实验六：Hive 编程实践

“大数据工程”课程实验报告		
题目：Hive 编程实践	学号姓名：郭加璐	日期：2024.5.11

实验环境：

虚拟机软件：VirtualBox 7.0.14

Linux 操作系统：Ubuntu Kylin 22.04.4，虚拟机名称 UbuntuRita

Java 版本：Oracle JDK 1.8

Java IDE：Eclipse

Hadoop：3.1.3

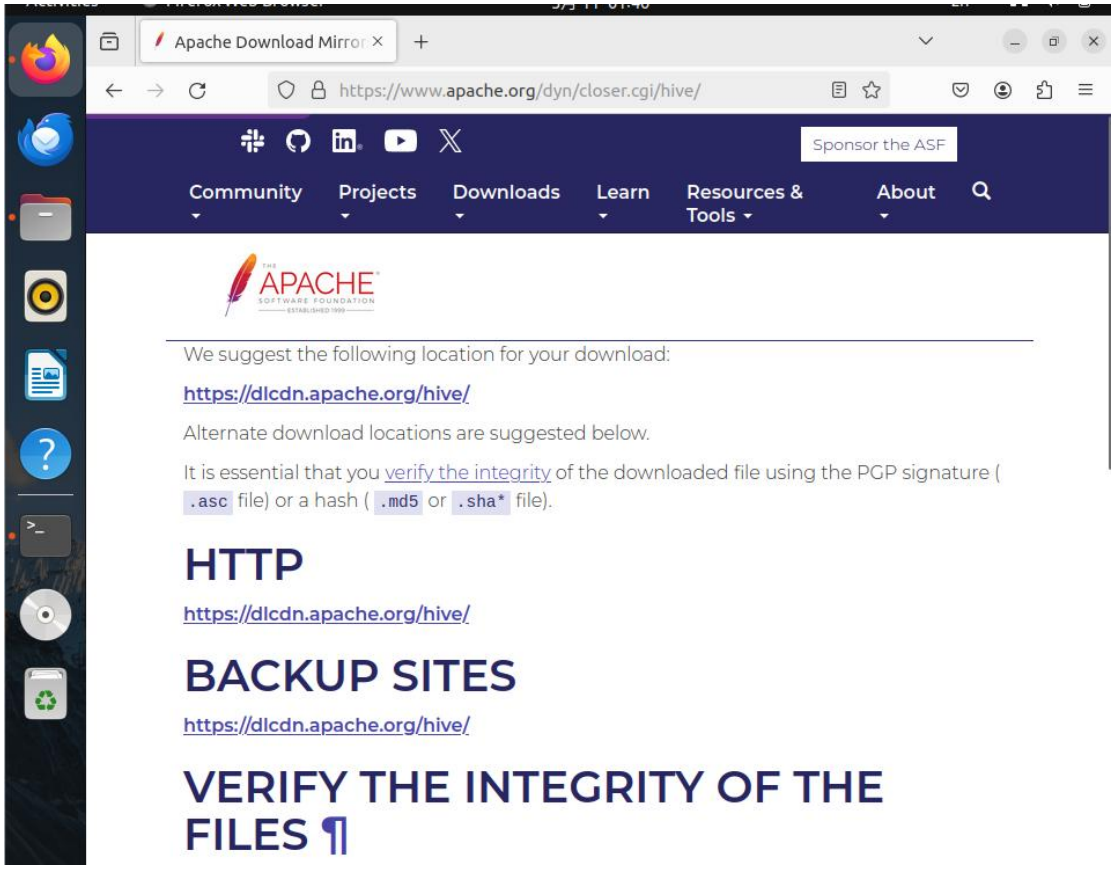
Hive：3.1.3

实验内容与完成情况：

一、安装部署 Hive

1. 下载 Hive

在 Apache 官网（https://www.apache.org/dyn/closer.cgi/hive/）下载 Hive3.1.3 安装包文件至~/Downloads 目录下：



Name	Last modified	Size	Description
Parent Directory		-	
hive-2.3.10/	2024-05-09 15:41	-	
hive-2.3.9/	2022-06-17 12:34	-	
hive-3.1.3/	2022-06-17 12:34	-	
hive-4.0.0/	2024-03-29 10:42	-	
hive-standalone-metastore-3.0.0/	2022-06-17 12:34	-	
hive-storage-2.7.3/	2022-06-17 12:34	-	
hive-storage-2.8.1/	2022-06-17 12:34	-	
stable-2/	2022-06-17 12:34	-	
KEYS	2024-04-20 16:41	114K	

Index of /hive/hive-3.1.3

Name	Last modified	Size	Description
Parent Directory		-	
apache-hive-3.1.3-bin.tar.gz	2022-04-08 17:42	312M	
apache-hive-3.1.3-bin.tar.gz.asc	2022-04-08 17:42	488	
apache-hive-3.1.3-bin.tar.gz.sha256	2022-04-08 17:42	95	
apache-hive-3.1.3-src.tar.gz	2022-04-08 17:42	25M	
apache-hive-3.1.3-src.tar.gz.asc	2022-04-08 17:42	488	
apache-hive-3.1.3-src.tar.gz.sha256	2022-04-08 17:42	95	

解压安装包至/usr/local 中，修改文件名和文件权限：

```
hadoop@UbuntuRita: ~/Downloads
hadoop@UbuntuRita:~$ cd Downloads
hadoop@UbuntuRita:~/Downloads$ ls
apache-hive-3.1.3-bin.tar.gz  eclipse-installer
hadoop@UbuntuRita:~/Downloads$ sudo tar -zxvf ./apache-hive-3.1.3-bin.tar.gz -C
/usr/local
[sudo] password for hadoop:
apache-hive-3.1.3-bin/LICENSE
apache-hive-3.1.3-bin/RELEASE_NOTES.txt
apache-hive-3.1.3-bin/NOTICE
apache-hive-3.1.3-bin/binary-package-licenses/com.thoughtworks.paranamer-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.codehaus.janino-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.jamon.jamon-runtime-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.mozilla.rhino-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.jruby-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/jline-LICENSE
apache-hive-3.1.3-bin/binary-package-licenses/org.antlr-LICENSE
```

```
hadoop@UbuntuRita:~/Downloads$ cd /usr/local/
hadoop@UbuntuRita:/usr/local$ sudo mv apache-hive-3.1.3-bin hive
hadoop@UbuntuRita:/usr/local$ sudo chown -R hadoop:hadoop hive
```

2. 配置 Hive 环境

使用 vim 打开.bashrc 文件，将 hive 命令加入到环境变量中：

```
hadoop@UbuntuRita:~$ vim ~/.bashrc
hadoop@UbuntuRita:~$ source ~/.bashrc
hadoop@UbuntuRita:~$
```

```
hadoop@UbuntuRita: ~
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_162
export JRE_HOME=${JAVA_HOME}/jre
export CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
export PATH=${JAVA_HOME}/bin:$PATH
export PATH=$PATH:/usr/local/hbase/bin
export HIVE_HOME=/usr/local/hive
export PATH=$PATH:$HIVE_HOME/bin
export HADOOP_HOME=/usr/local/hadoop
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
```

3. 修改 hive-site.xml

使用 vim 编辑器新建一个配置文件 hive-site.xml，添加配置信息：

```
hadoop@UbuntuRita:~$ cd /usr/local/hive/conf
hadoop@UbuntuRita:/usr/local/hive/conf$ mv hive-default.xml.template hive-defaul
t.xml
hadoop@UbuntuRita:/usr/local/hive/conf$ vim hive-site.xml

/bin/bash: line 1: q: command not found

shell returned 127

Press ENTER or type command to continue
hadoop@UbuntuRita:/usr/local/hive/conf$ ls
beeline-log4j2.properties.template    hive-site.xml
hive-default.xml                      ivysettings.xml
hive-env.sh.template                  llap-cli-log4j2.properties.template
hive-exec-log4j2.properties.template  llap-daemon-log4j2.properties.template
hive-log4j2.properties.template       parquet-logging.properties
hadoop@UbuntuRita:/usr/local/hive/conf$ vim hive-site.xml
hadoop@UbuntuRita:/usr/local/hive/conf$
```

```
hadoop@UbuntuRita: /usr/local/hive/conf
<configuration>
  <property>
    <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true&useSSL
=false</value>
    <description>JDBC connect string for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionDriverName</name>
    <value>com.mysql.jdbc.Driver</value>
    <description>Driver class name for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionUserName</name>
    <value>hive</value>
    <description>username to use against metastore database</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionPassword</name>
    <value>hive</value>
    <description>password to use against metastore database</description>
  </property>
</configuration>
:wq
```

4. 安装并配置 mysql

虚拟机上已安装 mysql，启动并登录 MySQL Shell：

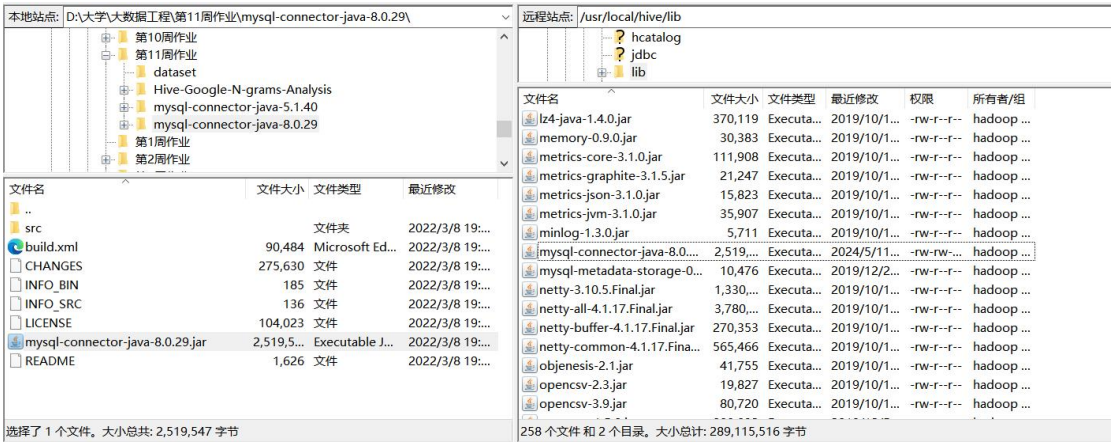
```
hadoop@UbuntuRita:~$ service mysql start
hadoop@UbuntuRita:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 11
Server version: 8.0.36-0ubuntu0.22.04.1 (Ubuntu)

Copyright (c) 2000, 2024, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

下载 JDBC 驱动程序（注意要和 mysql 版本匹配），复制到 Hive 的 lib 目录：



新建 Hive 数据库，配置 MySQL 允许 Hive 接入：

```
mysql> create user 'hive'@'localhost' identified by 'hive';
Query OK, 0 rows affected (0.45 sec)

mysql> grant all on *.* to 'hive'@'localhost';
Query OK, 0 rows affected (0.15 sec)
```

5. 升级元数据

这一步可能出现的报错及其解决方法详见“出现的问题”部分。

```
hadoop@UbuntuRita:/usr/local/hive$ ./bin/schematool -initSchema -dbType mysql --verbosecd
Metastore connection URL:        jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true&u
seSSL=false&allowPublicKeyRetrieval=true
Metastore Connection Driver :    com.mysql.jdbc.Driver
Metastore connection User:       hive
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.
jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver
 class is generally unnecessary.
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.mysql.sql
```

```
Initialization script completed
schemaTool completed
hadoop@UbuntuRita:/usr/local/hive$
```

6. 启动 Hive

需要首先启动 hadoop 集群，再启动 hive：

```
hadoop@UbuntuRita:~$ cd /usr/local/hadoop
hadoop@UbuntuRita:/usr/local/hadoop$ ./sbin/start-dfs.sh
```

```
hadoop@UbuntuRita:/usr/local/hadoop$ jps
7665 Jps
3153 SecondaryNameNode
3000 DataNode
2894 NameNode
```

```
hadoop@UbuntuRita:/usr/local/hadoop$ cd /usr/local/hive
hadoop@UbuntuRita:/usr/local/hive$ ./bin/hive
/usr/local/hadoop/libexec/hadoop-functions.sh: line 2360: HADOOP_ORG.APACHE.HADOOP.HBASE.UTIL.GET
JAVAPROPERTY_USER: invalid variable name
/usr/local/hadoop/libexec/hadoop-functions.sh: line 2455: HADOOP_ORG.APACHE.HADOOP.HBASE.UTIL.GET
JAVAPROPERTY_OPTS: invalid variable name
2024-05-11 15:46:52,611 INFO conf.HiveConf: Found configuration file file:/usr/local/hive/conf/hi
ve-site.xml
2024-05-11 15:46:53,529 WARN common.LogUtils: DEPRECATED: Ignoring hive-default.xml found on the
CLASSPATH at /usr/local/hive/conf/hive-default.xml
Hive Session ID = 1b6ef9ae-2e14-48a5-b885-50cbf329056b
2024-05-11 15:46:54,342 INFO SessionState: Hive Session ID = 1b6ef9ae-2e14-48a5-b885-50cbf329056b

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.3.jar!/hi
ve-log4j2.properties Async: true
2024-05-11 15:46:54,959 INFO SessionState:
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.3.jar!/hi
```

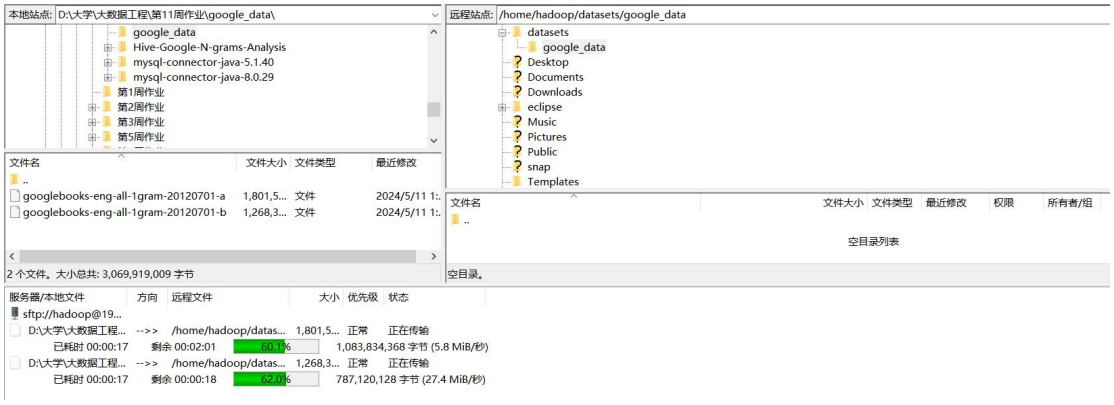
```
2024-05-11 15:47:16,421 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=get_t
ables_by_type: db=@hive#default pat=.*,type=MATERIALIZED_VIEW
2024-05-11 15:47:16,476 INFO metastore.HiveMetaStore: 1: get_multi_table : db=default tbls=
2024-05-11 15:47:16,476 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=get_m
ulti_table : db=default tbls=
2024-05-11 15:47:16,491 INFO metadata.HiveMaterializedViewsRegistry: Materialized views registry
has been initialized
hive>
```

启动成功。

二、上传数据并合并到一张 Hive 表中

1. 上传数据到 HDFS

首先，将两个数据文件传输到虚拟机：



在 HDFS 中创建新的文件夹/user/hive/ngdata，将两个数据文件上传到该文件夹下：

```
hadoop@UbuntuRita:/usr/local/hadoop$ ./bin/hdfs dfs -mkdir -p ./user/hive/ngdatahadoop@UbuntuRita:/
usr/local/hadoop$ ./bin/hdfs dfs -ls
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2024-04-01 17:27 MovieUserRatingsInfo
drwxr-xr-x   - hadoop supergroup          0 2024-03-13 00:18 input
drwxr-xr-x   - hadoop supergroup          0 2024-03-13 23:16 test
drwxr-xr-x   - hadoop supergroup          0 2024-05-11 21:18 user
drwxr-xr-x   - hadoop supergroup          0 2024-03-31 20:27 week5
hadoop@UbuntuRita:/usr/local/hadoop$ ./bin/hdfs dfs -put ~/datasets/google_data/googlebooks-eng-all
-1gram-20120701-a ./user/hive/ngdata
2024-05-11 21:18:45,788 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:46,367 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:46,694 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:47,545 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:48,653 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:52,475 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:18:55,549 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
```

```
hadoop@UbuntuRita:/usr/local/hadoop$ ./bin/hdfs dfs -put ~/datasets/google_data/googlebooks-eng-all
-1gram-20120701-b ./user/hive/ngdata
2024-05-11 21:21:13,811 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:21:17,831 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:21:22,091 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:21:26,878 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:21:45,311 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:22:01,396 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:22:06,035 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:22:10,961 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:22:35,508 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
2024-05-11 21:22:45,650 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTru
sted = false, remoteHostTrusted = false
```

```
hadoop@UbuntuRita:/usr/local/hadoop$ ./bin/hdfs dfs -ls ./user/hive/ngdata
Found 2 items
-rw-r--r--   1 hadoop supergroup 1801526075 2024-05-11 21:20 user/hive/ngdata/googlebooks-eng-all-1
gram-20120701-a
-rw-r--r--   1 hadoop supergroup 1268392934 2024-05-11 21:22 user/hive/ngdata/googlebooks-eng-all-1
gram-20120701-b
```

2. 创建 Hive 表

```
hive>
    > create table if not exists ngrams(bigram STRING,
    >     year INT,
    >     match_count INT,
    >     volume_count INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
2024-05-11 16:54:15,613 INFO conf.HiveConf: Using the default value passed in for log id: 51b3ee57-
da25-4237-8323-54b44d46ed56
2024-05-11 16:54:17,616 INFO ql.Driver: Compiling command(queryId=hadoop_20240511165416_d499cee9-92
56-442b-830f-ed9eefd884c8): create table if not exists ngrams(bigram STRING,
    year INT,
    match_count INT,
    volume_count INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
2024-05-11 16:54:18,464 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
```

3. 加载数据到 Hive 表

```
hive> load data inpath '/user/hive/ngdata' into table ngrams;
2024-05-11 17:04:32,158 INFO conf.HiveConf: Using the default value passed in for log id: 51b3ee57-
da25-4237-8323-54b44d46ed56
2024-05-11 17:04:32,158 INFO session.SessionState: Updating thread name to 51b3ee57-da25-4237-8323-
54b44d46ed56 main
2024-05-11 17:04:32,160 INFO ql.Driver: Compiling command(queryId=hadoop_20240511170432_c3545c53-9b
bd-47b8-8b07-08ec2a4f19a8): load data inpath '/user/hive/ngdata' into table ngrams
2024-05-11 17:04:32,173 INFO metastore.HiveMetaStoreClient: Mestastore configuration metastore.filt
er.hook changed from org.apache.hadoop.hive.metastore.DefaultMetaStoreFilterHookImpl to org.apache.
hadoop.hive.ql.security.authorization.plugin.AuthorizationMetaStoreFilterHook
2024-05-11 17:04:32,173 INFO metastore.HiveMetaStore: 0: Cleaning up thread local RawStore...
2024-05-11 17:04:32,173 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=Cleanin
g up thread local RawStore...
2024-05-11 17:04:32,173 INFO metastore.HiveMetaStore: 0: Done cleaning up thread local RawStore
2024-05-11 17:04:32,173 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=Done cl
eaning up thread local RawStore
2024-05-11 17:04:32,173 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-05-11 17:04:32,378 INFO metastore.HiveMetaStore: 0: Opening raw store with implementation clas
s:org.apache.hadoop.hive.metastore.ObjectStore
2024-05-11 17:04:32,378 WARN metastore.ObjectStore: datanucleus.autoStartMechanismMode is set to un
supported value null . Setting it to value: ignored
2024-05-11 17:04:32,378 INFO metastore.ObjectStore: ObjectStore, initialize called
2024-05-11 17:04:32,387 INFO metastore.MetaStoreDirectSql: Using direct SQL, underlying DB is MYSQL
2024-05-11 17:04:32,387 INFO metastore.ObjectStore: Initialized ObjectStore
2024-05-11 17:04:32,387 INFO metastore.RetryingMetaStoreClient: RetryingMetaStoreClient proxy=class
 org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient ugi=hadoop (auth:SIMPLE) retries=1 d
elay=1 lifetime=0
2024-05-11 17:04:32,388 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.ngrams
2024-05-11 17:04:32,388 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=get_tab
```

查看部分载入结果：

```
hive> select * from ngrams limit 30;
2024-05-11 17:06:37,351 INFO conf.HiveConf: Using the default value passed in for log id: 51b3ee57-
da25-4237-8323-54b44d46ed56
2024-05-11 17:06:37,351 INFO session.SessionState: Updating thread name to 51b3ee57-da25-4237-8323-
54b44d46ed56 main
2024-05-11 17:06:37,352 INFO ql.Driver: Compiling command(queryId=hadoop_20240511170637_983d3bd2-8d
ef-47a4-86d2-7db9416a2c2c): select * from ngrams limit 30
2024-05-11 17:06:37,362 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-05-11 17:06:37,362 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-05-11 17:06:37,362 INFO parse.CalcitePlanner: Completed phase 1 of Semantic Analysis
2024-05-11 17:06:37,362 INFO parse.CalcitePlanner: Get metadata for source tables
2024-05-11 17:06:37,362 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.ngrams
2024-05-11 17:06:37,362 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=get_tab
le : tbl=hive.default.ngrams
2024-05-11 17:06:37,403 INFO parse.CalcitePlanner: Get metadata for subqueries
2024-05-11 17:06:37,403 INFO parse.CalcitePlanner: Get metadata for destination tables
2024-05-11 17:06:37,413 INFO parse.CalcitePlanner: Completed getting MetaData in Semantic Analysis
```

```
sted = false, remoteHostTrusted = false
A'Aang_NOUN     1879    45      5
A'Aang_NOUN     1882    5       4
A'Aang_NOUN     1885    1       1
A'Aang_NOUN     1891    1       1
A'Aang_NOUN     1899    20      4
A'Aang_NOUN     1927    3       1
A'Aang_NOUN     1959    5       2
A'Aang_NOUN     1962    2       2
A'Aang_NOUN     1963    1       1
A'Aang_NOUN     1966    45      13
A'Aang_NOUN     1967    6       4
A'Aang_NOUN     1968    5       4
A'Aang_NOUN     1970    6       2
A'Aang_NOUN     1975    4       1
A'Aang_NOUN     2001    1       1
A'Aang_NOUN     2004    3       1
A'que_ADJ       1808    1       1
```

```
A'que_ADJ       1849    2       1
A'que_ADJ       1850    1       1
A'que_ADJ       1852    4       3
A'que_ADJ       1854    5       3
A'que_ADJ       1856    2       1
A'que_ADJ       1858    4       3
A'que_ADJ       1862    2       1
A'que_ADJ       1871    1       1
A'que_ADJ       1872    2       2
A'que_ADJ       1873    1       1
A'que_ADJ       1874    2       2
A'que_ADJ       1875    3       3
A'que_ADJ       1877    1       1
2024-05-11 17:06:37,584 INFO exec.TableScanOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERAT
OR_TS_0:30,
2024-05-11 17:06:37,584 INFO exec.SelectOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_
SEL_1:30,
```

三、计算每个独特的 bigram 的平均出现次数

使用下面的命令完成：通过 HiveQL 查询计算每个 bigram 出现的平均次数，将结果保存到新的 hive 表 bigram_averages 中。

```
hive> CREATE TABLE IF NOT EXISTS bigram_averages (
    >     bigram STRING,
    >     average_match_count FLOAT
    > )
    > STORED AS ORC;
2024-05-11 22:55:54,624 INFO conf.HiveConf: Using the default value passed in for log id: ccfca362-
0fe7-4478-926e-64720ee3fd13
2024-05-11 22:55:54,713 INFO ql.Driver: Compiling command(queryId=hadoop_20240511225554_835689f5-41
36-43c4-922b-1aa5d0be672a): CREATE TABLE IF NOT EXISTS bigram_averages (
    bigram STRING,
    average_match_count FLOAT
)
STORED AS ORC
2024-05-11 22:55:55,234 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-05-11 22:55:55,237 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-05-11 22:55:55,252 INFO sqlstd.SQLStdHiveAccessController: Created SQLStdHiveAccessController
for session context : HiveAuthzSessionContext [sessionString=ccfca362-0fe7-4478-926e-64720ee3fd13,
clientType=HIVECLI]
2024-05-11 22:55:55,254 WARN session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hiv
e.security.authorization.manager is set to instance of HiveAuthorizerFactory.
2024-05-11 22:55:55,254 INFO metastore.HiveMetaStoreClient: Mestastore configuration metastore.filt
er.hook changed from org.apache.hadoop.hive.metastore.DefaultMetaStoreFilterHookImpl to org.apache.
hadoop.hive.ql.security.authorization.plugin.AuthorizationMetaStoreFilterHook
2024-05-11 22:55:55,258 INFO metastore.HiveMetaStore: 0: Cleaning up thread local RawStore...
2024-05-11 22:55:55,259 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=Cleanin
g up thread local RawStore...
2024-05-11 22:55:55,259 INFO metastore.HiveMetaStore: 0: Done cleaning up thread local RawStore
2024-05-11 22:55:55,259 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=Done cl
eaning up thread local RawStore
2024-05-11 22:55:55,261 INFO metastore.HiveMetaStore: 0: Opening raw store with implementation clas
s:org.apache.hadoop.hive.metastore.ObjectStore
2024-05-11 22:55:55,261 WARN metastore.ObjectStore: datanucleus.autoStartMechanismMode is set to un
```

由于计算量过大，在本地虚拟机上无法实现计算，使用 ngrams 表中前 100000 条数据进行实验：

```
hive> INSERT INTO TABLE bigram_averages
    > SELECT
    >     bigram,
    >     SUM(match_count) / COUNT(DISTINCT year) AS average_match_count
    > FROM (
    >     SELECT bigram, match_count, year
    >     FROM ngrams
    >     LIMIT 100000
    > ) AS limited_ngrams
    > GROUP BY bigram;
2024-05-11 22:57:57,739 INFO conf.HiveConf: Using the default value passed in for log id: 72b818f1-
3a9e-45d4-ac1c-cdf79916e7a4
2024-05-11 22:57:57,822 INFO ql.Driver: Compiling command(queryId=hadoop_20240511225757_92b1b502-58
6b-49e8-91c2-c38245682a36): INSERT INTO TABLE bigram_averages
SELECT
    bigram,
    SUM(match_count) / COUNT(DISTINCT year) AS average_match_count
FROM (
    SELECT bigram, match_count, year
    FROM ngrams
    LIMIT 100000
) AS limited_ngrams
GROUP BY bigram
2024-05-11 22:57:58,332 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-05-11 22:57:58,335 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-05-11 22:57:58,350 INFO sqlstd.SQLStdHiveAccessController: Created SQLStdHiveAccessController
for session context : HiveAuthzSessionContext [sessionString=72b818f1-3a9e-45d4-ac1c-cdf79916e7a4,
clientType=HIVECLI]
```

通过 select * from bigram_averages limit 30; 查看结果：

```
sted = false, remoteHostTrusted = false
2024-05-11 22:48:14,558 INFO orc.OrcInputFormat: FooterCacheHitRatio: 0/1
2024-05-11 22:48:14,564 INFO orc.ReaderImpl: Reading ORC rows from hdfs://localhost:9000/user/hive/
warehouse/bigram_averages/000000_0 with {include: [true, true, true], offset: 3, length: 5232, sche
ma: struct<bigram:string,average_match_count:float>, includeAcidColumns: true}
A'3     12.3046875
A'ditya_NOUN    6.090909
A'ews   4.7647057
A'f_VERB        1.3870968
A'ishah_NOUN    120.94936
A.06    7.695652
A.140_NOUN      2.45
A.2A_NOUN       5.212766
A.3A_NOUN       3.483871
A.5.6_NOUN      9.488372
A.A.U_NOUN      1.7804878
A.B.Dick        3.3953488
A.D.139 1.6666666
A.F.A_NOUN      1.8863636
A.Gould_NOUN    2.0789473
A.I.I.C._NOUN   3.5
A.I.K.  5.3030305
A.IX_NUM        3.5833333
A.K.H.B.        4.9382715
A.L.B._NOUN     24.333334
A.L.s._NUM      10.965517
A.Mills_NOUN    2.8823528
A.R.P.S.        32.3625
A.S     54.46383
A.T.C   6.2739725
A.W.F._NOUN     39.907692
A.b_VERB        1.7428571
A.faecalis      4.5
A.flavus_ADJ    2.871795
A.s_.   2.4555554
2024-05-11 22:48:14,597 INFO exec.TableScanOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERAT
OR_TS_0:30,
2024-05-11 22:48:14,597 INFO exec.SelectOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_
SEL_1:30,
```

四、输出每年平均出现次数最高的 20 个 bigram

使用下面的命令完成：通过 HiveQL 查询每年平均出现次数最高的 20 个 bigram，将结果保存到新的 hive 表 top_bigrams 中。

创建 hive 表：

```
hive> CREATE TABLE IF NOT EXISTS top_bigrams (
    >     year INT,
    >     bigram STRING,
    >     average_match_count FLOAT
    > )
    > STORED AS ORC;
2024-05-11 23:03:33,540 INFO conf.HiveConf: Using the default value passed in for log id: 72b818f1-
3a9e-45d4-ac1c-cdf79916e7a4
2024-05-11 23:03:33,540 INFO session.SessionState: Updating thread name to 72b818f1-3a9e-45d4-ac1c-
cdf79916e7a4 main
2024-05-11 23:03:33,541 INFO ql.Driver: Compiling command(queryId=hadoop_20240511230333_a31bfb2f-3e
6f-49ec-ba84-e0567949a396): CREATE TABLE IF NOT EXISTS top_bigrams (
    year INT,
    bigram STRING,
    average_match_count FLOAT
)
STORED AS ORC
2024-05-11 23:03:33,557 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-05-11 23:03:33,557 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-05-11 23:03:33,557 INFO parse.CalcitePlanner: Creating table default.top_bigrams position=27
2024-05-11 23:03:33,557 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.top_bigrams
2024-05-11 23:03:33,557 INFO HiveMetaStore.audit: ugi=hadoop    ip=unknown-ip-addr      cmd=get_tab
le : tbl=hive.default.top_bigrams
2024-05-11 23:03:33,575 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-05-11 23:03:33,575 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties
:null)
2024-05-11 23:03:33,575 INFO ql.Driver: Completed compiling command(queryId=hadoop_20240511230333_a
31bfb2f-3e6f-49ec-ba84-e0567949a396); Time taken: 0.034 seconds
2024-05-11 23:03:33,575 INFO reexec.ReExecDriver: Execution #1 of query
```

将查询结果插入新建的表中（由于计算量过大，在本地虚拟机上无法实现计算，使用 ngrams 表中前 100000 条数据进行实验）：

```
hive> INSERT INTO TABLE top_bigrams
    > SELECT
    >     year,
    >     bigram,
    >     average_match_count
    > FROM (
    >     SELECT
    >         year,
    >         bigram,
    >         AVG(match_count) AS average_match_count,
    >         ROW_NUMBER() OVER (PARTITION BY year ORDER BY AVG(match_count) DESC) AS rn
    >     FROM (
    >         SELECT
    >             bigram,
    >             year,
    >             match_count
    >         FROM
    >             ngrams
    >         LIMIT 100000
    >     ) AS limited
    >     GROUP BY
    >         year, bigram
    > ) AS ranked
    > WHERE rn <= 20;
2024-05-11 23:15:56,644 INFO conf.HiveConf: Using the default value passed in for log id: 72b818f1-
3a9e-45d4-ac1c-cdf79916e7a4
2024-05-11 23:15:56,644 INFO session.SessionState: Updating thread name to 72b818f1-3a9e-45d4-ac1c-
cdf79916e7a4 main
2024-05-11 23:15:56,645 INFO ql.Driver: Compiling command(queryId=hadoop_20240511231556_ad8f3ea0-94
db-4469-9e66-883e795ec5fb): INSERT INTO TABLE top_bigrams
```

通过 select * from top_bigrams limit 30; 查看结果：

```
warehouse/top_bigrams/000000_0 with {include: [true, true, true, true], offset: 0, length: 8142, sc
hema: struct<year:int,bigram:string,average_match_count:float>, includeAcidColumns: true}
1678    B't_NOUN        1.0
1700    B'e_NOUN        1.0
1707    BEDFORD_X       1.0
1712    B.oman  1.0
1722    BEDFORDSHIRE_PRON       1.0
1722    BEDFORD_X       1.0
1728    BEHOLD_.        1.0
1729    BFM_NOUN        1.0
1730    BIS_NUM 1.0
1739    BEHOLD_.        2.0
1746    BEHOLD_.        1.0
1747    BEHOLD_.        2.0
1748    B.oman  1.0
1748    BEDFORD_X       1.0
1749    BEHOLD_.        1.0
1749    BARD_VERB       1.0
1752    BARD_VERB       1.0
1755    BETTY_DET       1.0
1757    B.c.-a.d._NOUN  3.0
1757    B't_NOUN        1.0
1757    BASTIONS_NOUN   1.0
1761    BEHOLD_.        1.0
1764    BEDFORD_X       2.0
1764    BARD_VERB       1.0
1766    BEDFORD_X       1.0
1767    BEHOLD_.        1.0
1769    BEDFORD_X       3.0
1773    BEDFORD_X       2.0
1776    B.oman  1.0
1776    BIHDS   1.0
2024-05-11 23:32:34,479 INFO exec.TableScanOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERAT
OR_TS_0:30,
2024-05-11 23:32:34,479 INFO exec.SelectOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_
SEL_1:30,
2024-05-11 23:32:34,479 INFO exec.LimitOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_L
IM_2:30,
2024-05-11 23:32:34,479 INFO exec.ListSinkOperator: RECORDS_OUT_OPERATOR_LIST_SINK_4:30, RECORDS_OU
T_INTERMEDIATE:0,
Time taken: 0.16 seconds, Fetched: 30 row(s)
```

出现的问题：

1. 在配置 MySQL 允许 Hive 接入时出现错误：

```
mysql> grant all on *.* to 'hive'@localhost identified by 'hive';
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that
corresponds to your MySQL server version for the right syntax to use near 'ident
ified by 'hive'' at line 1
```

2. 报错：SLF4J: Class path contains multiple SLF4J bindings.

```
hadoop@UbuntuRita:/usr/local/hive$ ./bin/schematool -initSchema -dbType mysql
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl
/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Exception in thread "main" java.lang.NoSuchMethodError: com.google.common.base.Preconditions.chec
kArgument(ZLjava/lang/String;Ljava/lang/Object;)V
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1357)
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1338)
        at org.apache.hadoop.mapred.JobConf.setJar(JobConf.java:518)
        at org.apache.hadoop.mapred.JobConf.setJarByClass(JobConf.java:536)
        at org.apache.hadoop.mapred.JobConf.<init>(JobConf.java:430)
        at org.apache.hadoop.hive.conf.HiveConf.initialize(HiveConf.java:5144)
        at org.apache.hadoop.hive.conf.HiveConf.<init>(HiveConf.java:5107)
        at org.apache.hive.beeline.HiveSchemaTool.<init>(HiveSchemaTool.java:96)
        at org.apache.hive.beeline.HiveSchemaTool.main(HiveSchemaTool.java:1473)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:318)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:232)
```

3. 报错：Exception in thread "main" java.lang.NoSuchMethodError: com.google.common.base.Preconditions.checkArgument(ZLjava/lang/String;Ljava/lang/Object;)V

```
hadoop@UbuntuRita:/usr/local/hive/lib$ cd ..
hadoop@UbuntuRita:/usr/local/hive$ ./bin/schematool -initSchema -dbType mysql
Exception in thread "main" java.lang.NoSuchMethodError: com.google.common.base.Preconditions.chec
kArgument(ZLjava/lang/String;Ljava/lang/Object;)V
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1357)
        at org.apache.hadoop.conf.Configuration.set(Configuration.java:1338)
        at org.apache.hadoop.mapred.JobConf.setJar(JobConf.java:518)
        at org.apache.hadoop.mapred.JobConf.setJarByClass(JobConf.java:536)
        at org.apache.hadoop.mapred.JobConf.<init>(JobConf.java:430)
        at org.apache.hadoop.hive.conf.HiveConf.initialize(HiveConf.java:5144)
        at org.apache.hadoop.hive.conf.HiveConf.<init>(HiveConf.java:5107)
        at org.apache.hive.beeline.HiveSchemaTool.<init>(HiveSchemaTool.java:96)
        at org.apache.hive.beeline.HiveSchemaTool.main(HiveSchemaTool.java:1473)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:318)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:232)
hadoop@UbuntuRita:/usr/local/hive$
```

4. 报错：Exception in thread "main" [com.ctc.wstx.exc.WstxLazyException] com.ctc.wstx.exc.WstxUnexpectedCharException: Unexpected character '=' (code 61); expected a semi-colon after the reference for entity 'useSSL' at [row,col,system-id]: [6,81,"file:/usr/local/hive/conf/hive-site.xml"]

```
hadoop@UbuntuRita:/usr/local/hive$ ./bin/schematool -initSchema -dbType mysql
Exception in thread "main" [com.ctc.wstx.exc.WstxLazyException] com.ctc.wstx.exc.WstxUnexpectedCh
arException: Unexpected character '=' (code 61); expected a semi-colon after the reference for en
tity 'useSSL'
 at [row,col,system-id]: [6,81,"file:/usr/local/hive/conf/hive-site.xml"]
        at com.ctc.wstx.exc.WstxLazyException.throwLazily(WstxLazyException.java:40)
        at com.ctc.wstx.sr.StreamScanner.throwLazyError(StreamScanner.java:724)
        at com.ctc.wstx.sr.BasicStreamReader.safeFinishToken(BasicStreamReader.java:3758)
        at com.ctc.wstx.sr.BasicStreamReader.getTextCharacters(BasicStreamReader.java:914)
        at org.apache.hadoop.conf.Configuration$Parser.parseNext(Configuration.java:3326)
        at org.apache.hadoop.conf.Configuration$Parser.parse(Configuration.java:3114)
        at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:3007)
        at org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2973)
        at org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2848)
        at org.apache.hadoop.conf.Configuration.get(Configuration.java:1460)
        at org.apache.hadoop.hive.conf.HiveConf.getVar(HiveConf.java:4999)
        at org.apache.hadoop.hive.conf.HiveConf.getVar(HiveConf.java:5072)
        at org.apache.hadoop.hive.conf.HiveConf.initialize(HiveConf.java:5159)
        at org.apache.hadoop.hive.conf.HiveConf.<init>(HiveConf.java:5107)
        at org.apache.hive.beeline.HiveSchemaTool.<init>(HiveSchemaTool.java:96)
        at org.apache.hive.beeline.HiveSchemaTool.main(HiveSchemaTool.java:1473)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:318)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:232)
Caused by: com.ctc.wstx.exc.WstxUnexpectedCharException: Unexpected character '=' (code 61); expe
cted a semi-colon after the reference for entity 'useSSL'
 at [row,col,system-id]: [6,81,"file:/usr/local/hive/conf/hive-site.xml"]
        at com.ctc.wstx.sr.StreamScanner.throwUnexpectedChar(StreamScanner.java:653)
        at com.ctc.wstx.sr.StreamScanner.parseEntityName(StreamScanner.java:2067)
        at com.ctc.wstx.sr.StreamScanner.fullyResolveEntity(StreamScanner.java:1525)
        at com.ctc.wstx.sr.BasicStreamReader.readTextSecondary(BasicStreamReader.java:4783)
        at com.ctc.wstx.sr.BasicStreamReader.finishToken(BasicStreamReader.java:3802)
        at com.ctc.wstx.sr.BasicStreamReader.safeFinishToken(BasicStreamReader.java:3756)
        ... 19 more
```

5. 报错：

```
hadoop@UbuntuRita:/usr/local/hive$ ./bin/schematool -initSchema -dbType mysql --verbose
Metastore connection URL:        jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true&u
seSSL=false
Metastore Connection Driver :    com.mysql.jdbc.Driver
Metastore connection User:       hive
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.
jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver
 class is generally unnecessary.
org.apache.hadoop.hive.metastore.HiveMetaException: Failed to get schema version.
Underlying cause: java.sql.SQLNonTransientConnectionException : Public Key Retrieval is not allow
ed
SQL Error code: 0
org.apache.hadoop.hive.metastore.HiveMetaException: Failed to get schema version.
        at org.apache.hadoop.hive.metastore.tools.HiveSchemaHelper.getConnectionToMetastore(HiveS
chemaHelper.java:94)
        at org.apache.hive.beeline.HiveSchemaTool.getConnectionToMetastore(HiveSchemaTool.java:16
9)
        at org.apache.hive.beeline.HiveSchemaTool.testConnectionToMetastore(HiveSchemaTool.java:4
75)
        at org.apache.hive.beeline.HiveSchemaTool.doInit(HiveSchemaTool.java:581)
        at org.apache.hive.beeline.HiveSchemaTool.doInit(HiveSchemaTool.java:567)
        at org.apache.hive.beeline.HiveSchemaTool.main(HiveSchemaTool.java:1517)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:318)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:232)
Caused by: java.sql.SQLNonTransientConnectionException: Public Key Retrieval is not allowed
        at com.mysql.cj.jdbc.exceptions.SQLError.createSQLException(SQLError.java:110)
        at com.mysql.cj.jdbc.exceptions.SQLExceptionsMapping.translateException(SQLExceptionsMapp
ing.java:122)
        at com.mysql.cj.jdbc.ConnectionImpl.createNewIO(ConnectionImpl.java:828)
        at com.mysql.cj.jdbc.ConnectionImpl.<init>(ConnectionImpl.java:448)
        at com.mysql.cj.jdbc.ConnectionImpl.getInstance(ConnectionImpl.java:241)
        at com.mysql.cj.jdbc.NonRegisteringDriver.connect(NonRegisteringDriver.java:198)
        at java.sql.DriverManager.getConnection(DriverManager.java:664)
        at java.sql.DriverManager.getConnection(DriverManager.java:247)
        at org.apache.hadoop.hive.metastore.tools.HiveSchemaHelper.getConnectionToMetastore(HiveS
chemaHelper.java:88)
```

6. 计算量太大导致运行错误。

解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：

1. 使用下面的命令解决：

CREATE USER 'hive'@'localhost' IDENTIFIED BY 'hive';

GRANT ALL ON *.* TO 'hive'@'localhost';

（参考：grant all on *.* to hive@localhost identified by 'hive'; ERROR 1064 (42000): You have an error in yo-CSDN 博客）

2. 使用下面命令解决：

```
hadoop@UbuntuRita:/usr/local/hive/lib$ mv log4j-slf4j-impl-2.17.1.jar log4j-slf4j-impl-2.17.1.jar
.bak
hadoop@UbuntuRita:/usr/local/hive/lib$
```

（参考：Hive 客户端启动报 SLF4J: Class path contains multiple SLF4J bindings. _hive slf4j: class path contains multiple slf4j bin-CSDN 博客）

3. 删除 hive 中低版本的 guava-19.0.jar 包，将 hadoop 中的 guava-27.0-jre.jar 复制到 hive 的 lib 目录下：

```
hadoop@UbuntuRita:/usr/local/hive/lib$ sudo rm -f guava-19.0.jar
[sudo] password for hadoop:
hadoop@UbuntuRita:/usr/local/hive/lib$ cd ..
hadoop@UbuntuRita:/usr/local/hive$ cd ..
hadoop@UbuntuRita:/usr/local$ cd hadoop
hadoop@UbuntuRita:/usr/local/hadoop$ cd share/hadoop/common/lib
hadoop@UbuntuRita:/usr/local/hadoop/share/hadoop/common/lib$
hadoop@UbuntuRita:/usr/local/hadoop/share/hadoop/common/lib$ ls |grep "guava"
guava-27.0-jre.jar
listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar
hadoop@UbuntuRita:/usr/local/hadoop/share/hadoop/common/lib$ sudo cp guava-27.0-jre.jar /usr/loca
l/hive/lib
```

（参考：Hive 启动报错 java.lang.NoSuchMethodError: com.google.common.base.Preconditions.checkArgument _ 启动 hive 提示 checkargument-CSDN 博客）

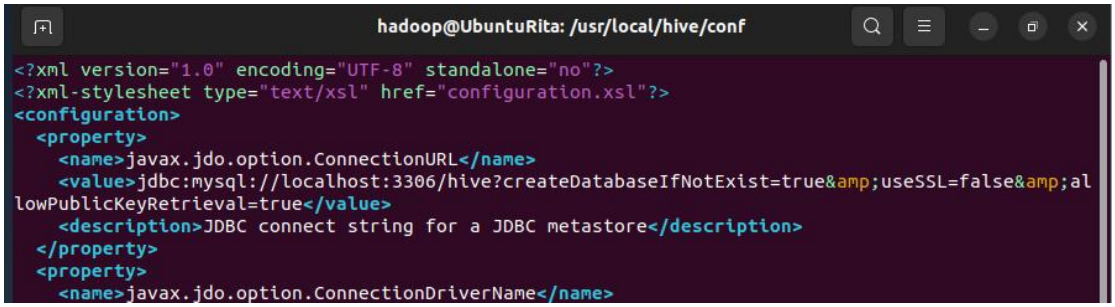
4. MySQL 连接的 URL 不能使用 & 字符，需要使用转义符 & 替代：

```
hadoop@UbuntuRita: /usr/local/hive/conf
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true&amp;useSSL=false</value
>
    <description>JDBC connect string for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionDriverName</name>
    <value>com.mysql.jdbc.Driver</value>
    <description>Driver class name for a JDBC metastore</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionUserName</name>
    <value>hive</value>
    <description>username to use against metastore database</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionPassword</name>
    <value>hive</value>
```

（参考：Hive 的安装与配置 _ exception in thread "main" [com.ctc.wstx.exc.wstxl-CSDN 博客）

5. 修改 hive-site.xml 中 JDBC 连接字符串，添加“allowPublicKeyRetrieval=true”，允许在使用 caching_sha2_password 认证方式时检索服务器的公钥。

```
hadoop@UbuntuRita:/usr/local/hive$ cd conf
hadoop@UbuntuRita:/usr/local/hive/conf$ ls
beeline-log4j2.properties.template    hive-site.xml
hive-default.xml                      ivysettings.xml
hive-env.sh.template                  llap-cli-log4j2.properties.template
hive-exec-log4j2.properties.template  llap-daemon-log4j2.properties.template
hive-log4j2.properties.template       parquet-logging.properties
hadoop@UbuntuRita:/usr/local/hive/conf$ vim hive-site.xml
```



6. 取数据集的前 100000 行数据进行计算。

实验过程中参考一下文章：

1. Hive3.1.3 安装和使用指南_厦大数据库实验室博客（xmu.edu.cn）

2. Ubuntu 下安装 Hive3.1.2 教程（附 MySQL 安装方法及安装包）_乌邦图安装 hive3.1.2-CSDN 博客

注：报告篇幅可根据实际题目情况进行调整。