CSP554—Big Data Technologies

Assignment #9

Worth: 15 points (1 point for each problem)

Exercises

Exercise 1) (1 point)

Create an HBase table with the following characteristics

Table Name: csp554Tbl

First column family: cf1

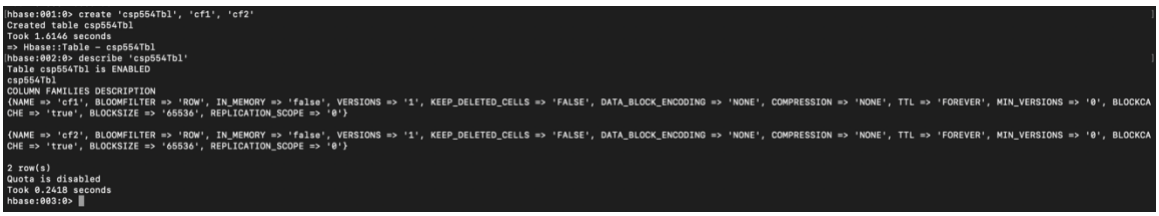
Second column family: cf2

Then execute the DESCRIBE command on the table and return command you wrote and the output as the results of this exercise.

Commands:

```
create 'csp554Tbl', 'cf1', 'cf2'
describe 'csp554Tbl'
```

OUTPUT:



Exercise 2)

Put the following data into the table created in exercise 1:

Row Key	Column Family	Column (Qualifier)	Value
Row1	cf1	name	Sam
Row2	cf1	name	Ahmed
Row1	cf2	job	Pilot
Row2	cf2	job	Doctor
Row1	cf2	level	LZ3
Row2	cf2	level	AR7

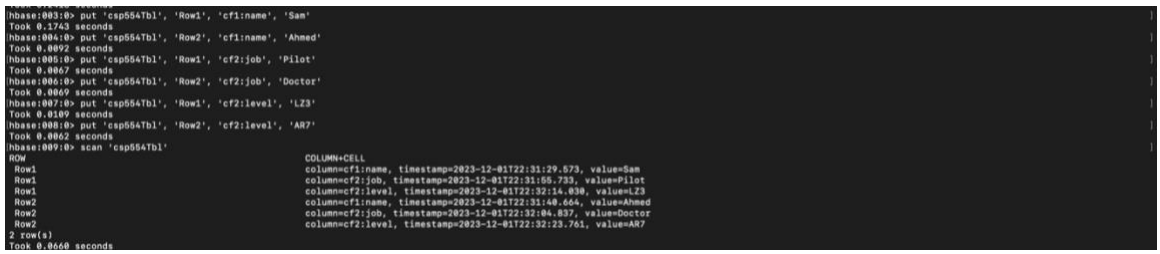
Execute the SCAN command on this table returning all rows, column families and columns. Provide the command and its result as the output of this exercise.

COMMANDS:

```
put 'csp554Tbl', 'Row1', 'cf1:name', 'Sam'
put 'csp554Tbl', 'Row2', 'cf1:name', 'Ahmed'
put 'csp554Tbl', 'Row1', 'cf2:job', 'Pilot'
put 'csp554Tbl', 'Row2', 'cf2:job', 'Doctor'
put 'csp554Tbl', 'Row1', 'cf2:level', 'LZ3'
put 'csp554Tbl', 'Row2', 'cf2:level', 'AR7'

scan 'csp554Tbl'
```

OUTPUT:



Exercise 3)

Using the above table write a command that will get the value associated with row (Row1), column family (cf2) and column/qualifier (level). Provide the command and its result as the output of this exercise.

COMMANDS:

```
get 'csp554Tbl', 'Row1', {COLUMN => 'cf2:level'}
```

OUTPUT:

```
Took 0.0068 seconds
hbase:010:0> get 'csp554Tbl', 'Row1', {COLUMN => 'cf2:level'}
COLUMN                                  CELL
 cf2:level                              timestamp=2023-12-01T22:32:14.030, value=LZ3
1 row(s)
Took 0.0332 seconds
```

Exercise 4)

Using the above table write command that will get the value associated with row (Row2), column family (cf1) and column/qualifier (name). Provide the command and its result as the output of this exercise.

COMMANDS:

```
get 'csp554Tbl', 'Row2', {COLUMN => 'cf1:name'}
```

OUTPUT:

```
hbase:011:0> get 'csp554Tbl', 'Row2', {COLUMN => 'cf1:name'}
COLUMN                                  CELL
 cf1:name                               timestamp=2023-12-01T22:31:40.664, value=Ahmed
1 row(s)
Took 0.0208 seconds
```

Exercise 5)

Using the above table write a SCAN command that will return information about only two rows using the LIMIT modifier. Provide the command and its result as the output of this exercise.

COMMAND:

```
scan 'csp554Tbl', {'LIMIT' => 2}
```

OUTPUT:

```
hbase:012:0> scan 'csp554Tbl', {'LIMIT' => 2}
ROW                                     COLUMN+CELL
 Row1                                   column=cf1:name, timestamp=2023-12-01T22:31:29.573, value=Sam
 Row1                                   column=cf2:job, timestamp=2023-12-01T22:31:55.733, value=Pilot
 Row1                                   column=cf2:level, timestamp=2023-12-01T22:32:14.030, value=LZ3
 Row2                                   column=cf1:name, timestamp=2023-12-01T22:31:40.664, value=Ahmed
 Row2                                   column=cf2:job, timestamp=2023-12-01T22:32:04.837, value=Doctor
 Row2                                   column=cf2:level, timestamp=2023-12-01T22:32:23.761, value=AR7
2 row(s)
Took 0.0318 seconds
hbase:013:0>
```

Cassandra

Exercises:

Exercise 1) (1 point)

Read the article "A Big Data Modeling Methodology for Apache Cassandra" and provide a ½ page summary including your comments and impressions.

- Summary

Apache Cassandra is a premier scalable, highly available, and transactional distributed database. It has maintained some of the world's largest databases on clusters with tens of thousands of nodes distributed across several data centres. Several reasons, including its scalable and fault-tolerant peer-to-peer architecture, adaptive and flexible data model derived from the BigTable data model, declarative nature, and user-friendliness, have contributed to Cassandra's widespread adoption in big data applications. Cassandra data management use cases include product catalogues and playlists, sensor data and the Internet of Things, messaging and social networking, recommendation, customization, fraud detection, and a variety of other time series data-related applications.

In this paper, a rigorous query-driven data modeling approach for Apache Cassandra is presented. In a variety of areas, including query-driven schema design, data nesting, and data duplication, the methodology was demonstrated to be significantly different from the conventional relational data modelling approach. In order to move from conceptual data models that are independent of technology to logical data models that are specific to Cassandra, this article describes mapping rules and mapping patterns and elaborates on the core data modeling concepts for Cassandra. Critical large data applications may scale to millions of transactions per second using the Cassandra Query Language (CQL) and very efficient write and read access pathways, and they can easily tolerate node failures and even data center failures.

It also emphasises the purpose of physical data modelling and recommends Chebotko Diagrams, a cutting-edge visualisation tool for recording complex logical and physical data models. It also introduces KDM, a powerful data modelling tool that automates some of the most complex, error-prone, and time-consuming data modelling procedures, such as conceptual-to-logical mapping, logical-to-physical mapping, and CQL development.

Exercise 2) (1 point)

Step A – Start an EMR cluster

Start up an EMR/Hadoop cluster as previously, but instead of choosing the "Core Hadoop" configuration chose the "Spark" configuration (see below), otherwise proceed as before.

Step B – Install the Cassandra database software and start it

Open up a terminal connection to your EMR master node. Over the course of this exercise, you will need to open up three separate terminal connections to your EMR master node. This is the first, which we will call Cass-Term.

Enter the following two commands:

```
wget https://archive.apache.org/dist/cassandra/3.11.2/apache-cassandra-3.11.2-bin.tar.gz
```

tar -xzvf apache-cassandra-3.11.2-bin.tar.gz

Note, this will create a new directory (apache-cassandra-3.11.2) holding the Cassandra software release.

Then enter this command to start Cassandra (lots of diagnostic messages will appear):

apache-cassandra-3.11.2/bin/cassandra &

Step C – Run the Cassandra interactive command line interface

Open a second terminal connection to the EMR master node. Going forward we will call this terminal connection: Cqlsh-Term.

Enter the following into this terminal to start the command line interface csqlsh:

apache-cassandra-3.11.2/bin/cqlsh

Step D – Prepare to edit your Cassandra code

Open a third terminal connection to the EMR master node. Going forward we will call this terminal connection: Edit-Term.

You will use this terminal window to run the 'vi' editor to create your Cassandra code files. See the "Free Books and Chapters" section of our blackboard site for information on how to use the 'vi' editor.

As an alternative you could edit your Cassandra code files on your PC/MAC using a text editor like "notepad" or "textedit" "and then 'scp' them to the EMR mater node.

a) Create a file in your working (home) directory called init.cql using your Edit-term (or using your PC/MAC and then scp it to the EMR master node) and enter the following command. Use your IIT id as the name of your keyspace… For example, if your id is A1234567, then replace <IIT id> below with that value:

CREATE KEYSPACE <IIT id> WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };

For example, you might write:

CREATE KEYSPACE A1234567 WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };

b) Then execute this file in the CQL shell using the Cqlsh-Term as follows…

source ‘./init.cql’;

c) To check if your script file has created a keyspace execute the following in the CQL shell:

describe keyspaces;

d) At this point you have created a keyspace unique to you. So, make that keyspace the default by entering the following into the CQL shell:

USE <IIT id>;

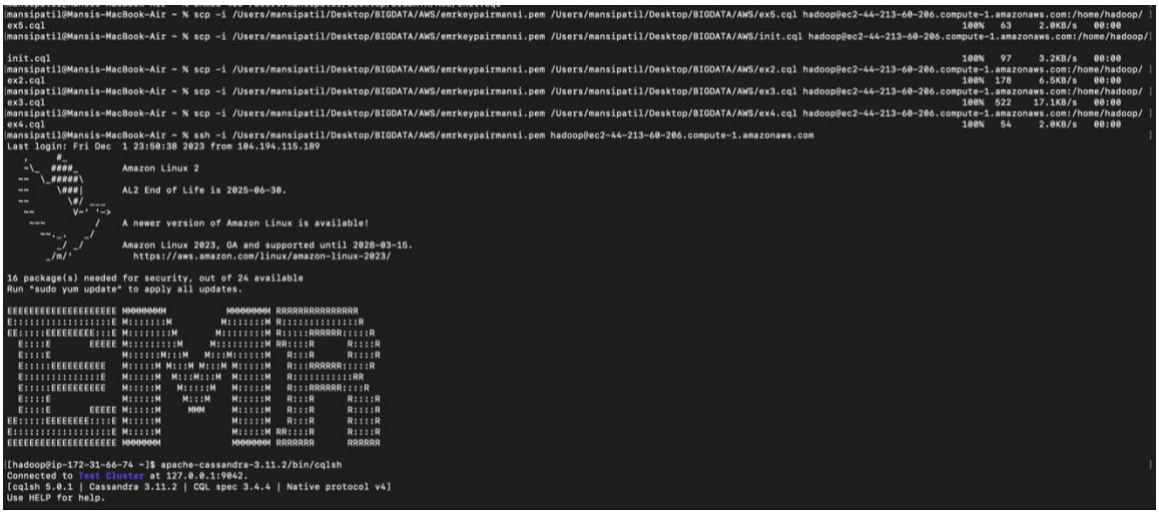
For example,

USE A1234567;

Now create a file in your working directory called ex2.cql using the Edit-Term (or PC/MAC and scp). In this file write the command to create a table named ‘Music’ with the following characteristics:

Attribute Name	Attribute Type	Primary Key / Cluster Key
artistName	text	Primary Key
albumName	text	Cluster Key
numberSold	int	Non Key Column
Cost	int	Non Key Column

Execute ex2.cql in the CQL shell. Then execute the shell command ‘DESCRIBE TABLE Music;’ and include the output as the result of this exercise.




```
cqlsh:a20549858> source './ex2.cql';
cqlsh:a20549858> DESCRIBE TABLE MUSIC;

CREATE TABLE a20549858.music (
    artistname text,
    albumname text,
    cost int,
    numbersold int,
    PRIMARY KEY (artistname, albumname)
) WITH CLUSTERING ORDER BY (albumname DESC)
    AND bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
    AND compression = {'chunk_length_in_kb': '64', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND crc_check_chance = 1.0
    AND dclocal_read_repair_chance = 0.1
    AND default_time_to_live = 0
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair_chance = 0.0
    AND speculative_retry = '99PERCENTILE';
```

Exercise 3) (1 point)

Now create a file in your working directory called ex3.cql using the Edit-Term. In this file write the commands to insert the following records into table ‘Music’…

artistName	albumName	numberSold	cost
Mozart	Greatest Hits	100000	10
Taylor Swift	Fearless	2300000	15
Black Sabbath	Paranoid	534000	12
Katy Perry	Prism	800000	16
Katy Perry	Teenage Dream	750000	14

a) Execute ex3.cql. Provide the content of this file as the result of this exercise.

OUTPUT:

```
[hadoop@ip-172-31-66-74 ~]$ cat ex3.cql
INSERT INTO MUSIC(artistname, albumname, numbersold, cost) VALUES('Mozart','Greatest Hits',100000,10);
INSERT INTO MUSIC(artistname, albumname, numbersold, cost) VALUES('Taylor Swift','Fearless',2300000,15);
INSERT INTO MUSIC(artistname, albumname, numbersold, cost) VALUES('Black Sabbath','Paranoid',534000,12);
INSERT INTO MUSIC(artistname, albumname, numbersold, cost) VALUES('Katy Perry','Prism',800000,16);
INSERT INTO MUSIC(artistname, albumname, numbersold, cost) VALUES('Katy Perry','Teenage Dream',750000,14);[hadoop@ip-172-31-66-74 ~]$
```

b) Execute the command 'SELECT * FROM Music;' and provide the output of this command as another result of the exercise.

OUTPUT:

```
cqlsh:a20549858> source './ex3.cql';
cqlsh:a20549858> SELECT * FROM Music;

 artistname   | albumname     | cost | numbersold
--------------+---------------+------+------------
       Mozart | Greatest Hits |   10 |     100000
 Black Sabbath |      Paranoid |   12 |     534000
  Taylor Swift |      Fearless |   15 |    2300000
    Katy Perry | Teenage Dream |   14 |     750000
    Katy Perry |         Prism |   16 |     800000

(5 rows)
```

Exercise 4)

Now create a file in your working directory called ex4.cql using the Edit-Term. In this file write the commands to query and output only Katy Perry songs. Execute ex4.cql. Provide the content of this file and output of executing this file as the result of this exercise.

OUTPUT:

```
[hadoop@ip-172-31-66-74 ~]$ cat ex4.cql
SELECT * FROM Music where artistname in('Katy Perry');[hadoop@ip-172-31-66-74 ~]$
```

```
cqlsh:a20549858> source './ex4.cql';

 artistname | albumname     | cost | numbersold
------------+---------------+------+------------
 Katy Perry | Teenage Dream |   14 |     750000
 Katy Perry |         Prism |   16 |     800000
```

Exercise 5)

Now create a file in your working directory called ex5.cql using the Edit-Term. In this file write the commands to query only albums that have sold 700000 copies or more. Execute ex5.cql. Provide the content of this file and the output of executing this file as the result of this exercise.

OUTPUT:

```
[[hadoop@ip-172-31-66-74 ~]$ cat ex5.cql
 SELECT * FROM MUSIC WHERE numbersold >= 700000 ALLOW FILTERING;[
```

```
cqlsh:a20549858> source './ex5.cql';

 artistname   | albumname     | cost | numbersold
--------------+---------------+------+------------
 Taylor Swift |      Fearless |   15 |    2300000
   Katy Perry | Teenage Dream |   14 |     750000
   Katy Perry |         Prism |   16 |     800000

(3 rows)
```

MongoDB

Note, the files named “demo*.js” (also included in the mongoex.tar file) provide examples of how to operate in the unicorn collection. These are a VERY good idea to review and understand and will present you with information helpful in completing the assignment. Also, try them out by typing something like

load(./demo1.js’);

Exercises:

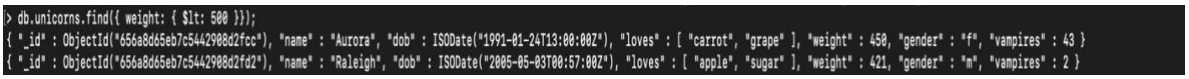
Exercise 1) (1 point)

Write a command that finds all unicorns having weight less than 500 pounds. Include the code you executed and some sample output as the result of this exercise. Recall you can place the command, if you choose, into a file, say 'ex1.js' and execute it with the load command as above and similarly for the following exercises.

COMMAND:

```
db.unicorns.find({ weight: { $lt: 500 }});
```

OUTPUT:



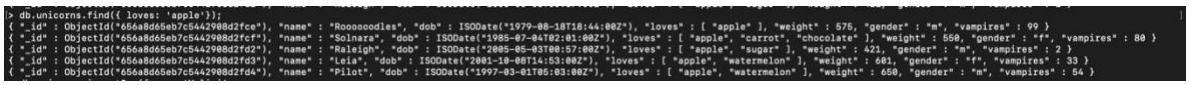
Exercise 2)

Write a command that finds all unicorns who love apples. Hint, search for "apple". Include the code you executed and some sample output as the result of this exercise.

COMMAND:

```
db.unicorns.find({ loves: 'apple'});
```

OUTPUT:



Exercise 3) (1 point)

Write a command that adds a unicorn with the following attributes to the collection. Note dob means "Date of Birth."

Attribute	Value(s)
name	Malini
dob	11/03/2008
loves	pears, grapes
weight	450
gender	F
vampires	23
horns	1

Include the code you executed to insert this unicorn into the collection along with the output of a find command showing it is in the collection.

COMMAND:

```
db.unicorns.insertOne({name: 'Malini',
dob: new Date(2008, 11, 03),
loves: ['pears', 'grapes'],
weight: 450,
gender: 'f',
vampires: 23,
horns: 1});

db.unicorns.find();
```

OUTPUT:

Exercise 4) (1 point)

Write a command that updates the above record to add apricots to the list of things Malini loves. Include the code you executed and some sample output showing the addition.

COMMAND:

```
db.unicorns.updateOne(
{ name: 'Malini' },
{ $push: { loves: 'apricots' } }
);
db.unicorns.find({ name: 'Malini' });
```

Exercise 5)

Write a command that deletes all unicorns with weight more than 600 pounds. Include the code you executed and some sample output as the result of this exercise.

COMMAND:

```
db.unicorns.find();
db.unicorns.deleteMany({ weight: { $gt: 600 }});

db.unicorns.find();
```

OUTPUT: