Big Data

Homework 12

Ashmita Gupta (A20512498)

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.

Ans:

The paper explores various aspects of data modeling, including traditional and Cassandra data modeling, conceptual and logical data modeling, application workflow, query-driven mapping, and physical data modeling. In the context of Cassandra Data Model, it explains that a CQL table is a grouping of divisions with rows having similar structures.

Cassandra Data Model:

A CQL table can be considered a grouping of divisions containing rows with similar structures. A partition key is distinct from each partition in a table, whereas a clustering key is distinct from each row within a partition. A primary key is a combination of a partition key and a clustering key that uniquely identifies a database row. A table schema is a collection of columns that contains a primary key. Each column's data type is either primitive (int, text, etc.), complex (set, list, or map), or counter. CQL, which has a syntax similar to SQL, is used to express queries over tables. CQL does not support binary operations like joins and instead relies on a set of query predicates rules to ensure efficiency and scalability.

Conceptual data modeling and application workflow:

Understanding the data to be maintained and how a data-driven application needs to access it is required when designing a Cassandra database schema. The ER diagram depicts the former Application workflow diagrams, which define data access patterns for application tasks and capture the latter.

Query driven mapping Data Modeling Principles:

The four data modeling principles listed below serve as a foundation for translating conceptual data models into logical data models.

DMP1 (Know your data): The first step in successful database design is to understand the data, which is recorded using a conceptual data model.

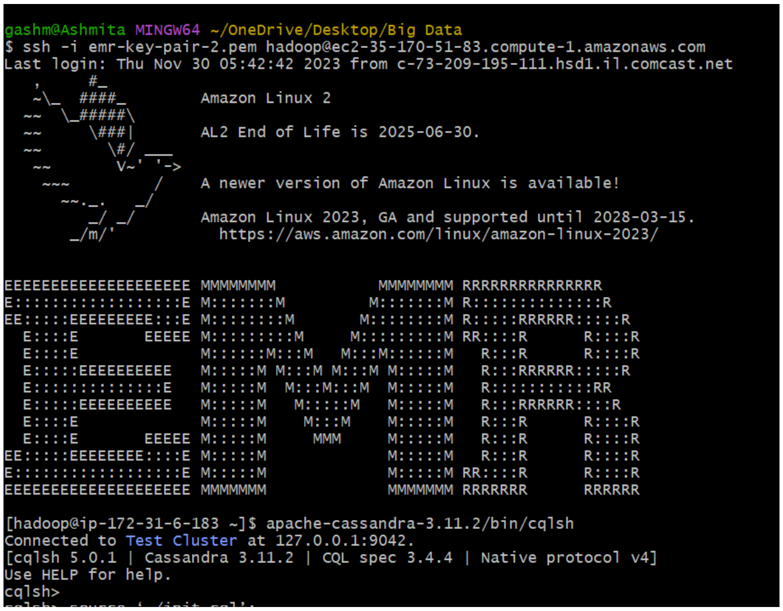
DMP2 (Know your Questions): Knowing your queries captured by an application process is the second key to a successful database design.

DMP3 (Data Nesting): Data nesting is the third key to a successful database design.

DMP4 (Data Duplication): Data duplication is the fourth key to a successful database design.

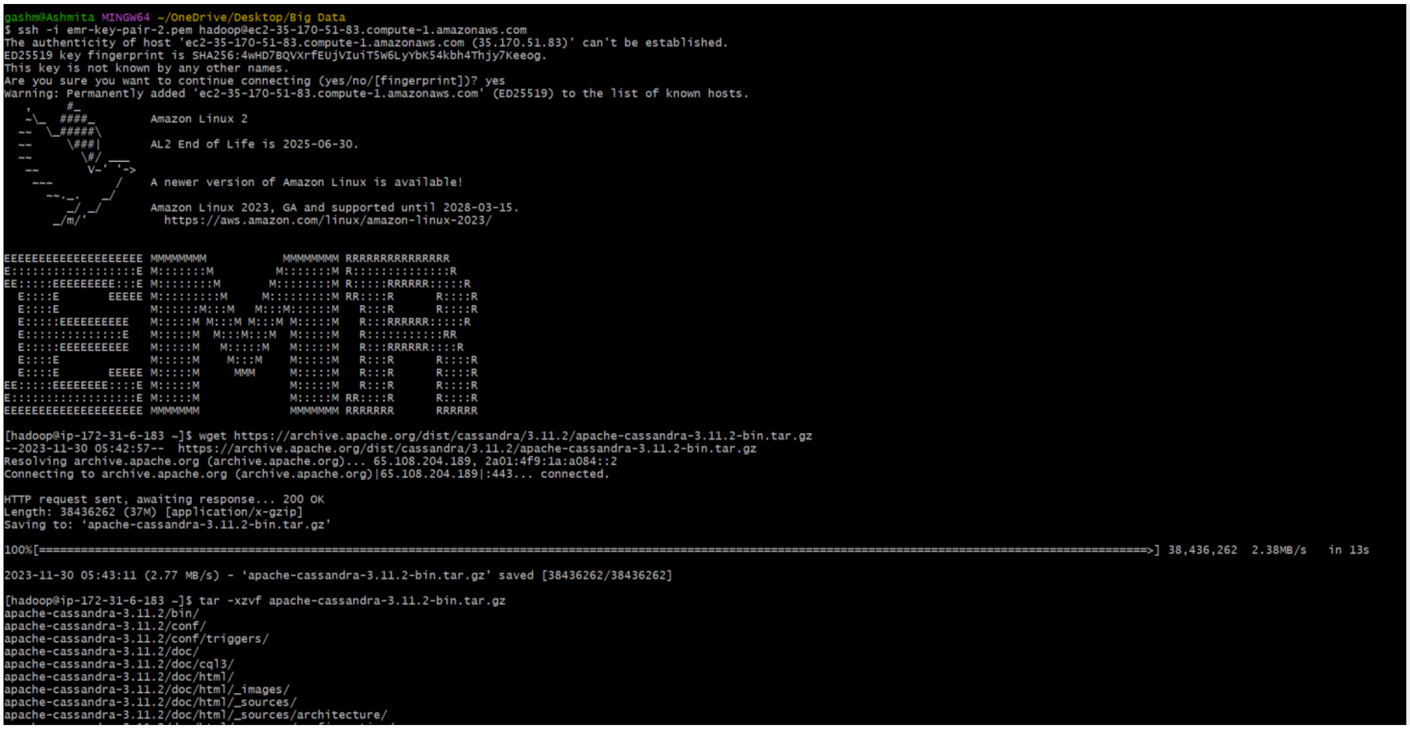
Exercise 2 (3 points):

1) Starting EMR cluster



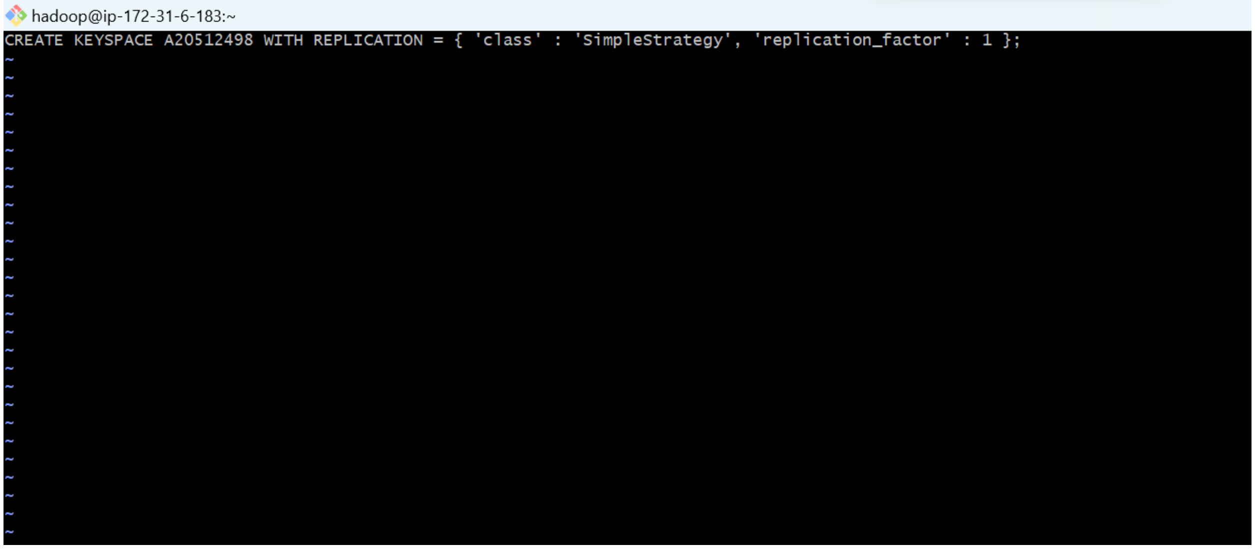
```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ ssh -i emr-key-pair-2.pem hadoop@ec2-35-170-51-83.compute-1.amazonaws.com
Last login: Thu Nov 30 05:42:42 2023 from c-73-209-195-111.hsd1.il.comcast.net
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-6-183 ~]$ apache-cassandra-3.11.2/bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.2 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
cqlsh>
```

2) Installing Cassandra
Using command:
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
apache-cassandra-3.11.2/bin/cqlsh



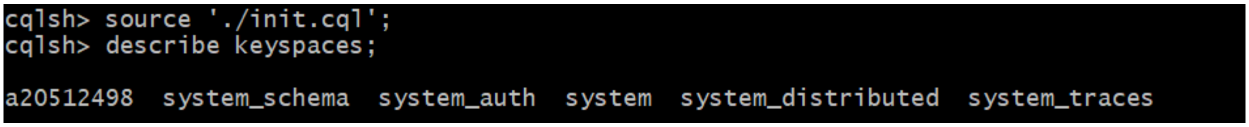
```
gashm@Ashmita MINGW64 ~/OneDrive/Desktop/Big Data
$ ssh -i emr-key-pair-2.pem hadoop@ec2-35-170-51-83.compute-1.amazonaws.com
The authenticity of host 'ec2-35-170-51-83.compute-1.amazonaws.com (35.170.51.83)' can't be established.
ED25519 key fingerprint is SHA256:4wHD7BQVXrfEUjVIuiT5W6LyYbK54kbh4Thjy7Keeog.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-35-170-51-83.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-6-183 ~]$ wget https://archive.apache.org/dist/cassandra/3.11.2/apache-cassandra-3.11.2-bin.tar.gz
--2023-11-30 05:42:57--  https://archive.apache.org/dist/cassandra/3.11.2/apache-cassandra-3.11.2-bin.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.

HTTP request sent, awaiting response... 200 OK
Length: 38436262 (37M) [application/x-gzip]
Saving to: ‘apache-cassandra-3.11.2-bin.tar.gz’

100%[======================================================================================================================================================>] 38,436,262  2.38MB/s   in 13s

2023-11-30 05:43:11 (2.77 MB/s) - ‘apache-cassandra-3.11.2-bin.tar.gz’ saved [38436262/38436262]

[hadoop@ip-172-31-6-183 ~]$ tar -xzvf apache-cassandra-3.11.2-bin.tar.gz
apache-cassandra-3.11.2/bin/
apache-cassandra-3.11.2/conf/
apache-cassandra-3.11.2/conf/triggers/
apache-cassandra-3.11.2/doc/
apache-cassandra-3.11.2/doc/cql3/
apache-cassandra-3.11.2/doc/html/
apache-cassandra-3.11.2/doc/html/_images/
apache-cassandra-3.11.2/doc/html/_sources/
apache-cassandra-3.11.2/doc/html/_sources/architecture/
```

3) Creating keyspace



hadoop@ip-172-31-6-183:~

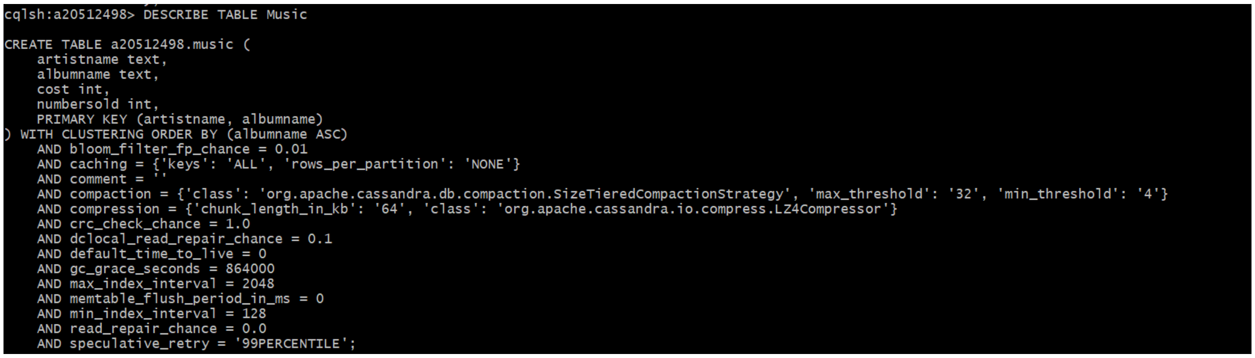
```
CREATE KEYSPACE A20512498 WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
```



```
cqlsh> source './init.cql';
cqlsh> describe keyspaces;

a20512498  system_schema  system_auth  system  system_distributed  system_traces
```

4) Executing the below command:

source './ex2.cql'; DESCRIBE TABLE Music;

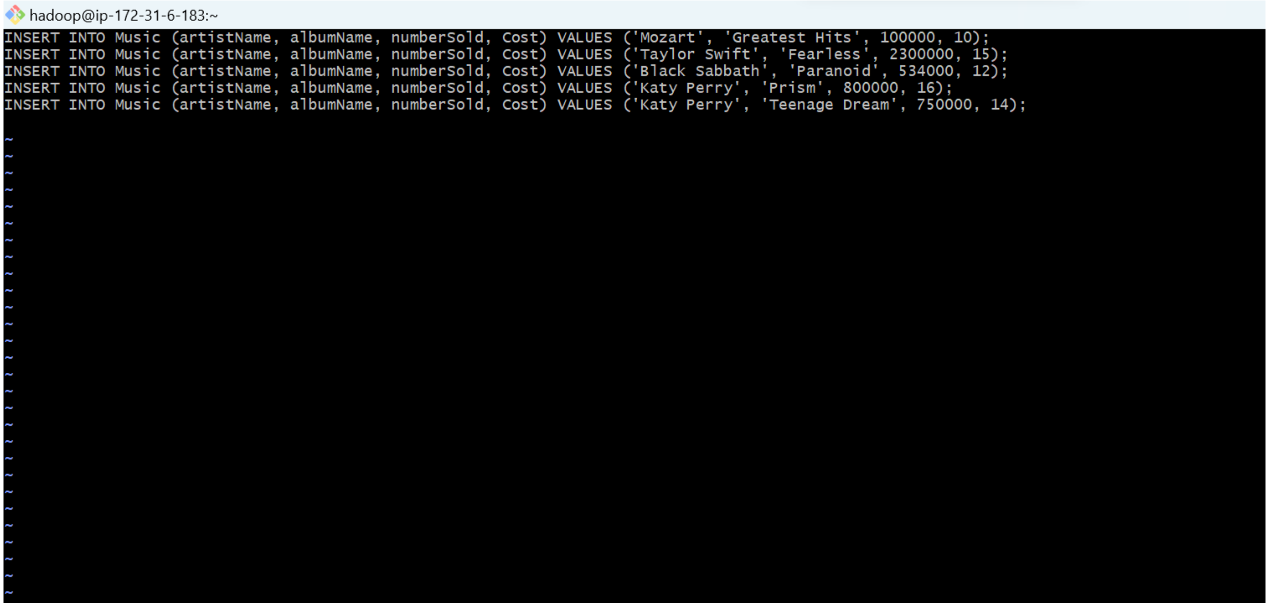


```
cqlsh:a20512498> DESCRIBE TABLE Music

CREATE TABLE a20512498.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.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';
```

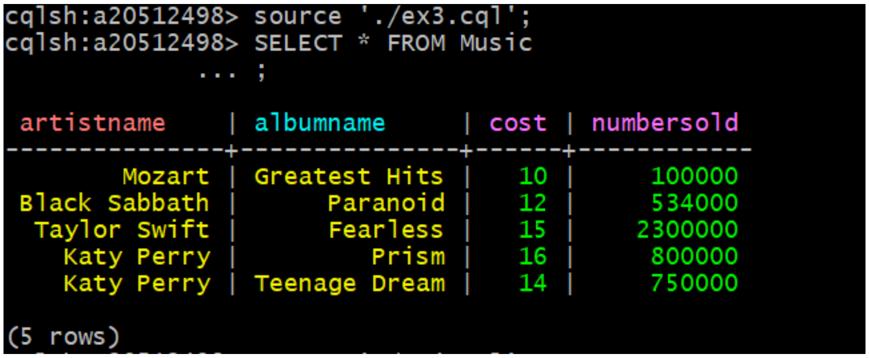
Exercise 3) (3 points)

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

vi 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);

~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
```

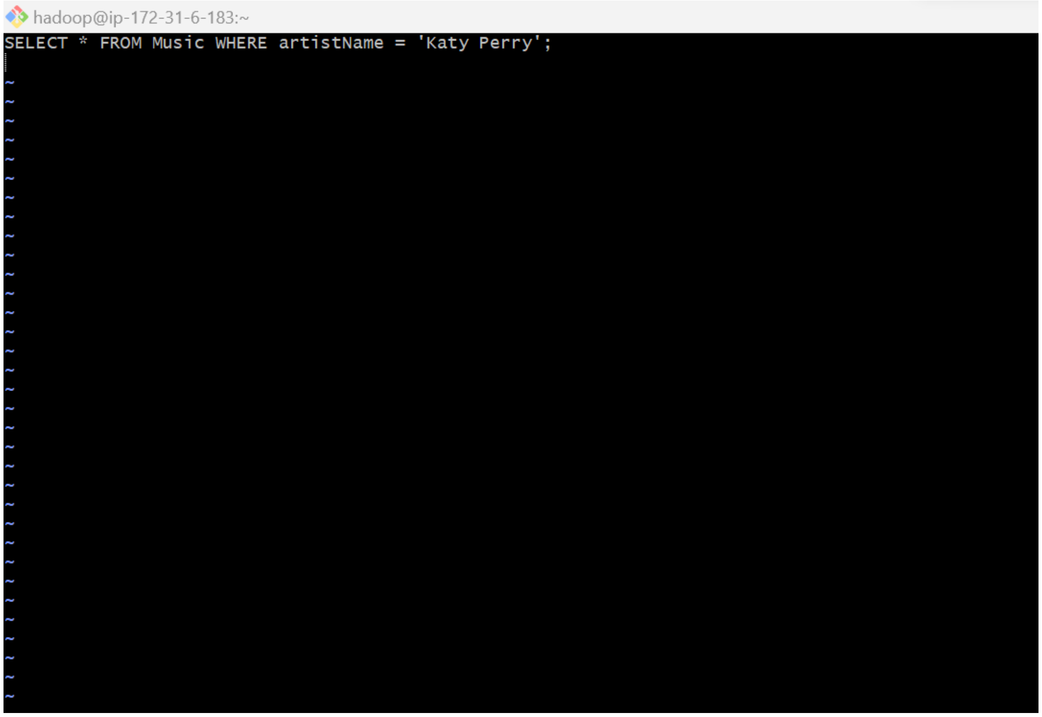


```
cqlsh:a20512498> source './ex3.cql';
cqlsh:a20512498> SELECT * FROM Music
             ... ;

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

(5 rows)
```

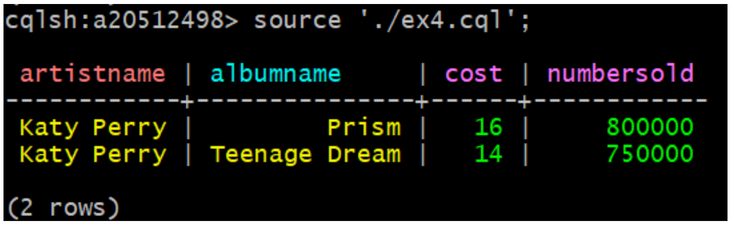
Exercise 4) (2 points)

Execute ex4.cql. Provide the content of this file as the result of this exercise

vi ex4.cql



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

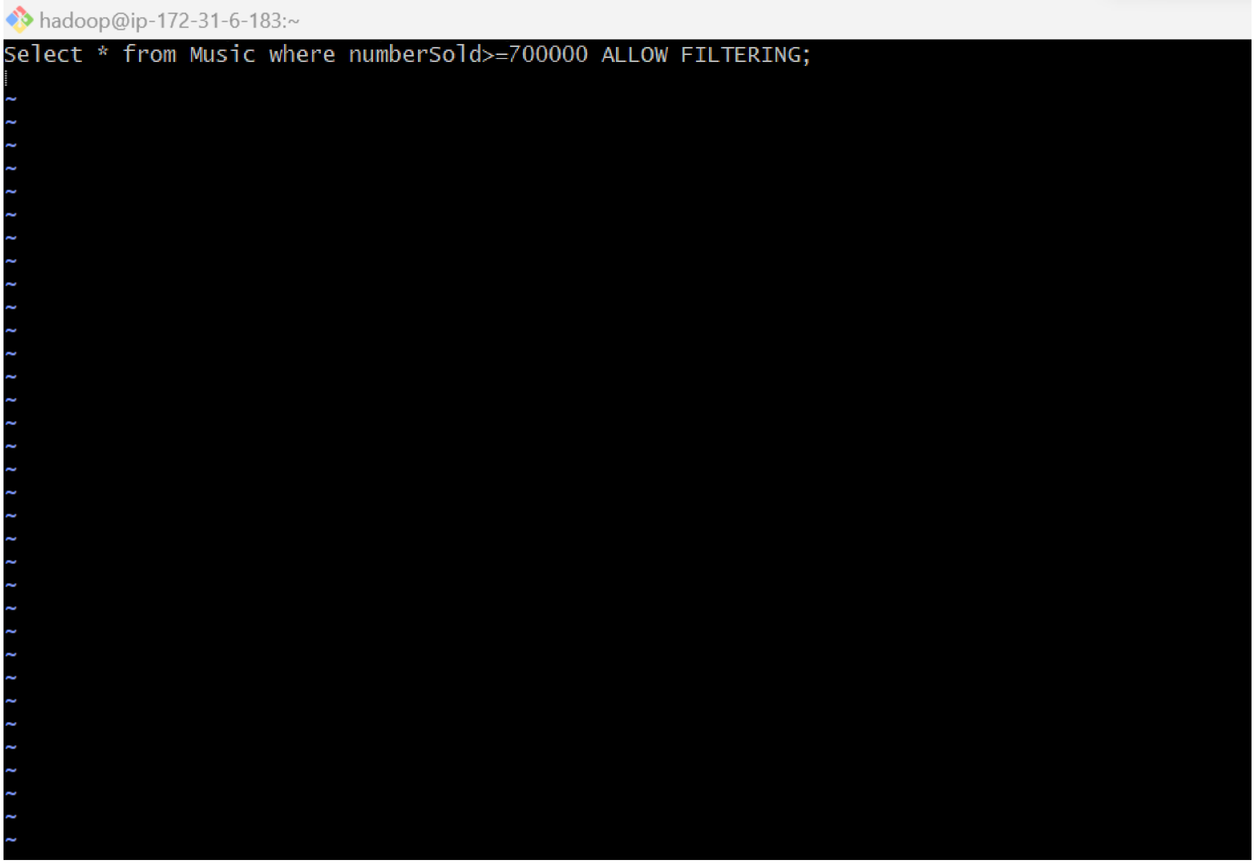


```
cqlsh:a20512498> source './ex4.cql';

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

(2 rows)
```

Exercise 5) (2 points)

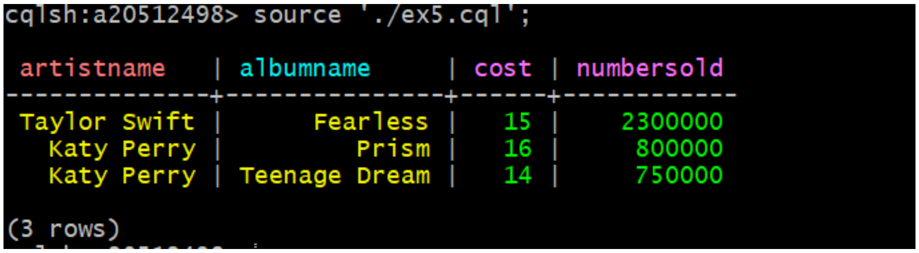
Execute ex5.cql. Provide the content of this file as the result of this exercise

vi ex5.cql



hadoop@ip-172-31-6-183:~

```
Select * from Music where numberSold>=700000 ALLOW FILTERING;

~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
```



```
cqlsh:a20512498> 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)
```