**实验3-统计某电商网站买家收藏商品数量**

现有某电商网站用户对商品的收藏数据，记录了用户收藏的商品id以及收藏日期，名为buyer\_favorite1。buyer\_favorite1包含：买家id，商品id，收藏日期这三个字段，数据以“\t”分割，样本数据及格式如下：

1. 买家id   商品id    收藏日期
2. 10181   1000481   2010-04-04 16:54:31
3. 20001   1001597   2010-04-07 15:07:52
4. 20001   1001560   2010-04-07 15:08:27
5. 20042   1001368   2010-04-08 08:20:30
6. 20067   1002061   2010-04-08 16:45:33
7. 20056   1003289   2010-04-12 10:50:55
8. 20056   1003290   2010-04-12 11:57:35
9. 20056   1003292   2010-04-12 12:05:29
10. 20054   1002420   2010-04-14 15:24:12
11. 20055   1001679   2010-04-14 19:46:04
12. 20054   1010675   2010-04-14 15:23:53
13. 20054   1002429   2010-04-14 17:52:45
14. 20076   1002427   2010-04-14 19:35:39
15. 20054   1003326   2010-04-20 12:54:44
16. 20056   1002420   2010-04-15 11:24:49
17. 20064   1002422   2010-04-15 11:35:54
18. 20056   1003066   2010-04-15 11:43:01
19. 20056   1003055   2010-04-15 11:43:06
20. 20056   1010183   2010-04-15 11:45:24
21. 20056   1002422   2010-04-15 11:45:49
22. 20056   1003100   2010-04-15 11:45:54
23. 20056   1003094   2010-04-15 11:45:57
24. 20056   1003064   2010-04-15 11:46:04
25. 20056   1010178   2010-04-15 16:15:20
26. 20076   1003101   2010-04-15 16:37:27
27. 20076   1003103   2010-04-15 16:37:05
28. 20076   1003100   2010-04-15 16:37:18
29. 20076   1003066   2010-04-15 16:37:31
30. 20054   1003103   2010-04-15 16:40:14
31. 20054   1003100   2010-04-15 16:40:16

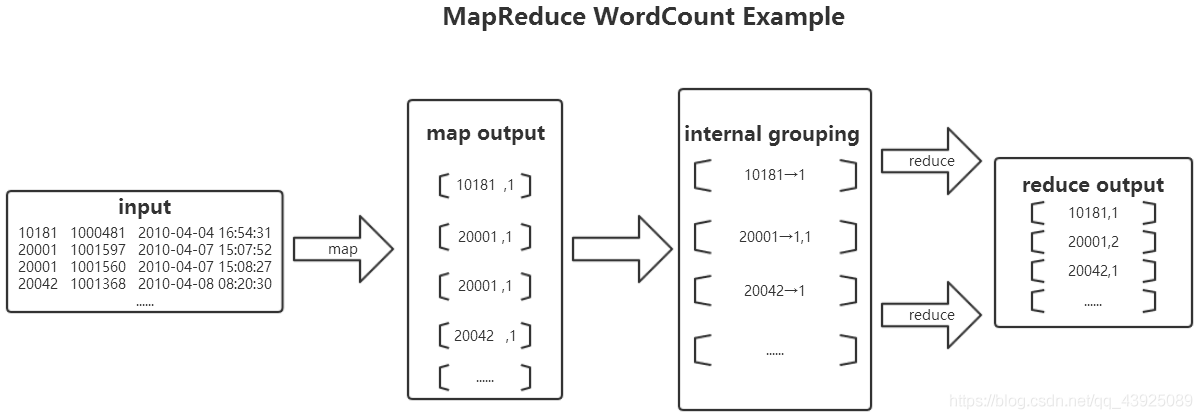
要求编写MapReduce程序，统计每个买家收藏商品数量，并撰写实验报告。

实验思路

大致思路是将hdfs上的文本作为输入，MapReduce通过InputFormat会将文本进行切片处理，并将每行的首字母相对于文本文件的首地址的偏移量作为输入键值对的key，文本内容作为输入键值对的value，经过在map函数处理，输出中间结果<word,1>的形式，并在reduce函数中完成对每个单词的词频统计。整个程序代码主要包括两部分：Mapper部分和Reducer部分。

在map函数里有三个参数，前面两个Object key,Text value就是输入的key和value，第三个参数Context context是可以记录输入的key和value。例如context.write(word,one)；此外context还会记录map运算的状态。map阶段采用Hadoop的默认的作业输入方式，把输入的value用StringTokenizer()方法截取出的买家id字段设置为key，设置value为1，然后直接输出<key,value>。

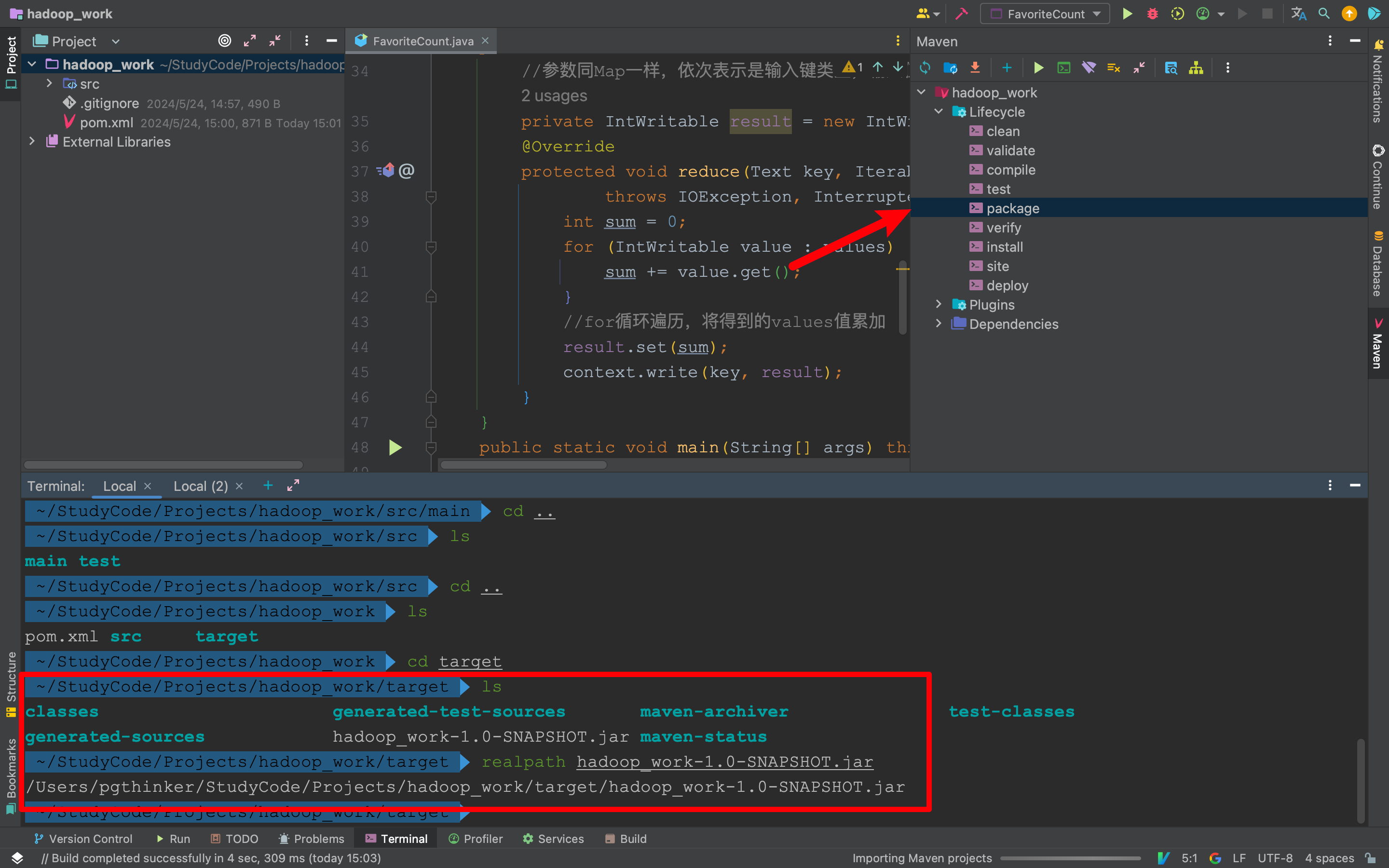
map输出的<key,value>先要经过shuffle过程把相同key值的所有value聚集起来形成<key,values>后交给reduce端。reduce端接收到<key,values>之后，将输入的key直接复制给输出的key,用for循环遍历values并求和，求和结果就是key值代表的单词出现的总次，将其设置为value，直接输出<key,value>。



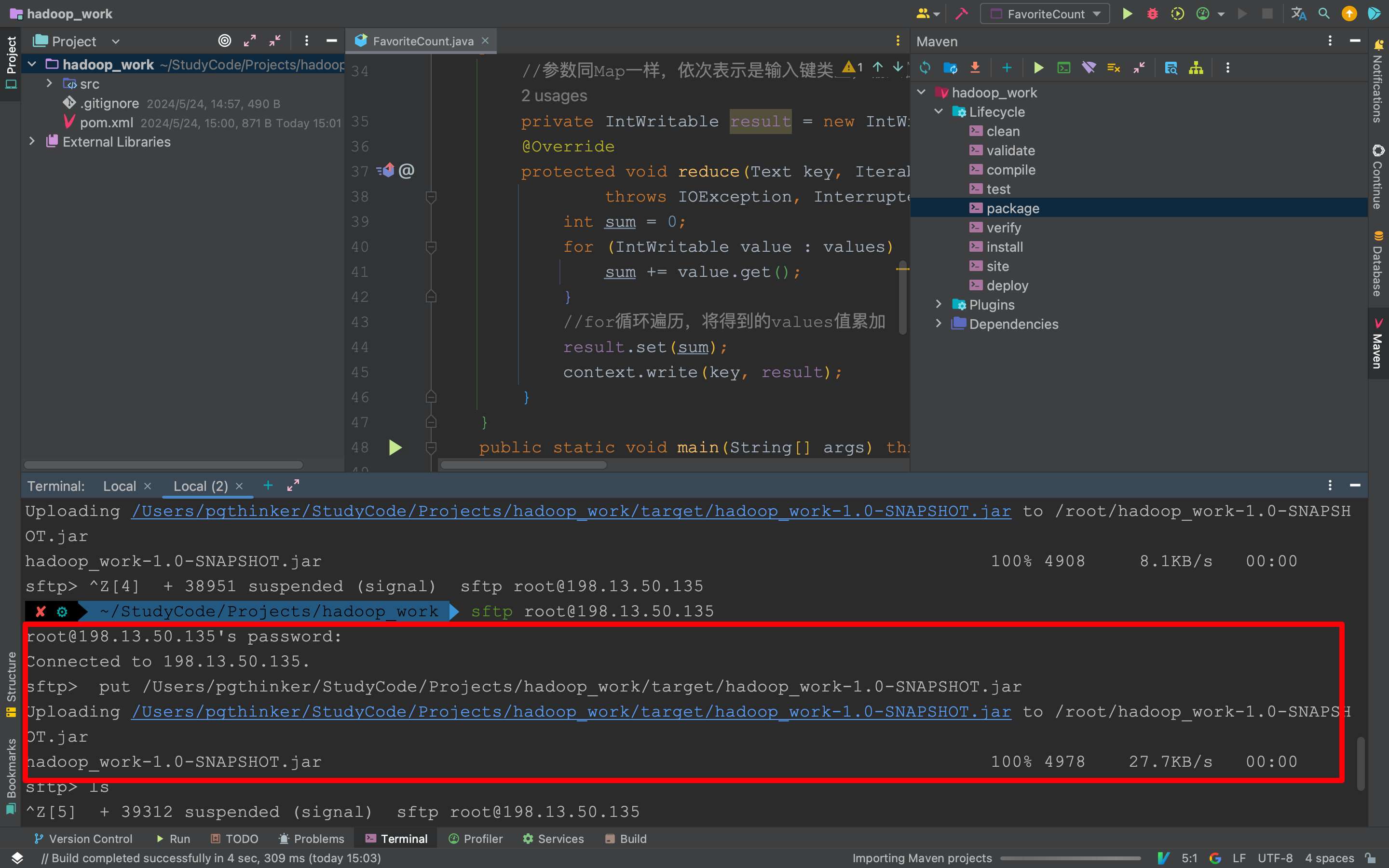
实现代码

运行步骤:

1. 将程序打包成jar包



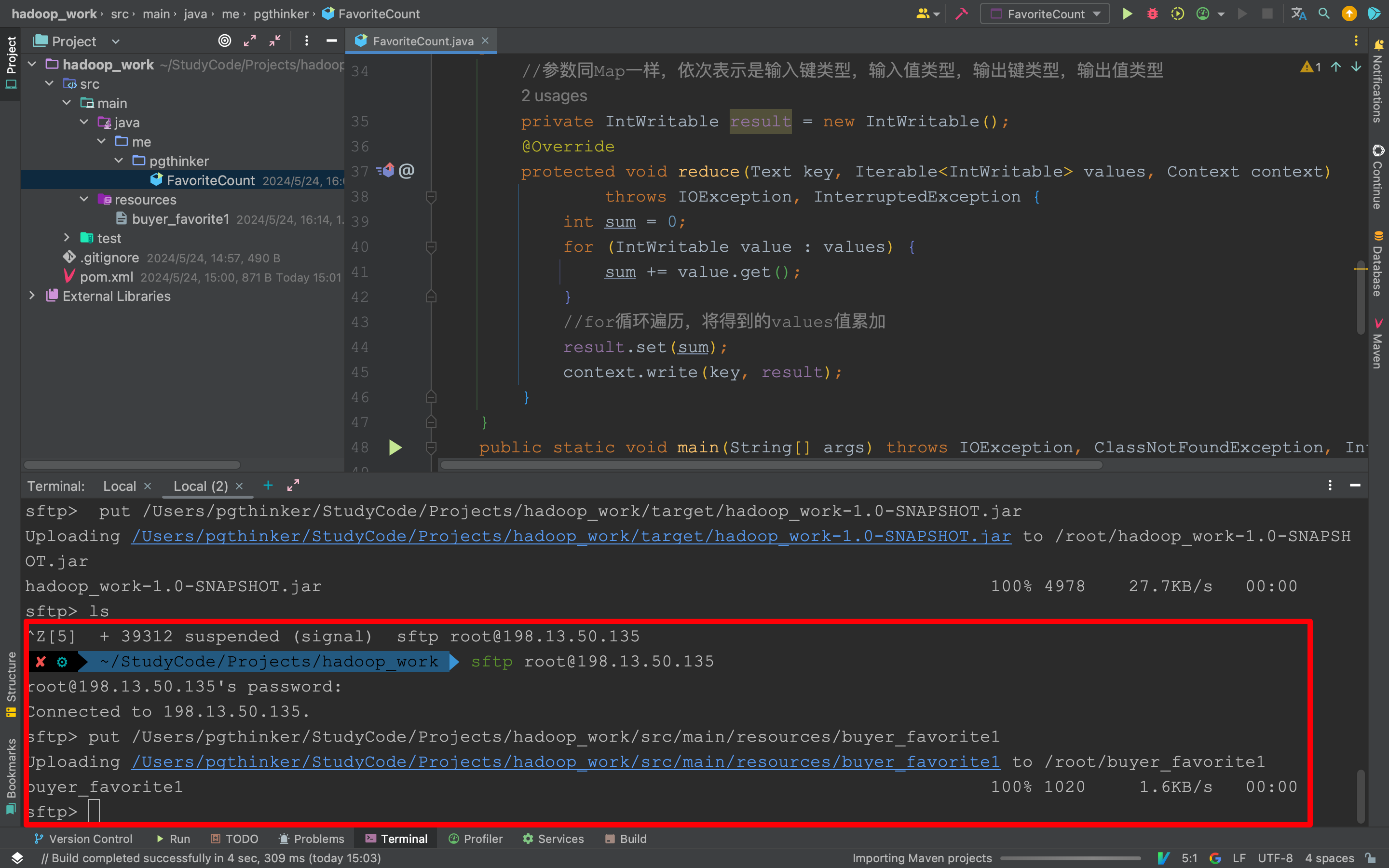
1. 将jar包上传至云上



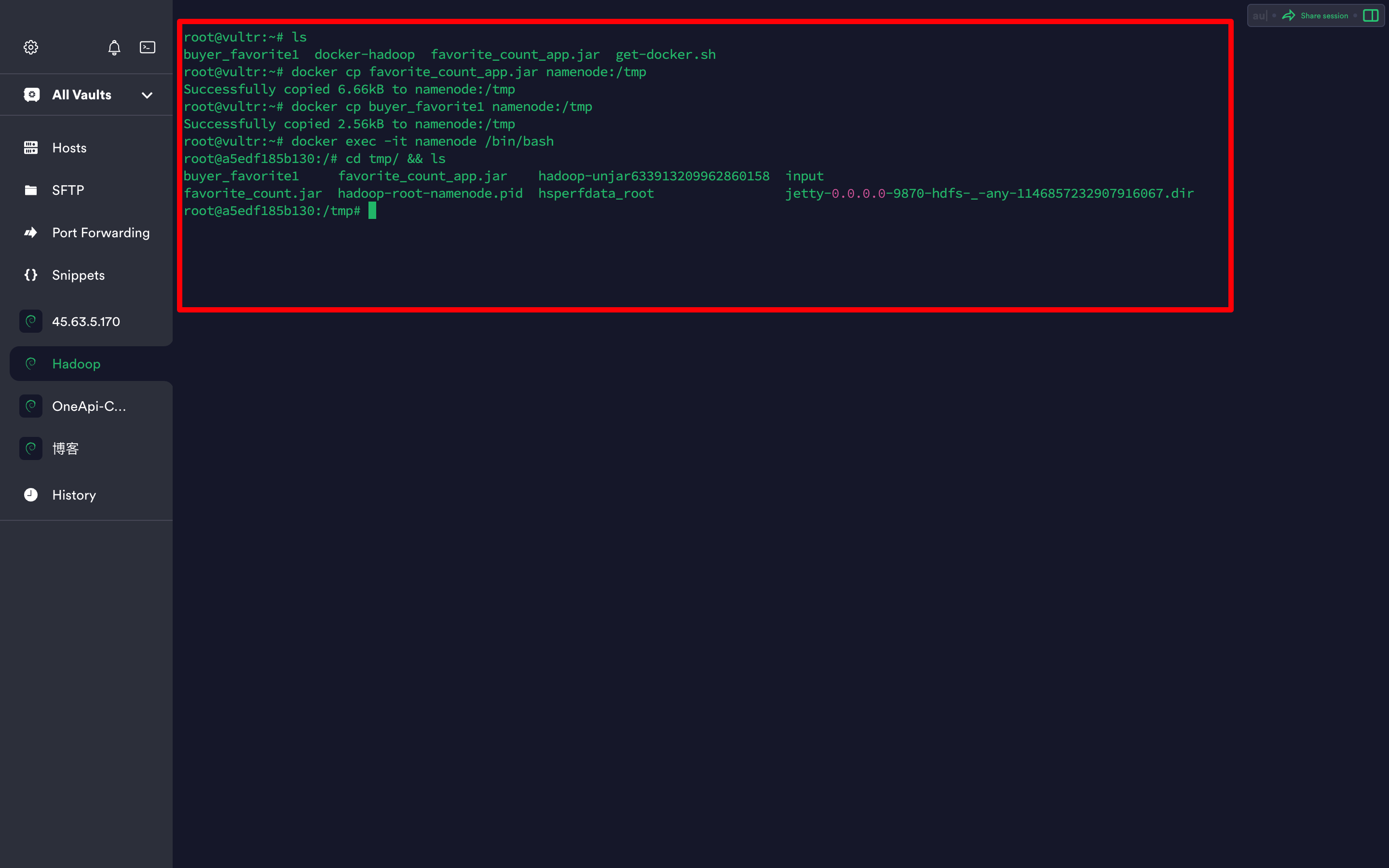
1. 上传buyer\_favorite1文件至Hadoop的hdfs上

由于我的Hadoop部署在云上，因此需要先将本地的buyer\_favorite1文件上传到云服务器上，再在云服务器上将buyer\_favorite1上传到Hadoop的HDFS上。

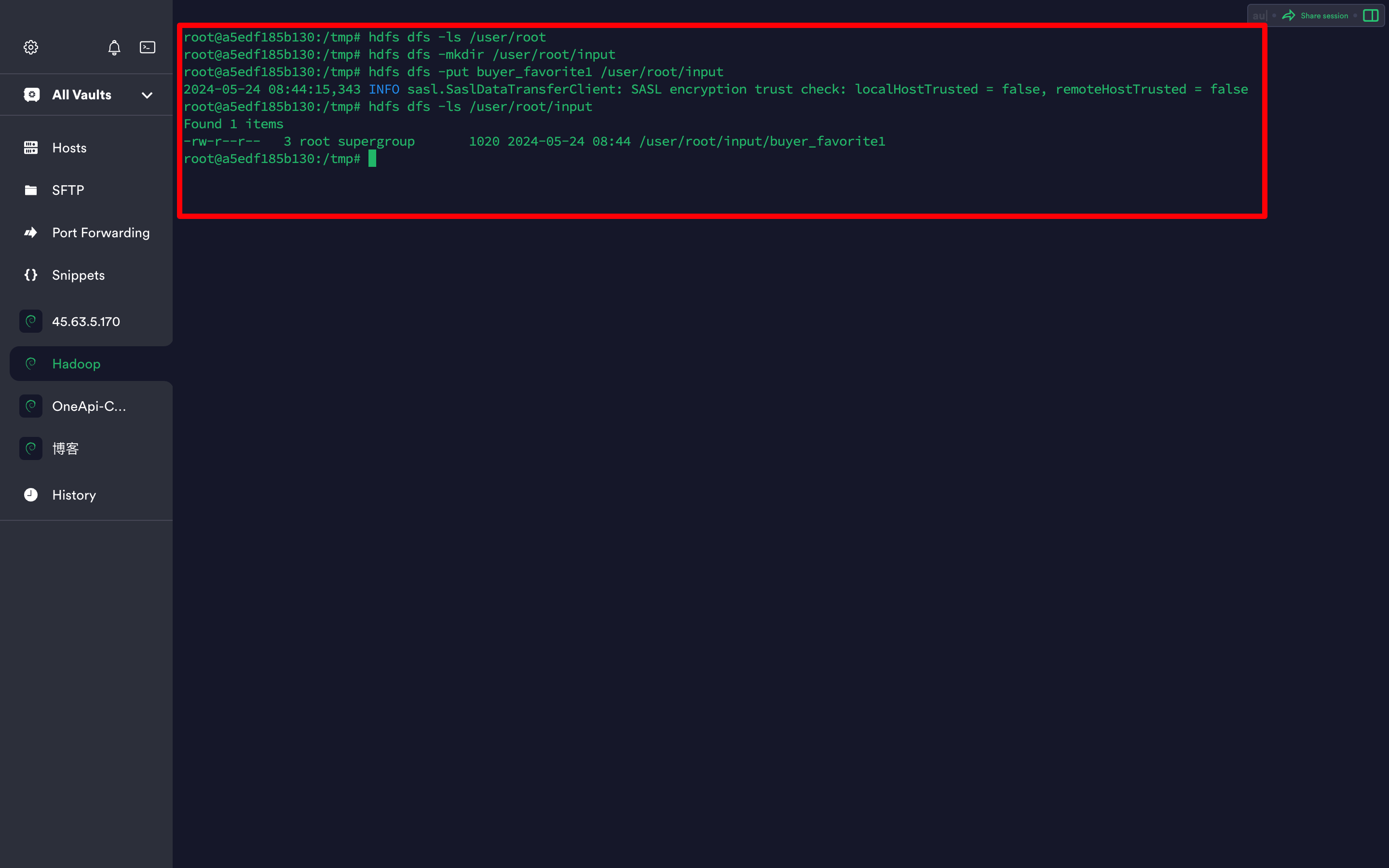
1. 使用sfpt工具上传文件至云上



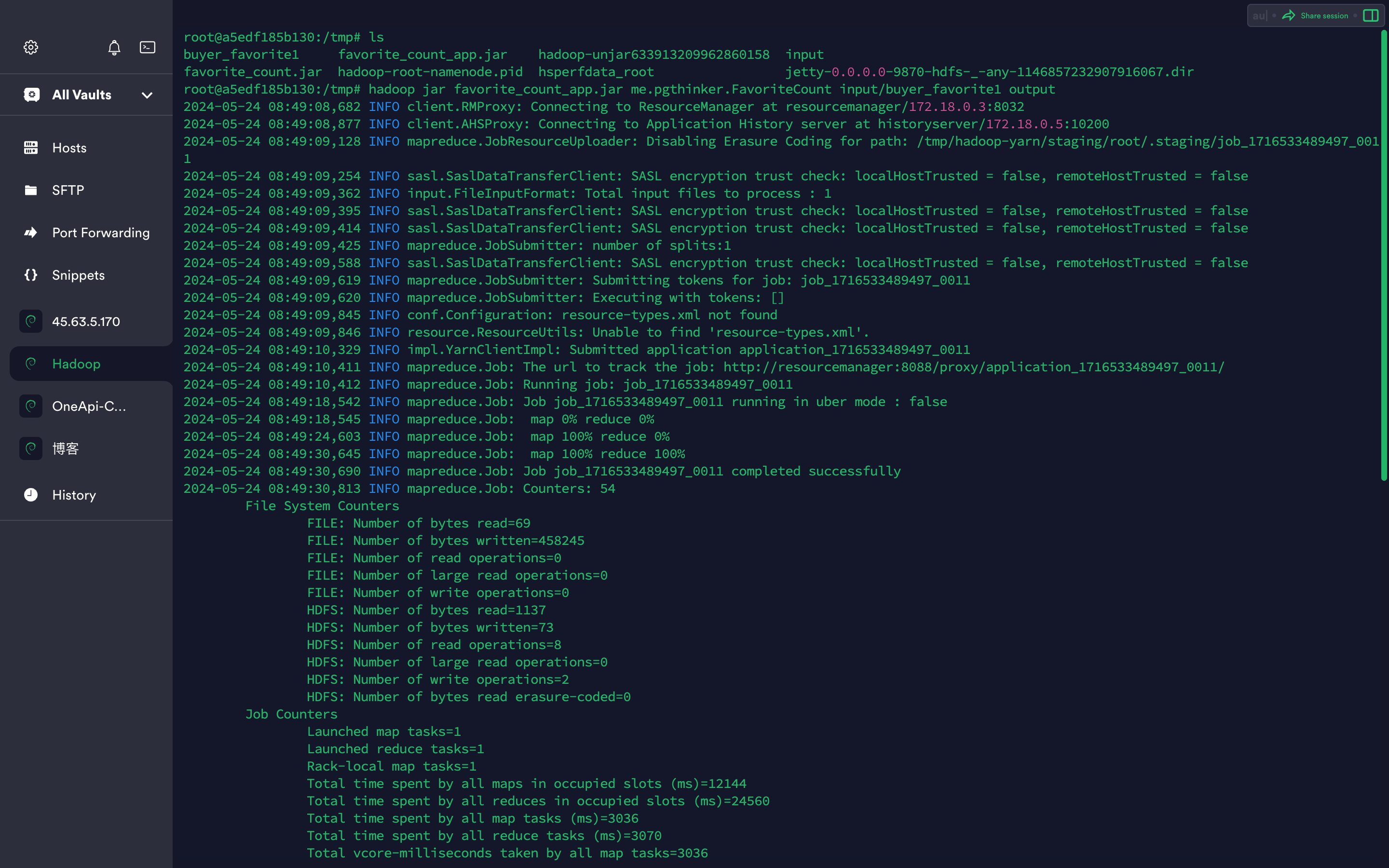
1. 使用docker命令将文件拷贝至namenode容器中



1. 使用hdfs命令创建input目录并上传文件



1. 使用hadoop命令指定输入输出目录运行jar包程序



1. 使用hdfs查看输出结果

