实验题目

数据获取和预处理：Sqoop

实验要求

【作业要求】：

1、安装sqoop

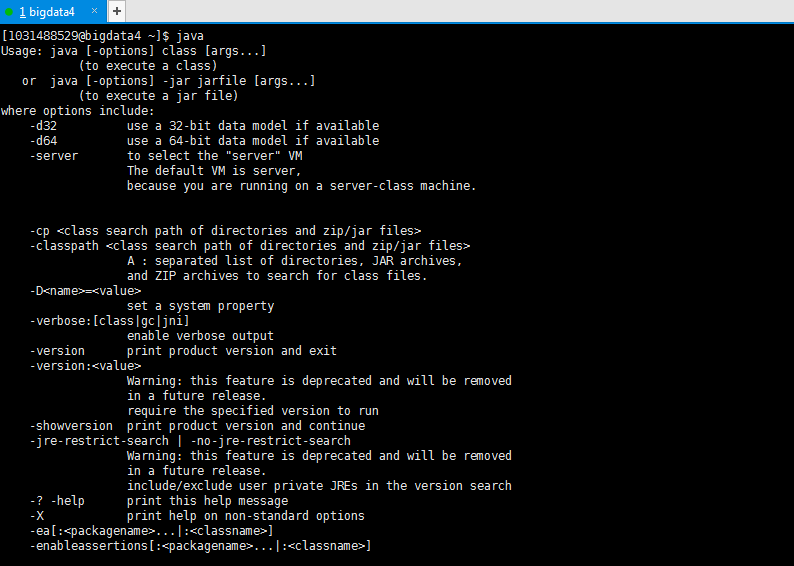
2、配置课程视频中给定的数据库连接，通过sqoop将数据导入到hive表中

3、查询hive表product表中拥有多少个课程

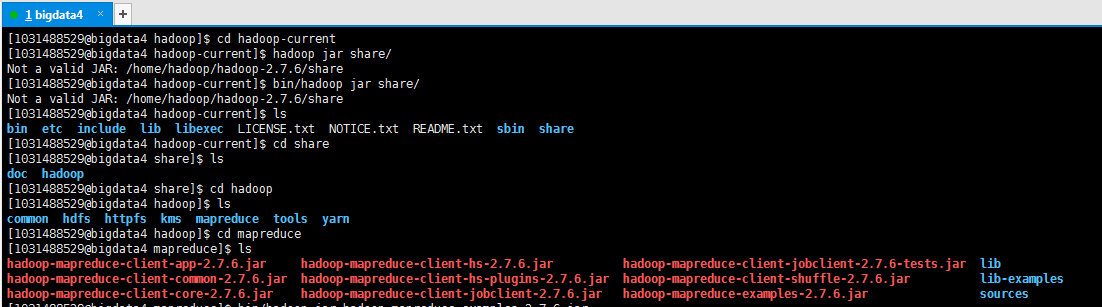
操作步骤

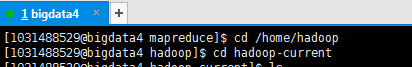
1，安装sqoop

1. 验证java依赖：ok

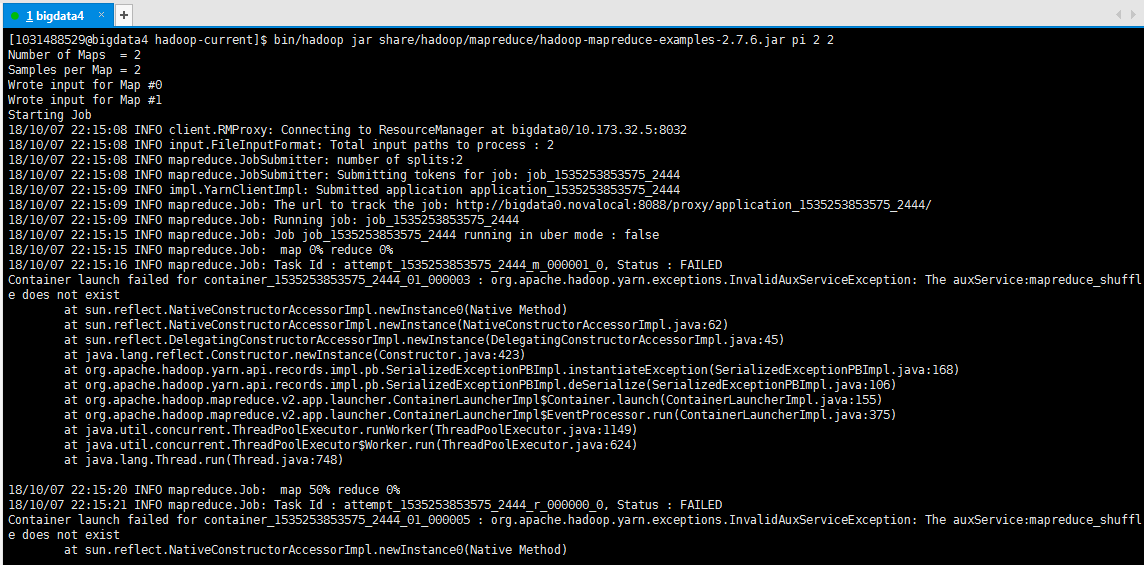


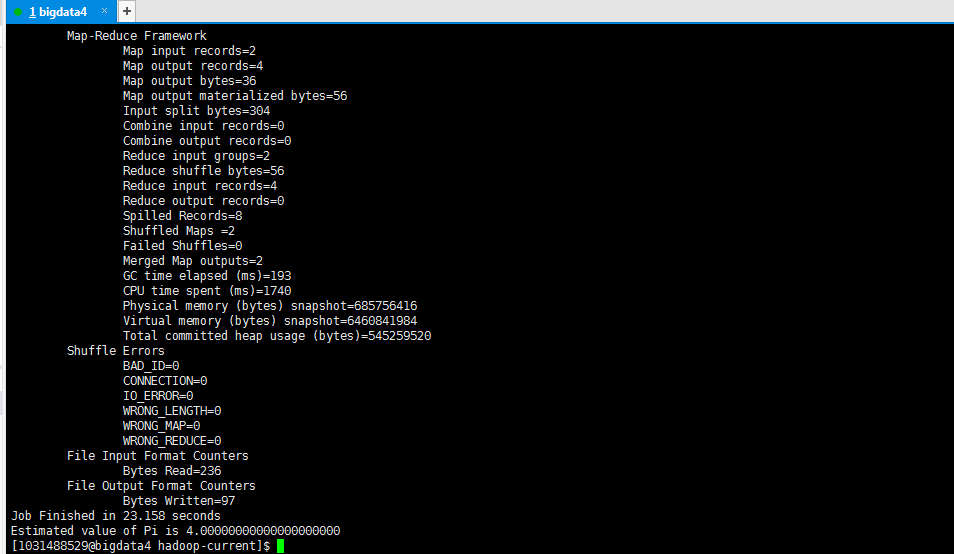
1. 验证hadoop依赖：ok





[1031488529@bigdata4 hadoop-current]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.6.jar pi 2 2

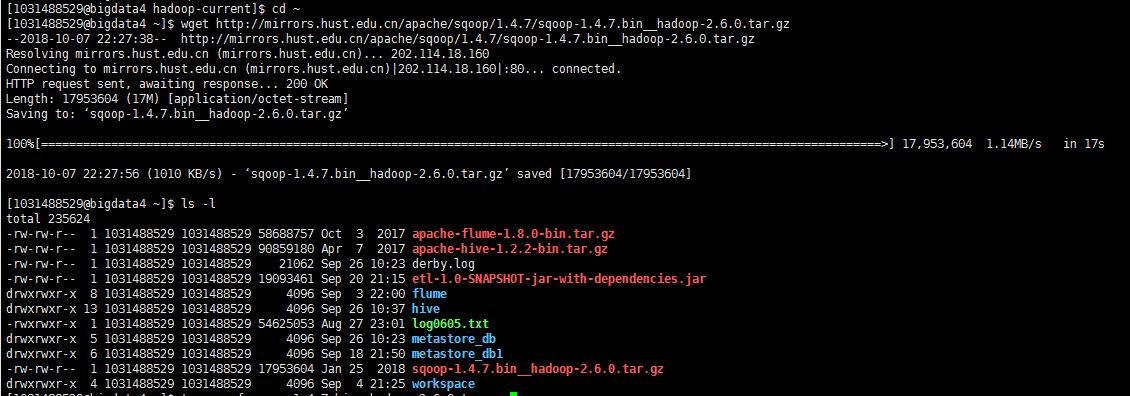




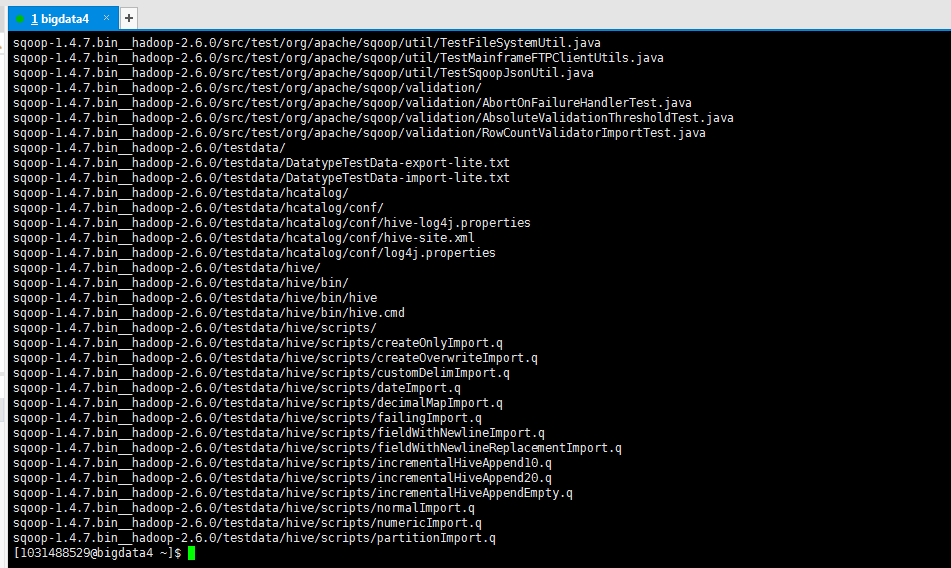
1. 下载sqoop安装包：

连接地址: <http://mirrors.hust.edu.cn/apache/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz>

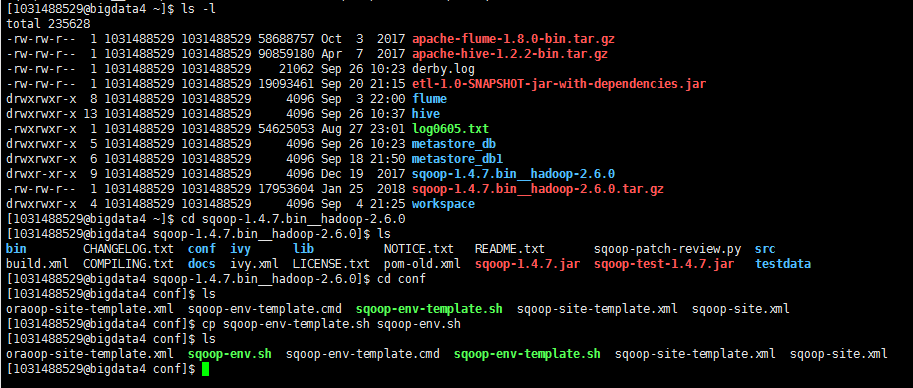
cd ~进入自己的根目录再下载：

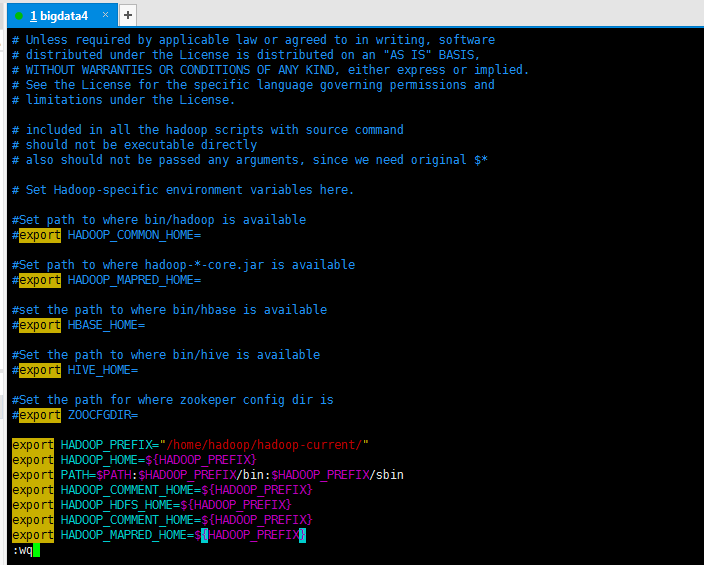


1. [1031488529@bigdata4 ~]$ tar -zxvf sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz 解压tar包

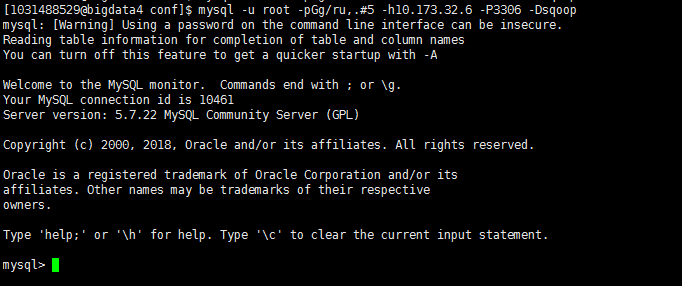


1. 配置sqoop-env.sh

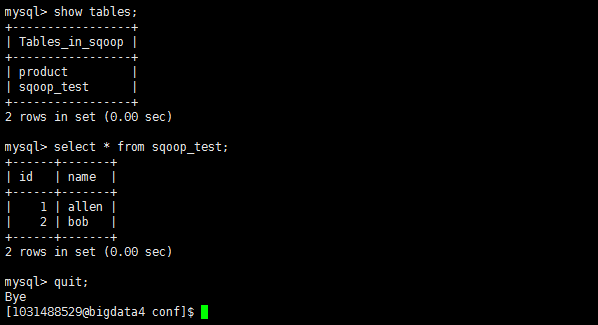




1. 连接mysql:



其中有product表，后续把这个表导入到hive的表中；

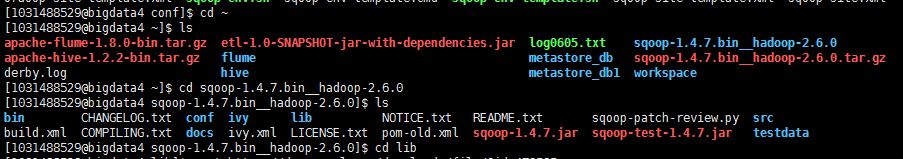


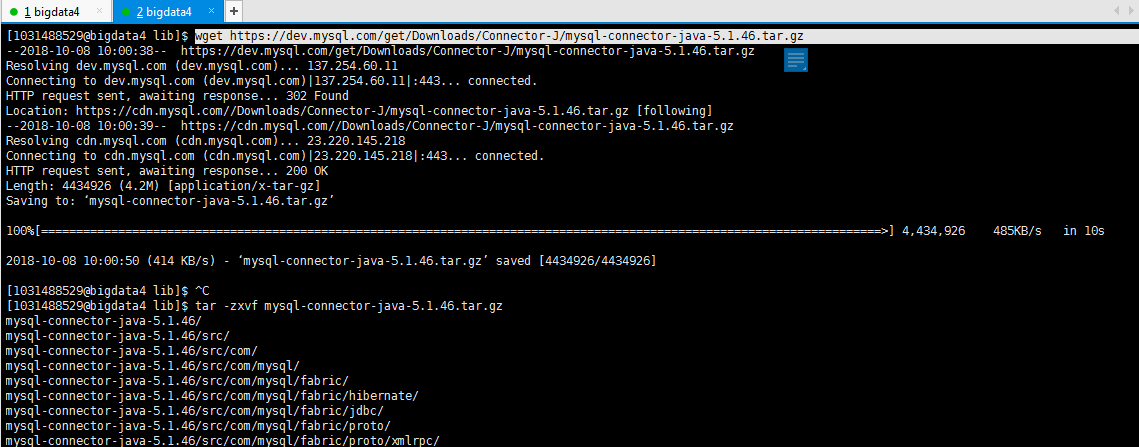
1. 下载jdbc:

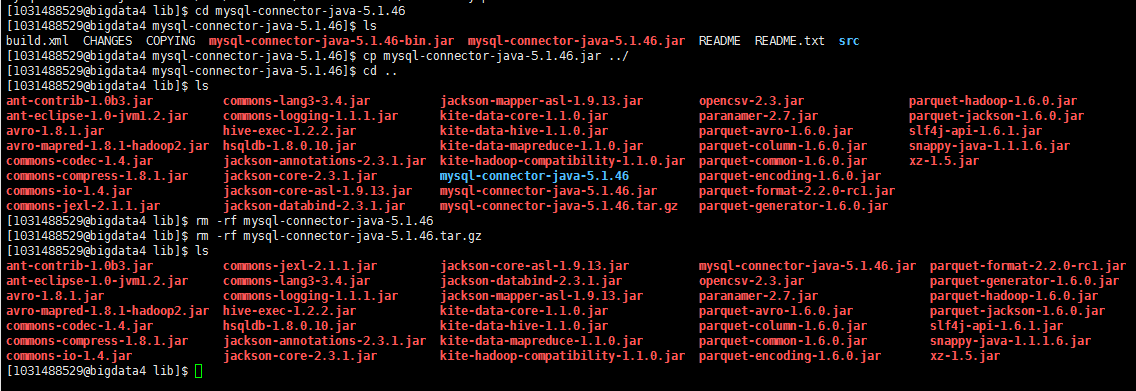
进入到自己的sqoop的lib目录再下载

[1031488529@bigdata4 sqoop-1.4.7.bin\_\_hadoop-2.6.0]$ cd lib

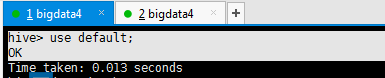
[1031488529@bigdata4 lib]$ wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-5.1.46.tar.gz

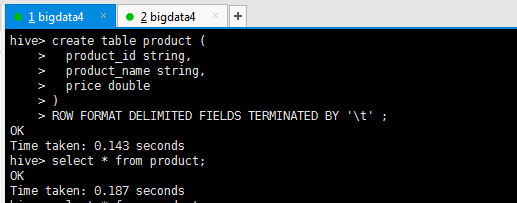






1. Sqoop从mysql导数据到Hive表
2. hive建内部表product，初始无数据

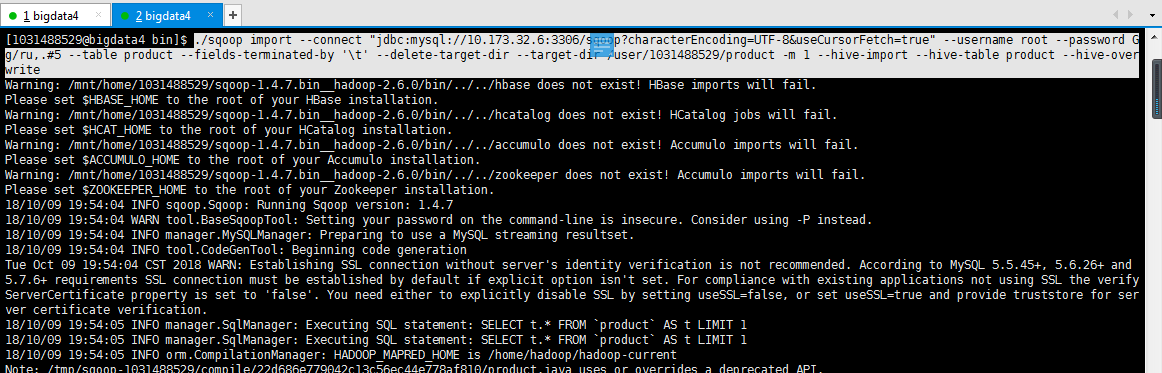


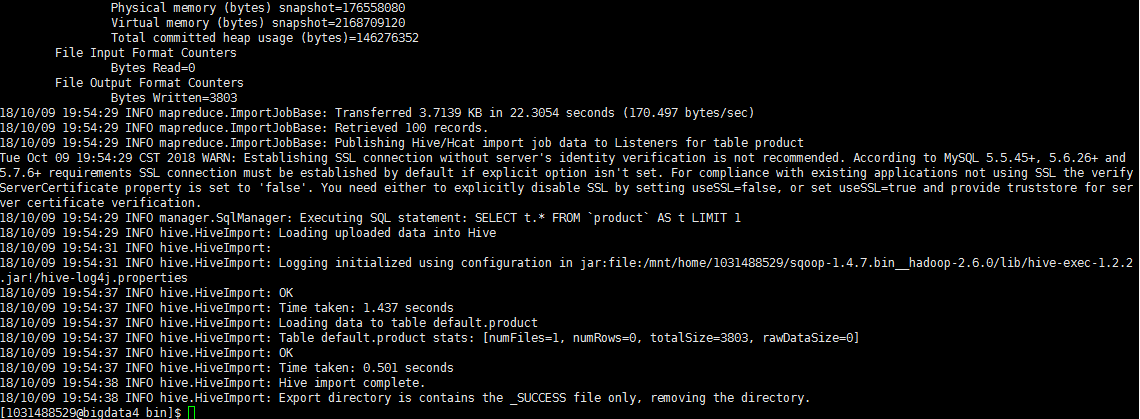


1. 执行sqoop导入命令

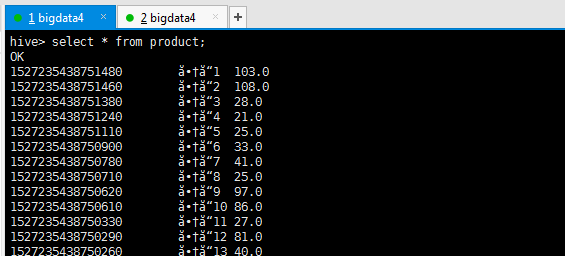
先进入sqoop-1.4.7.bin\_\_hadoop-2.6.0/bin目录，再执行命令：

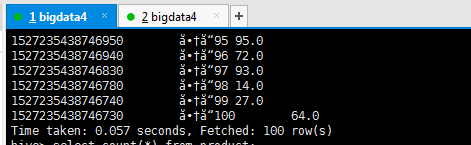
./sqoop import --connect "jdbc:mysql://10.173.32.6:3306/sqoop?characterEncoding=UTF-8&useCursorFetch=true" --username root --password Gg/ru,.#5 --table product --fields-terminated-by '\t' --delete-target-dir --target-dir /user/1031488529/product -m 1 --hive-import --hive-table product --hive-overwrite





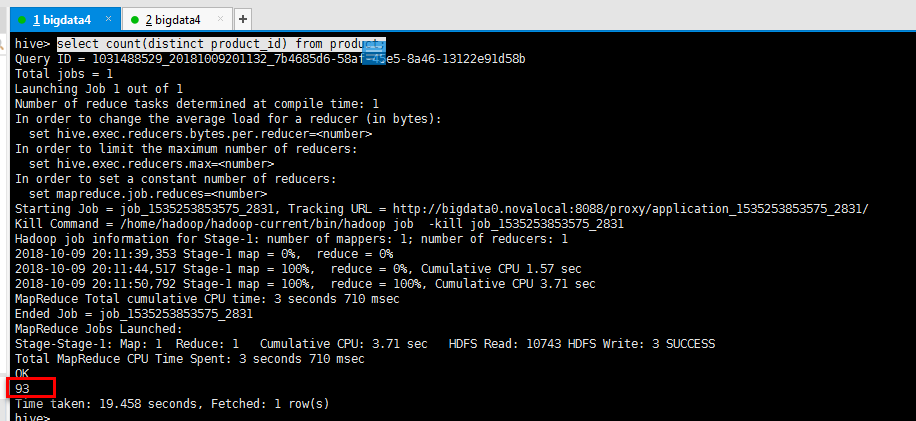
1. 验证hive的product表中已经有100条数据：



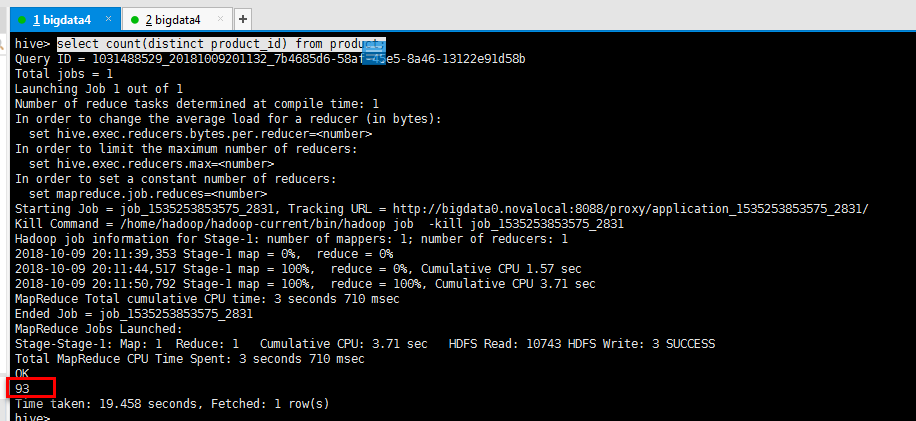


1. 查询hive表product表中拥有多少个课程

按product\_id统计去重复计数，得到结果为93，即有93个课程；



实验结果

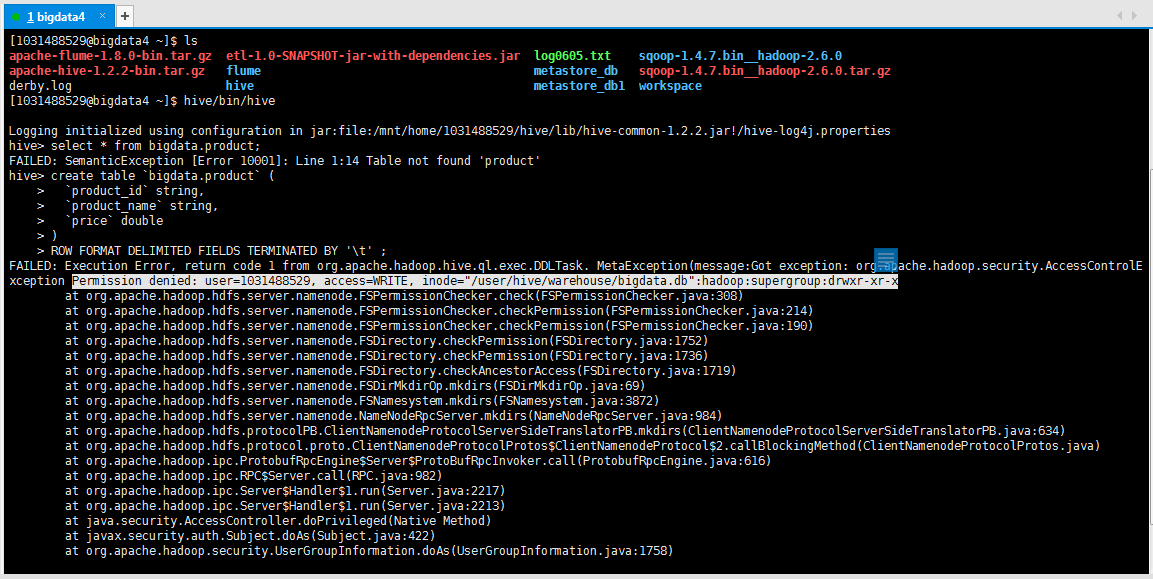


总结反思

1，mysql-connector版本最好使用5.1.46版本，版本太高会有些不兼容的风险；曾无意中安装过8.0.12版本，报了一些奇怪的异常，换回5.1.46版本就好了；

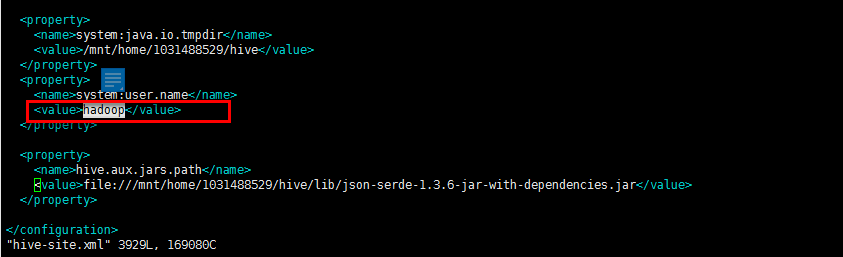
2，创建内部表报错：

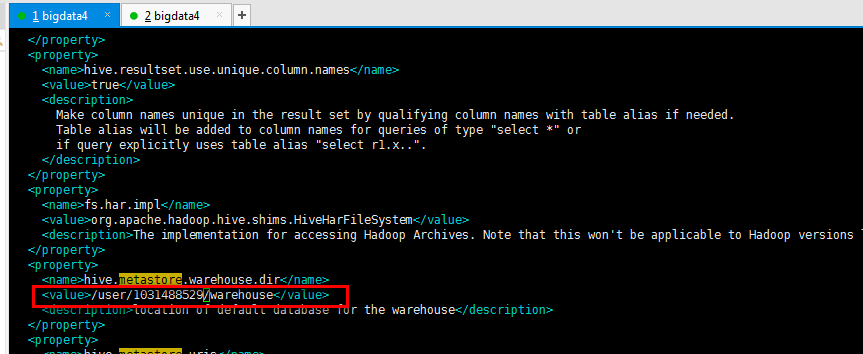
Permission denied: user=1031488529, access=WRITE, inode="/user/hive/warehouse/bigdata.db":hadoop:supergroup:drwxr-xr-x



原因是配置hive时，使用了默认的username和warehouse路径

**解决办法是修改hive-site.xml，把system:user.name下的账号名从hadoop改为自己的账号1031488529，再把hive.metastore.warehouse.dir的值改为自己的hdfs目录/user/1031488529/warehouse，然后重建hive；**





3，sqoop导入到hive表注意事项：

1. 需要先创建hive表，此时hive表不应该创建为外部表，如果使用外部表，则不需要在sqoop import语句指定hive表，只需指定输出目录，和字段分隔符，再另外创建外部表，使用匹配的字段名、字段类型、分隔符，并且location指定到sqoop输出目录即可；创建内部表，数据直接输出到内部表；
2. 需要指定split-by或m参数，否则报错，如果指定split-by参数，则要求参数指定的字段是数值型，如果字段不是数值型，只能使用m参数，比如-m 1；
3. 最好用--fields-terminated-by参数指定字段分隔符，以和hive表行格式匹配；hive表的行格式不要用上个章节的json格式，最好用’\t’，即tab分隔符；
4. 最好指定--hive-overwrite，导入时覆盖原数据，如果原来表里有数据的话；
5. 最好指定--delete-target-dir --target-dir /user/1031488529/product，否则会输出到另一个路径，未深入了解原理的情况下，不确定会输出到哪里，可能会报错；
6. 对sqoop的理解：

Sqoop就是简单说就是做数据同步的，可以在关系型数据库、hdfs和hive数据之间导数据；Sqoop 依赖MapReduce程序，环境上依赖JAVA、JDBC和hadoop；sqoop项对统的数据仓库最明显的优势是性能，通过MapReduce、数据切分、分区表等，解决性能瓶颈；