目录

[1. 实验一：hadoop集群运行 2](#_Toc22631)

[1.1. 实验目的 2](#_Toc28230)

[1.2. 实验要求 2](#_Toc20548)

[1.3. 实验环境 2](#_Toc8106)

[1.4. 实验过程 2](#_Toc28868)

[1.4.1. 实验任务一：配置Hadoop格式化 2](#_Toc8512)

[1.4.2. 实验任务二：查看Java进程 3](#_Toc16223)

[1.4.3. 实验任务三：查看HDFS的报告 4](#_Toc17789)

[1.4.4. 实验任务四：使用浏览器查看节点状态 5](#_Toc1062)

[1.4.5. 实验任务五：停止Hadoop 8](#_Toc15826)

# 实验一：hadoop集群运行

## 实验目的

完成本实验，您应该能够：

* 掌握hadoop的运行状态
* 掌握hadoop文件系统格式化配置
* 掌握hadoop java运行状态查看
* 掌握hadoop hdfs报告查看
* 掌握hadoop节点状态查看
* 掌握停止hadoop进程操作

## 实验要求

* 熟悉如何查看hadoop的运行状态
* 熟悉停止hadoop进程的操作

## 实验环境

本实验所需之主要资源环境如表1-1所示。

表1-1 资源环境

|  |  |
| --- | --- |
| **服务器集群** | 3个以上节点，节点间网络互通，各节点最低配置：双核CPU、8GB内存、100G硬盘 |
| **运行环境** | CentOS 7.4 |
| **大数据平台** | H3C教学与实践管理系统 |
| **服务和组件** | 完成前面章节的实验，其他服务及组件根据实验需求安装 |

## 实验过程

### 实验任务一：配置Hadoop格式化

#### 步骤一：NameNode格式化

将NameNode上的数据清零，第一次启动HDFS时要进行格式化，以后启动无需再格式化，否则会缺失DataNode进程。另外，只要运行过HDFS，Hadoop的工作目录（本书设置为/usr/local/src/hadoop/tmp）就会有数据，如果需要重新格式化，则在格式化之前一定要先删除工作目录下的数据，否则格式化时会出问题。

执行如下命令，格式化NameNode

注意:格式化只需执行一次,请勿再次格式化!!!

[root@master ~]# su - hadoop

[hadoop@master ~]# cd /usr/local/src/hadoop/

#格式化需注意 – 符号, 必须为英文 – 符号,若为中文 – 符号,namenode启动时进程不会显示

[hadoop@master hadoop]$ bin/hdfs namenode -format

结果：出现status 0 则格式化成功 ,

**20/08/26 16:16:57 INFO util.ExitUtil: Exiting with status 0**

**20/08/26 16:16:57 INFO namenode.NameNode: SHUTDOWN\_MSG:**

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**SHUTDOWN\_MSG: Shutting down NameNode at master/192.168.110.44**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**==================================================================**

**注意：如果status显示不正常.请执行以下步骤.重新格式化namenode**

删除/usr/local/src/hadoop/dfs下的data和name文件夹

[hadoop@master hadoop]$ rm -rf /usr/local/src/hadoop/dfs/\*

删除/usr/local/src/hadoop/tmp下的namesecondary文件夹

[hadoop@master hadoop]$ rm -rf /usr/local/src/hadoop/tmp/\*

注意：删除完master节点的数据后，还需删除slave1和slave2节点的数据

[hadoop@slave1 hadoop]$ rm -rf /usr/local/src/hadoop/dfs/\*

[hadoop@slave1 hadoop]$ rm -rf /usr/local/src/hadoop/tmp/\*

[hadoop@slave2 hadoop]$ rm -rf /usr/local/src/hadoop/dfs/\*

[hadoop@slave2 hadoop]$ rm -rf /usr/local/src/hadoop/tmp/\*

再次执行namenode格式化

[hadoop@master hadoop]$ bin/hdfs namenode -format

**==================================================================**

#### 步骤二：启动NameNode

执行如下命令，启动NameNode：

[hadoop@master hadoop]$ hadoop-daemon.sh start namenode

**starting namenode, logging to /opt/module/hadoop-2.7.1/logs/hadoop-hadoop-namenode-master.out**

### 实验任务二：查看Java进程

启动完成后，可以使用JPS命令查看是否成功。JPS命令是Java提供的一个显示当前所有Java进程pid的命令。

[hadoop@master hadoop]$ jps

**3557 NameNode**

**3624 Jps**

#### 步骤一：slave启动DataNode

执行如下命令，启动DataNode：

#**slave1**

[hadoop@slave1 ~]$ cd /usr/local/src/hadoop/

[hadoop@slave1 hadoop]$ hadoop-daemon.sh start datanode

**starting datanode, logging to /opt/module/hadoop-2.7.1/logs/hadoop-hadoop-datanode-master.out**

#slave2

[hadoop@slave2 ~]$ cd /usr/local/src/hadoop/

[hadoop@slave2 hadoop]$ hadoop-daemon.sh start datanode

**starting datanode, logging to /opt/module/hadoop-2.7.1/logs/hadoop-hadoop-datanode-master.out**

[hadoop@slave1 hadoop]$ jps

**3557 DataNode**

**3725 Jps**

[hadoop@slave2 hadoop]$ jps

**3557 DataNode**

**3725 Jps**

#### 步骤二：启动SecondaryNameNode

执行如下命令，启动SecondaryNameNode：

[hadoop@master hadoop]$ hadoop-daemon.sh start secondarynamenode

**starting secondarynamenode, logging to /opt/module/hadoop-2.7.1/logs/hadoop-hadoop-secondarynamenode-master.out**

[hadoop@master hadoop]$ jps

**34257 NameNode**

**34449 SecondaryNameNode**

**34494 Jps**

查看到有NameNode和SecondaryNameNode两个进程，就表明HDFS启动成功。

#### 步骤三：查看HDFS数据存放位置：

执行如下命令，查看Hadoop工作目录：

[hadoop@master hadoop]$ ll dfs/

**总用量 0**

**drwx------ 3 hadoop hadoop 21 8月 14 15:26 data**

**drwxr-xr-x 3 hadoop hadoop 40 8月 14 14:57 name**

**[hadoop@master hadoop]$ ll ./tmp/dfs**

**总用量 0**

**drwxrwxr-x. 3 hadoop hadoop 21 5月 2 16:34 namesecondary**

可以看出HDFS的数据保存在/usr/local/src/hadoop/dfs目录下，NameNode、DataNode和/usr/local/src/hadoop/tmp/目录下,SecondaryNameNode各有一个目录存放数据。

### 实验任务三：查看HDFS的报告

[hadoop@master sbin]$ hdfs dfsadmin -report

**Configured Capacity: 8202977280 (7.64 GB)**

**Present Capacity: 4421812224 (4.12 GB)**

**DFS Remaining: 4046110720 (3.77 GB)**

**DFS Used: 375701504 (358.30 MB)**

**DFS Used%: 8.50%**

**Under replicated blocks: 88**

**Blocks with corrupt replicas: 0**

**Missing blocks: 0**

**-------------------------------------------------**

**Live datanodes (2):**

**Name: 192.168.1.7:50010 (slave1)**

**Hostname: slave1**

**Decommission Status : Normal**

**Configured Capacity: 4101488640 (3.82 GB)**

**DFS Used: 187850752 (179.15 MB)**

**Non DFS Used: 2109939712 (1.97 GB)**

**DFS Remaining: 1803698176 (1.68 GB)**

**DFS Used%: 4.58%**

**DFS Remaining%: 43.98%**

**Configured Cache Capacity: 0 (0 B)**

**Cache Used: 0 (0 B)**

**Cache Remaining: 0 (0 B)**

**Cache Used%: 100.00%**

**Cache Remaining%: 0.00%**

**Xceivers: 1**

**Last contact: Mon May 04 18:32:32 CST 2020**

**Name: 192.168.1.8:50010 (slave2)**

**Hostname: slave2**

**Decommission Status : Normal**

**Configured Capacity: 4101488640 (3.82 GB)**

**DFS Used: 187850752 (179.15 MB)**

**Non DFS Used: 1671225344 (1.56 GB)**

**DFS Remaining: 2242412544 (2.09 GB)**

**DFS Used%: 4.58%**

**DFS Remaining%: 54.67%**

**Configured Cache Capacity: 0 (0 B)**

**Cache Used: 0 (0 B)**

**Cache Remaining: 0 (0 B)**

**Cache Used%: 100.00%**

**Cache Remaining%: 0.00%**

**Xceivers: 1**

**Last contact: Mon May 04 18:32:32 CST 2020**

### 实验任务四：使用浏览器查看节点状态

打开虚拟机内的浏览器，在地址栏输入http://master:50070，进入页面可以查看NameNode和DataNode信息，如图5-2所示。

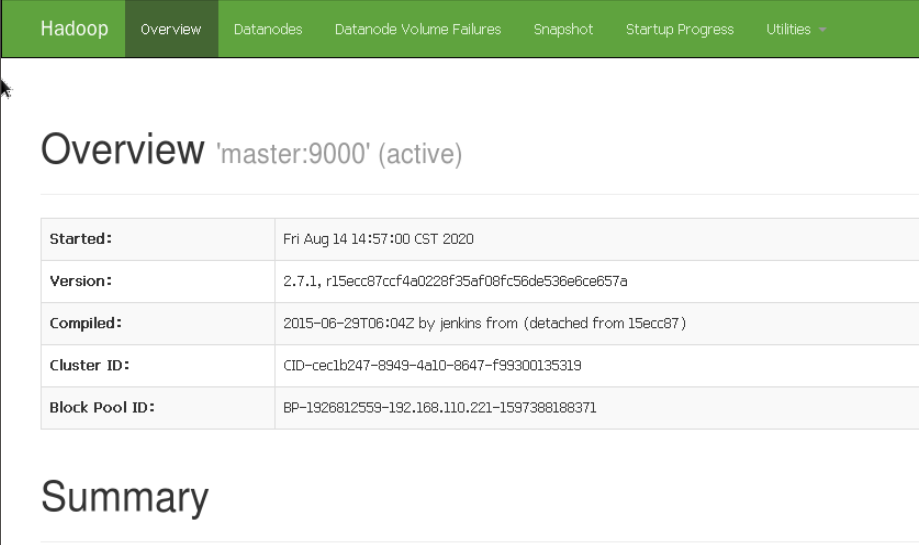


图5-2 通过Web查看HDFS

在浏览器的地址栏输入http://master:50090，进入页面可以查看SecondaryNameNode信息，如图5-3所示。

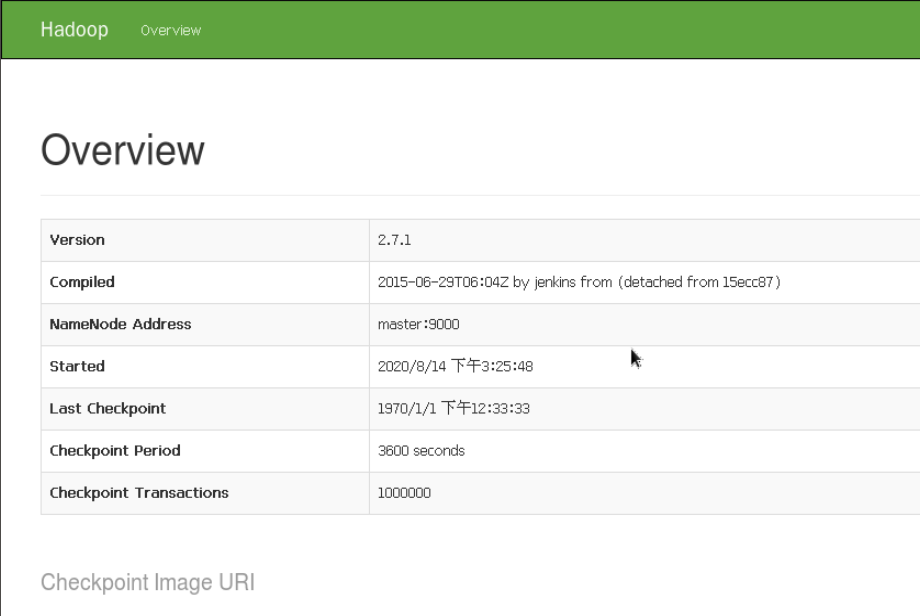


图5-3通过Web查看SecondaryName

可以使用start-dfs.sh命令启动HDFS。第一次连接到slave1和slave2会提示是否继续连接。没配置SSH免密登陆则还需输入slave1和slave2主机密码

[hadoop@master hadoop]$ stop-dfs.sh

[hadoop@master hadoop]$ start-dfs.sh

运行测试：

下面运行WordCount官方案例，统计data.txt文件中单词的出现频度。这个案例可以用来统计年度十大热销产品、年度风云人物、年度最热名词等。

#### 步骤一：在HDFS文件系统中创建数据输入目录

确保dfs和yarn都启动成功

[hadoop@master hadoop]$ start-yarn.sh

[hadoop@master hadoop]$ jps

**34257 NameNode**

**34449 SecondaryNameNode**

**34494 Jps**

**32847 ResourceManager**

如果是第一次运行MapReduce程序，需要先在HDFS文件系统中创建数据输入目录，存放输入数据。这里指定/input目录为输入数据的存放目录。

执行如下命令，在HDFS文件系统中创建/input目录：

[hadoop@master hadoop]$ hdfs dfs -mkdir /input

[hadoop@master hadoop]$ hdfs dfs -ls /

**Found 1 items**

**drwxr-xr-x - hadoop supergroup 0 2020-05-02 22:26 /input**

此处创建的/input目录是在HDFS文件系统中，只能用HDFS命令查看和操作。

#### 步骤二：将输入数据文件复制到HDFS的/input目录中

测试用数据文件仍然是上一节所用的测试数据文件~/input/data.txt，内容如下所示。

[hadoop@master hadoop]$ cat ~/input/data.txt

**Hello World**

**Hello Hadoop**

**Hello Huasan**

执行如下命令，将输入数据文件复制到HDFS的/input目录中：

[hadoop@master hadoop]$ hdfs dfs -put ~/input/data.txt /input

确认文件已复制到HDFS的/input目录：

[hadoop@master hadoop]$ hdfs dfs -ls /input

**Found 1 items**

**-rw-r--r-- 1 hadoop supergroup 38 2020-05-02 22:32 /input/data.txt**

#### 步骤三：运行WordCount案例，计算数据文件中各单词的频度。

运行MapReduce命令需要指定数据输出目录，该目录为HDFS文件系统中的目录，会自动生成。如果在执行MapReduce命令前，该目录已经存在，则执行MapReduce命令会出错。例如MapReduce命令指定数据输出目录为/output，/output目录在HDFS文件系统中已经存在，则执行相应的MapReduce命令就会出错。所以如果不是第一次运行MapReduce，就要先查看HDFS中的文件，是否存在/output目录。如果已经存在/output目录，就要先删除/output目录，再执行上述命令。

自动创建的/output目录在HDFS文件系统中，使用HDFS命令查看和操作。

[hadoop@master hadoop]$ hdfs dfs -mkdir /output

先执行如下命令查看HDFS中的文件：

[hadoop@master hadoop]$ hdfs dfs -ls /

**Found 3 items**

**drwxr-xr-x - hadoop supergroup 0 2020-05-02 22:32 /input**

**drwxr-xr-x - hadoop supergroup 0 2020-05-02 22:49 /output**

上述目录中/input目录是输入数据存放的目录，/output目录是输出数据存放的目录。执行如下命令，删除/output目录。

[hadoop@master hadoop]$ hdfs dfs -rm -r -f /output

**20/05/03 09:43:43 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.**

**Deleted /output**

执行如下命令运行WordCount案例：

[hadoop@master hadoop]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount /input/data.txt /output

MapReduce程序运行过程中的输出信息如下所示：

**20/05/02 22:39:41 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032**

**20/05/02 22:39:43 INFO input.FileInputFormat: Total input paths to process : 1**

**20/05/02 22:39:43 INFO mapreduce.JobSubmitter: number of splits:1**

**20/05/02 22:39:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1588469277215\_0001**

**…… 省略 ……**

**20/05/02 22:40:32 INFO mapreduce.Job: map 0% reduce 0%**

**20/05/02 22:41:07 INFO mapreduce.Job: map 100% reduce 0%**

**20/05/02 22:41:25 INFO mapreduce.Job: map 100% reduce 100%**

**20/05/02 22:41:27 INFO mapreduce.Job: Job job\_1588469277215\_0001 completed successfully**

**…… 省略 ……**

由上述信息可知MapReduce程序提交了一个作业，作业先进行Map，再进行Reduce操作。

MapReduce作业运行过程也可以在YARN集群网页中查看。在浏览器的地址栏输入：http://master:8088，页面如图5-4所示，可以看到MapReduce程序刚刚完成了一个作业。

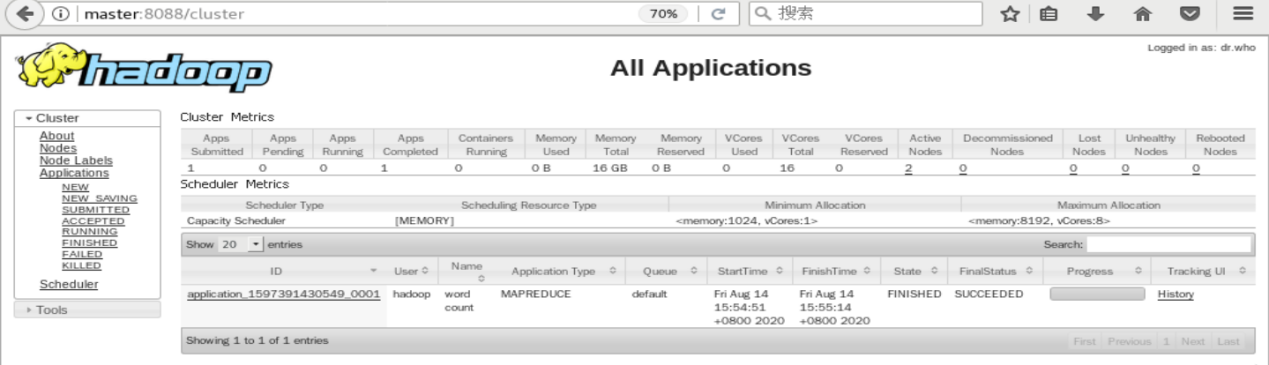


图5-4 通过Web查看作业

除了可以用HDFS命令查看HDFS文件系统中的内容，也可使用网页查看HDFS文件系统。在浏览器的地址栏输入http://master:50070，进入页面，在Utilities菜单中选择Browse the file system，可以查看HDFS文件系统内容。如图5-5所示，查看HDFS的根目录，可以看到HDFS根目录中有三个目录，input、output和tmp。

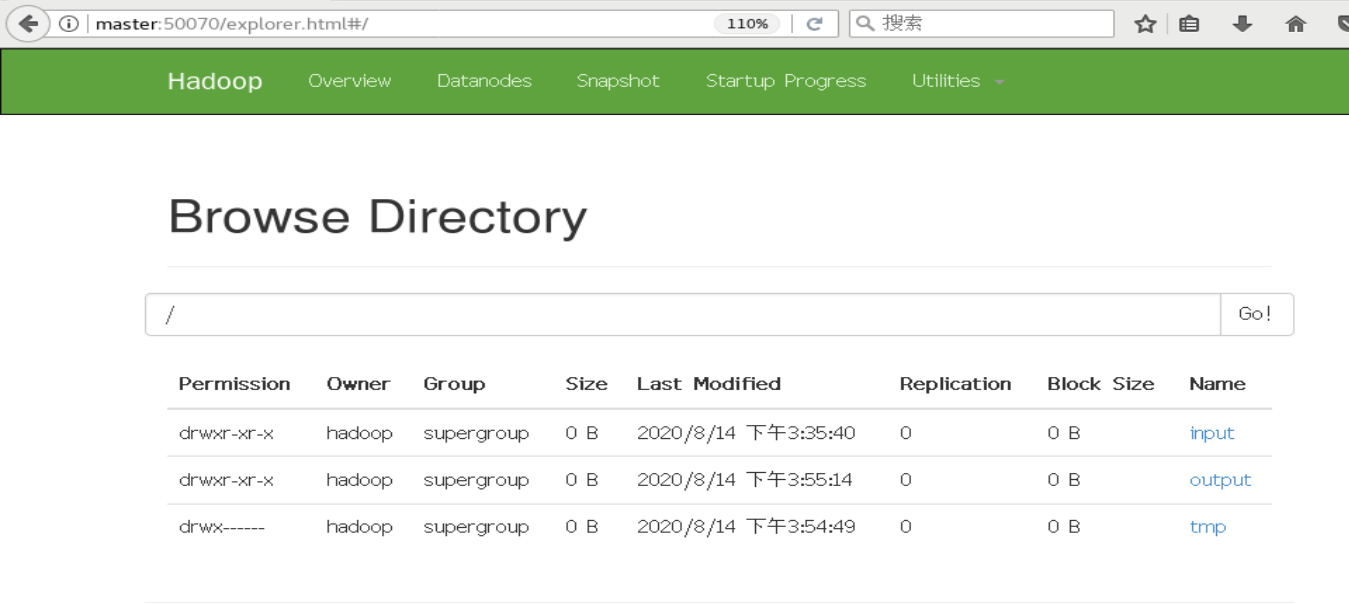


图5-5 通过Web查看HDFS

查看output目录，如图5-6所示，发现有两个文件。文件\_SUCCESS表示处理成功，处理的结果存放在part-r-00000文件中。在页面上不能直接查看文件内容，需要下载到本地系统才行。

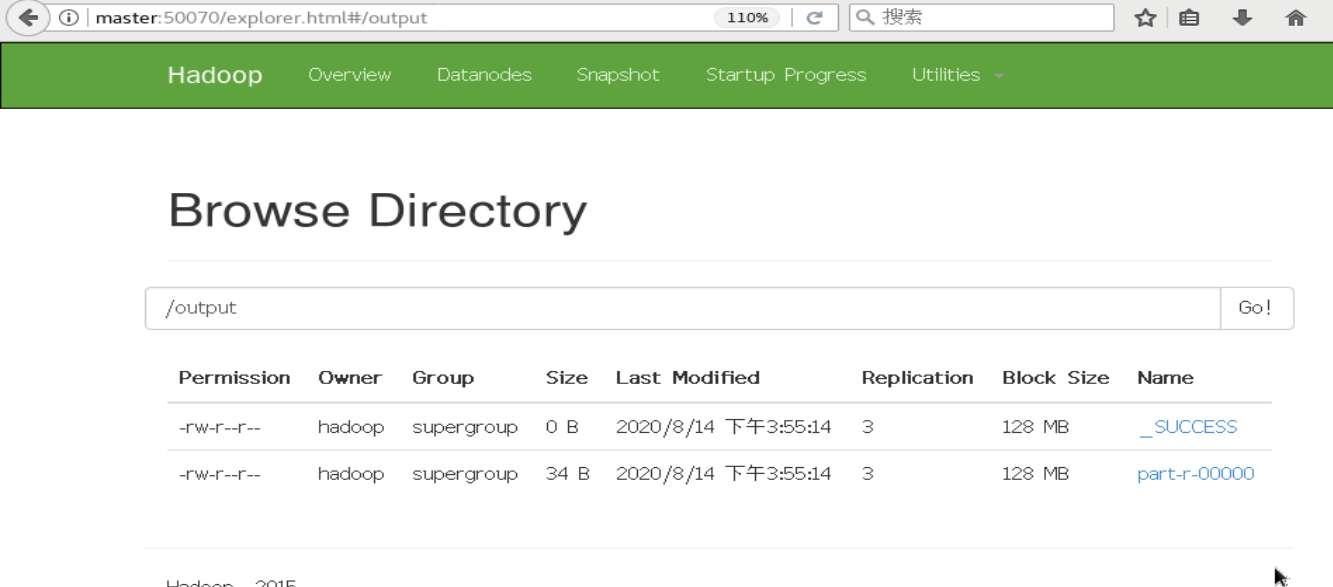


图5-6 通过Web查看输出结果

可以使用HDFS命令直接查看part-r-00000文件内容，结果如下所示：

[hadoop@master hadoop]$ hdfs dfs -cat /output/part-r-00000

**Hadoop 1**

**Hello 3**

**Huasan 1**

**World 1**

可以看出统计结果正确，说明Hadoop运行正常。

### 实验任务五：停止Hadoop

#### 步骤一：使用stop-all.sh停止hadoop服务

stop-all命令停止与hadoop有关的NameNode ，SecondaryNameNode，DataNode，NodeManager与ResourceManager进程，既是实现stop-dfs.sh与stop-yarn.sh功能

[hadoop@master hadoop]$ stop-all.sh

[hadoop@master hadoop]$ jps

**81557 Jps**

[hadoop@slave1 hadoop]$ jps

**111994 Jps**

[hadoop@slave2 hadoop]$ jps

**110644 Jps**

#### 步骤二：使用start-all启动hadoop服务

与stop-all相反，start-all启动所有与hadoop有关的服务

[hadoop@master hadoop]$ start-all.sh