BDMH: Assignment I

Group 45

Harsh Bandhey 2017234
Ridam Pal PhD19201

Question 1

Docker Run
Commands:

```
docker pull raptor4/harsh17234_ridamp_a1
docker run -it  raptor4/harsh17234_ridamp_a1
```

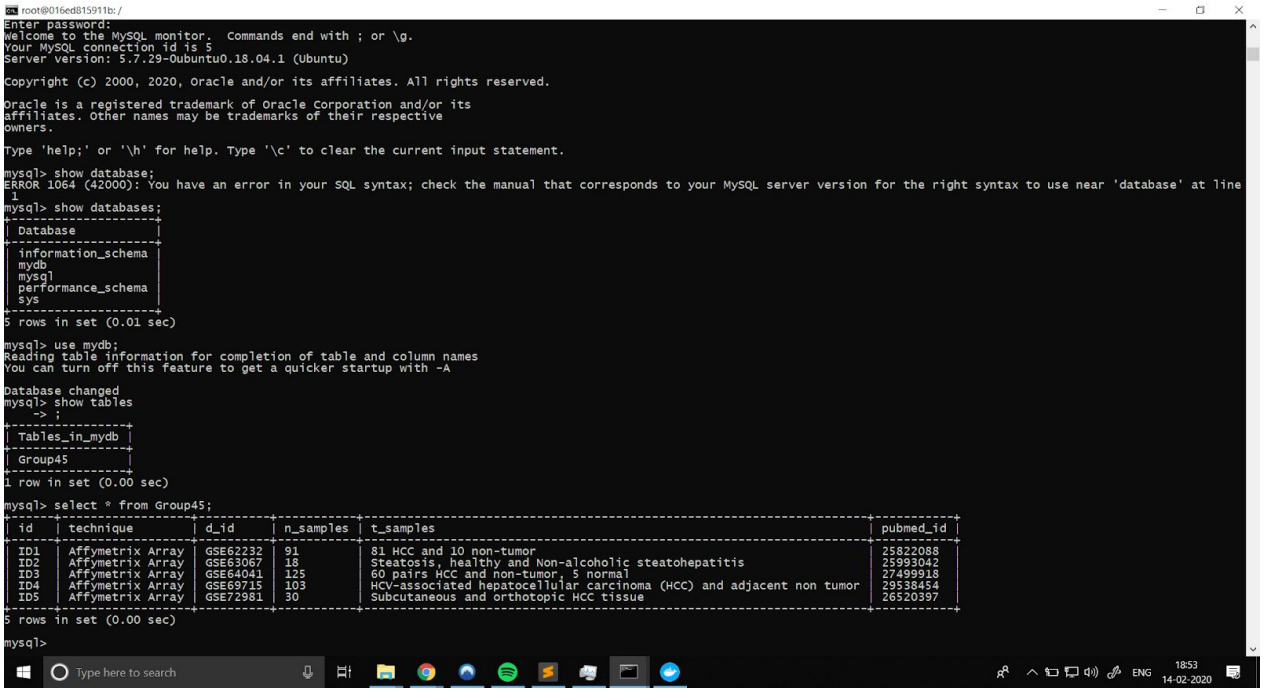
For a non-persistant non-ported container image. Use -v and -p flags accordingly for mounting volumes and ports.
Volume mount points for

MongoDB /data/db
MySQL /var/lib/mysql

Usage:

Run “mongo”, “hive” and “mysql -u root -proot” for accessing MongoDB, hive and MySQL shells respectively.

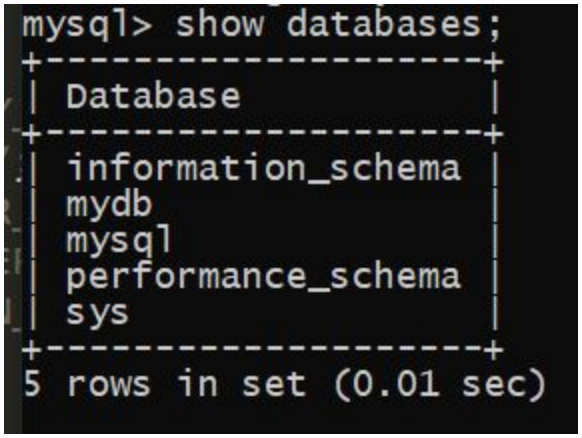
MySQL Question 1.



MySQL commands and demos:

(With pre-existing data from Group45.csv)

Show databases command is used for showing all the databases stored within it.



Use command inherits the database we need to access from the list of databases.

Show tables command used to show all the tables present in the database.

```
mysql> use mydb;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables
    -> ;
+----------------+
| Tables_in_mydb |
+----------------+
| Group45        |
+----------------+
1 row in set (0.00 sec)
```

1) **Select**

Select command query states about selecting all the rows from the table Group45

```
mysql> select * from Group45;
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
| id   | technique       | d_id    | n_samples | t_samples                                                             | pubmed_id |
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
| ID1  | Affymetrix Array | GSE62232 | 91        | 81 HCC and 10 non-tumor                                               | 25822088  |
| ID2  | Affymetrix Array | GSE63067 | 18        | Steatosis, healthy and Non-alcoholic steatohepatitis                  | 25993042  |
| ID3  | Affymetrix Array | GSE64041 | 125       | 60 pairs HCC and non-tumor, 5 normal                                  | 27499918  |
| ID4  | Affymetrix Array | GSE69715 | 103       | HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor  | 29538454  |
| ID5  | Affymetrix Array | GSE72981 | 30        | Subcutaneous and orthotopic HCC tissue                                | 26520397  |
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
5 rows in set (0.00 sec)
```

This query is selecting all the rows from table Group45 **where** ‘id=ID1’.

```
mysql> select * from Group45 where id="ID1";
+------+-----------------+---------+-----------+-------------------------+-----------+
| id   | technique       | d_id    | n_samples | t_samples               | pubmed_id |
+------+-----------------+---------+-----------+-------------------------+-----------+
| ID1  | Affymetrix Array | GSE62232 | 91        | 81 HCC and 10 non-tumor | 25822088  |
+------+-----------------+---------+-----------+-------------------------+-----------+
1 row in set (0.00 sec)
```

2) **Insert**

This query states about inserting rows into the table Group45

```
mysql> INSERT INTO `Group45` VALUES ('ID1','Affymetrix Array','GSE62232','91','81 HCC and 10 non-tumor','25822088');
Query OK, 1 row affected (0.01 sec)

mysql> mysql> INSERT INTO `Group45` VALUES ('ID1','Affymetrix Array','GSE62232','91','81 HCC and 10 non-tumor','25822088');    select * from Group45;
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
| id   | technique       | d_id    | n_samples | t_samples                                                             | pubmed_id |
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
| ID1  | Affymetrix Array | GSE62232 | 91        | 81 HCC and 10 non-tumor                                               | 25822088  |
| ID2  | Affymetrix Array | GSE63067 | 18        | Steatosis, healthy and Non-alcoholic steatohepatitis                  | 25993042  |
| ID3  | Affymetrix Array | GSE64041 | 125       | 60 pairs HCC and non-tumor, 5 normal                                  | 27499918  |
| ID4  | Affymetrix Array | GSE69715 | 103       | HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor  | 29538454  |
| ID5  | Affymetrix Array | GSE72981 | 30        | Subcutaneous and orthotopic HCC tissue                                | 26520397  |
| ID1  | Affymetrix Array | GSE62232 | 91        | 81 HCC and 10 non-tumor                                               | 25822088  |
+------+-----------------+---------+-----------+-----------------------------------------------------------------------+-----------+
6 rows in set (0.08 sec)

mysql>
```

3) **Delete and Drop**

Delete command is used for deleting the rows from the table.

Drop command is used for removing the table from the database.

```
mysql> DELETE FROM Group45;
Query OK, 6 rows affected (0.01 sec)

mysql> DROP TABLE Group45;
Query OK, 0 rows affected (0.03 sec)

mysql>
```

4) **Create Table**

Create table command is used for creating the Table45 hence defining the data types of each entry along with each column name.

```
mysql> CREATE TABLE `Group45` (
    ->   `id` varchar(25) COLLATE utf8_unicode_ci DEFAULT NULL,
    ->   `technique` varchar(255) COLLATE utf8_unicode_ci DEFAULT NULL,
    ->   `d_id` varchar(255) COLLATE utf8_unicode_ci DEFAULT NULL,
    ->   `n_samples` varchar(25) COLLATE utf8_unicode_ci DEFAULT NULL,
    ->   `t_samples` varchar(255) COLLATE utf8_unicode_ci DEFAULT NULL,
    ->   `pubmed_id` varchar(25) COLLATE utf8_unicode_ci DEFAULT NULL
    -> );
Query OK, 0 rows affected (0.03 sec)

mysql>
```

5) **Update**

Update command helps in updating the specific row **where** ‘id=ID1’ with the updated value.

```
mysql> update Group45 set t_samples="Edit" where id="ID1";
Query OK, 0 rows affected (0.01 sec)
Rows matched: 1  Changed: 0  Warnings: 0

mysql> select * from Group45 where id="ID1";
+------+------------------+----------+-----------+-----------+-----------+
| id   | technique        | d_id     | n_samples | t_samples | pubmed_id |
+------+------------------+----------+-----------+-----------+-----------+
| ID1  | Affymetrix Array | GSE62232 | 91        | Edit      | 25822088  |
+------+------------------+----------+-----------+-----------+-----------+
1 row in set (0.00 sec)
```

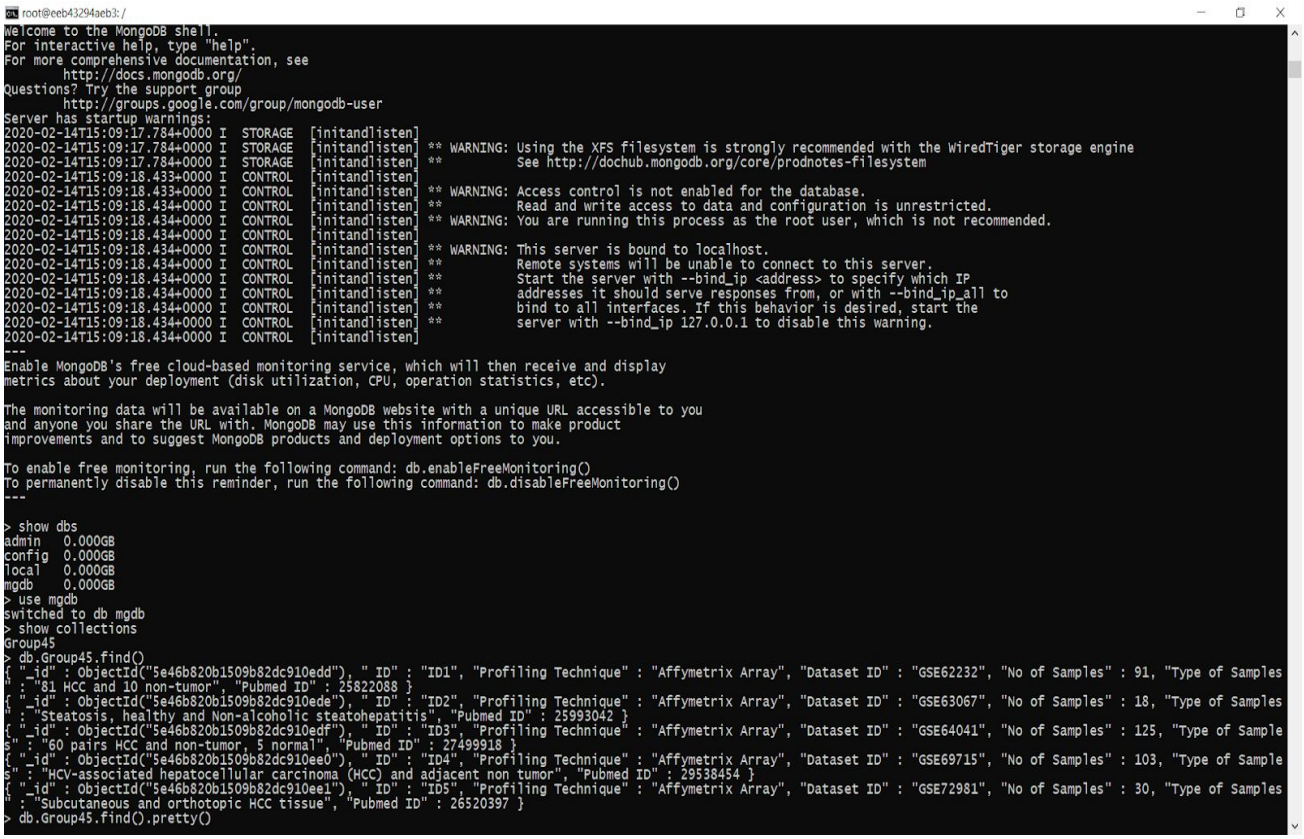
Question 2.

MongoDB commands and demos:

(With pre-existing data from Group45.csv)

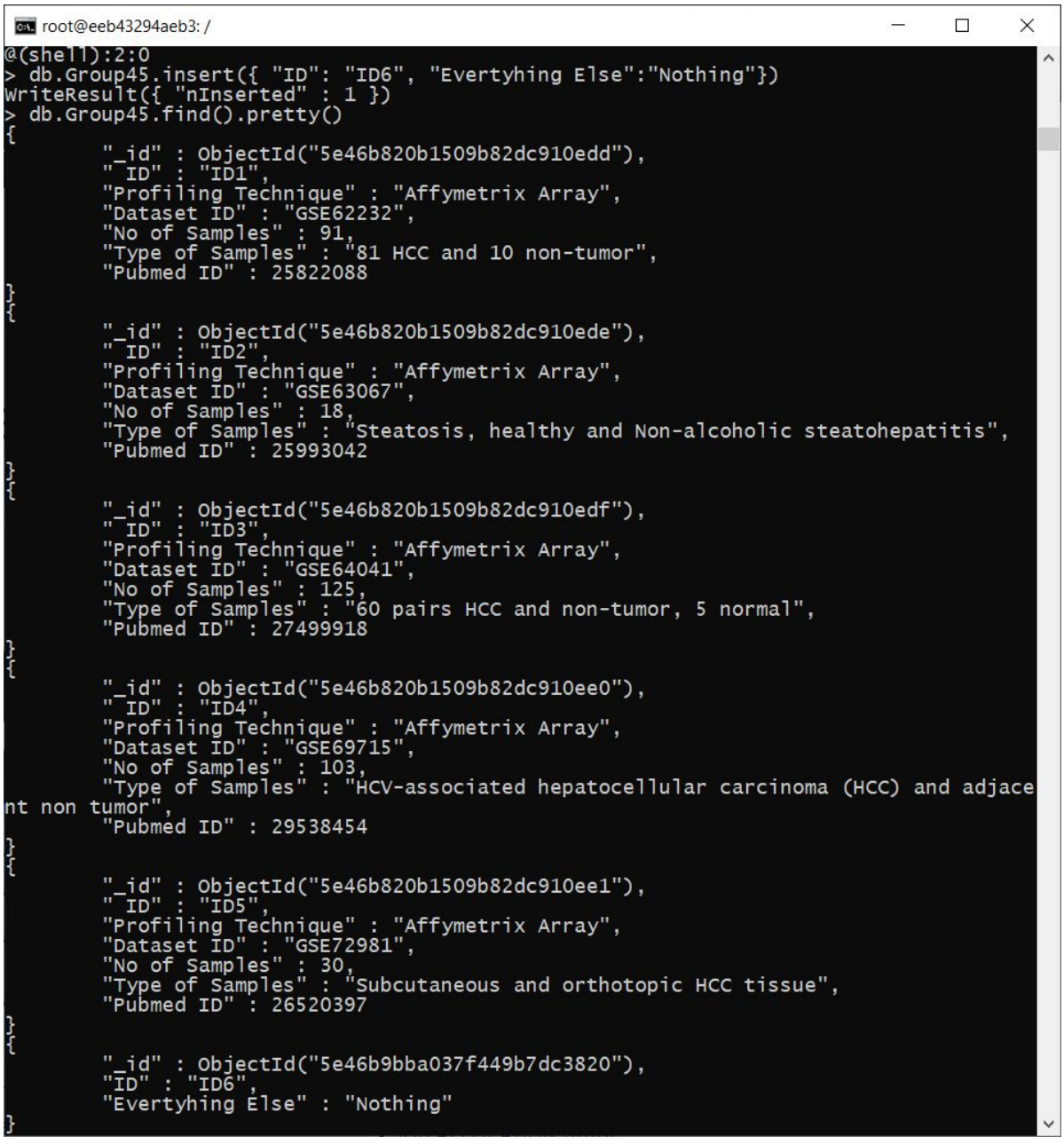
Show dbs command is used for showing all the databases stored within it.

MongoDB Question 2.



1. **find()** method is used in query for displaying the documents in a random and unstructured way.
2. **Insert ()** method is used in query for inserting rows in the database.
3. **Update()** method is used in query for updating the document in the database.
4. **Remove()** method is used in query for removing rows in the database.
5. **Drop()** method common is used to drop the whole collection.

The query for each command has been shown in the images below where all the queries related to each method have been shown.



```
> db.Group45.update({ "ID": "ID6"},{$set:{'title':'Extra Tittle'}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Group45.find({"ID":"ID6"}).pretty()
{
        "_id" : ObjectId("5e46b9bba037f449b7dc3820"),
        "ID" : "ID6",
        "Evertyhing Else" : "Nothing",
        "title" : "Extra Tittle"
}
>
```

Select root@eeb43294aeb3: /

```
@(shell):1:1
> db.Group45.remove({"ID":"ID6"})
WriteResult({ "nRemoved" : 1 })
> db.Group45.find({"ID":"ID6"}).pretty()
> db.Group45.find().pretty()
{
        "_id" : ObjectId("5e46b820b1509b82dc910edd"),
        " ID" : "ID1",
        "Profiling Technique" : "Affymetrix Array",
        "Dataset ID" : "GSE62232",
        "No of Samples" : 91,
        "Type of Samples" : "81 HCC and 10 non-tumor",
        "Pubmed ID" : 25822088
}
{
        "_id" : ObjectId("5e46b820b1509b82dc910ede"),
        " ID" : "ID2",
        "Profiling Technique" : "Affymetrix Array",
        "Dataset ID" : "GSE63067",
        "No of Samples" : 18,
        "Type of Samples" : "Steatosis, healthy and Non-alcoholic steatohepatitis",
        "Pubmed ID" : 25993042
}
{
        "_id" : ObjectId("5e46b820b1509b82dc910edf"),
        " ID" : "ID3",
        "Profiling Technique" : "Affymetrix Array",
        "Dataset ID" : "GSE64041",
        "No of Samples" : 125,
        "Type of Samples" : "60 pairs HCC and non-tumor, 5 normal",
        "Pubmed ID" : 27499918
}
{
        "_id" : ObjectId("5e46b820b1509b82dc910ee0"),
        " ID" : "ID4",
        "Profiling Technique" : "Affymetrix Array",
        "Dataset ID" : "GSE69715",
        "No of Samples" : 103,
        "Type of Samples" : "HCV-associated hepatocellular carcinoma (HCC) and adjace
nt non tumor",
        "Pubmed ID" : 29538454
}
{
        "_id" : ObjectId("5e46b820b1509b82dc910ee1"),
        " ID" : "ID5",
        "Profiling Technique" : "Affymetrix Array",
        "Dataset ID" : "GSE72981",
        "No of Samples" : 30,
        "Type of Samples" : "Subcutaneous and orthotopic HCC tissue",
        "Pubmed ID" : 26520397
}
>
```

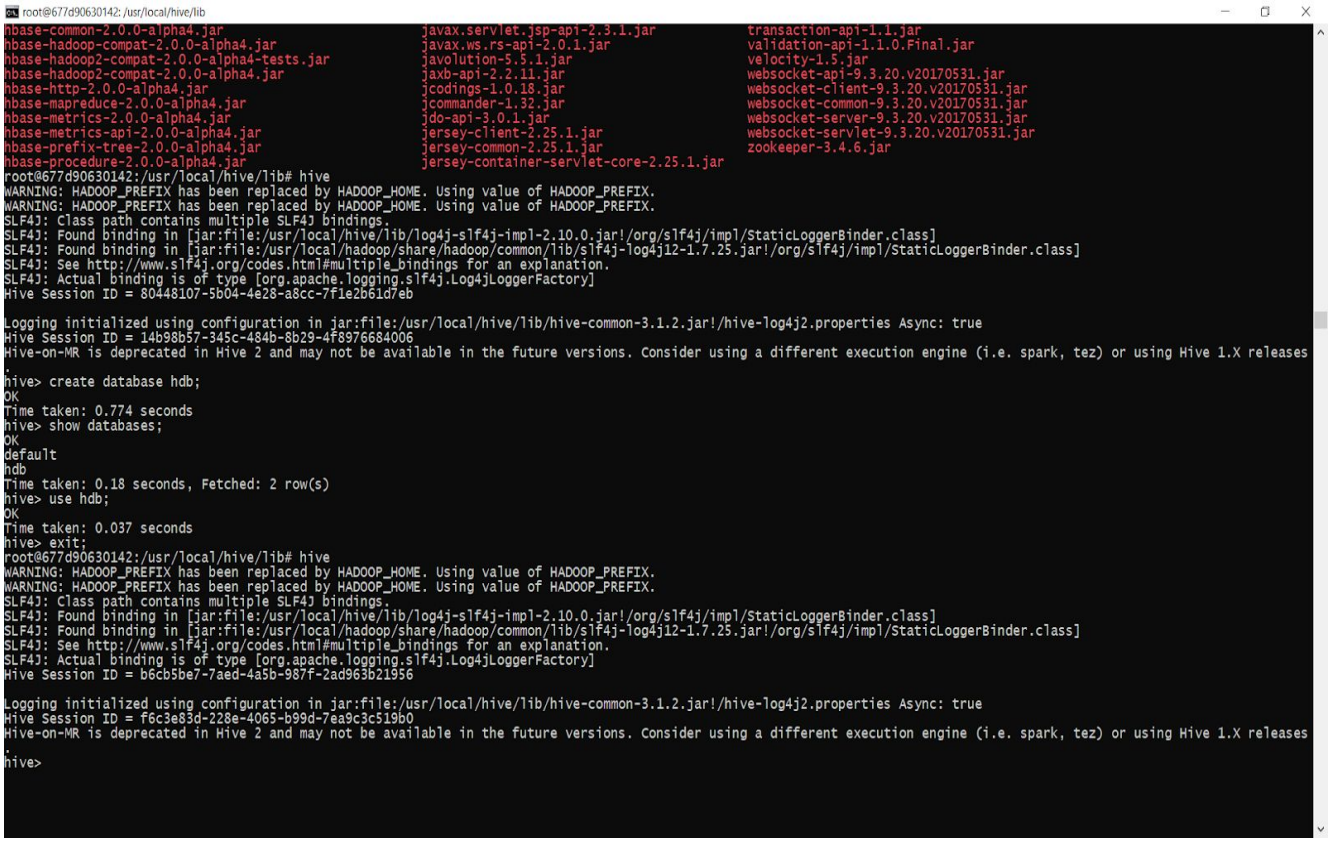
Hive Question 3.

Hive commands and demos:

(With pre-existing data from Group45.csv)

Show databases command is used for showing all the databases stored within it.

1. **Create databases** command is used for creating the database.
   Create table command is used for creating the table within the databases.
2. **Select** command is used to select a set of rows from the table.
3. **Alter** command is used in the query to edit/update columns of the table.
4. **Drop command** is used to drop the table from the database although the database still persists.
   The query for all the commands has been shown in the images attached below.



```
hive> create database hdb;
OK
Time taken: 0.774 seconds
hive> show databases;
OK
default
hdb
Time taken: 0.18 seconds, Fetched: 2 row(s
hive> use hdb;
OK
Time taken: 0.037 seconds
```

```
.
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Group45(
    >   id String,
    >   technique String,
    >   d_id String,
    >   n_samples String,
    >   t_samples String,
    >   pubmed_id String
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE
    > LOCATION '/user/root/hdb';
OK
Time taken: 1.34 seconds
hive>
```

```
hive> Select * from Group45
    > ;
OK
 ID     Profiling Technique     Dataset ID      No of Samples   Type of Samples Pubmed ID
ID1     Affymetrix Array        GSE62232        91      81 HCC and 10 non-tumor 25822088
ID2     Affymetrix Array        GSE63067        18      Steatosis, healthy and Non-alcoholic steatohepatitis    25993042
ID3     Affymetrix Array        GSE64041        125     60 pairs HCC and non-tumor, 5 normal    27499918
ID4     Affymetrix Array        GSE69715        103     HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor    29538454
ID5     Affymetrix Array        GSE72981        30      Subcutaneous and orthotopic HCC tissue  26520397
Time taken: 2.048 seconds, Fetched: 6 row(s)
hive>
```

```
    > ;
hive> ALTER TABLE Group45 REPLACE COLUMNS(id String, technique String, d_id String, n_samples String, t_samples String);
OK
Time taken: 0.123 seconds
hive> Select * from Group45
    > ;
OK
 ID     Profiling Technique     Dataset ID      No of Samples   Type of Samples
ID1     Affymetrix Array        GSE62232        91      81 HCC and 10 non-tumor
ID2     Affymetrix Array        GSE63067        18      Steatosis, healthy and Non-alcoholic steatohepatitis
ID3     Affymetrix Array        GSE64041        125     60 pairs HCC and non-tumor, 5 normal
ID4     Affymetrix Array        GSE69715        103     HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor
ID5     Affymetrix Array        GSE72981        30      Subcutaneous and orthotopic HCC tissue
Time taken: 0.121 seconds, Fetched: 6 row(s)
hive>
```

```
hive> DROP TABLE IF EXISTS Group45;
OK
Time taken: 0.404 seconds
hive>
```

JSON Question 4.

Using pandas and json library we convert the Group45.csv to json format.

```
root@677d90630142:/usr/local/hive/lib# cd /
root@677d90630142:/# python3 /code/answer3.py
python3: can't open file '/code/answer3.py': [Errno 2] No such file or directory
root@677d90630142:/# cd /code/
root@677d90630142:/code# ls
Group45.csv  answer1.py  answer2.py  answer4.py
root@677d90630142:/code# python3 answer4.py
[{' ID': 'ID1', 'Profiling Technique': 'Affymetrix Array', 'Dataset ID': 'GSE62232', 'No of Samples': 91, 'Type of Samples': '81 HCC and 10 non-tumor', 'Pubmed ID': 2582208
8}, {' ID': 'ID2', 'Profiling Technique': 'Affymetrix Array', 'Dataset ID': 'GSE63067', 'No of Samples': 18, 'Type of Samples': 'Steatosis, healthy and Non-alcoholic steato
hepatitis', 'Pubmed ID': 25993042}, {' ID': 'ID3', 'Profiling Technique': 'Affymetrix Array', 'Dataset ID': 'GSE64041', 'No of Samples': 125, 'Type of Samples': '60 pairs H
CC and non-tumor, 5 normal', 'Pubmed ID': 27499918}, {' ID': 'ID4', 'Profiling Technique': 'Affymetrix Array', 'Dataset ID': 'GSE69715', 'No of Samples': 103, 'Type of Samp
les': 'HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor', 'Pubmed ID': 29538454}, {' ID': 'ID5', 'Profiling Technique': 'Affymetrix Array', 'Dataset ID'
: 'GSE72981', 'No of Samples': 30, 'Type of Samples': 'Subcutaneous and orthotopic HCC tissue', 'Pubmed ID': 26520397}]
```


```
{
    " ID": "ID1",
    "Dataset ID": "GSE62232",
    "No of Samples": 91,
    "Profiling Technique": "Affymetrix Array",
    "Pubmed ID": 25822088,
    "Type of Samples": "81 HCC and 10 non-tumor"
},
{
    " ID": "ID2",
    "Dataset ID": "GSE63067",
    "No of Samples": 18,
    "Profiling Technique": "Affymetrix Array",
    "Pubmed ID": 25993042,
    "Type of Samples": "Steatosis, healthy and Non-alcoholic steatohepatitis"
},
{
    " ID": "ID3",
    "Dataset ID": "GSE64041",
    "No of Samples": 125,
    "Profiling Technique": "Affymetrix Array",
    "Pubmed ID": 27499918,
    "Type of Samples": "60 pairs HCC and non-tumor, 5 normal"
},
{
    " ID": "ID4",
    "Dataset ID": "GSE69715",
    "No of Samples": 103,
    "Profiling Technique": "Affymetrix Array",
    "Pubmed ID": 29538454,
    "Type of Samples": "HCV-associated hepatocellular carcinoma (HCC) and adjacent non tumor"
},
{
    " ID": "ID5",
    "Dataset ID": "GSE72981",
    "No of Samples": 30,
    "Profiling Technique": "Affymetrix Array",
    "Pubmed ID": 26520397,
    "Type of Samples": "Subcutaneous and orthotopic HCC tissue"
}
```