CS595—Big Data Technologies

Assignment #10

Worth: 12 points

Due by the start of the next class period

Assignments should be uploaded via the Blackboard portal.

Exercise 1) (4 points)

Read the article "A Big Data Modeling Methodology for Apache Cassandra" available on the blackboard in the 'Articles' section. Provide a ½ page summary including your comments and impressions.

Apache Cassandra is a newly created distributed database of choice when it comes to big data management with requirements of zero downtime, linear scalability, and seamless multiple data center deployment.

Cassandra data management use cases include product catalogs and playlists, sensor data and Internet of Things, messaging and social networking, recommendation, personalization, fraud detection, and numerous other applications that deal with time series data. The wide adoption of Cassandra in big data applications is attributed to, among other things, its scalable and fault-tolerant peer-to-peer architecture, versatile and flexible data model that evolved from the BigTable data model, declarative and user-friendly Cassandra Query Language (CQL), and very efficient write and read access paths that enable critical big data applications to stay always on, scale to millions of transactions per second, and handle node and even entire data center failures with ease. One of the biggest challenges that new projects face when adopting Cassandra is data modeling that has significant differences from traditional data modeling approaches used in the past.

Exercise 2) (2 points)

For this and the following exercises you will use and instance of the Cassandra database that I have set up for you in the Azure cloud. Note, as I am paying for this myself, I will only keep the database available until next Thursday.

To access the database do the following:

a) Access the databaseVM via ssh…

   ssh –p 22 cass1@13.90.97.198

b) When prompted for a password use…

   Unix79127912

c) Once logged in to the Cassandra VM create a working directory for yourself. The name of the directory should be your IIT id (mine is A20155104)

d) This will be your working directory into which you will place your CQL command files

e) Open a second terminal window and ssh into the VM again as above.

f) Change to the working directory you created previously.

g) Start the Casandra shell by entering the following…

   cqlsh –u cassandra

h) When prompted for a password, enter the following…

   PGAIGPNtUin2 <- the 4th letter is a capital I,

i) Now create a file in your working directory called init.cql and enter the following commands. Use your IIT id as the name of your keyspace…

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

j) Then execute this file in the CQL shell as follows…

   source ‘./init.cql’

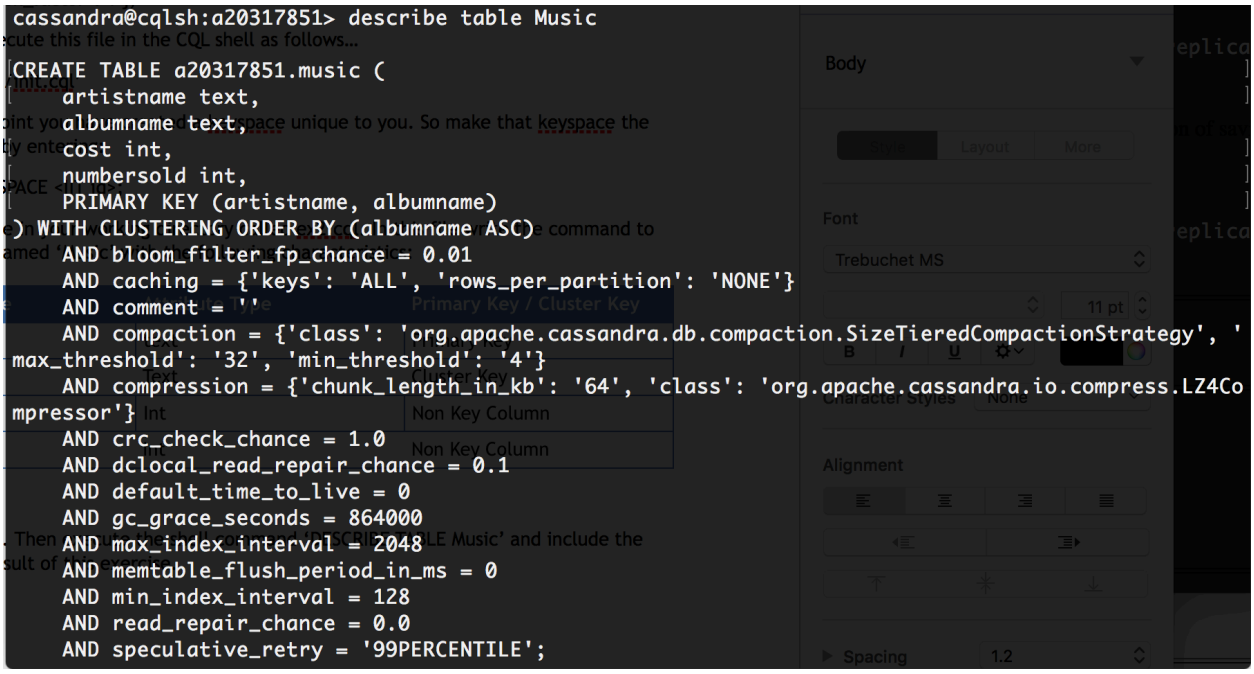
k) At this point you have created a keyspace unique to you. So make that keyspace the default by entering:

   USE KEYSPACE <IIT id>;

Now create a file in your working directory called ex2.cql. 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. Then execute the shell command ‘DESCRIBE TABLE Music’ and include the output as the result of this exercise.



```
cassandra@cqlsh:a20317851> describe table Music

CREATE TABLE a20317851.music (
    artistname text,
    albumname text,
    cost int,
    numbersold int,
    PRIMARY KEY (artistname, albumname)
) WITH CLUSTERING ORDER BY (albumname ASC)
    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.LZ4Co
mpressor'}
    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) (2 points)

Now create a file in your working directory called ex3.cql. 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.

```
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);
```

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

Exercise 4) (2 points)

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

```
cassandra@cqlsh:a20317851> source './ex4.cql'

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

(2 rows)
```

```
SELECT * FROM Music WHERE artistName='Katy Perry';
```

Exercise 5) (2 points)

Now create a file in your working directory called ex5.cql. 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 result of executing this file as the result of this exercise.

```
SELECT * FROM Music WHERE numberSold>=700000 ALLOW
FILTERING;
```

```
cassandra@cqlsh:a20317851> source './ex5.cql'

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

(3 rows)
```