大数据平台技术课程实验报告

**实验二：HDFS操作实践**

姓名：\_\_\_\_\_\_杨佳森

学号：\_\_\_\_\_ 2112080106

专业：\_\_\_ 数据科学与大数据技术

年级：\_ 2021

主讲教师：\_ 刘春 \_

实验时间： 3月3日14：00 至 18：00

实验内容与要求：

[1](#_Toc43287683) HDFS基本知识总结

2 HDFS接口的操作实践

2.1 HDFS Shell 实践

2.2 HDFS Web客户端

2.3 HDFS Java API 实践

**特别提醒：**

**（1）基本知识点的总结请使用自己的语言，根据自己的理解去总结，就像写课程笔记一样，不要去网上抄写**

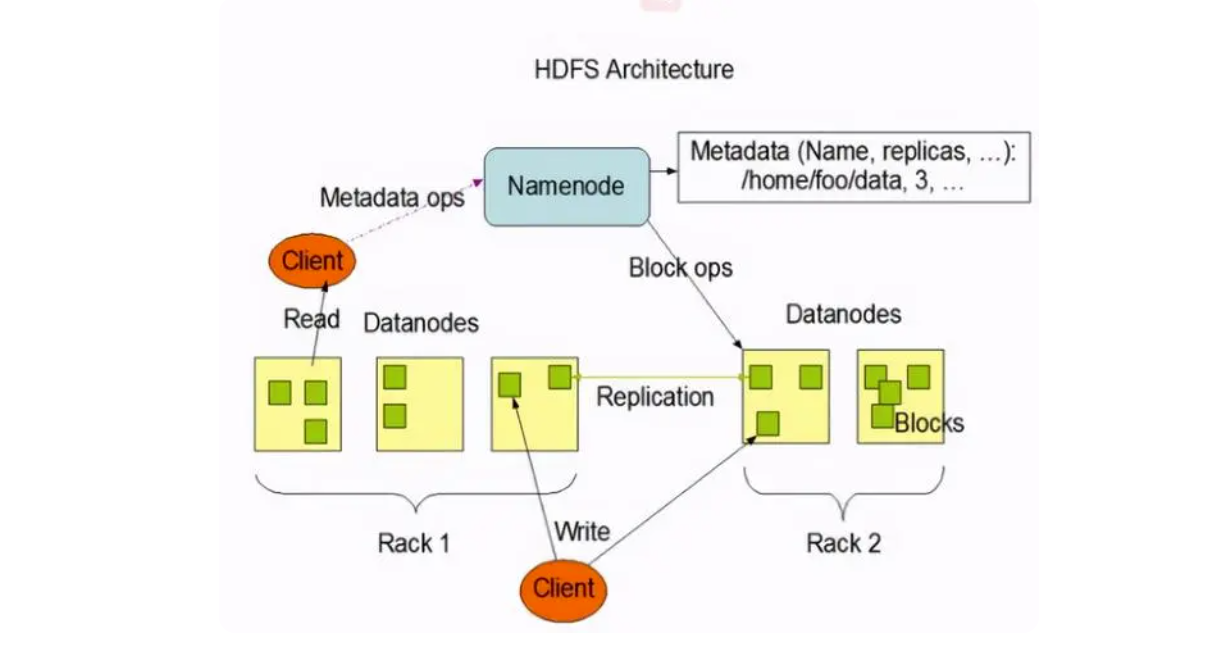
**（2）实践需要有截图以及相应的文字介绍，欢迎大家把自己在实践过程中碰到的问题以及解决方法也记录下来**

1. **HDFS基本知识的总结**

HDFS（Hadoop Distributed File System），它是一个文件系统，用于存储文件，通过目 录树来定位文件；其次，它是分布式的，由很多服务器联合起来实现其功能，集群中的服务器有各自的角色。适合一次写入，多次读出的场景。一个文件经过创建、写入和关闭 之后就不需要改变。

HDFS的优点：高容错性 数据自动保存多个副本。它通过增加副本的形式，提高容错性, 某一个副本丢失以后，它可以自动恢复。适合处理大数据 数据规模：能够处理数据规模达到GB、TB、甚至PB级别的数据； 某一个副本丢失以后，它可以自动恢复。文件规模：能够处理百万规模以上的文件数量，数量相当之大。可构建在廉价机器上，通过多副本机制，提高可靠性。

HDFS的缺点：不适合低延时数据访问，比如毫秒级的存储数据，是做不到的。无法高效对大量小文件进行存储。存储大量小文件的话，它会占用NameNode大量的内存来存储文件目录和块信息。这样是不可取的，因为NameNode的内存总是有限的；小文件存储的寻址时间会超过读取时间，它违反了HDFS的设计目标。不支持并发写入、文件随机修改。 一个文件只能有一个写，不允许多个线程同时写；仅支持数据append（追加），不支持文件的随机修改。



  1. NameNode：管理整个文件系统的元数据，如命名空间、数据块（Block）映射信息、副本策略及处理客户端读写请求。

       2. DateNode：管理每个数据块，如存储实际的数据块，处理客户端对数据块的读/写操作。

       3. Client：通过 CLI 或 API 来操作 DataNode（读 / 写操作）和 NameNode（获取文件位置信息）。

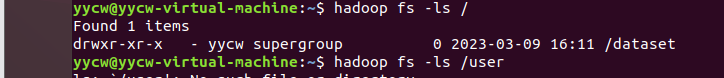
       4. Secondary NameNode：辅助 NameNode 分担工作量，定期合并 fsimage（命名空间镜像） 和 fsedits（修改日志） 并推送给 NameNode。

       5. Block：默认最基本的存储单为是128M的数据块。1. 区别：不同于普通文件系统的是，HDFS中，如果一个文件小于一个数据块的大小，并不占用整个数据块存储空间。2. 目的：减少硬盘寻道时间（减少最小化寻址开销），减少 NameNode 内存消耗，加快数据传输速度。

1. **HDFS的操作实践**

**2.1 HDFS Shell 实践**

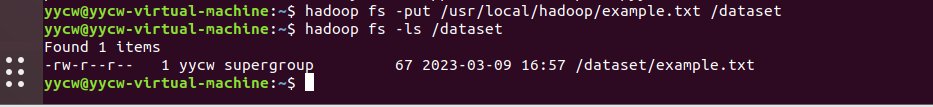
（1）列举一个目录的所有文件 hadoop fs -ls 目录的路径

****

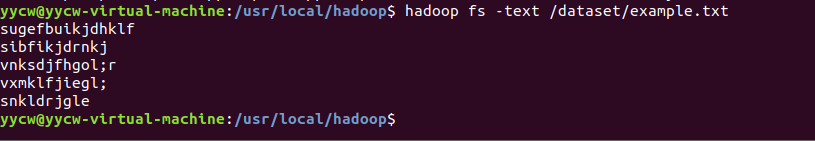
（2）创建文件夹 hadoop fs -mkdir 文件夹的路径和名称

（3）将本地文件上传至HDFS hadoop fs -put 本地文件路径 目标路径

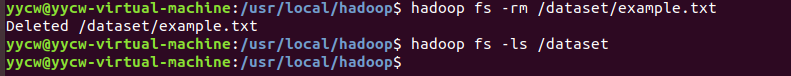
（4）将文件从HDFS下载到本地文件系统 hadoop fs -get HDFS文件路径 本地存放路径

****

（5）查看文件的内容 hadoop fs -text HDFS文件的路径

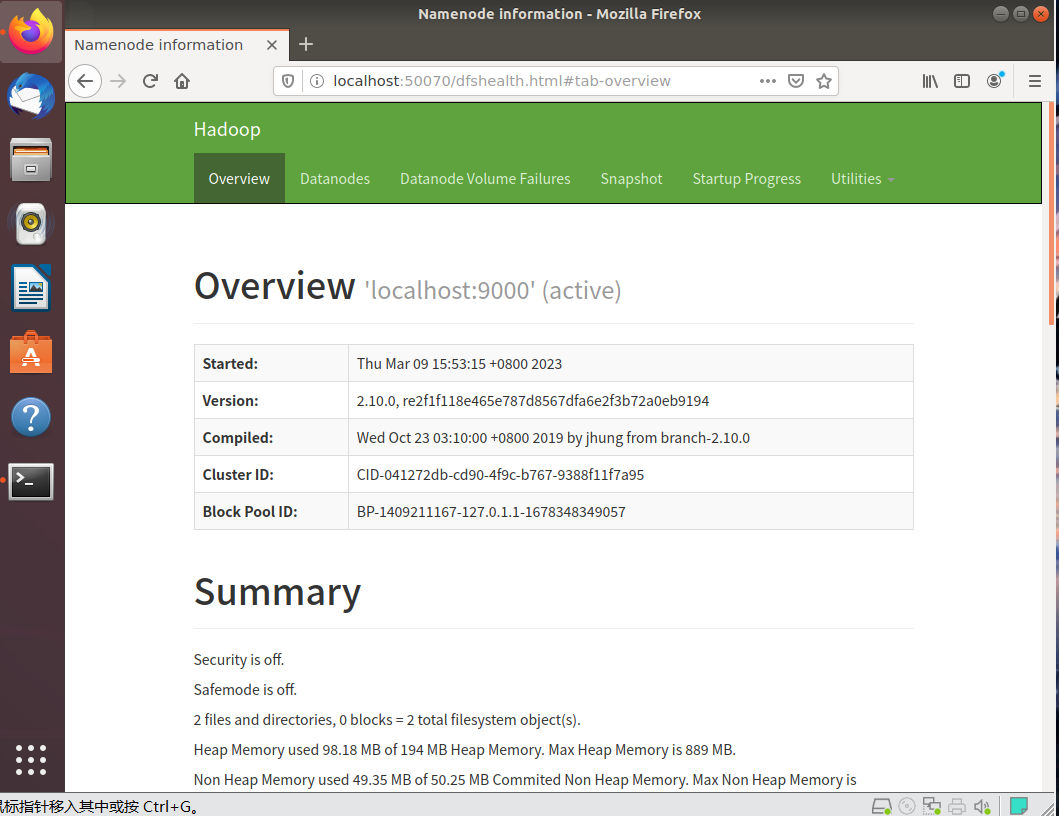
****

（6）删除文件或者文件夹 hadoop fs -rm

****

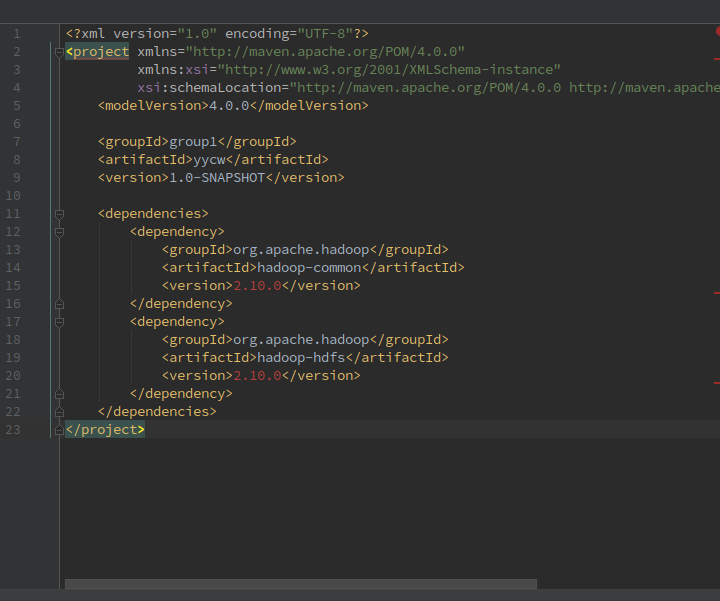
**2.2 HDFS Web客户端**

在火狐浏览器里输入<http://localhost:50070>

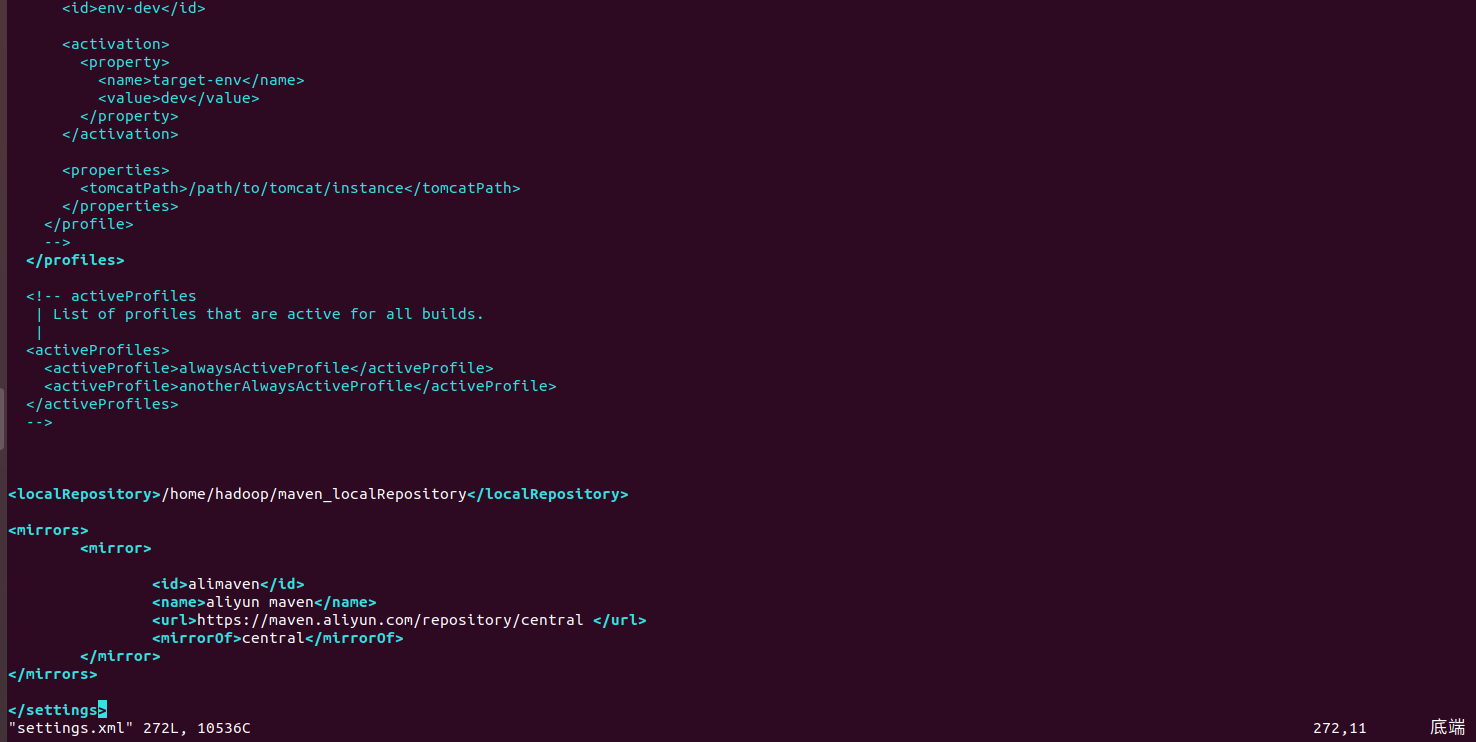
****

* 1. **HDFS Java API 实践**

**在pom.xml文件中配置好jar包依赖信息：**

****

在Hadoop安装文件“/etc/hadoop”路径下的hdfs-site.xml文件添加配置信息：



在Maven项目中建立一个名为HDFSapp的Java类，通过Java API来实现针对HDFS的目录创建、文件创建、上传及下载文件、删除文件操作：

