [root@Hadoop01 ~]#hdfs namenode -format
```

8.启动集群

8.1在主节点启动所有HDFS服务进程，确保集群能进行文件的分布式存储和管理

```
1  在此之前一定要配置免密登录
2  [root@Hadoop01 ~]#start-dfs.sh
```

8.2在主节点启动所有YARN服务进程，确保集群能调度资源并运行计算任务

```
1  [root@Hadoop01 ~]#start-yarn.sh
```

8.3 使用jps命令查看进程

```
1 [root@Hadoop01 /export/servers/hadoop-2.7.4/etc/hadoop]#jps
2 58194 ResourceManager # YARN 主节点进程：负责集群资源（CPU、内存）的管理和调度
3 59355 Jps # jps 命令自身的进程（无实际业务意义）
4 57820 NameNode # HDFS 主节点进程：管理文件系统元数据，协调 DataNode 工作
5 58333 NodeManager # YARN 从节点进程：管理本节点资源，监控容器运行状态
6 57982 DataNode # HDFS 从节点进程：实际存储数据块，执行数据读写操作
7
8 [root@Hadoop02 /export/servers]#jps
9 5106 NodeManager
10 5218 Jps
11 4996 SecondaryNameNode # HDFS 辅助进程：定期合并 NameNode 的元数据（FSImage 和
EditLog），减轻 NameNode 负担
12 4858 DataNode
13
14 [root@Hadoop03 /export/software]#jps
15 57505 DataNode
16 57748 Jps
17 57645 NodeManager
```

9 windows访问hadoop集群

9.1添加hosts解析

修改C:\Windows\System32\drivers\etc打开hosts文件添加以下内容

```
1 10.0.0.139 hadoop01
2 10.0.0.140 hadoop02
3 10.0.0.142 hadoop03
```

9.2浏览器查看集群状态

访问<http://hadoop01:8088>



Cluster

- About
- Nodes
- Node Labels
- Applications
 - NEW
 - NEW SAVING
 - SUBMITTED
 - ACCEPTED
 - RUNNING
 - FINISHED
 - FAILED
 - KILLED
- Scheduler

Tools

Nodes of the cluster

Logged in as: dr.who

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	24 GB	0 B	0	24	0	3	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024,VCores:1>	<memory:8192,VCores:8>

Show 20 entries

Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Mem Used	Mem Avail	VCores Used	VCores Avail	Version
/default-rack		RUNNING	hadoop02:44287	hadoop02:8042	Sun Jul 27 18:10:44 +0800 2025		0	0 B	8 GB	0	8	2.7.4
/default-rack		RUNNING	hadoop01:35335	hadoop01:8042	Sun Jul 27 18:10:45 +0800 2025		0	0 B	8 GB	0	8	2.7.4
/default-rack		RUNNING	hadoop03:38175	hadoop03:8042	Sun Jul 27 18:10:46 +0800 2025		0	0 B	8 GB	0	8	2.7.4

Showing 1 to 3 of 3 entries

First Previous 1 Next Last