CS595—Big Data Technologies

Assignment #12

Worth: 12 points

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.

The given paper discusses the first query-driven big data modeling methodology for Apache Cassandra, defines important data modeling principles, mapping rules, and mapping patterns to guide logical data modeling, presents visual diagrams for Cassandra logical and physical data models, and demonstrates a data modeling tool that automates the entire data modeling process. Many web-scale companies have adopted Cassandra for online transaction processing, it ensures linear scalability, seamless multiple data center deployment and zero downtime.

This work deals with:

1. Query-driven methodology for Apache Cassandra.
2. Defines important Data Modelling rules, patterns and guidelines.
3. Visual representations for Cassandra logical and physical data models.
4. Exhibits an automation tool for data modelling.

The data modelling techniques highlighted in this paper were drastically different from traditional relational data modelling. It explained the role of physical data modeling and proposed a novel visualization technique, called Chebotko Diagrams, which can be used to capture complex logical and physical data models. Chebotko diagrams presents a database schema design as a combination of individual table schemas and query-driven application workflow transitions. Some of the advantages of Chebotko Diagrams, when compared to regular CQL schema definition scripts, include improved overall readability, superior intelligibility for complex data models, and better expressivity featuring both table schemas and their supported application queries.

It is query driven. This paper defines and establishes different principles for Cassandra like data nesting, data duplication, mapping rules, mapping patterns. It also talks about transition from technology independent conceptual models to Cassandra specific logical data models.

In the end it presented a powerful data modeling tool, called KDM, which automates some of the most complex, error-prone, and time-consuming data modeling tasks, including conceptual-to-logical mapping, logical-to-physical mapping, and CQL generation. The tool works by using mapping patterns and proprietary algorithms to automate the most complex, error prone and time-consuming data modelling tasks. For expert users, KDM supports several advanced features, such as automatic schema

generation in the presence of type hierarchies, n-array relationship types, explicit roles, and alternative keys.

Exercise 2) (2 points)

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.

CREATE TABLE Music (

artistName text,

albumName text,

numberSold int,

cost int,

PRIMARY KEY (artistName,albumName))

WITH CLUSTERING ORDER BY (albumName DESC);

```
bitnami@cass1: ~/A20429225
Password:
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.1 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
cassandra@cqlsh> touch init.cql
   ...
   ...
   ... ;
SyntaxException: line 1:0 no viable alternative at input 'touch' ([touch]...)
cassandra@cqlsh> source 'init.cql';
cassandra@cqlsh> USE KEYSPACE A20429225;
Improper USE command.
cassandra@cqlsh> USE A20429225;
cassandra@cqlsh:a20429225> source 'init.cql';
init.cql:2:AlreadyExists: Keyspace 'a20429225' already exists
cassandra@cqlsh:a20429225> USE A20429225;
cassandra@cqlsh:a20429225> source 'ex2.cql';
Could not open 'ex2.cql': [Errno 2] No such file or directory: 'ex2.cql'
cassandra@cqlsh:a20429225> source 'ex2.cql';
Could not open 'ex2.cql': [Errno 2] No such file or directory: 'ex2.cql'
cassandra@cqlsh:a20429225> ls
            ...
            ... ;
SyntaxException: line 1:0 no viable alternative at input 'ls' ([ls]...)
cassandra@cqlsh:a20429225> source 'ex2.cql';
cassandra@cqlsh:a20429225> DESCRIBE TABLE Music

CREATE TABLE a20429225.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';

cassandra@cqlsh:a20429225>
```

Exercise 3) (2 points)

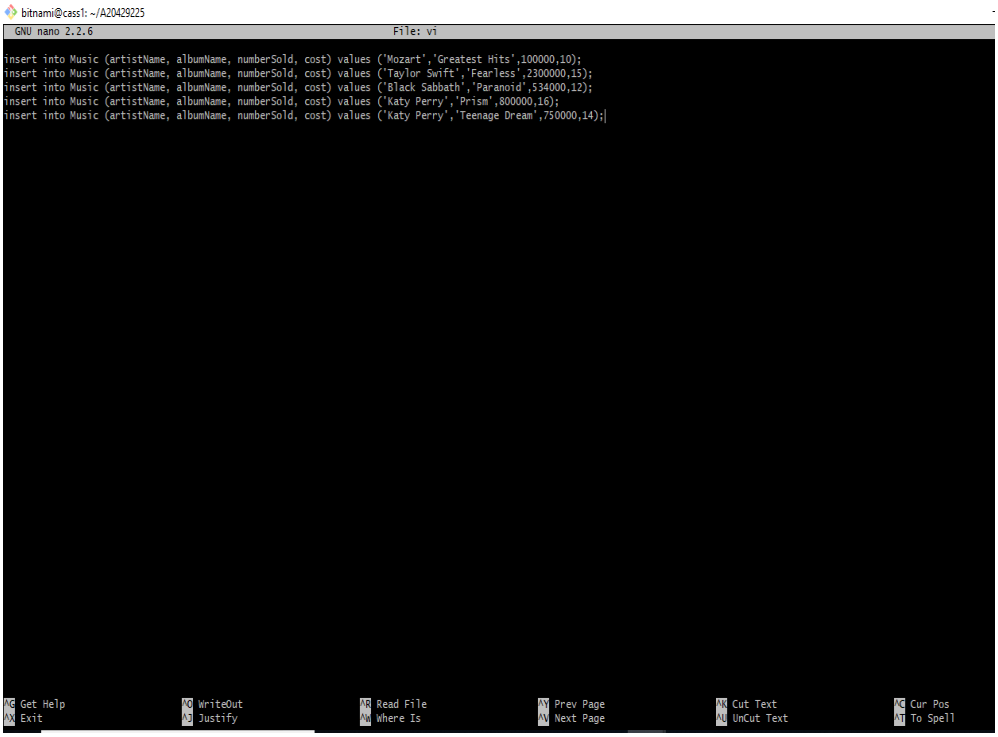
Now create a file in your working directory called ex3.cql. In this file write the commands to insert the following records into table 'Music'…

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', 'GreatestHits', 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);

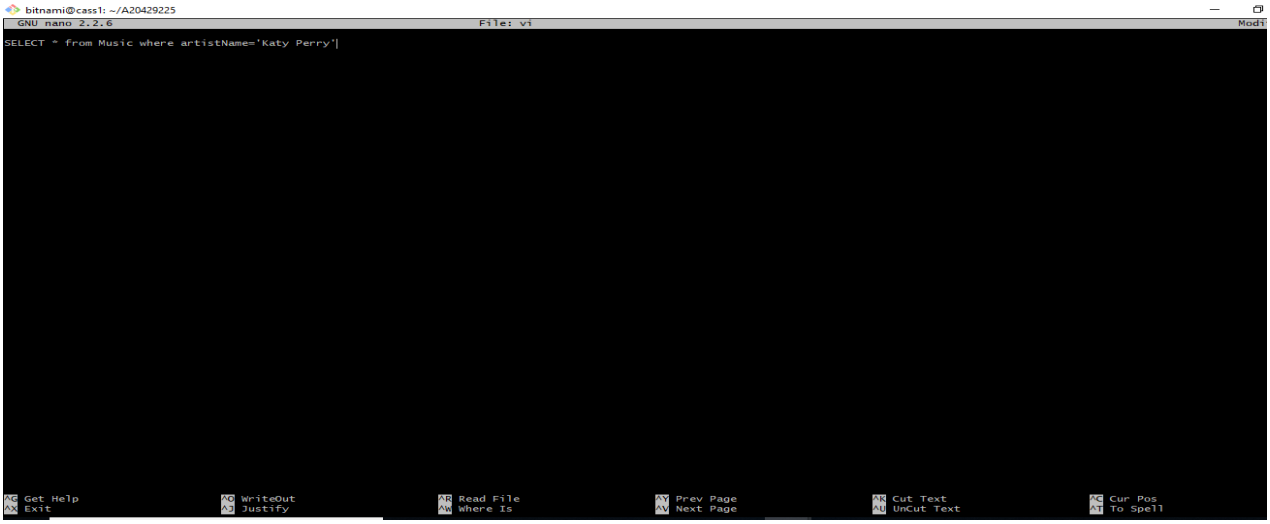


b) Execute the command 'SELECT * FROM Music;' and provide the output of this command as another result of the exercise.

```
cassandra@cqlsh:a20429225> source 'ex3.cql';
cassandra@cqlsh:a20429225> SELECT * FROM Music;

 artistname | albumname | cost | numbersold
------------+-----------+------+------------

(0 rows)
cassandra@cqlsh:a20429225> 'SELECT * FROM Music;'
            ...
            ... ;
SyntaxException: line 1:0 no viable alternative at input 'SELECT * FROM Music;' (['SELECT * FROM Music]...)
cassandra@cqlsh:a20429225> source 'ex3.cql';
cassandra@cqlsh:a20429225> 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)
cassandra@cqlsh:a20429225>
```

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.




```
(5 rows)
cassandra@cqlsh:a20429225> source 'ex4.cql';
ex4.cql:2:Incomplete statement at end of file
cassandra@cqlsh:a20429225> source 'ex4.cql';

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

(2 rows)
cassandra@cqlsh:a20429225>
```

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;

SELECT albumName from Music where numberSold >=700000 ALLOW FILTERING;

```
(2 rows)
cassandra@cqlsh:a20429225> 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)

 albumname
---------------
      Fearless
 Teenage Dream
         Prism

(3 rows)
cassandra@cqlsh:a20429225>
```