Big Data Technologies
Assignment - 4

```
narth@Sumanth MINGW64 ~/downloads/sai-bigdata-assign4
$ ssh -i saikey.pem hadoop@ec2-3-133-145-122.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-3-133-145-122.us-east-2.compute.amazonaws.com (3.1
33.145.122)' can't be established.
ED25519 key fingerprint is SHA256:2ZmUHK7NH+OF9vV3x8OQZwtiw8ZckiJCsCglRyfRNXA.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-133-145-122.us-east-2.compute.amazonaws.com' (
ED25519) to the list of known hosts.
Last login: Wed Sep 27 02:52:32 2023

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

```
[hadoop@ip-172-31-4-182 ~]$ java TestDataGen
Magic Number = 49384
[hadoop@ip-172-31-4-182 ~]$ ls
foodplaces49384.txt  foodratings49384.txt  hql.zip  TestDataGen.class
[hadoop@ip-172-31-4-182 ~]$ unzip hql.zip
Archive:  hql.zip
   creating: hql/
  inflating: __MACOSX/._hql
  inflating: hql/salaries3.hql
  inflating: __MACOSX/hql/._salaries3.hql
  inflating: hql/salaries.hql
  inflating: __MACOSX/hql/._salaries.hql
  inflating: hql/salaries2.hql
  inflating: __MACOSX/hql/._salaries2.hql
  inflating: hql/demoreadme.txt
  inflating: __MACOSX/hql/._demoreadme.txt
  inflating: hql/loadsalaries.hql
  inflating: __MACOSX/hql/._loadsalaries.hql
  inflating: hql/basicsetup.hql
  inflating: __MACOSX/hql/._basicsetup.hql
  inflating: hql/partsetup.hql
  inflating: __MACOSX/hql/._partsetup.hql
  inflating: hql/Salaries.tsv
  inflating: __MACOSX/hql/._Salaries.tsv
```

```
narth@Sumanth MINGW64 ~/downloads/sai-bigdata-assign4
$ scp -i saikey.pem TestDataGen.class hadoop@ec2-3-133-145-122.us-east-2.compute
.amazonaws.com~

narth@Sumanth MINGW64 ~/downloads/sai-bigdata-assign4
$ scp -i saikey.pem TestDataGen.class hadoop@ec2-3-133-145-122.us-east-2.compute
.amazonaws.com:/home/hadoop
TestDataGen.class                             100% 2189    60.3KB/s   00:00

narth@Sumanth MINGW64 ~/downloads/sai-bigdata-assign4
$ scp -i saikey.pem hql.zip hadoop@ec2-3-133-145-122.us-east-2.compute.amazonaws
.com:/home/hadoop
hql.zip                                       100%  402KB   1.5MB/s   00:00
```

```
[hadoop@ip-172-31-4-182 ~]$ cd /home/hadoop/hql
[hadoop@ip-172-31-4-182 hql]$  beeline -u jdbc:hive2://localhost:10000/ -n hadoop -d
 org.apache.hive.jdbc.HiveDriver --showDbInPrompt
Connecting to jdbc:hive2://localhost:10000/
Connected to: Apache Hive (version 3.1.3-amzn-5)
Driver: Hive JDBC (version 3.1.3-amzn-5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3-amzn-5 by Apache Hive
```

```
0: jdbc:hive2://localhost:10000/ (default)> source ./basicsetup.hql;
No rows affected (0.025 seconds)
No rows affected (0.008 seconds)
0: jdbc:hive2://localhost:10000/ (default)> source ./partsetup.hql;
No rows affected (0.01 seconds)
No rows affected (0.007 seconds)
No rows affected (0.007 seconds)
No rows affected (0.012 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (default)> source ./salaries.hql;
INFO  : Compiling command(queryId=hive_20230927030221_5ef52885-55a1-455a-9c84-9e259e
19306f): CREATE DATABASE IF NOT EXISTS cs595
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030221_5ef52885-55a1-455a-9
c84-9e259e19306f); Time taken: 1.022 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030221_5ef52885-55a1-455a-9c84-9e259e
19306f): CREATE DATABASE IF NOT EXISTS cs595
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030221_5ef52885-55a1-455a-9
c84-9e259e19306f); Time taken: 0.724 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (1.855 seconds)
INFO  : Compiling command(queryId=hive_20230927030223_aaf57124-9815-4c2a-a0aa-66686e
fbde99): use cs595
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030223_aaf57124-9815-4c2a-a
0aa-66686efbde99); Time taken: 0.011 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030223_aaf57124-9815-4c2a-a0aa-66686e
fbde99): use cs595
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030223_aaf57124-9815-4c2a-a
0aa-66686efbde99); Time taken: 0.011 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.047 seconds)
INFO  : Compiling command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8939-8e4d5b
00e381): CREATE TABLE IF NOT EXISTS cs595.salaries (
name STRING,
jobTitle STRING,
agencyID STRING,
agency STRING,
hireDate STRING,
annualSalary DOUBLE,
grossPay DOUBLE)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8
939-8e4d5b00e381); Time taken: 0.374 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8939-8e4d5b
00e381): CREATE TABLE IF NOT EXISTS cs595.salaries (
name STRING,
```

```
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8
939-8e4d5b00e381); Time taken: 0.374 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8939-8e4d5b
00e381): CREATE TABLE IF NOT EXISTS cs595.salaries (
name STRING,
jobTitle STRING,
agencyID STRING,
agency STRING,
hireDate STRING,
annualSalary DOUBLE,
grossPay DOUBLE)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030223_06135bec-b08d-4d2a-8
939-8e4d5b00e381); Time taken: 0.353 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.771 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (cs595)> source ./loadsalaries.hql;
INFO  : Compiling command(queryId=hive_20230927030302_3b4af9f9-a429-4ae3-b10a-83bace
8ade27): LOAD DATA LOCAL INPATH '/home/hadoop/hql/Salaries.tsv' OVERWRITE INTO TABLE
 cs595.salaries
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030302_3b4af9f9-a429-4ae3-b
10a-83bace8ade27); Time taken: 0.218 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030302_3b4af9f9-a429-4ae3-b10a-83bace
8ade27): LOAD DATA LOCAL INPATH '/home/hadoop/hql/Salaries.tsv' OVERWRITE INTO TABLE
 cs595.salaries
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table cs595.salaries from file:/home/hadoop/hql/Salaries.tsv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table cs595.salaries stats: [numFiles=1, numRows=0, totalSize=1538148, rawDa
taSize=0]
INFO  : Completed executing command(queryId=hive_20230927030302_3b4af9f9-a429-4ae3-b
10a-83bace8ade27); Time taken: 0.688 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.934 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (cs595)> source ./salaries2.hql;
INFO  : Compiling command(queryId=hive_20230927030315_89037689-c38b-4960-919e-ebe194
40da4b): CREATE TABLE IF NOT EXISTS cs595.salpart (
name STRING,
agencyID STRING,
agency STRING,
hireDate STRING,
annualSalary DOUBLE,
grossPay DOUBLE)
PARTITIONED BY (jobTitle STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030315_89037689-c38b-4960-9
19e-ebe19440da4b); Time taken: 0.016 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030315_89037689-c38b-4960-919e-ebe194
40da4b): CREATE TABLE IF NOT EXISTS cs595.salpart (
name STRING,
agencyID STRING,
agency STRING,
hireDate STRING,
annualSalary DOUBLE,
grossPay DOUBLE)
PARTITIONED BY (jobTitle STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030315_89037689-c38b-4960-9
19e-ebe19440da4b); Time taken: 0.04 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.072 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (cs595)> source ./salaries3.hql;
INFO  : Compiling command(queryId=hive_20230927030324_ca2b446c-a2bd-411b-81fe-fdb385
39dc87): INSERT OVERWRITE TABLE salpart
PARTITION (jobtitle)
SELECT name, agencyid, agency, hiredate, annualsalary, grosspay, jobtitle
FROM salaries
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:stri
ng, comment:null), FieldSchema(name:agencyid, type:string, comment:null), FieldSchem
a(name:agency, type:string, comment:null), FieldSchema(name:hiredate, type:string, c
omment:null), FieldSchema(name:annualsalary, type:double, comment:null), FieldSchema
(name:grosspay, type:double, comment:null), FieldSchema(name:jobtitle, type:string,
comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030324_ca2b446c-a2bd-411b-8
1fe-fdb38539dc87); Time taken: 3.398 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030324_ca2b446c-a2bd-411b-81fe-fdb385
39dc87): INSERT OVERWRITE TABLE salpart
PARTITION (jobtitle)
SELECT name, agencyid, agency, hiredate, annualsalary, grosspay, jobtitle
FROM salaries
INFO  : Query ID = hive_20230927030324_ca2b446c-a2bd-411b-81fe-fdb38539dc87
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927030324_ca2b446c-a2bd-41
1b-81fe-fdb38539dc87
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: INSERT OVERWRITE TABLE salpart
PA...salaries (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0001)

INFO  : Map 1: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
```

```
ows=2, totalSize=190, rawDataSize=188]
INFO  : Partition {jobtitle=SUPERINTENDENT OPERATIONS} stats: [numFiles=1, numRows=1
, totalSize=94, rawDataSize=93]
INFO  : Partition {jobtitle=LIQUOUR BOARD ASST EXE SECRETA} stats: [numFiles=1, numR
ows=1, totalSize=90, rawDataSize=89]
INFO  : Partition {jobtitle=PROCESS SERVER, SHERIFF} stats: [numFiles=1, numRows=8,
totalSize=699, rawDataSize=691]
INFO  : Partition {jobtitle=POLICE OFFICER} stats: [numFiles=1, numRows=1758, totalS
ize=153974, rawDataSize=152216]
INFO  : Partition {jobtitle=ENGINEER I} stats: [numFiles=1, numRows=48, totalSize=44
07, rawDataSize=4359]
INFO  : Partition {jobtitle=BPD 10} stats: [numFiles=1, numRows=1, totalSize=78, raw
DataSize=77]
INFO  : Partition {jobtitle=DIRECTOR PUBLIC WORKS} stats: [numFiles=1, numRows=1, to
talSize=89, rawDataSize=88]
INFO  : Partition {jobtitle=Administrative Services} stats: [numFiles=1, numRows=10,
 totalSize=880, rawDataSize=870]
INFO  : Partition {jobtitle=Retired Chief Judge Ophans' Co} stats: [numFiles=1, numR
ows=1, totalSize=92, rawDataSize=91]
INFO  : Partition {jobtitle=CHIEF JUDGE ORPHANS' COURT} stats: [numFiles=1, numRows=
1, totalSize=83, rawDataSize=82]
INFO  : Partition {jobtitle=POLLUTION CONTROL ANALYST SUPV} stats: [numFiles=1, numR
ows=5, totalSize=457, rawDataSize=452]
INFO  : Partition {jobtitle=Operations Officer V} stats: [numFiles=1, numRows=47, to
talSize=4264, rawDataSize=4217]
INFO  : Partition {jobtitle=CONTRACT COOR CONVENTION} stats: [numFiles=1, numRows=2,
 totalSize=183, rawDataSize=181]
INFO  : Partition {jobtitle=BPD 1} stats: [numFiles=1, numRows=1, totalSize=77, rawD
ataSize=76]
INFO  : Partition {jobtitle=TIRE MAINTENANCE WORKER I} stats: [numFiles=1, numRows=7
, totalSize=595, rawDataSize=588]
INFO  : Partition {jobtitle=BPD 3} stats: [numFiles=1, numRows=1, totalSize=77, rawD
ataSize=76]
INFO  : Partition {jobtitle=DEPUTY SHERIFF SERGEANT} stats: [numFiles=1, numRows=9,
totalSize=779, rawDataSize=770]
INFO  : Partition {jobtitle=UTILITY INVESTIGATOR SUPV} stats: [numFiles=1, numRows=3
, totalSize=279, rawDataSize=276]
INFO  : Completed executing command(queryId=hive_20230927030324_ca2b446c-a2bd-411b-8
1fe-fdb38539dc87); Time taken: 183.413 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (186.836 seconds)
```

EXERCISE - 1

```
0: jdbc:hive2://localhost:10000/ (cs595)> create database MyDb;
INFO  : Compiling command(queryId=hive_20230927030702_6fb679a6-4df3-4e58-92fe-fbe78c
0fa4bc): create database MyDb
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030702_6fb679a6-4df3-4e58-9
2fe-fbe78c0fa4bc); Time taken: 0.003 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030702_6fb679a6-4df3-4e58-92fe-fbe78c
0fa4bc): create database MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030702_6fb679a6-4df3-4e58-9
2fe-fbe78c0fa4bc); Time taken: 0.048 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.067 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (cs595)> USE MyDb;
INFO  : Compiling command(queryId=hive_20230927030723_d984a3dd-69b7-457f-b81e-7b1372
b22514): USE MyDb
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927030723_d984a3dd-69b7-457f-b
81e-7b1372b22514); Time taken: 0.008 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927030723_d984a3dd-69b7-457f-b81e-7b1372
b22514): USE MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927030723_d984a3dd-69b7-457f-b
81e-7b1372b22514); Time taken: 0.006 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.03 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> create table if not exists MyDb.foodratings
(
. . . . . .> name STRING COMMENT 'Food name',
. . . . . .> food1 INT COMMENT 'Ratings food1',
. . . . . .> food2 INT COMMENT 'Ratings food2',
. . . . . .> food3 INT COMMENT 'Ratings food3',
. . . . . .> food4 INT COMMENT 'Ratings food4',
. . . . . .>  id INT COMMENT 'Food rating'
. . . . . .> )
. . . . . .> COMMENT 'Food rating'
. . . . . .> ROW FORMAT DELIMITED FIELDS TERMINATED BY '
,'
. . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20230927031126_fa1219a8-fcf0-40ba-9846-7ad15e
3bb72a): create table if not exists MyDb.foodratings(
name STRING COMMENT 'Food name',
food1 INT COMMENT 'Ratings food1',
food2 INT COMMENT 'Ratings food2',
food3 INT COMMENT 'Ratings food3',
food4 INT COMMENT 'Ratings food4',
id INT COMMENT 'Food rating'
)
COMMENT 'Food rating'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031126_fa1219a8-fcf0-40ba-9
846-7ad15e3bb72a); Time taken: 0.043 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031126_fa1219a8-fcf0-40ba-9846-7ad15e
3bb72a): create table if not exists MyDb.foodratings(
name STRING COMMENT 'Food name',
food1 INT COMMENT 'Ratings food1',
food2 INT COMMENT 'Ratings food2',
food3 INT COMMENT 'Ratings food3',
food4 INT COMMENT 'Ratings food4',
id INT COMMENT 'Food rating'
)
COMMENT 'Food rating'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927031126_fa1219a8-fcf0-40ba-9
846-7ad15e3bb72a); Time taken: 0.067 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.141 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> describe formatted MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20230927031154_fda0294d-8406-4935-b047-60d1a4
6140ec): describe formatted MyDb.foodratings
```

```
 | NULL                                                |
| Retention:                    | 0
 | NULL                                                |
| Location:                     | hdfs://ip-172-31-4-182.us-east-2.compute.internal:
8020/user/hive/warehouse/mydb.db/foodratings | NULL
            |
| Table Type:                   | MANAGED_TABLE
 | NULL                                                |
| Table Parameters:             | NULL
 | NULL                                                |
|                               | COLUMN_STATS_ACCURATE
 | {\"BASIC_STATS\":\"true\",\"COLUMN_STATS\":{\"food1\":\"true\",\"food2\":\"true\"
,\"food3\":\"true\",\"food4\":\"true\",\"id\":\"true\",\"name\":\"true\"}} |
|                               | bucketing_version
 | 2                                                   |
|                               | comment
 | Food rating                                         |
|                               | numFiles
 | 0                                                   |
|                               | numRows
 | 0                                                   |
|                               | rawDataSize
 | 0                                                   |
|                               | totalSize
 | 0                                                   |
|                               | transient_lastDdlTime
 | 1695784286                                          |
|                               | NULL
 | NULL                                                |
| # Storage Information         | NULL
 | NULL                                                |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
 | NULL                                                |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat
 | NULL                                                |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutp
utFormat | NULL                                                 |
| Compressed:                   | No
 | NULL                                                |
| Num Buckets:                  | -1
 | NULL                                                |
| Bucket Columns:               | []
 | NULL                                                |
| Sort Columns:                 | []
 | NULL                                                |
| Storage Desc Params:          | NULL
 | NULL                                                |
|                               | field.delim
 | ,                                                   |
|                               | serialization.format
 | ,                                                   |
+-------------------------------+-----------------------------------------------
-+-------------------------------------------------+
38 rows selected (0.155 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> create table if not exists MyDb.foodplaces(

. . . . . . . . . . . . . . . . . . . . .>  id INT,
. . . . . . . . . . . . . . . . . . . . .>  place String
. . . . . . . . . . . . . . . . . . . . .>  )
. . . . . . . . . . . . . . . . . . . . .>  ROW FORMAT DELIMITED FIELDS TERMINATED BY
; ;
 ','
. . . . . . . . . . . . . . . . . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20230927031322_343cca1a-6248-4a02-9c65-408c97
d5a31f): create table if not exists MyDb.foodplaces(
id INT,
place String
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031322_343cca1a-6248-4a02-9
c65-408c97d5a31f); Time taken: 0.008 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031322_343cca1a-6248-4a02-9c65-408c97
d5a31f): create table if not exists MyDb.foodplaces(
id INT,
place String
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927031322_343cca1a-6248-4a02-9
c65-408c97d5a31f); Time taken: 0.027 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.047 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> describe formatted MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20230927031350_942ba8ab-8d3e-4077-8e8c-2941c8
7be337): describe formatted MyDb.foodplaces
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:
string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment
:from deserializer), FieldSchema(name:comment, type:string, comment:from deserialize
r)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031350_942ba8ab-8d3e-4077-8
e8c-2941c87be337); Time taken: 0.017 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031350_942ba8ab-8d3e-4077-8e8c-2941c8
7be337): describe formatted MyDb.foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927031350_942ba8ab-8d3e-4077-8
e8c-2941c87be337); Time taken: 0.017 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------------+----------------------------------------------------
-+----------------------------------------------------+
|           col_name            |                     data_type
 |                     comment                        |
+-------------------------------+----------------------------------------------------
-+----------------------------------------------------+
| # col_name                    | data_type
 | comment                                            |
| id                            | int
 |                                                    |
| place                         | string
 |                                                    |
|                               | NULL
 | NULL                                               |
| # Detailed Table Information  | NULL
 | NULL                                               |
| Database:                     | mydb
 | NULL                                               |
| OwnerType:                    | USER
 | NULL                                               |
| Owner:                        | hadoop
 | NULL                                               |
| CreateTime:                   | Wed Sep 27 03:13:22 UTC 2023
 | NULL                                               |
| LastAccessTime:               | UNKNOWN
 | NULL                                               |
| Retention:                    | 0
 | NULL                                               |
| Location:                     | hdfs://ip-172-31-4-182.us-east-2.compute.internal:
8020/user/hive/warehouse/mydb.db/foodplaces | NULL
            |
| Table Type:                   | MANAGED_TABLE
```

```
| 2                                                  |
|                               | numFiles           |
| 0                                                  |
|                               | numRows            |
| 0                                                  |
|                               | rawDataSize        |
| 0                                                  |
|                               | totalSize          |
| 0                                                  |
|                               | transient_lastDdlTime |
| 1695784402                                         |
|                               | NULL               |
| NULL                                               |
| # Storage Information         | NULL               |
| NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe |
| NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat |
| NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutp
utFormat | NULL                                            |
| Compressed:                   | No                 |
| NULL                                               |
| Num Buckets:                  | -1                 |
| NULL                                               |
| Bucket Columns:               | []                 |
| NULL                                               |
| Sort Columns:                 | []                 |
| NULL                                               |
| Storage Desc Params:          | NULL               |
| NULL                                               |
|                               | field.delim        |
| ,                                                  |
|                               | serialization.format |
| ,                                                  |
+-------------------------------+----------------------------------------------------
-+----------------------------------------------------+
33 rows selected (0.053 seconds)
```

EXERCISE - 2

```
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodra
tings49384.txt'
. . . . . .> INTO TABLE MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20230927031522_168af00f-4219-43f2-936c-43dae7
f5f6f0): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings49384.txt'
INTO TABLE MyDb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031522_168af00f-4219-43f2-9
36c-43dae7f5f6f0); Time taken: 0.012 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031522_168af00f-4219-43f2-936c-43dae7
f5f6f0): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings49384.txt'
INTO TABLE MyDb.foodratings
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratings from file:/home/hadoop/foodratings493
84.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table mydb.foodratings stats: [numFiles=1, numRows=0, totalSize=17507, rawDa
taSize=0]
INFO  : Completed executing command(queryId=hive_20230927031522_168af00f-4219-43f2-9
36c-43dae7f5f6f0); Time taken: 0.148 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.176 seconds)
```

Command and Output:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> select min(food3) as min, max(food3) as max
 , avg(food3) as average
. . . . . . . . . . . . . . . . . . . . .> from MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20230927031629_0f784ab3-387d-41bc-b71f-fbf7a9
213c0c): select min(food3) as min, max(food3) as max , avg(food3) as average
from MyDb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int,
comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:avera
ge, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031629_0f784ab3-387d-41bc-b
71f-fbf7a9213c0c); Time taken: 0.541 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031629_0f784ab3-387d-41bc-b71f-fbf7a9
213c0c): select min(food3) as min, max(food3) as max , avg(food3) as average
from MyDb.foodratings
INFO  : Query ID = hive_20230927031629_0f784ab3-387d-41bc-b71f-fbf7a9213c0c
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927031629_0f784ab3-387d-41
bc-b71f-fbf7a9213c0c
INFO  : Session is already open
INFO  : Dag name: select min(food3) as min,...MyDb.foodratings (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0002)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/1      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20230927031629_0f784ab3-387d-41bc-b
71f-fbf7a9213c0c); Time taken: 14.21 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+------+----------+
| min  | max  | average  |
+------+------+----------+
| 1    | 50   | 24.65    |
+------+------+----------+
1 row selected (14.798 seconds)
```

Magic Number: 49384

EXERCISE - 3

Command and Output:

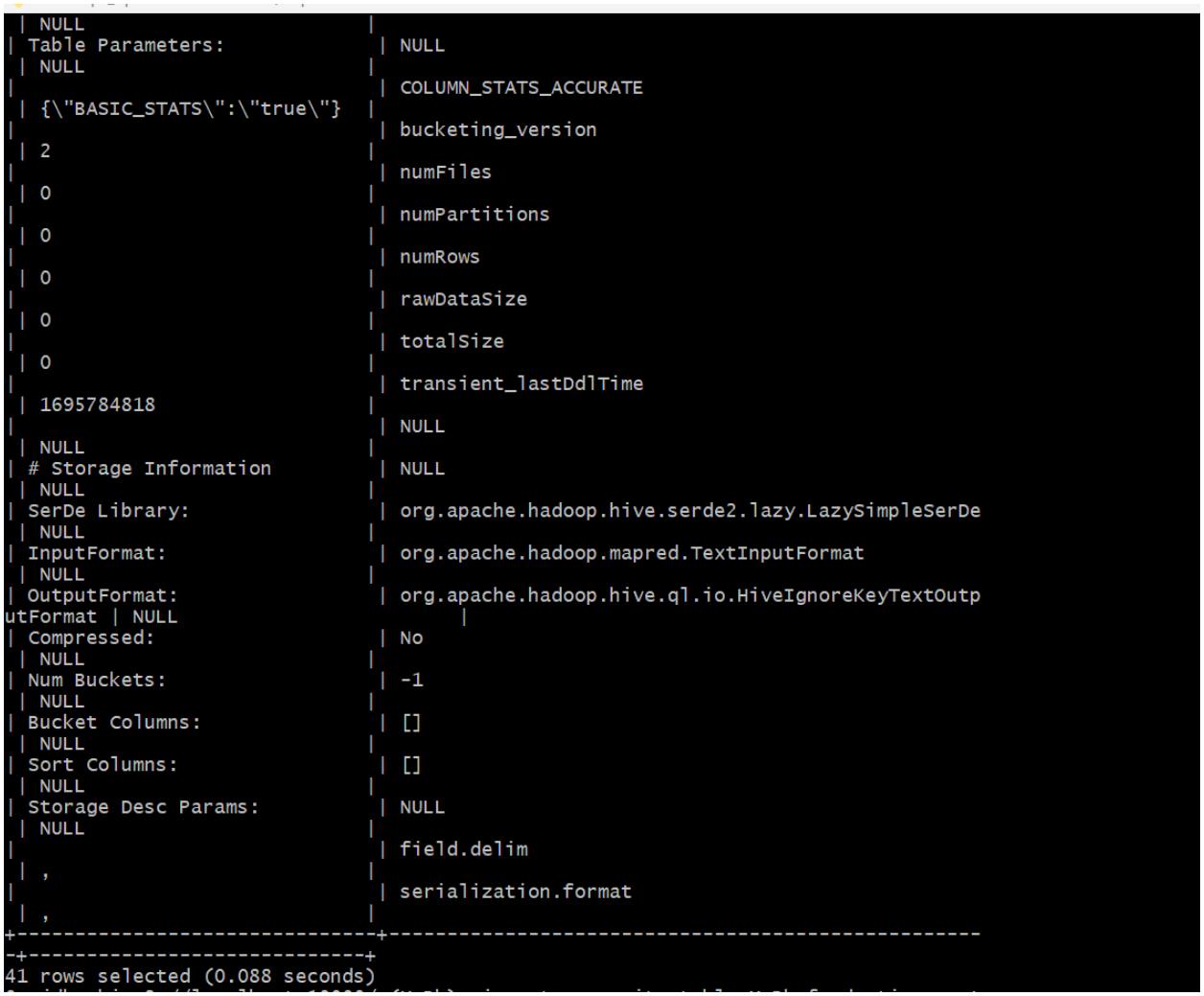
```
0: jdbc:hive2://localhost:10000/ (MyDb)> select name, min(food1) as min, max(food1)
as max, avg(food1) as average
. . . . . .> from MyDb.foodratings Group by name;
INFO  : Compiling command(queryId=hive_20230927031728_131ba808-e6b0-448a-b4a5-3800ab
db9a59): select name, min(food1) as min, max(food1) as max, avg(food1) as average
from MyDb.foodratings Group by name
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:stri
ng, comment:null), FieldSchema(name:min, type:int, comment:null), FieldSchema(name:m
ax, type:int, comment:null), FieldSchema(name:average, type:double, comment:null)],
properties:null)
INFO  : Completed compiling command(queryId=hive_20230927031728_131ba808-e6b0-448a-b
4a5-3800abdb9a59); Time taken: 0.094 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927031728_131ba808-e6b0-448a-b4a5-3800ab
db9a59): select name, min(food1) as min, max(food1) as max, avg(food1) as average
from MyDb.foodratings Group by name
INFO  : Query ID = hive_20230927031728_131ba808-e6b0-448a-b4a5-3800abdb9a59
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927031728_131ba808-e6b0-44
8a-b4a5-3800abdb9a59
INFO  : Session is already open
INFO  : Dag name: select name, min(food1) as min, max(f...name (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0002)

INFO  : Map 1: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20230927031728_131ba808-e6b0-448a-b
4a5-3800abdb9a59); Time taken: 5.32 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+------+------+--------------------+
| name  | min  | max  |       average      |
+-------+------+------+--------------------+
| Joy   | 1    | 50   | 25.693121693121693 |
| Jill  | 1    | 50   | 25.795             |
| Joe   | 1    | 50   | 25.695238095238096 |
| Mel   | 1    | 50   | 28.190721649484537 |
| Sam   | 1    | 49   | 24.647342995169083 |
+-------+------+------+--------------------+
5 rows selected (5.448 seconds)
```

EXERCISE - 4

```
0: jdbc:hive2://localhost:10000/ (MyDb)> create table if not exists MyDb.foodratings
part(
. . . . . .> food1 INT,
. . . . . .> food2 INT,
. . . . . .> food3 INT,
. . . . . .> food4 INT,
. . . . . .> id INT
. . . . . .>  )
. . . . . .>  PARTITIONED BY(name STRING)
. . . . . .> ROW FORMAT DELIMITED FIELDS TERMINATED BY '
,'
. . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20230927032018_fd0d4434-2dc1-49bd-8021-b496ef
d46efd): create table if not exists MyDb.foodratingspart(
food1 INT,
food2 INT,
food3 INT,
food4 INT,
id INT
)
PARTITIONED BY(name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927032018_fd0d4434-2dc1-49bd-8
021-b496efd46efd); Time taken: 0.014 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032018_fd0d4434-2dc1-49bd-8021-b496ef
d46efd): create table if not exists MyDb.foodratingspart(
food1 INT,
food2 INT,
food3 INT,
food4 INT,
id INT
)
PARTITIONED BY(name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927032018_fd0d4434-2dc1-49bd-8
021-b496efd46efd); Time taken: 0.023 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.049 seconds)
```

```
No rows affected (0.045 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> describe formatted MyDb.foodratingspart;
INFO  : Compiling command(queryId=hive_20230927032101_a0c3a3ec-0ed5-423e-884f-ff97c4
c63305): describe formatted MyDb.foodratingspart
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:
string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment
:from deserializer), FieldSchema(name:comment, type:string, comment:from deserialize
r)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927032101_a0c3a3ec-0ed5-423e-8
84f-ff97c4c63305); Time taken: 0.017 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032101_a0c3a3ec-0ed5-423e-884f-ff97c4
c63305): describe formatted MyDb.foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230927032101_a0c3a3ec-0ed5-423e-8
84f-ff97c4c63305); Time taken: 0.044 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------------+-----------------------------------------------
-+-----------------------------+
|           col_name            |                     data_type
 |           comment             |
+-------------------------------+-----------------------------------------------
-+-----------------------------+
| # col_name                    | data_type
 | comment                     |
| food1                         | int
 |                             |
| food2                         | int
 |                             |
| food3                         | int
 |                             |
| food4                         | int
 |                             |
| id                            | int
 |                             |
|                               | NULL
 | NULL                        |
| # Partition Information       | NULL
 | NULL                        |
| # col_name                    | data_type
 | comment                     |
| name                          | string
 |                             |
|                               | NULL
 | NULL                        |
| # Detailed Table Information  | NULL
 | NULL                        |
| Database:                     | mydb
 | NULL                        |
| OwnerType:                    | USER
```



EXERCISE - 5

The names of food consumers are an easy way to organize information because there are typically fewer of them than there are restaurants. As a result, researching a specific critic is quicker than researching a particular place. It's not a good idea to use a restaurant's identification number, though, as there can be several of them, each with a different division. Finding data will therefore grow more challenging and time-consuming. So keep that in mind. A smaller division can help with quicker data search results.

EXERCISE - 6

```
0: jdbc:hive2://localhost:10000/ (MyDb)> insert overwrite table MyDb.foodratingspart

. . . . . .> partition (name)
. . . . . .> select food1, food2, food3, food4, id, name

. . . . . .> from MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20230927032215_816682b1-b7f7-425d-b775-e29d97
487fb0): insert overwrite table MyDb.foodratingspart
partition (name)
select food1, food2, food3, food4, id, name
from MyDb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:food1, type:int
, comment:null), FieldSchema(name:food2, type:int, comment:null), FieldSchema(name:f
ood3, type:int, comment:null), FieldSchema(name:food4, type:int, comment:null), Fiel
dSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, commen
t:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927032215_816682b1-b7f7-425d-b
775-e29d97487fb0); Time taken: 0.156 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032215_816682b1-b7f7-425d-b775-e29d97
487fb0): insert overwrite table MyDb.foodratingspart
partition (name)
select food1, food2, food3, food4, id, name
from MyDb.foodratings
INFO  : Query ID = hive_20230927032215_816682b1-b7f7-425d-b775-e29d97487fb0
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927032215_816682b1-b7f7-42
5d-b775-e29d97487fb0
INFO  : Session is already open
INFO  : Dag name: insert overwrite table My...MyDb.foodratings (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0002)

INFO  : Map 1: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs:/
/ip-172-31-4-182.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodrat
ingspart/.hive-staging_hive_2023-09-27_03-22-15_085_4162509450975083351-1/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.201 seconds
INFO  :          Time taken for adding to write entity : 0.0 seconds
```

```
INFO  : Map 1: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs:/
/ip-172-31-4-182.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodrat
ingspart/.hive-staging_hive_2023-09-27_03-22-15_085_4162509450975083351-1/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.201 seconds
INFO  :          Time taken for adding to write entity : 0.0 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Executing stats task
INFO  : Partition {name=Mel} stats: [numFiles=1, numRows=194, totalSize=2596, rawDat
aSize=2402]
INFO  : Partition {name=Jill} stats: [numFiles=1, numRows=200, totalSize=2656, rawDa
taSize=2456]
INFO  : Partition {name=Joe} stats: [numFiles=1, numRows=210, totalSize=2762, rawDat
aSize=2552]
INFO  : Partition {name=Sam} stats: [numFiles=1, numRows=207, totalSize=2765, rawDat
aSize=2558]
INFO  : Partition {name=Joy} stats: [numFiles=1, numRows=189, totalSize=2528, rawDat
aSize=2339]
INFO  : Completed executing command(queryId=hive_20230927032215_816682b1-b7f7-425d-b
775-e29d97487fb0); Time taken: 6.749 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (6.914 seconds)
```

Command and Output:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> select min(food2) as min, max(food2) as max
 , avg(food2) as average
. . . . . .> from MyDb.foodratingspart
. . . . . .> where name = 'Mel' or name = 'Jill';
INFO  : Compiling command(queryId=hive_20230927032320_a8c27873-c356-4dd1-a559-cbed6f
b1e73b): select min(food2) as min, max(food2) as max , avg(food2) as average
from MyDb.foodratingspart
where name = 'Mel' or name = 'Jill'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int,
comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:avera
ge, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230927032320_a8c27873-c356-4dd1-a
559-cbed6fb1e73b); Time taken: 0.486 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032320_a8c27873-c356-4dd1-a559-cbed6f
b1e73b): select min(food2) as min, max(food2) as max , avg(food2) as average
from MyDb.foodratingspart
where name = 'Mel' or name = 'Jill'
INFO  : Query ID = hive_20230927032320_a8c27873-c356-4dd1-a559-cbed6fb1e73b
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927032320_a8c27873-c356-4d
d1-a559-cbed6fb1e73b
INFO  : Session is already open
INFO  : Dag name: select min(food2) as min, max(food2...'Jill' (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0002)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/1      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20230927032320_a8c27873-c356-4dd1-a
559-cbed6fb1e73b); Time taken: 5.173 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+------+---------------------+
| min  | max  |       average       |
+------+------+---------------------+
| 1    | 50   | 25.890862944162436  |
+------+------+---------------------+
1 row selected (5.69 seconds)
```

EXERCISE - 7

```
0: jdbc:hive2://localhost:10000/ (MyDb)> load data local inpath '/home/hadoop/foodpl
aces49384.txt'
. . . . . .> overwrite into table MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20230927032415_5264867e-c451-4098-8e03-0d8ff1
94b9e2): load data local inpath '/home/hadoop/foodplaces49384.txt'
overwrite into table MyDb.foodplaces
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230927032415_5264867e-c451-4098-8
e03-0d8ff194b9e2); Time taken: 0.017 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032415_5264867e-c451-4098-8e03-0d8ff1
94b9e2): load data local inpath '/home/hadoop/foodplaces49384.txt'
overwrite into table MyDb.foodplaces
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodplaces from file:/home/hadoop/foodplaces49384
.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table mydb.foodplaces stats: [numFiles=1, numRows=0, totalSize=59, rawDataSi
ze=0]
INFO  : Completed executing command(queryId=hive_20230927032415_5264867e-c451-4098-8
e03-0d8ff194b9e2); Time taken: 0.098 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.126 seconds)
```

Command and Output:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> select p.place, avg(r.food4) as average
. . . . . .> from MyDb.foodratings r
. . . . . .> JOIN MyDb.foodplaces p
. . . . . .> on p.id=r.id
. . . . . .> where p.place = 'Soup Bowl'
. . . . . .> Group by p.place;
INFO  : Compiling command(queryId=hive_20230927032537_e4063e14-3315-423a-a84c-ab04da
9743e5): select p.place, avg(r.food4) as average
from MyDb.foodratings r
JOIN MyDb.foodplaces p
on p.id=r.id
where p.place = 'Soup Bowl'
Group by p.place
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:p.place, type:s
tring, comment:null), FieldSchema(name:average, type:double, comment:null)], propert
ies:null)
INFO  : Completed compiling command(queryId=hive_20230927032537_e4063e14-3315-423a-a
84c-ab04da9743e5); Time taken: 0.393 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230927032537_e4063e14-3315-423a-a84c-ab04da
9743e5): select p.place, avg(r.food4) as average
from MyDb.foodratings r
JOIN MyDb.foodplaces p
on p.id=r.id
where p.place = 'Soup Bowl'
Group by p.place
INFO  : Query ID = hive_20230927032537_e4063e14-3315-423a-a84c-ab04da9743e5
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230927032537_e4063e14-3315-42
3a-a84c-ab04da9743e5
INFO  : Session is already open
INFO  : Dag name: select p.place, avg(r.food4) as av...p.place (Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_169578280
4851_0002)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/2
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1(+1)/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2
INFO  : Completed executing command(queryId=hive_20230927032537_e4063e14-3315-423a-a
84c-ab04da9743e5); Time taken: 8.264 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
```

```
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1(+1)/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2
INFO  : Completed executing command(queryId=hive_20230927032537_e4063e14-3315-423a-a
84c-ab04da9743e5); Time taken: 8.264 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+--------------------+
|  p.place   |      average       |
+------------+--------------------+
| Soup Bowl  | 26.15740740740741  |
+------------+--------------------+
1 row selected (8.674 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

EXERCISE - 8

a) Selecting a row format permits you to quickly and easily insert the most recent information. It functions well when a whole row of data needs to be accessible or processed at once since each entry needs to have access to one or more columns as well as one or more entries. However, when utilizing a column format, every piece of data not needed for a particular query is removed. A huge dataset is simple to read because you don't have to spend time sorting through pointless data. When similar types of data are combined, the dataset uses less storage space since the number of null results in columns with few items is decreased.

b) Make sure the data can be broken up into various, slowly manageable bits when utilizing a column-based format. For this, the term "splitability" is relevant. It is simpler to divide the data into distinct jobs if the query only has to look at one field at a time. Data processing is more productive as a result because various techniques can be used to perform each task. This means that the data should be divided into as many pieces as feasible, with each component passing through a different stage of processing. Operation distribution among multiple steps as possible will provide optimal performance.

c) Data is organized by column rather than row when it is saved in a columnar format. Because all related information is saved together, working with data is made simpler. The procedure can be sped up and made more effective by avoiding the columns you don't need to look at. This facilitates computing and data compression. Furthermore, it makes it simple for the computer to ignore irrelevant entries, which helps save time and resources.

d) For evaluating a lot of data with numerous columns, the checkerboard data format is great. Data input and reading are intended to be rapid and simple. This is done, among other things, by including details about the different data kinds and compression techniques at the conclusion of the file. The genuine data can be obtained more quickly because you don't need to sort through every piece of metadata. Apache Impala is a database that is used by Hadoop machine

learning. Parquet is widely utilized because of how well it manages database systems with several columns. This makes it straightforward to quickly review the data, even while working with several different questions all at once.

Submitted by:
Sailavanya Narthu
A20516764
snarthu@hawk.iit.edu