# HDFS文件系统实战

## 实验介绍

### 关于本实验

HDFS是Hadoop大数据平台中的分布式文件系统，为上层应用或其他大数据组件提供数据存储，如Hive，Mapreduce，Spark，HBase等。在HDFS shell客户端我们可以实现对分布式文件系统的操作和管理等。掌握HDFS的使用对我们更好的理解和掌握大数据大有裨益

### 实验目的

* 掌握HDFS常用操作。
* 掌握HDFS访问OBS对象存储服务操作。

### 内容描述

* 子实验1进行前置实验，启动Hadoop集群，验证Hadoop启动状态等；
* 子实验2介绍了HDFS文件系统常用操作，例如上传、查看、下载、新建等操作；
* 子实验3介绍了如何将华为公有云对象存储服务OBS集成进鲲鹏Hadoop中

## 实验任务1

### 启动Hadoop集群

在node1节点执行以下命令：

> start-dfs.sh ; start-yarn.sh

返回信息中有以下内容，表示hadoop集群启动成功：

Starting namenodes on [node1]

Starting secondary namenodes [node1]

starting yarn daemons

备注：关闭Hadoop集群的命令为：

> stop-dfs.sh && stop-yarn.sh

### 验证Hadoop状态

使用jps命令在node1-4中查看java进程ll

在node1中可以查看到 NameNode，SecondaryNameNode，ResourceManager

进程，在node2-4中可以查看到 NodeManager 和 Datanode 进程，表示hadoop集群状态正常。

> jps

**1538 WrapperSimpleApp**

**5732 SecondaryNameNode**

**5508 NameNode**

6205 Jps

**5918 ResourceManager**

> jps

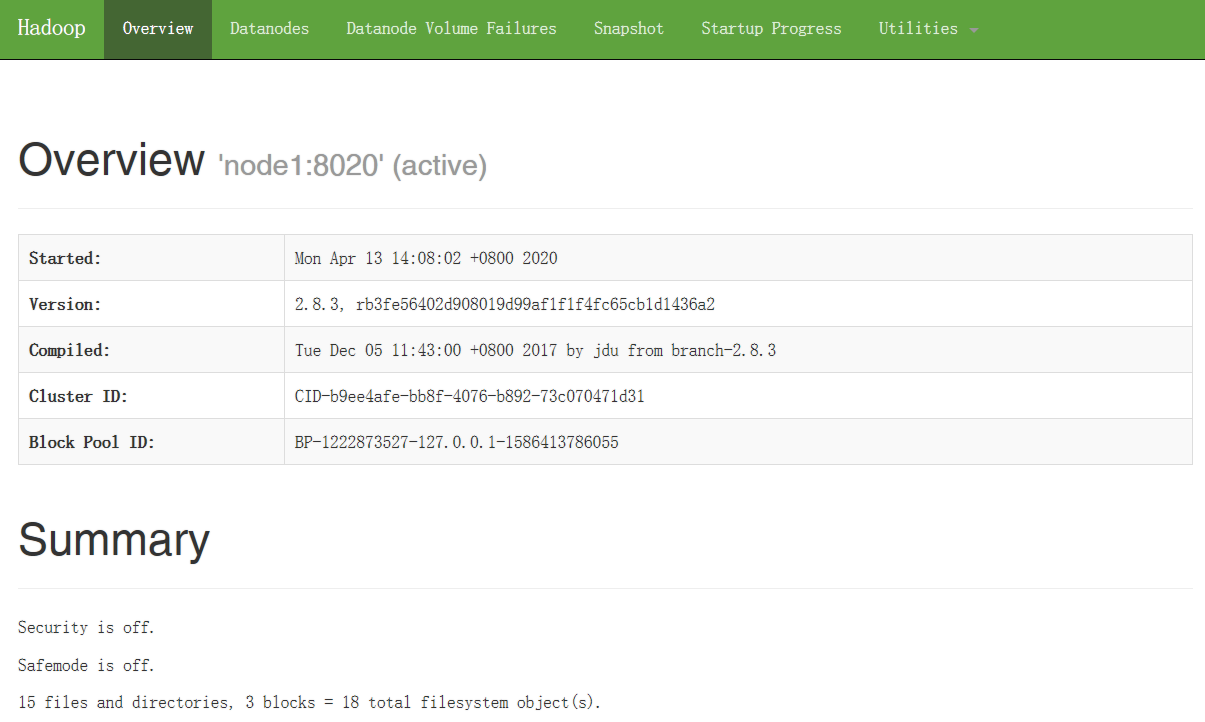
3026 Jps

**2740 DataNode**

1515 WrapperSimpleApp

**2862 NodeManager**

访问<http://node1弹性公网IP:50070>，可以登录namenode的web界面



HDFS的WebUI界面8088

## 实验任务2

### 常用命令操作

学习最基础的HDFS操作命令，可以为今后学习更复杂的命令和服务做好必备知识铺垫。

-help 功能： 查看命令使用说明

> hdfs dfs -help

Usage: hadoop fs [generic options]

[-appendToFile <localsrc> ... <dst>]

[-cat [-ignoreCrc] <src> ...]

[-checksum <src> ...]

[-chgrp [-R] GROUP PATH...]

[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]

[-chown [-R] [OWNER][:[GROUP]] PATH...]

[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]

[-createSnapshot <snapshotDir> [<snapshotName>]]

[-deleteSnapshot <snapshotDir> <snapshotName>]

[-df [-h] [<path> ...]]

[-du [-s] [-h] [-x] <path> ...]

[-expunge]

[-find <path> ... <expression> ...]

[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-getfacl [-R] <path>]

[-getfattr [-R] {-n name | -d} [-e en] <path>]

[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]

……

-ls功能：显示目录信息（因为目前HDFS上没有内容，所以无法查看到内容）：

> hdfs dfs -ls /

20/10/15 17:05:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

20/10/15 17:05:40 INFO obs.OBSFileSystem: This Filesystem GC-ful, clear resource.

-mkdir功能：在HDFS文件系统上创建目录

> hdfs dfs -mkdir -p /tmp/test\_hdfs

> hdfs dfs -ls /tmp

Found 2 items

drwxrwxrwt - root root 0 2020-04-09 17:09 /tmp/logs

drwxr-xr-x - root supergroup 0 2020-04-09 17:25 /tmp/test\_hdfs

-put功能：上传本地文件到HDFS指定目录

# 使用truncate命令创建一个100MB的zip测试文件

> truncate -s 100MB testfile.zip

# 使用HDFS的put命令将zip测试文件上传至指定文件夹中

> hdfs dfs -put testfile.zip /tmp/test\_hdfs

20/04/09 17:30:38 INFO obs.OBSFileSystem: This Filesystem GC-ful, clear resource.

# 验证上传结果

> hdfs dfs -ls /tmp/test\_hdfs

Found 1 items

-rw-r--r-- 3 root supergroup 100000000 2020-04-09 17:30 /tmp/test\_hdfs/testfile.zip

-get功能：从hdfs下载文件到本地，与copyToLocal功能相同

# 拷贝/tmp/test\_hdfs/testfile.zip文件至本地目录

> hdfs dfs -get /tmp/test\_hdfs/testfile.zip /home

20/04/09 17:35:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

20/04/09 17:35:36 INFO obs.OBSFileSystem: This Filesystem GC-ful, clear resource.

> ll /home

total 2821812

drwxr-xr-x 3 root root 4096 Jan 23 22:39 extend\_tools

-rw-r--r-- 1 root root 2789493563 Apr 9 10:45 extend\_tools.tar.gz

drwx------ 2 root root 16384 Apr 9 10:28 lost+found

drwxr-xr-x 4 root root 4096 Apr 9 11:05 modules

drwxr-xr-x 3 root root 4096 Apr 9 10:35 nm

**-rw-r--r-- 1 root root 100000000 Apr 9 17:35 testfile.zip**

drwxr-xr-x 2 root root 4096 Apr 9 10:35 test\_tools

-moveFromLocal 功能：从本地剪切粘贴到HDFS

# 使用touch命令创建一个txt文件用于剪切测试，再添加一句话进去

> touch testshear.txt

> echo this is a share test! >> testshear.txt

> cat testshear.txt

this is a share test!

> ls

# 可以看到有以下文件

testfile.zip **testshear.txt**

# 使用moveFromLocal将testshear.txt文件移动到HDFS文件系统的/ tmp/test\_hdfs目录下

> hdfs dfs -moveFromLocal testshear.txt /tmp/test\_hdfs

# 可以看到本地的testshear.txt文件已消失

> ls

testfile.zip

# 被剪切至HDFS文件系统中

> hdfs dfs -ls /tmp/test\_hdfs

Found 2 items

-rw-r--r-- 3 root supergroup 100000000 2020-04-09 17:30 /tmp/test\_hdfs/testfile.zip

**-rw-r--r-- 3 root supergroup 100000000 2020-04-09 17:40 /tmp/test\_hdfs/testshear.txt**

-cat 功能：显示文件内容

> hdfs dfs -cat /tmp/test\_hdfs/testshear.txt

**this is a share test!**

- appendToFile 功能：在文件末尾追加数据。

# 在当前路径新建一个文件appendtext.txt

> vim appendtext.txt

# 其内容为：

10,Spark

11,Storm

12,Kafka

13,Flink

14,ELK

15,kunpeng bigdata pro

# 将appendtext.txt中的内容追加到user01.txt末尾：

> hdfs dfs -appendToFile ./appendtext.txt /tmp/test\_hdfs/testshear.txt

# 查看追加结果：

> hdfs dfs -cat /tmp/test\_hdfs/testshear.txt

**this is a share test!**

**10,Spark**

**11,Storm**

**12,Kafka**

**13,Flink**

**14,ELK**

**15,kunpeng bigdata pro**

-chmod 功能：更改文件所属权限。

# 查看文件当前的权限

> hdfs dfs -ls /tmp/test\_hdfs/testshear.txt

-rw-r--r-- 3 root supergroup 88 2020-04-13 09:07 /tmp/test\_hdfs/testshear.txt

# 使用chmod命令修改文件权限为755

> hdfs dfs -chmod 755 /tmp/test\_hdfs/testshear.txt

# 再次查看，发现文件权限已经被修改

> hdfs dfs -ls /tmp/test\_hdfs/testshear.txt

-rwxr-xr-x 3 root supergroup 88 2020-04-13 09:07 /tmp/test\_hdfs/testshear.txt

除了-chmod以外，还有-chgrp、-chown等功能，与Linux文件系统中的用法一样，都是修改文件所属权限。

-cp 功能：实现文件的拷贝。

# 将testshear.txt拷贝到/tmp下：

> hdfs dfs -cp /tmp/test\_hdfs/testshear.txt /tmp

# 查看拷贝结果

> hdfs dfs -ls /tmp

Found 3 items

drwxrwxrwt - root root 0 2020-04-09 17:09 /tmp/logs

drwxr-xr-x - root supergroup 0 2020-04-09 17:50 /tmp/test\_hdfs

**-rw-r--r-- 3 root supergroup 88 2020-04-13 09:11 /tmp/testshear.txt**

-mv 功能：移动文件。

将testshear.txt移动到/user目录下

> hdfs dfs -mkdir /user

> hdfs dfs -mv /tmp/testshear.txt /user

> hdfs dfs -ls /user

Found 1 items

-rw-r--r-- 3 root supergroup 88 2020-04-13 09:11 /user/testshear.txt

-getmerge 功能：合并下载一个文件夹中的多个文件

# 准备/tmp/merge\_test文件夹，及file01.txt和file02.txt两个文件

> hdfs dfs -mkdir /tmp/merge\_test

# 在文件夹里上传两个文件file01.txt和file02.txt

> vim file01.txt

# file01.txt内容分别为：

1

2

3

4

5

> vim file02.txt

# file02.txt内容分别为：

6

7

8

9

10

> hdfs dfs -put file01.txt file02.txt /tmp/merge\_test

> hdfs dfs -ls /tmp/merge\_test

Found 2 items

-rw-r--r-- 3 root supergroup 10 2020-04-13 09:17 /tmp/merge\_test/file01.txt

-rw-r--r-- 3 root supergroup 11 2020-04-13 09:17 /tmp/merge\_test/file02.txt

> hdfs dfs -cat /tmp/merge\_test/file01.txt

1

2

3

4

5

> hdfs dfs -cat /tmp/merge\_test/file02.txt

6

7

8

9

10

# 合并下载/tmp/merge\_test文件夹下的所有文件

> hdfs dfs -getmerge /tmp/merge\_test Download.txt

# 查看下载结果

> cat Download.txt

1

2

3

4

5

6

7

8

9

10

-rm 功能：删除文件或文件夹。

# 删除hdfs中的文件

> hdfs dfs -rm /tmp/merge\_test/file01.txt

Deleted /tmp/merge\_test/file01.txt

> hdfs dfs -ls /tmp/merge\_test

Found 1 items

-rw-r--r-- 3 root supergroup 11 2020-04-13 09:17 /tmp/merge\_test/file02.txt

# 删除hdfs中的文件文件夹

> hdfs dfs -rm -R /tmp/merge\_test

Deleted /tmp/merge\_test

> hdfs dfs -ls /tmp

Found 2 items

drwxrwxrwt - root root 0 2020-04-09 17:09 /tmp/logs

drwxr-xr-x - root supergroup 0 2020-04-13 09:22 /tmp/test\_hdfs

-rmdir 功能：删除空目录

> hdfs dfs -mkdir /test

> hdfs dfs -ls /

> hdfs dfs -rmdir /test

> hdfs dfs -ls /

-df 功能：统计文件系统的可用空间信息。

> hdfs dfs -df -h /

Filesystem Size Used Available Use%

hdfs://node1:8020 294.9 G 576.9 M 277.1 G 0%

-du 功能：统计文件或文件夹的大小信息。

> hdfs dfs -du -s -h /tmp/test\_hdfs

95.4 M /tmp/test\_hdfs

> hdfs dfs -du -h /tmp/test\_hdfs

88 /tmp/test\_hdfs/testshear.txt

95.4 M /tmp/test\_hdfs/testshear.zip

-count 功能：统计一个指定目录下的文件数量。

> hdfs dfs -count -h /tmp/test\_hdfs

1 2 95.4 M /tmp/test\_hdfs

第一列1表示指定目录下文件夹的数量，第二列3表示指定目录下文件的个数。190.7 M表示指定目录下所有文件占用的磁盘容量（不包括hdfs副本）。

-setrep 功能：设置HDFS文件副本数量

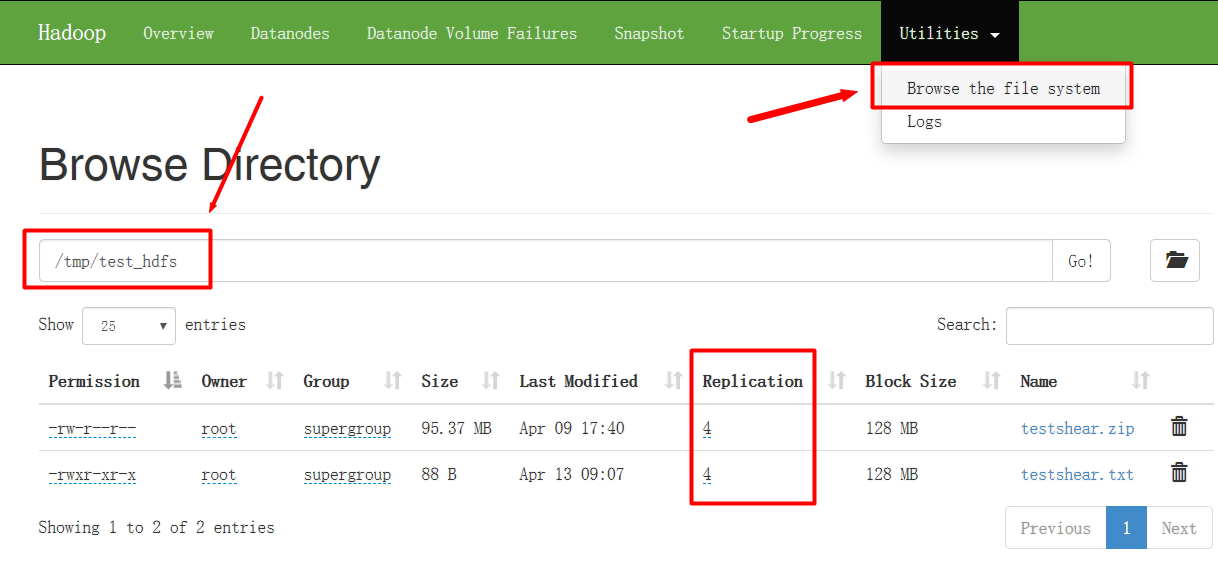
> hadoop fs -setrep 4 /tmp/test\_hdfs

Replication 4 set: /tmp/test\_hdfs/testshear.txt

Replication 4 set: /tmp/test\_hdfs/testshear.zip

登录namenode的web界面查看详细信息，<http://node1弹性公网IP:50070>

“Utilities”-“Browse the file system”,在“Browse Directory”中输入想查看的目录，点击“Go！”或按回车键



namenode的web界面

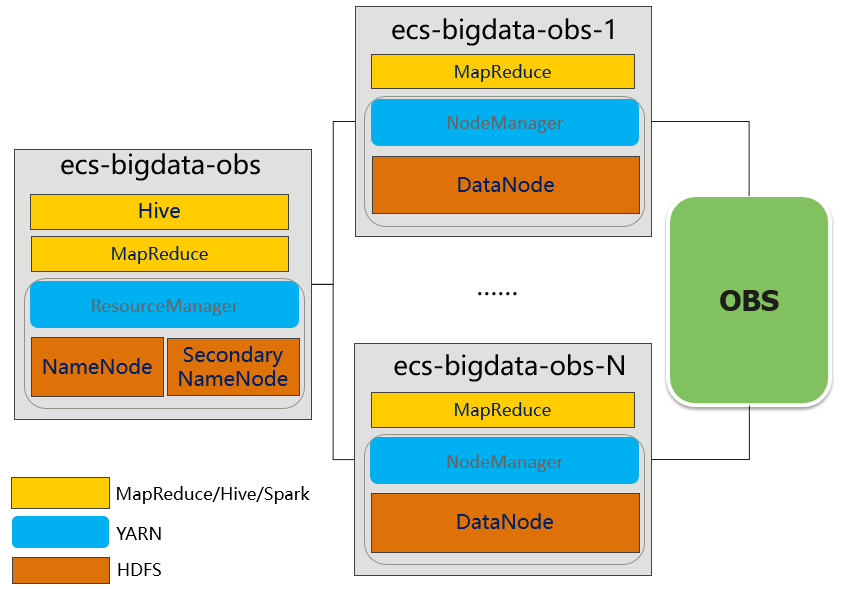
注意：

1.如果setrep命令设置的副本数大于Datanode的数量，虽然在NameNode的元数据中是显示setrep命令设置的副本数，但实际上HDFS最多只存储与Datanode节点数相等的副本数。

2. setrep命令不适用OBS作为底层存储

## 实验任务3

华为公有云对象存储服务（Object Storage Service，OBS）是一个基于对象的海量存储服务，为客户提供**海量、安全、高可靠、低成本的数据存储能力**。鲲鹏BigDataPro解决方案将OBS融入到大数据生态圈，增强大数据集群的存储能力。



Hadoop对接OBS部署视图

OBSFileSystem客户端封装提供HDFS接口对接OBS，负责HDFS与OBS之间的文件语义转换，使OBS服务可以融入大数据生态，实现大数据计算MapReduce、Hive、Spark等组件与OBS对象存储服务对接，为大数据计算提供“数据湖”存储。

### 准备工作

登录华为公有云，选择OBS对象存储服务



华为公有云OBS服务

在OBS控制台中选择要对接hadoop的OBS桶



OBS控制台

如果当前账号下没有OBS桶，请自行创建新OBS桶，自定义桶名称，其它参数默认即可



记录桶名称、Endpoint、区域等基本信息



OBS桶基本信息

记录AK和SK

访问密钥即AK/SK（Access Key ID/Secret Access Key），是通过开发工具（API、CLI、SDK）访问华为云时的身份凭证。

注意：每个访问密钥仅能在创建时下载一次，如果已经获取过，则不需要重新获取。

获取方式如下图，打开“我的凭证”-“访问秘钥”-“新建访问密钥”



我的凭证

检查core-site.xml配置文件

查看node1-4各节点core-site.xml配置文件中fs.obs.access.key、fs.obs.secret.key、fs.obs.endpoint 3项内容，是否按照实际情况填写

<name>fs.obs.access.key</name>

<value>Access Key ID</value>

</property>

<property>

<name>fs.obs.secret.key</name>

<value>Secret Access Key</value>

</property>

<property>

<name>fs.obs.endpoint</name>

<value>obs.cn-north-4.myhuaweicloud.com</value>

</property>

如果core-site.xml配置文件有修改，则需要重新启动鲲鹏hadoop集群

### 访问OBS操作

创建文件夹，并查看OBS桶（XX-bigdatapro01需修改为自己的桶名）

> hdfs dfs -mkdir obs://XX-bigdatapro01/OBS\_test

> hdfs dfs -ls obs://XX-bigdatapro01/

Found 1 items

drwxrwxrwx - root root 0 2020-04-13 11:47 obs://XX-bigdatapro01/OBS\_test

可以查看到创建的文件夹则表示OBS对接鲲鹏hadoop成功！

公有云控制台验证

可以查看到OBS\_test文件已创建在OBS桶中



OBS桶控制台

上传大文件

# 使用truncate命令创建一个5GB的zip测试文件

> truncate -s 5GB testfile.zip.gz

# 使用HDFS的put命令将zip测试文件上传至OBS桶的指定文件夹中

> hdfs dfs -put testfile.zip.gz obs://XX-bigdatapro01/OBS\_test

20/04/09 17:30:38 INFO obs.OBSFileSystem: This Filesystem GC-ful, clear resource.

# 验证上传结果

> hdfs dfs -ls obs://XX-bigdatapro01/OBS\_test

Found 1 items

-rw-rw-rw- 1 root root 10000000000 2020-04-13 11:59 obs://XX-bigdatapro01/OBS\_test/testfile.zip.gz

根据输出的INFO日志信息可以观察到，使用OBS做HDFS底层数据存储时，上传和下载速度还是非常快的。

OBS桶控制台验证



OBS桶控制台

删除OBS的文件

> hdfs dfs -rm -r obs://XX-bigdatapro01/OBS\_test

Deleted obs://XX-bigdatapro01/OBS\_test

其它操作命令与“2.3.1”章节介绍方法一致，学员可以自行体验

*--- 结束！*

## 实验小结

本实验主要讲述HDFS的常见操作以及访问OBS数据湖，通过本实验，学员可掌握鲲鹏Hadoop的HDFS使用和管理操作。