

# 第 4 章 Hadoop 命令快速上手

## ★本章导读★

本章将对 Hadoop 集群有更深层次的了解，内容涉及 Hadoop 集群的两大核心功能：分布式存储以及并行计算引擎。通过对集群的存储资源和计算资源的了解，加深对 Hadoop 集群的数据存储以及数据计算处理调配知识的理解。最后帮助读者快速上手熟练掌握常用的 Hadoop 基本命令。

## ★知识要点★

通过本章内容的学习，读者将掌握以下知识：

- Hadoop 集群基本信息的查询
- Hadoop 集群计算资源的协同调配
- 能熟练使用 HDFS 常用命令

## 4.1 Hadoop 集群信息查询

Hadoop 集群最重要的两大核心功能：分布式存储和并行计算引擎。首先需要更深入了解 Hadoop 集群的存储资源和计算资源情况，才有助于接下来实现 MapReduce 程序的开发以及向集群提交计算任务。本小节主要从 Hadoop 集群存储系统和计算资源两大方面进行介绍。

### 4.1.1 集群节点信息查询

在上一章节中，启动 Hadoop 集群之后可以通过 Web 端访问 Hadoop 集群，默认是通过主节点（NameNode）的 50070 端口来访问，例如本书主节点 IP 地址为 192.168.75.100，主机名为 master，则在浏览器中输入 <http://192.168.75.100:50070> 或 <http://master:50070> 来访问。接下来详细介绍 Web 端所展示的集群节点等详细信息及其他信息的含义。

Hadoop 集群 Web 端完整页面如图 4-1 所示。

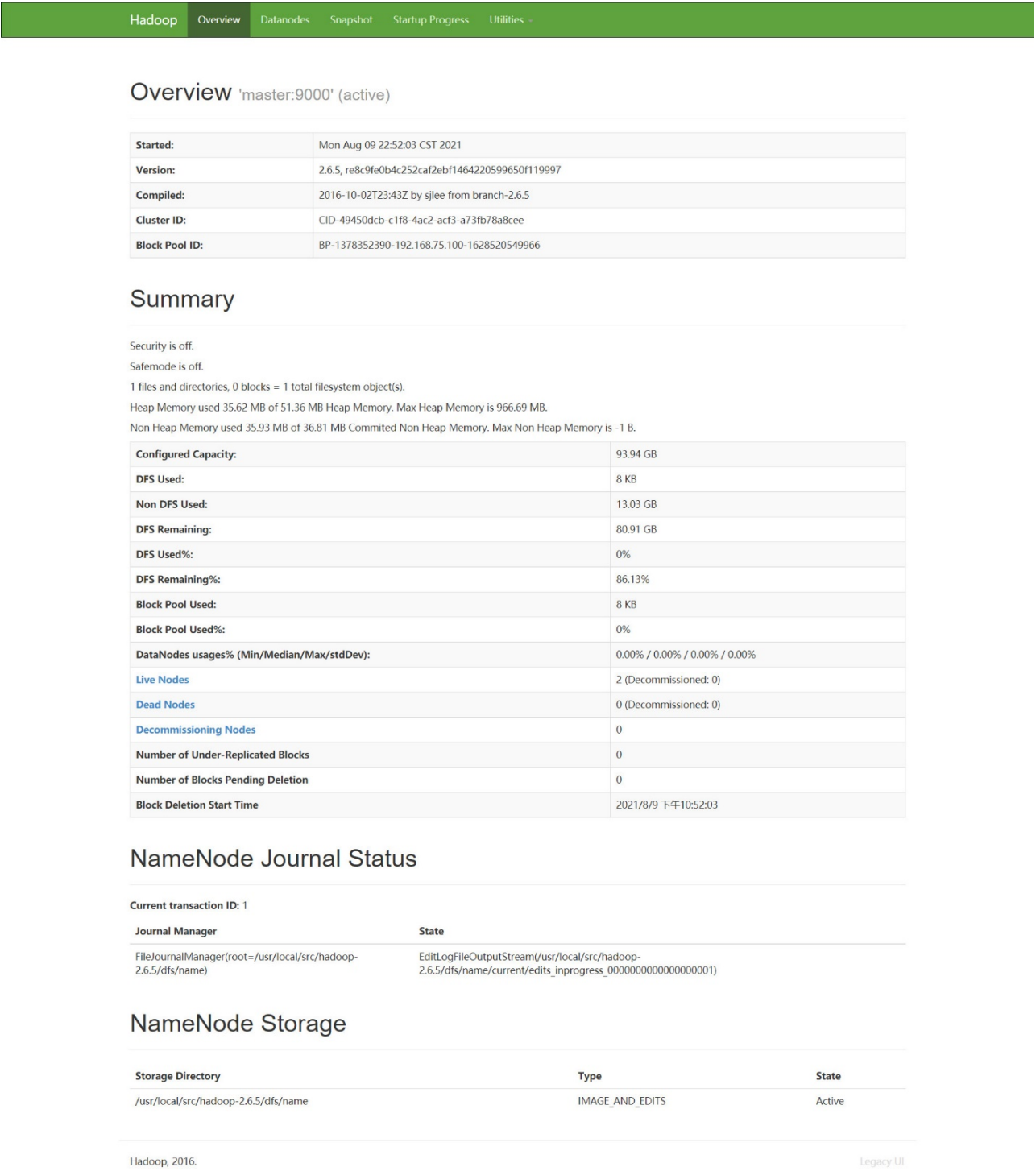


表 4-1 Hadoop 集群 Web 页面参数

名称	含义
Overview	NameNode 的启动时间、版本号、编译版本等基本信息。
Summary	提供当前集群环境的信息，如 DataNode 节点的基本存储信息、DataNode 的信息、DataNode 的活跃状态等。
NameNode Storage	提供 NameNode 的信息，最后的 State 表示此节点为活动节点，可以正常提供服务。

在图 4-1 中，几个比较重要的参数信息如下。

1. **Configured Capacity:** 代表一阶配置的文件系统存储总容量为 93.94GB。
2. **DFS Remaining:** 代表可使用的存储容量为 80.91GB。
3. **DFS Used:** 代表已经使用的存储容量为 8KB。
4. **Non DFS Used:** 代表被非 DFS 的应用所占用的存储容量为 13.03GB，例如本地目录中的 Linux 文件。

当然，Hadoop 集群存储相关信息即可以通过 HDFS 监控页面进行查询，也可以通过命令的方式进行查询，在 Linux 终端输入如下命令即可查询。

```
hadoop dfsadmin -report
```

查询结果如下所示。

```
[root@master ~]# hadoop dfsadmin -report
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

```
Configured Capacity: 100865679360 (93.94 GB)
Present Capacity: 86872088576 (80.91 GB)
DFS Remaining: 86872080384 (80.91 GB)
DFS Used: 8192 (8 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
```

```
-----
Live datanodes (2):
```

```
Name: 192.168.75.101:50010 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 50432839680 (46.97 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 7035662336 (6.55 GB)
DFS Remaining: 43397173248 (40.42 GB)
DFS Used%: 0.00%
DFS Remaining%: 86.05%
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: Wed Aug 25 23:00:02 CST 2021

Name: 192.168.75.102:50010 (slave2)
Hostname: slave2
Decommission Status : Normal
Configured Capacity: 50432839680 (46.97 GB)
DFS Used: 4096 (4 KB)
Non DFS Used: 6958071808 (6.48 GB)
DFS Remaining: 43474763776 (40.49 GB)
DFS Used%: 0.00%
DFS Remaining%: 86.20%
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: Wed Aug 25 23:00:02 CST 2021

```

在上述代码输出信息中，存储使用情况等信息详情如下。

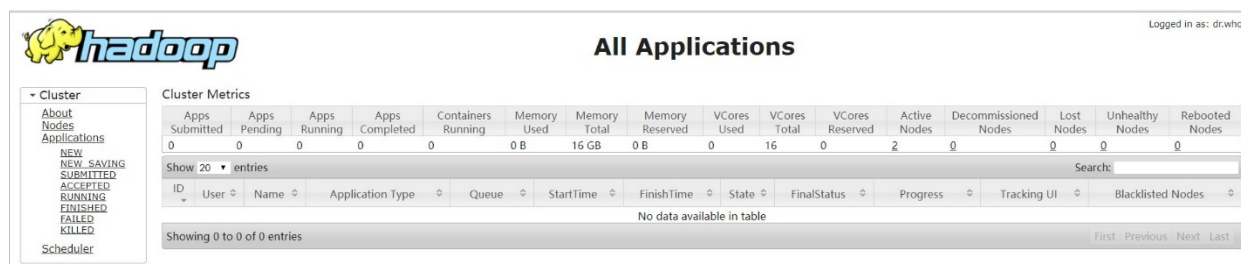
1. 文件系统的总容量为 93.94GB。
2. 已经使用的存储容量为 8KB。
3. 可使用的存储容量为 80.91GB。
4. 各 DataNode 的可使用存储容量及所占比例，本书中 DataNode 为 slave1 和 slave2。

通过上述信息，可以方便的帮助我们了解当前 Hadoop 集群的存储系统情况。例如发现有宕机节点（Dead Nodes）或者其他异常情况，便于我们采取相应的措施。

### 4.1.3 集群计算资源信息查询

Hadoop 集群的计算资源，也就是前面章节提到的资源管理器 ResourceManager，Hadoop 集群的资源均是通过资源管理器进行统一管理的。通过 ResourceManager 的监控服务，可以方便地查询当前集群上的计算资源，方便做出调整。资源管理器的 Web 端默认是 ResourceManager 所在的节点机器的 8088 端口访问的。

本书集群 Web 地址为 “<http://master/cluster/nodes>” 或 “<http://192.168.75.100:8088/cluster/nodes>”



The screenshot shows the Hadoop Yarn Application page. At the top, there's a header with the Hadoop logo and the title "All Applications". Below the header, there's a sidebar with navigation links: "Cluster", "About", "Nodes", "Applications", "NEW", "NEW SAVING", "SUBMITTED", "ACCEPTED", "RUNNING", "FINISHED", "FAILED", "KILLED", and "Scheduler". The main content area displays "Cluster Metrics" with a table showing various metrics: Apps Submitted (0), Apps Pending (0), Apps Running (0), Apps Completed (0), Containers Running (0), Memory Used (0 B), Memory Total (16 GB), Memory Reserved (0 B), VCoers Used (0), VCoers Total (16), VCoers Reserved (0), Active Nodes (2), Decommissioned Nodes (0), Lost Nodes (0), Unhealthy Nodes (0), and Rebooted Nodes (0). Below the metrics table, there's a search bar and a table header with columns: ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, FinalStatus, Progress, Tracking UI, and Blacklisted Nodes. The table currently shows "Showing 0 to 0 of 0 entries" and "No data available in table".

图 4-2 Yarn Application 页面

图 4-2 中展示的当前集群的计算资源信息如下所示。

1. **Active Nodes:** 在线的计算节点，本书集群有 2 个。
2. **Memory Total:** 可使用的内存容量，本书集群中为 16GB。
3. **Vcores Total:** 可使用的 CPU 核数，本书集群中有 16 个。
4. **Containers Running:** 执行计算任务的容量数量，无任务时其值为 0。
5. **Memory Used:** 实际内存使用数量，本书集群此时为 0B。

根据上述信息，可以了解当前集群的计算资源及其使用情况，包括集群在线的节点、可使用的 CPU 核数以及内存大小。

## 4.2 HDFS 操作灵活应用

本小节中，我们将介绍如何使用命令操作 HDFS，例如在 HDFS 上创建文件夹、查看 HDFS 目录存储文件以及上传/下载文件等等。通过命令操作 HDFS，加深对 Hadoop 存储系统 HDFS 的了解。

### 4.2.1 HDFS 目录操作命令

#### 1. HDFS 创建单层目录

创建目录可使用“`hdfs dfs -mkdir <path>`”的命令进行创建，在集群服务器终端，直接输出“`hdfs dfs`”，然后回车，就可以看到“`hdfs dfs`”相关的命令帮助。接下来我们通过命令在 HDFS 上创建 `/user/data` 的目录，具体实现命令如代码下所示。

```
[root@master ~]# hdfs dfs -mkdir /user
```

执行该命令之后，在 HDFS 文件目录系统中可以看到我们所创建的 `user` 文件夹，如图 4-3 所示。

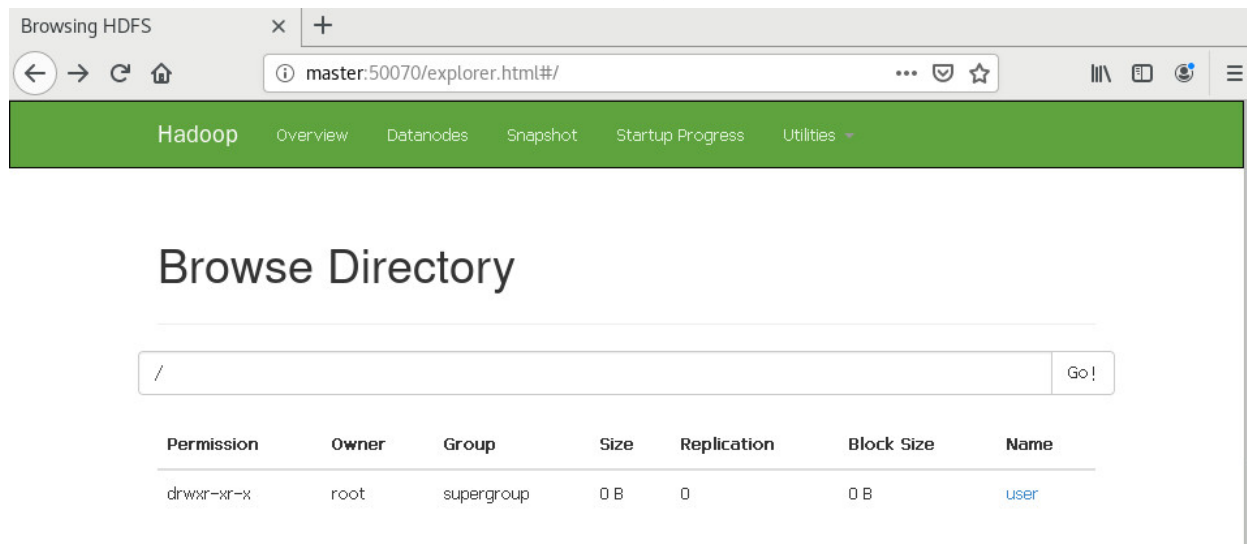


图 4-3 HDFS 目录页面

#### 2. HDFS 创建层级目录

“`hdfs dfs -mkdir <path>`”只能逐级创建目录，即如果父目录不存在，则该命令会报错，例如执行“`hdfs dfs -mkdir /user/data/demo`”本例中的不存在的父目录即 `/user/data`。此时，需要加上参数 `-p`，可以实现层级目录创建。如果执行“`hdfs dfs -mkdir /user`”则不需要添加参数 `-p`，如下所示。

```
[root@master ~]# hdfs dfs -mkdir -p /user/data/demo
```

执行该命令之后，可以在 HDFS 文件目录系统中看到在 `/user/data` 文件夹下创建的 `demo` 文件夹，如图 4-4 所示。

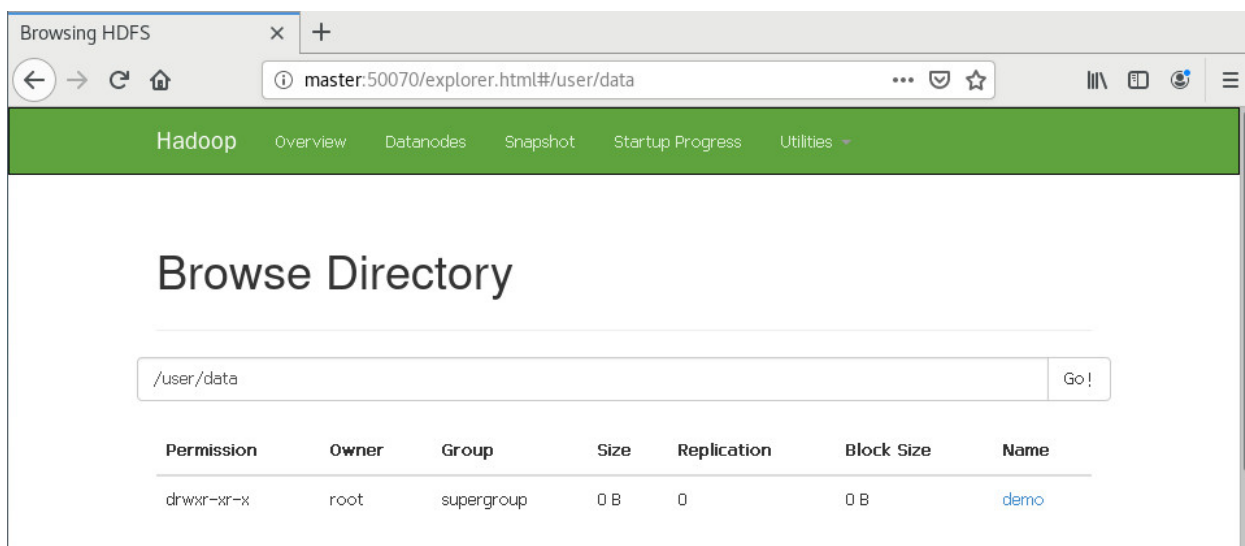


图 4-4 HDFS 层级目录

### 3. 删除 HDFS 目录

执行“`hdfs dfs -rm <path>`”可以删除不存在子目录的文件夹，但当该文件夹下存在文件或者子目录时，可以选择加入“-r”、“-rf”等参数，具体使用方法可以在终端输入“`hdfs dfs -rm`”，然后回车，就可以看到相关参数使用介绍。

接下来，我们通过命令来删除/user/data/demo 文件夹，执行代码如下。

```
[root@master ~]# hdfs dfs -rm -r /user/data/demo
```

执行该命令之后，可以从 HDFS 目录系统中看到，/user/data 下的 demo 文件夹已经不存在，如图 4-5 所示。

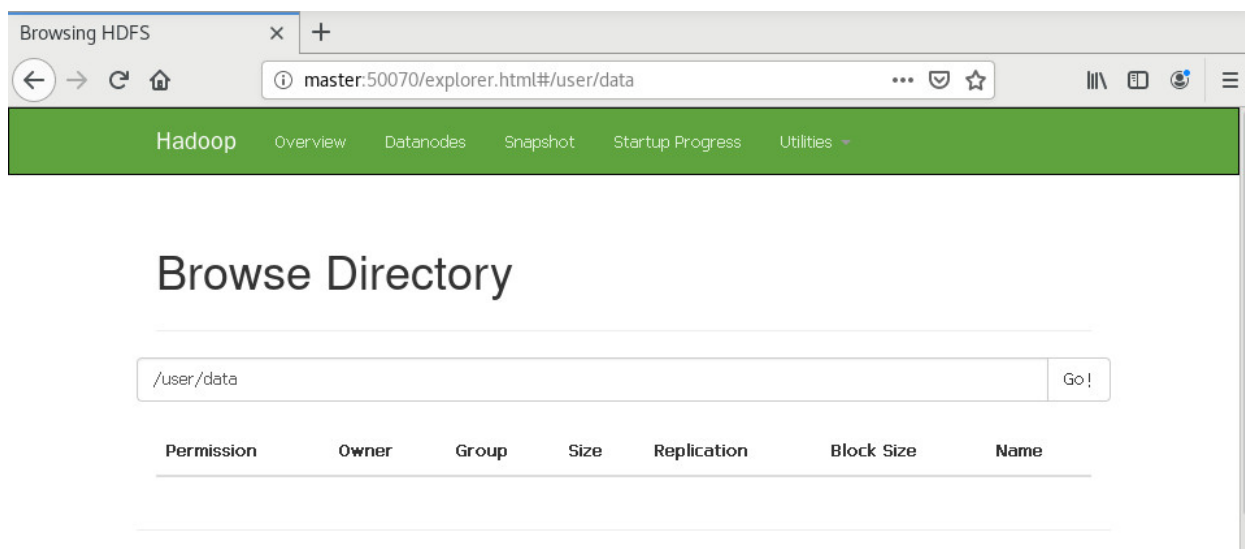


图 4-5 HDFS 删除层级目录结果

## 4.2.2 HDFS 文件操作命令

### 1. 上传文件

在上一小节创建 HDFS 目录之后，我们可以通过命令操作，将本地文件上传至 HDFS。HDFS 上传命令主要有如下几个，如表 4-2 所示。

表 4-2 HDFS 上传文件命令

名称	含义
<code>hdfs dfs [-copyFromLocal [-f] [-p] [-l] &lt;localsrc&gt;...&lt;dst&gt;]</code>	将文件从本地文件系统复制到 HDFS 文件系统，主要参数<localsrc>为本地文件路径，<dst>为 HDFS 目标路径。
<code>hdfs dfs [-moveFromLocal [-f] [-p] [-l] &lt;localsrc&gt;...&lt;dst&gt;]</code>	将文件从本地文件系统移动到 HDFS 文件系统，主要参数<localsrc>为本地文件路径，<dst>为 HDFS 目标路径。
<code>hdfs dfs [-put [-f] [-p] [-l] &lt;localsrc&gt;...&lt;dst&gt;]</code>	将文件从本地文件系统上传到 HDFS 文件系统，主要参数<localsrc>为本地文件路径，<dst>为 HDFS 目标路径。

我们分别使用上述 3 个命令，将本地的 test.txt 文件上传至 HDFS。值得注意的是，第 2 个命令执行完成后，本地文件将会被删除，所以将第 2 个命令放在最后执行，代码如下所示。

```
[root@master Desktop]# cat test.txt
Test for HDFS
copy
move
put

[root@master Desktop]# hdfs dfs -copyFromLocal test.txt /user/data/
[root@master Desktop]# hdfs dfs -put test.txt /user/data/
[root@master Desktop]# hdfs dfs -put test.txt /user/data/test-put.txt
[root@master Desktop]# hdfs dfs -moveFromLocal test.txt /user/data/test-move.txt

[root@master Desktop]# cat test.txt
cat: test.txt: 没有那个文件或目录
```

执行该命令之后，可以在 HDFS 文件目录系统的/user/data 目录下看到我们所上传的文件，如图 4-6 所示。

Hadoop Overview Datanodes Snapshot Startup Progress Utilities						
Browse Directory						
<input type="text" value="/user/data"/>						<input type="button" value="Go !"/>
Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	root	supergroup	28 B	2	128 MB	<a href="#">test-move.txt</a>
-rw-r--r--	root	supergroup	28 B	2	128 MB	<a href="#">test-put.txt</a>
-rw-r--r--	root	supergroup	28 B	2	128 MB	<a href="#">test.txt</a>

图 4-6 HDFS 上传文件结果

## 2. 下载文件

同样，我们可以使用命令的方式从 HDFS 上下载我们所需要的文件。HDFS 下载文件相关的命令如表 4-3 所示。



表 4-3 下载 HDFS 文件命令

名称	含义
<code>hdfs dfs [-copyToLocal [-p] [-ignoreCre] [-crc] &lt;src&gt;...&lt;localsrc&gt;]</code>	将文件从 HDFS 文件系统复制到本地文件系统,主要参数<src>为 HDFS 文件系统路径, <localsrc>为本地文件系统路径。
<code>hdfs dfs [-get [-p] [-ignoreCre] [-crc] &lt;src&gt;...&lt;localsrc&gt;]</code>	获取 HDFS 文件系统上指定路径的文件到文件系统,主要参数<src>为 HDFS 文件系统路径, <localsrc>为本地文件系统路径。

我们分别使用上述 2 个命令,将 HDFS 文件系统/user/data 目录下的 test-put.txt 和 test-move.txt 文件下载到本地/data/hdfs\_demo 路径下,代码如下所示。

```
[root@master Desktop]# hdfs dfs -copyToLocal /user/data/test-put.txt ~/data/hdfs_demo/ test-put.txt
[root@master Desktop]# hdfs dfs -get /user/data/test-move.txt ~/data/hdfs_demo/ test-move.txt

[root@master Desktop]# ls /data/hdfs_demo
test-move.txt test-put.txt
```

### 3. 查看文件

当用户想查看 HDFS 上某个文件时,可以使用“`hdfs dfs -cat <filePath>`”进行查看。HDFS 查看文件相关的命令如表 4-4 所示。

表 4-4 查看 HDFS 文件命令

名称	含义
<code>hdfs dfs [-cat [-ignoreCre] &lt;file&gt;...]</code>	查看 HDFS 指定<file>文件内容。
<code>hdfs dfs [-tail [-ignoreCre] [-f] &lt;file&gt;]</code>	查看 HDFS 指定<file>文件最后 1024 字节内容。

我们分别使用上述 2 个命令,查看 HDFS 文件系统/user/data/test.txt 文件内容,代码如下所示。

```
[root@master hdfs_demo]# hdfs dfs -cat /user/data/test.txt
Test for HDFS
copy
move
put

[root@master hdfs_demo]# hdfs dfs -tail /user/data/test.txt
Test for HDFS
copy
move
```

## 4.2.3 HDSF Web 操作 HDFS

通过 HDFS Web 页面可以进行简单的文件存储信息查看以及下载文件。我们打开 Web 页面,进入到/user/data 目录下,查看 test.txt 文件存储信息,并下载该文件,如图 4-7 所示。





图 4-7 下载 HDFS 文件到本地  
当单击【Download】之后，即可等待下载完成，如图 4-8 所示。

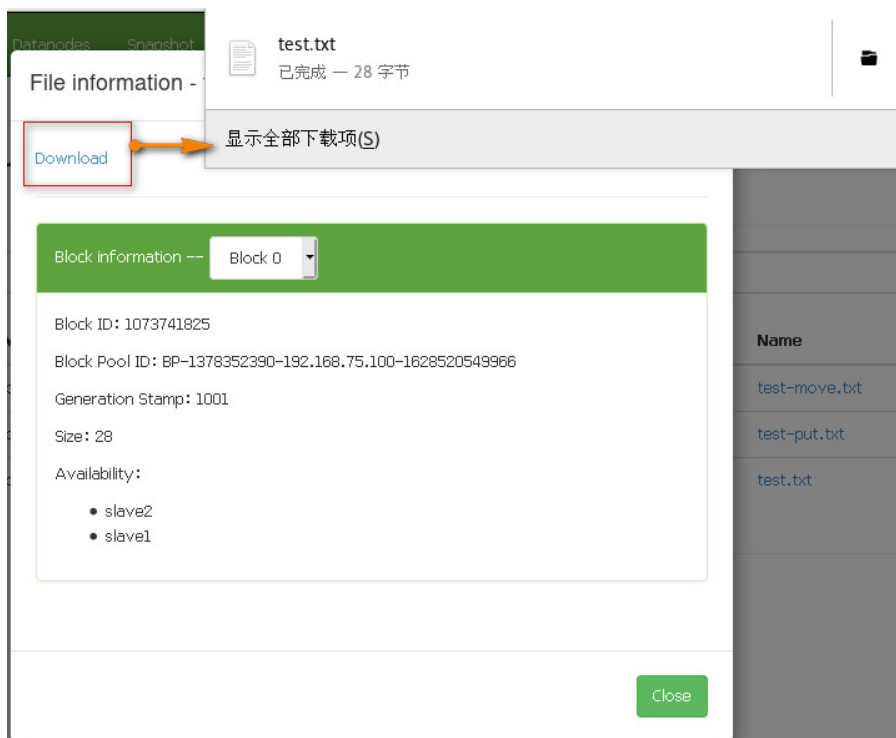


图 4-8 下载完成展示

## ★回顾思考★

### 01 HDFS 上传文件命令有哪些？

答：包括如下：

- (1) `hdfs dfs [-copyFromLocal [-f] [-p] [-l] <localsrc>...<dst>]`
- (2) `hdfs dfs [-moveFromLocal [-f] [-p] [-l] <localsrc>...<dst>]`
- (3) `hdfs dfs [-put [-f] [-p] [-l] <localsrc>...<dst>]`

### 02 HDFS 查看文件命令有哪些？

答：包括如下：

- (1) `hdfs dfs [-cat [-ignoreCrc] <file>...]`
- (2) `hdfs dfs [-tail [-ignoreCrc] [-f] <file>]`

### 03 当上传一个 200MB 的文件至 HDFS 会如何存储？

答：由于 HDFS 集群有 2 个副本节点（默认为 3 个），所以该文件会被存储在 2 个不同的数据节点机器上。又因为 Hadoop 2.x 版本默认设置的 HDFS 文件块大小为 128MB，该文件大小为 200MB，大于 128MB 且小于 256MB，所以该文件会被存储在 2 个文件块 Block 中。

## ★练习★

### 一、选择题

1. 以下不是 HDFS 上传命令参数的选项是（ ）
  - A. -f
  - B. -p
  - C. -s'
  - D. -l''
2. 以下不是 HDFS 下载参数的选项是（ ）
  - A. -p
  - B. -ignoreCrc
  - C. -l
  - D. -crc
3. 上传当前文件夹下 demo.txt 到 HDFS 的 /user/data/ 目录下正确的命令是（ ）
  - A. `hdfs -dfs -put demo.txt /user/data/demo.txt`
  - B. `hdfs dfs -put demo.txt /user/data/demo.txt`
  - C. `hdfs -dfs -copyToLocal demo.txt /user/data/demo.txt`
  - D. `hdfs -dfs copyFromLocal demo.txt /user/data/demo.txt`

### 二、填空题

1. HDFS 命令创建层级文件夹时不需要加入参数 -p。 ( )

2. 使用-r参数可以实现递归删除多层级的目录文件夹。

( )

### 三、实战练习

1. 在本地创建 linux.txt 文件，写入“Hello Hadoop”，并上传至 HDFS 的/user/data 目录下。
2. 使用 IDEA 工具创建 Scala 工程并运行。

## 本章小结

本章详细介绍了 Web 页面展示的关于 Hadoop 集群核心的两大功能：分布式存储系统和计算资源。对所展示的信息主要参数含义进行了详细介绍和实例展示，以便于读者更好的理解 Hadoop 的存储机制和计算资源情况。

本章通过实例的方式讲解了如何通过命令操作 HDFS 文件系统，包括上传、下载、查看文件等。加深读者对 Hadoop 存储系统和计算资源的了解，在接下来提交 MapReduce 任务的时候有更深刻的理解。