EX-5 Create tables in Hive and write queries to access the data in the table

AIM:

To create tables in Hive and write queries to access the data in the table.

PROCEDURE:

Installation of Hive:

1. Installing Apache Derby
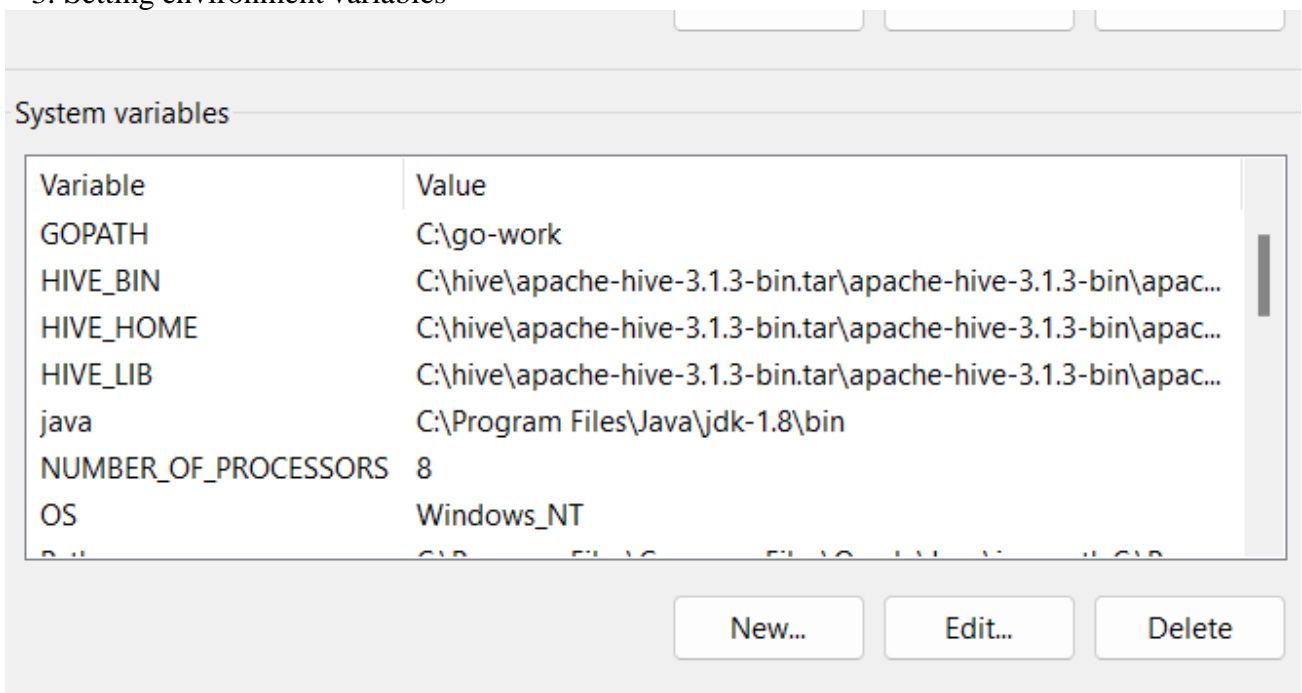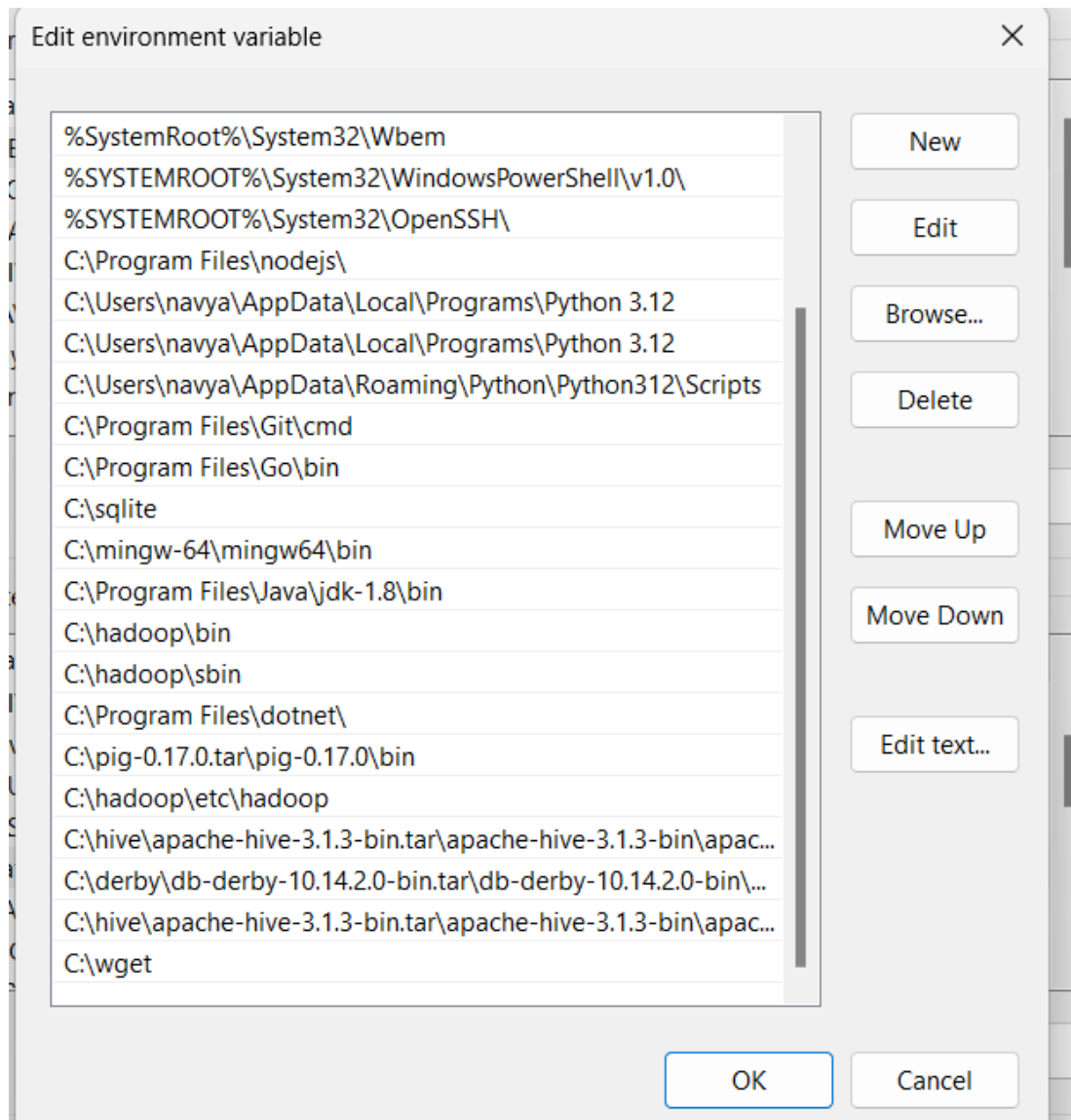
Install Apache Derby 10.14.2.0

https://db.apache.org/derby/derby_downloads.html#For+Java+8+and+Higher

2. Downloading Apache Hive binaries

https://downloads.apache.org/hive/hive-3.1.2/

3. Setting environment variables

Edit environment variable

%SystemRoot%\System32\Wbem	New
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\	
%SYSTEMROOT%\System32\OpenSSH\	Edit
C:\Program Files\nodejs\	
C:\Users\navya\AppData\Local\Programs\Python 3.12	Browse...
C:\Users\navya\AppData\Local\Programs\Python 3.12	
C:\Users\navya\AppData\Roaming\Python\Python312\Scripts	Delete
C:\Program Files\Git\cmd	
C:\Program Files\Go\bin	
C:\sqlite	Move Up
C:\mingw-64\mingw64\bin	
C:\Program Files\Java\jdk-1.8\bin	Move Down
C:\hadoop\bin	
C:\hadoop\sbin	
C:\Program Files\dotnet\	
C:\pig-0.17.0.tar\pig-0.17.0\bin	Edit text...
C:\hadoop\etc\hadoop	
C:\hive\apache-hive-3.1.3-bin.tar\apache-hive-3.1.3-bin\apac...	
C:\derby\db-derby-10.14.2.0-bin.tar\db-derby-10.14.2.0-bin\...	
C:\hive\apache-hive-3.1.3-bin.tar\apache-hive-3.1.3-bin\apac...	
C:\wget	

OK   Cancel

3.1. Copy Derby libraries

we should go to the Derby libraries directory (db-derby-10.14.2.0\lib) and copy all *.jar files. Then, we should paste them within the Hive libraries directory.

3.2. Configuring hive-site.xml and Hive's Bin folder

Refer following link to download the file. Also download the guava file. Put hive-site.xml file to hive's conf location and replace hive's current guava file with this one in lib location. Also download the bin folder from link and replace the existing hive's bin folder.

https://1drv.ms/f/s!ArSg3Xpur4Grmw0SDqW0g44T7HYU?e=wDsoBn

4. Starting Hadoop Services

start-all.cmd

5. Derby Network Server: Run the following command in separate window to open Derby

```
StartNetworkServer -h 0.0.0.0
```

6. Starting Apache Hive

Go to Apache Hive's bin location with cd command and run the following command:

```
 hive --service  schematool -dbType derby -initSchema
```

7. Open Hive shell by typing:

```
 hive
```

Create a Database:

Start by creating a database. Open the Hive CLI and follow the steps below:

1. Use the **CREATE DATABASE** statement to create a new database:

```
CREATE DATABASE sample;
```

2. Verify the database is present:

```
SHOW DATABASES;
```

3. Switch to the new database:

```
USE sample;
```

Create a Table in Hive

```
CREATE TABLE students (name STRING, roll INT, dept STRING);
```

Add Data

Run the **LOAD DATA LOCAL INPATH** command:

```
LOAD DATA INPATH
'/C:/Users/mercy/OneDrive/Documents/DataAnalytics/Hive/students.csv' WRITE INTO
TABLE students;
```

List Hive Tables and Data

To show all tables in a selected database, use the following statement:

```
SHOW TABLES;
```

To show table column names and data types, run:

```
DESC students;
```

To display table data, use a **SELECT** statement. For example, to select everything in a table, run:

SELECT * FROM students;

OUTPUT:

```
hive> CREATE TABLE samplehiv (id INT, name STRING, age INT);
2024-09-03 10:01:35,031 INFO conf.HiveConf: Using the default value passed in for log id: 2456a1f2-959e-4268-b38b-1d570d125a1d
2024-09-03 10:01:35,031 INFO session.SessionState: Updating thread name to 2456a1f2-959e-4268-b38b-1d570d125a1d main
2024-09-03 10:01:35,033 INFO ql.Driver: Compiling command(queryId=navya_20240903100135_d6cbb6d6-10f9-4dde-9936-497ad9466af8): CREATE TABLE samplehiv (id INT, name STRING, age INT)
2024-09-03 10:01:35,093 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-03 10:01:35,097 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-09-03 10:01:35,115 INFO parse.CalcitePlanner: Creating table default.samplehiv position=13
2024-09-03 10:01:35,136 INFO metastore.HiveMetaStore: 0: get_database: @hive#default
2024-09-03 10:01:35,136 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_database: @hive#default
2024-09-03 10:01:35,165 INFO ql.Driver: Semantic Analysis Completed (retrial = false)
2024-09-03 10:01:35,172 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties:null)
2024-09-03 10:01:35,172 INFO ql.Driver: Completed compiling command(queryId=navya_20240903100135_d6cbb6d6-10f9-4dde-9936-497ad9466af8); Time taken: 0.139 seconds
2024-09-03 10:01:35,173 INFO reexec.ReExecDriver: Execution #1 of query
2024-09-03 10:01:35,177 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-03 10:01:35,178 INFO ql.Driver: Executing command(queryId=navya_20240903100135_d6cbb6d6-10f9-4dde-9936-497ad9466af8): CREATE TABLE samplehiv (id INT, name STRING, age INT)
2024-09-03 10:01:35,186 INFO ql.Driver: Starting task [Stage-0:DDL] in serial mode
2024-09-03 10:01:35,484 INFO metastore.HiveMetaStore: 0: create_table: Table(tableName:samplehiv, dbName:default, owner:navya, createTime:1725337895, lastAccessTime:0, retention:0, sd:StorageDesc
ldSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, comment:null), FieldSchema(name:age, type:int, comment:null)], location:null, inputFormat:org.apache.hadoop.mapred.
outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.serde2.lazy.LazyS
eters:{serialization.format=1}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), storedAsSubDirectories:fa
```

```
hive> INSERT INTO TABLE samplehiv VALUES (1, 'navya', 20);
2024-09-03 10:03:06,762 INFO conf.HiveConf: Using the default value passed in for log id: 2456a1f2-959e-4268-b38b-1d570d125a1d
2024-09-03 10:03:06,762 INFO session.SessionState: Updating thread name to 2456a1f2-959e-4268-b38b-1d570d125a1d main
2024-09-03 10:03:06,762 INFO ql.Driver: Compiling command(queryId=navya_20240903100306_4b7d1705-f4f8-4530-b86c-fc3908337889): INSERT INTO TABLE samplehiv VALUES (1, 'navya', 20)
2024-09-03 10:03:06,778 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-03 10:03:06,778 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-09-03 10:03:06,778 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.samplehiv
2024-09-03 10:03:06,778 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_table : tbl=hive.default.samplehiv
2024-09-03 10:03:07,061 INFO parse.CalcitePlanner: Completed phase 1 of Semantic Analysis
2024-09-03 10:03:07,061 INFO parse.CalcitePlanner: Get metadata for source tables
2024-09-03 10:03:07,061 INFO parse.CalcitePlanner: Get metadata for subqueries
2024-09-03 10:03:07,061 INFO parse.CalcitePlanner: Get metadata for destination tables
2024-09-03 10:03:07,077 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.samplehiv
2024-09-03 10:03:07,077 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_table : tbl=hive.default.samplehiv
2024-09-03 10:03:07,110 INFO parse.CalcitePlanner: Completed getting MetaData in Semantic Analysis
2024-09-03 10:03:07,974 INFO ql.Context: New scratch dir is hdfs://localhost:9000/tmp/hive/navya/2456a1f2-959e-4268-b38b-1d570d125a1d/hive_2024-09-03_10-03-06_762_841370934897622218-1
2024-09-03 10:03:08,684 INFO metastore.HiveMetaStore: 0: get_not_null_constraints : tbl=hive._dummy_database._dummy_table
2024-09-03 10:03:08,684 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_not_null_constraints : tbl=hive._dummy_database._dummy_table
```

```
2024-09-03 10:05:20,062 INFO session.SessionState: Resetting thread name to  main
hive> SELECT * FROM samplehiv;
2024-09-03 10:05:34,210 INFO conf.HiveConf: Using the default value passed in for log id: 2456a1f2-959e-4268-b38b-1d570d125a1d
2024-09-03 10:05:34,210 INFO session.SessionState: Updating thread name to 2456a1f2-959e-4268-b38b-1d570d125a1d main
2024-09-03 10:05:34,213 INFO ql.Driver: Compiling command(queryId=navya_20240903100534_e550bd6b-2e84-4818-ad7f-1be0b9ed36fa): SELECT * FROM samplehiv
2024-09-03 10:05:34,225 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-03 10:05:34,225 INFO parse.CalcitePlanner: Starting Semantic Analysis
2024-09-03 10:05:34,225 INFO parse.CalcitePlanner: Completed phase 1 of Semantic Analysis
2024-09-03 10:05:34,225 INFO parse.CalcitePlanner: Get metadata for source tables
2024-09-03 10:05:34,225 INFO metastore.HiveMetaStore: 0: get_table : tbl=hive.default.samplehiv
2024-09-03 10:05:34,225 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_table : tbl=hive.default.samplehiv
2024-09-03 10:05:34,241 INFO parse.CalcitePlanner: Get metadata for subqueries
2024-09-03 10:05:34,241 INFO parse.CalcitePlanner: Get metadata for destination tables
2024-09-03 10:05:34,258 INFO ql.Context: New scratch dir is hdfs://localhost:9000/tmp/hive/navya/2456a1f2-959e-4268-b38b-1d570d125a1d/hive_2024-09-03_10-05-34_225_5001655230486780851-1
2024-09-03 10:05:34,258 INFO parse.CalcitePlanner: Completed getting MetaData in Semantic Analysis
2024-09-03 10:05:34,258 INFO results.QueryResultsCache: Initializing query results cache at /tmp/hive/_resultscache_
2024-09-03 10:05:34,272 INFO results.QueryResultsCache: Query results cache: cacheDirectory /tmp/hive/_resultscache_/results-1ee53f07-aa05-496e-8ea9-b2154f1f8fe4, maxCacheSize 2147483648,
maxEntryLifetime 3600000
2024-09-03 10:05:34,272 INFO metastore.HiveMetaStore: 0: get_not_null_constraints : tbl=hive.default.samplehiv
2024-09-03 10:05:34,272 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_not_null_constraints : tbl=hive.default.samplehiv
2024-09-03 10:05:34,272 INFO metastore.HiveMetaStore: 0: get_primary_keys : tbl=hive.default.samplehiv
2024-09-03 10:05:34,272 INFO HiveMetaStore.audit: ugi=navya    ip=unknown-ip-addr    cmd=get_primary_keys : tbl=hive.default.samplehiv
2024-09-03 10:05:34,272 INFO metastore.HiveMetaStore: 0: get primary keys : tbl=hive.default.samplehiv
```

```
OK
2024-09-03 10:05:34,399 INFO ql.Driver: OK
2024-09-03 10:05:34,399 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
2024-09-03 10:05:34,415 INFO mapred.FileInputFormat: Total input files to process : 3
2024-09-03 10:05:34,458 INFO exec.TableScanOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_TS_0:3,
2024-09-03 10:05:34,458 INFO exec.SelectOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_SEL_1:3,
2024-09-03 10:05:34,463 INFO exec.ListSinkOperator: RECORDS_OUT_INTERMEDIATE:0, RECORDS_OUT_OPERATOR_LIST_SINK_3:3,
1       navya   20
2       rathi   21
3       oviya   20
Time taken: 0.186 seconds, Fetched: 3 row(s)
2024-09-03 10:05:34,463 INFO CliDriver: Time taken: 0.186 seconds, Fetched: 3 row(s)
2024-09-03 10:05:34,463 INFO conf.HiveConf: Using the default value passed in for log id: 2456a1f2-959e-4268-b38b-1d570d125a1d
2024-09-03 10:05:34,463 INFO session.SessionState: Resetting thread name to  main
hive> _
```

RESULT:

Thus to create tables in Hive and write queries to access the data in the table is completed
successfully.