实验二　大数据分析平台中HDFS的使用

（一）实验目的

1. 理解HDFS在Hadoop体系结构中的角色；
2. 理解HDFS存在的原因；
3. 理解HDFS体系架构；
4. 理解HDFS读写数据过程；
5. 熟练使用HDFS常用的Shell命令。

（二）实验要求

1. 在HDFS中进行目录操作；
2. 在HDFS中进行文件操作；
3. 从本机中上传文件到HDFS；
4. 从HDFS下载文件到本机。

（三）实验环境

1. 大数据分析实验系统（FSDP）；
2. CentOS 6.7；
3. Hadoop 2.7.1。

（四）实验步骤

1. HDFS预备知识；
2. HDFS读写数据的过程；
3. HDFS的目录和文件操作。

1、HDFS预备知识

分布式文件系统（Distributed File System）是指文件系统管理的物理存储资源不一定直接连接在本地节点，而是通过计算机网络与节点相连。

HDFS（Hadoop分布式文件系统，Hadoop Distributed File System）是一种适合运行在通用硬件上的分布式文件系统，它是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。

HDFS为大数据分析平台的其他所有组件提供了最基本的存储功能。它具有高容错、高可靠、可扩展、高吞吐率等特征，为大数据存储和处理提供了强大的底层存储架构。

HDFS采用主/从（master/slave）式体系结构，从最终用户的角度来看，它就像传统的文件系统，可通过目录路径对文件执行增删改查操作。由于其分布式存储的性质，HDFS拥有一个NameNode和一些DataNode，NameNode管理文件系统的元数据，DataNode存储实际的数据。

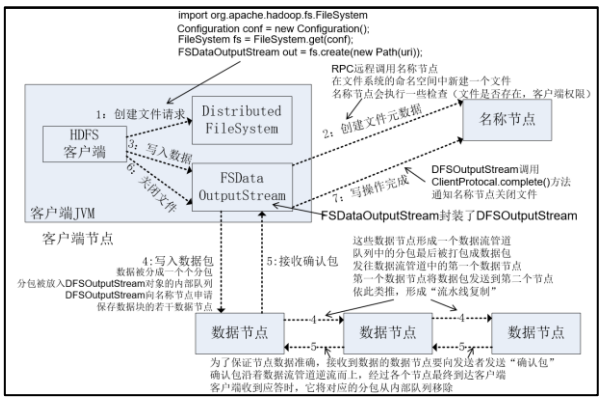
HDFS提供高吞吐量应用程序访问功能，适合带有大型数据集的场景，具体包括：

- 数据密集型并行计算：数据量大，但是计算相对简单的并行处理，如大规模Web信息搜索；
- 计算密集型并行计算：数据量相对不是很大，但是计算较为复杂的并行处理，如3D建模与渲染、气象预报、科学计算等；
- 数据密集型与计算密集型混合的计算，如3D电影渲染等。

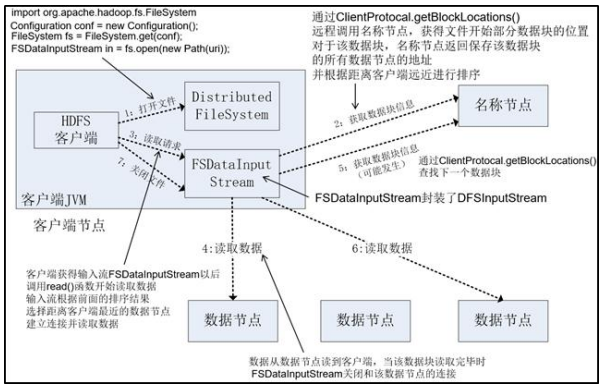
2、HDFS读写数据的过程

普通文件系统主要用于随机读写以及与用户进行交互，而HDFS则是为了满足批量数据处理的要求而设计的，因此为了提高数据吞吐率，HDFS放松了一些POSIX的要求，从而能够以流方式来访问文件系统数据。

HDFS读数据的过程



HDFS写数据的过程



实验二　大数据分析平台中HDFS的使用

3、HDFS的目录和文件操作

(1)在HDFS中创建目录

```
[test@fsmanager~]$ hadoop fs  -mkdir /user/test
[test@fsmanager ~]$ hadoop fs -ls /user/test
[test@fsmanager ~]$
```

(2)在此用户目录下创建text、ab文件夹，并查看文件列表

```
[test@fsmanager ~]$ hadoop  fs  -mkdir  text ab
[test@fsmanager ~]$ hadoop fs -ls /user/test
Found 2 items
drwxr-xr-x   - test hdfs 0 2018-04-09 11:40
/user/test/text
drwxr-xr-x   - test hdfs 0 2018-04-09 11:40
/user/test/ab
[test@fsmanager ~]$
```

(3)将 ~ /.bashrc文件上传到HDFS的text文件夹，并查看test

```
[test@fsmanager ~]$ cd ~
[test@fsmanager ~]$ hadoop fs -put .bashrc
text
[test@fsmanager ~]$ hadoop fs -ls text
Found 1 items
-rw-r--r--   2 test hdfs 124 2018-04-09 11:45
text/.bashrc
[test@fsmanager ~]$
```

(4)将HDFS文件夹text下载到本地

```
[test@fsmanager ~]$ hadoop fs -get text ./
18/04/09 11:46:23 WARN hdfs.DFSClient:
DFSInputStream has been closed already
[test@fsmanager ~]$ ls
abc  a.tar  text
[test@fsmanager ~]$
```

(5)删除HDFS中的文件. bashrc

```
[test@fsmanager ~]$hadoop fs -rm text/.bashrc
18/04/09 11:47:17 INFO fs.TrashPolicyDefault:
Namenode trash configuration: Deletion interval
= 0 minutes, Emptier interval = 0 minutes.
Deleted text/.bashrc
[test@fsmanager ~]$ hadoop fs -ls text
[test@fsmanager ~]$
```

(6)在HDFS中查看文件内容

```
命令：hadoop fs -text filepath 例如：
[test@fsmanager ~]$ hadoop fs -put .bashrc  .
[test@fsmanager ~]$ hadoop fs -text ./.bashrc
# .bashrc
# Source global definitions
fi
# User specific aliases and functions
```

(7)在HDFS中创建并删除目录test123

```
[test@fsmanager ~]$ hadoop fs -mkdir  text123
[test@fsmanager ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x - test hdfs 0 2018-04-09 11:47 text
drwxr-xr-x - test hdfs 0 2018-04-09 11:51
text123
[test@fsmanager ~]$ hadoop fs -rm -r text123
18/04/09 11:51:55 INFO fs.TrashPolicyDefault:
Namenode trash configuration: Deletion interval
= 0 minutes, Emptier interval = 0 minutes.
Deleted text123
[test@fsmanager ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x - test hdfs 0 2018-04-09 11:47 text
```

(8)查看HDFS中文件内容

```
命令格式：hadoop fs -mv 源文件路径 目标文件路径
[test@fsmanager ~]$ hadoop fs -ls text  text1
Found 1 items
drwxr-xr-x   - test hdfs 0 2018-04-09 11:54
text/hello
Found 1 items
drwxr-xr-x   - test hdfs 0 2018-04-09 11:55
text1/hello1
[test@fsmanager ~]$ hadoop fs -mv text/hello
text1
[test@fsmanager ~]$ hadoop fs -mv text1/hello1
text
[test@fsmanager ~]$ hadoop fs -ls text text1
Found 1 items
drwxr-xr-x   - test hdfs 0 2018-04-09 11:55
text/hello1
Found 1 items
drwxr-xr-x   - test hdfs 0 2018-04-09 11:54
text1/hello
```

(9)查看HDFS文件或目录占用空间

```
[test@fsmanager ~]$ hadoop fs -du -h /user/test
124  /user/test/.bashrc
0    /user/test/text
```

HDFS在使用过程中有以下限制：

- HDFS不适合大量小文件的存储。由于namenode将文件系统的元数据存放在内存中，因此存储的文件数目受限于NameNode的内存大小；
- HDFS适用于高吞吐量，而不适用于低时间延迟访问的应用场景；
- HDFS流式读取的方式，不适合多用户写入一个文件（一个文件同时只能被一个客户端写），以及任意位置写入（不支持随机写）；
- HDFS更加适合一次写入，读取多次的应用场景。