Name – Rishabh Jain || CSP554 || A20495530

Assignment 12

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.

Solution)

Summary
The paper covers traditional data modeling, Cassandra data modeling, conceptual and logical data modeling and application workflow, query driven mapping from a conceptual to a logical data model, and physical data modeling.

Cassandra Data Model:
A CQL table can be conceived of as a collection of divisions that include rows with similar structures. A partition key is unique to each partition in a table, and a clustering key is unique to each row within a partition. A primary key is a combination of a partition key and a clustering key that allows a database row to be uniquely identified. A table schema is a set of columns that includes a primary key. The data type for each column is either primitive (int, text, etc.), complex (set, list, or map), or counter.

CQL, which has a SQL-like syntax, is used to express queries over tables. CQL does not support binary operations such as joins and has a set of query predicates rules that ensure efficiency and scalability.

Conceptual data modelling 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, 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): Understanding the data, which is recorded using a conceptual data model, is the first step in successful database design.
DMP2 (Know your Questions): The second key to a successful database design is knowing your queries, which are captured by an application process.
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.

Mapping Rule: - Five mapping rules that facilitate a query-driven move from a conceptual data model to a logical data model are listed below.
MR1 -> Entity and relationship types map to tables, while entities and relationships map to table rows in MR1 (Entities and Relationships).
MR2 -> (Equality Search Attributes): Equality search attributes map to the prefix columns of a table primary key in a query predicate.
MR3 -> (Inequality Search Attributes): A table clustering key column maps to an inequality search attribute utilized in a query predicate.

Name – Rishabh Jain || CSP554 || A20495530

MR4 -> (Ordering Attributes): Ordering attributes, which are supplied in a query, map to clustering key columns in the query's chosen ascending or descending clustering order.

MR5 -> (Key Attributes): Primary key columns are mapped to key attribute types.

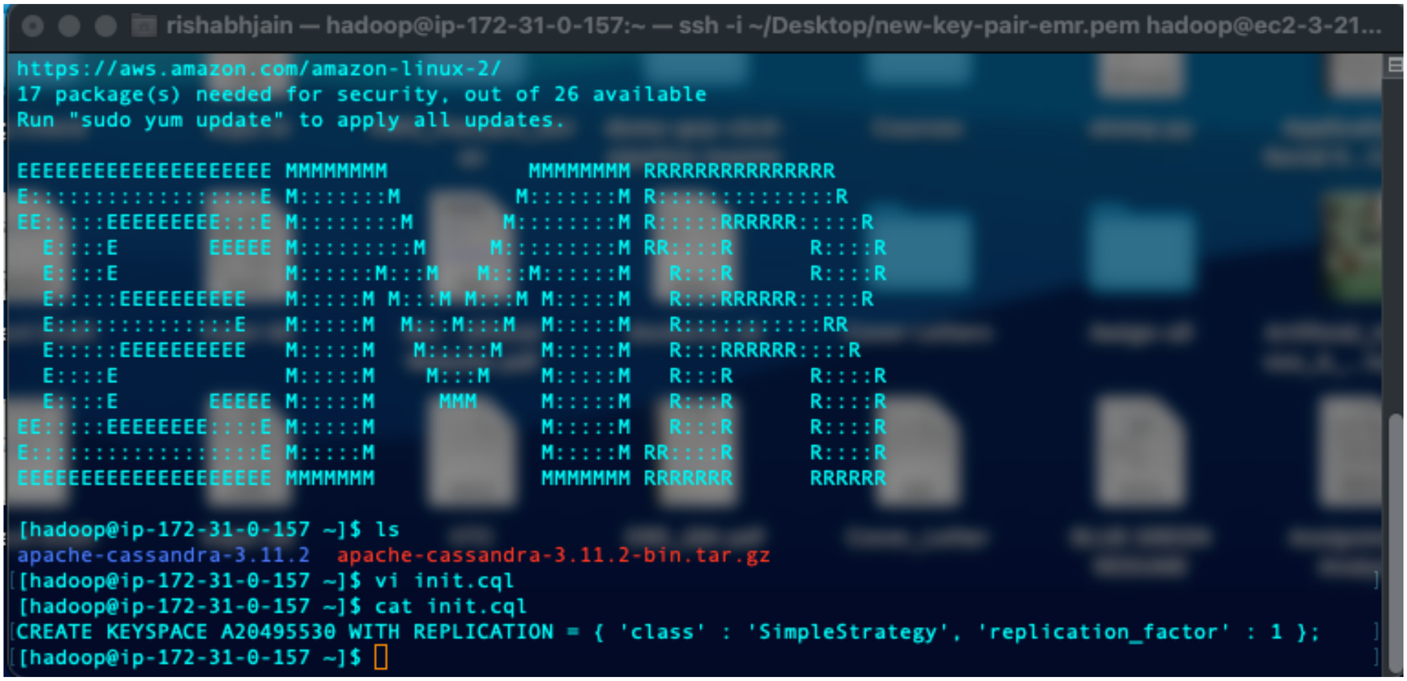
Mapping Patterns: Mapping Patterns serve as the basis for automating Cassandra database schema design.

Physical Data Modeling

The final step is the analysis and optimization of a logical data model to produce a physical data model.

Exercise 2) (3 points)

a)

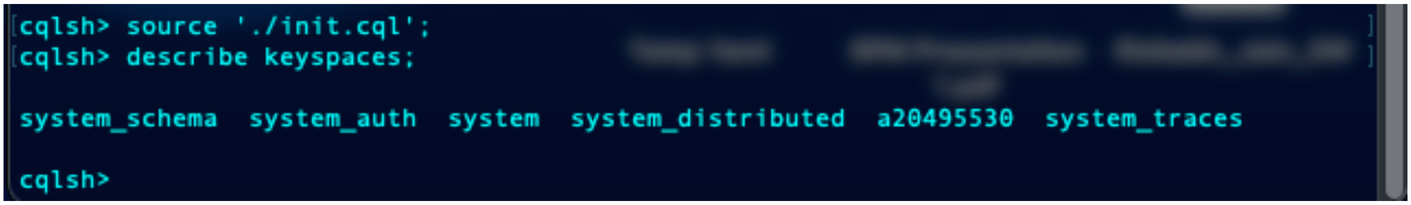

rishabhjain — hadoop@ip-172-31-0-157:~ — ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-3-21...

```
https://aws.amazon.com/amazon-linux-2/
17 package(s) needed for security, out of 26 available
Run "sudo yum update" to apply all updates.

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-172-31-0-157 ~]$ ls
apache-cassandra-3.11.2  apache-cassandra-3.11.2-bin.tar.gz
[hadoop@ip-172-31-0-157 ~]$ vi init.cql
[hadoop@ip-172-31-0-157 ~]$ cat init.cql
CREATE KEYSPACE A20495530 WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };
[hadoop@ip-172-31-0-157 ~]$ 
```

b)

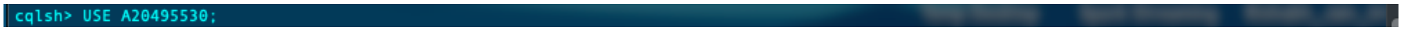
source ‘./init.cql’;

c)



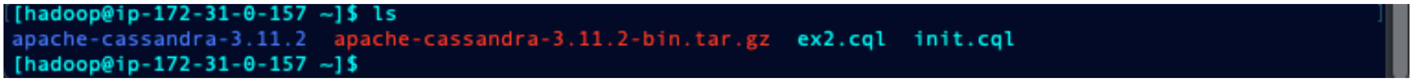
```
cqlsh> source './init.cql';
cqlsh> describe keyspaces;

system_schema  system_auth  system  system_distributed  a20495530  system_traces

cqlsh>
```

d)

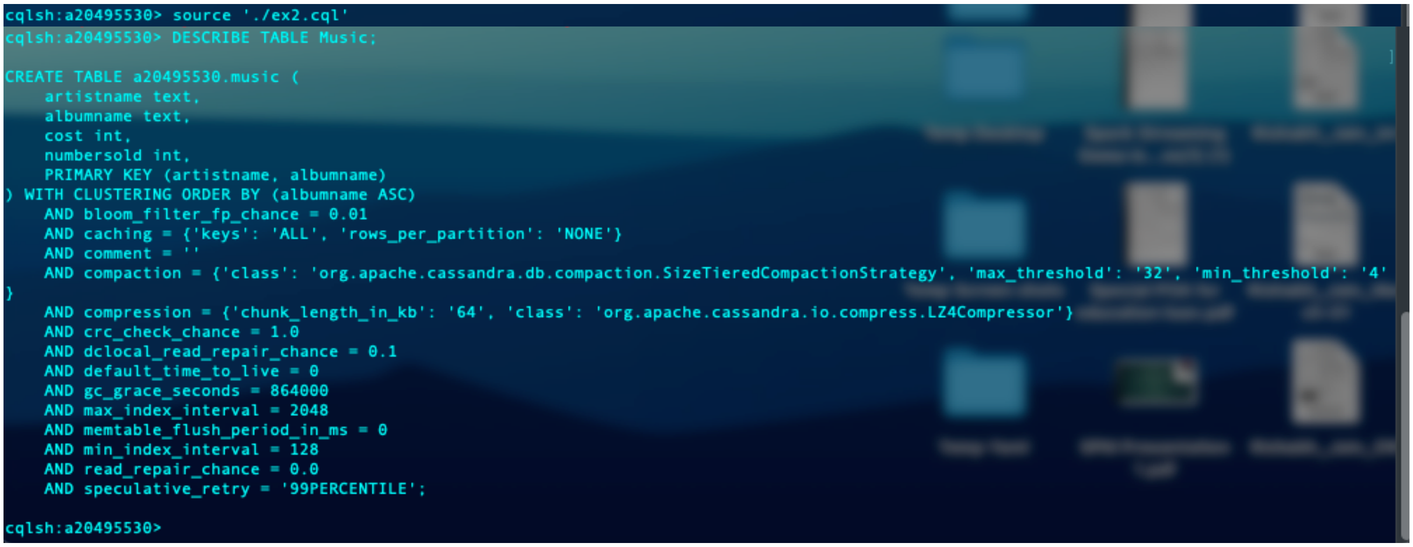


```
cqlsh> USE A20495530;
```

Name – Rishabh Jain || CSP554 || A20495530



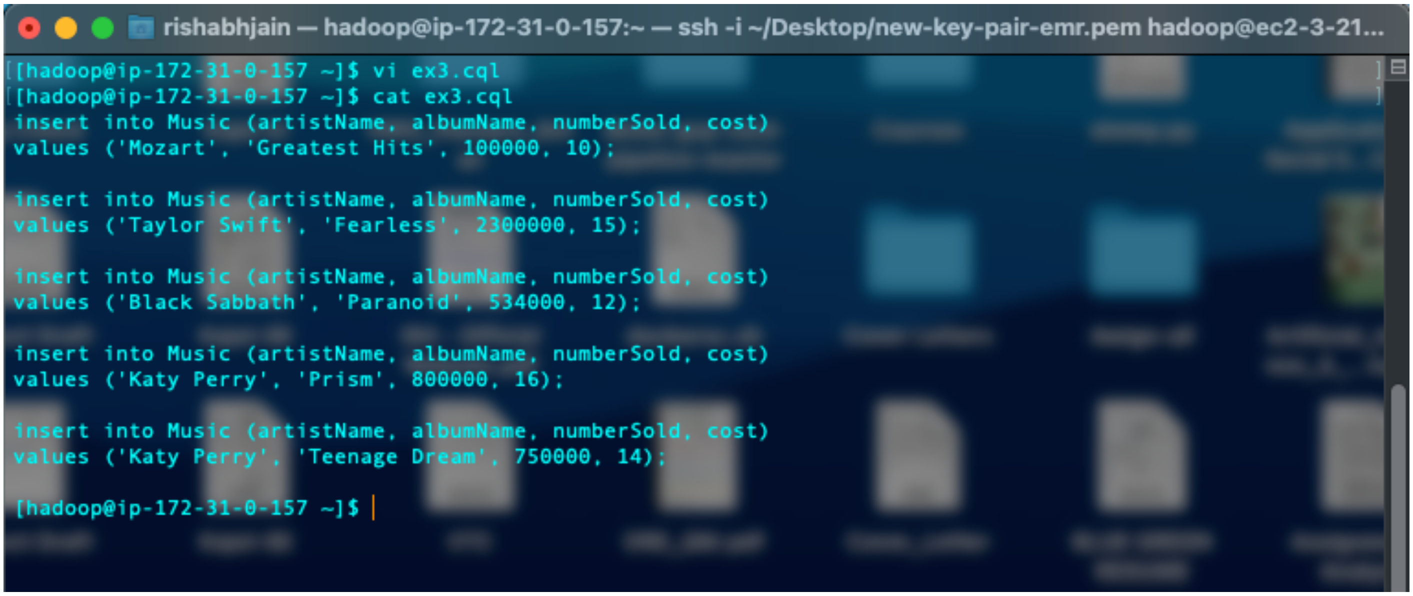
```
[hadoop@ip-172-31-0-157 ~]$ ls
apache-cassandra-3.11.2  apache-cassandra-3.11.2-bin.tar.gz  ex2.cql  init.cql
[hadoop@ip-172-31-0-157 ~]$
```



```
cqlsh:a20495530> source './ex2.cql'
cqlsh:a20495530> DESCRIBE TABLE Music;

CREATE TABLE a20495530.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';

cqlsh:a20495530>
```

Exercise 3) (3 points)

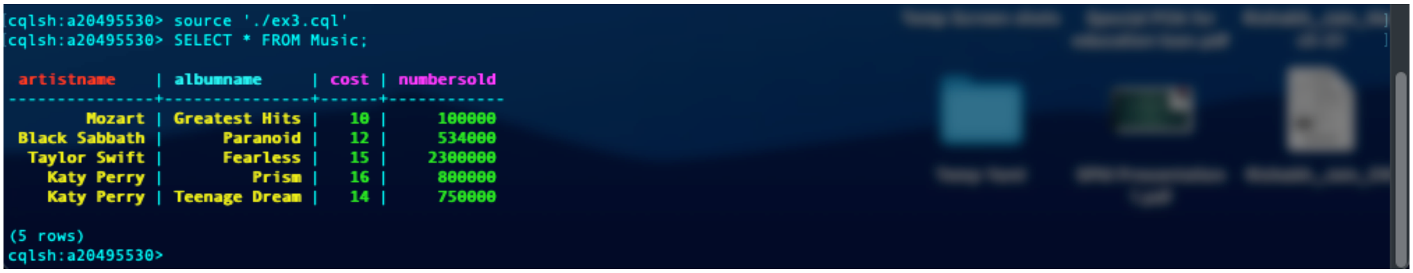
a)



rishabhjain — hadoop@ip-172-31-0-157:~ — ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-3-21...

```
[hadoop@ip-172-31-0-157 ~]$ vi ex3.cql
[hadoop@ip-172-31-0-157 ~]$ 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-172-31-0-157 ~]$ |
```

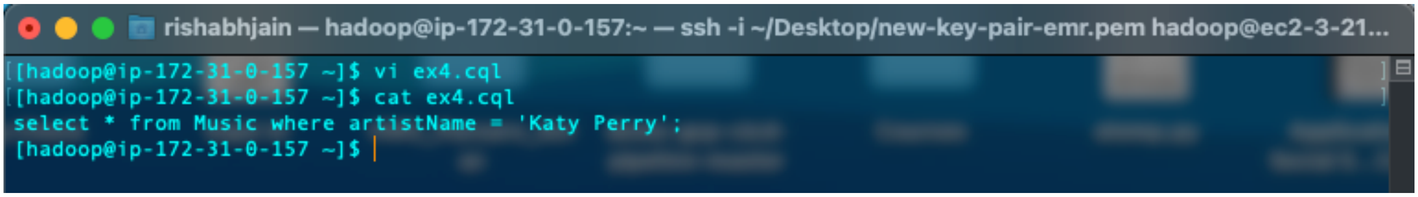
b)



```
cqlsh:a20495530> source './ex3.cql'
cqlsh:a20495530> 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)
cqlsh:a20495530>
```

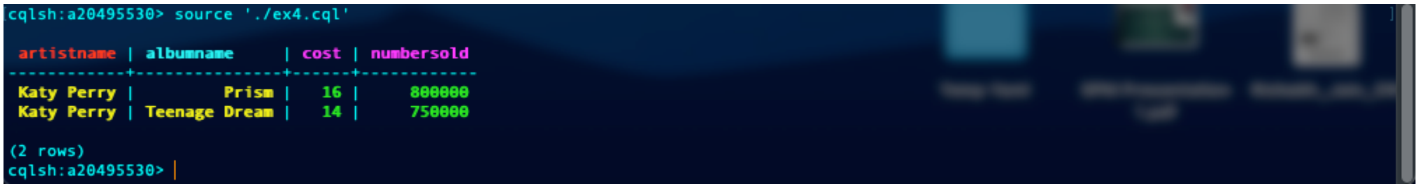
Name – Rishabh Jain || CSP554 || A20495530

Exercise 4) (2 points)



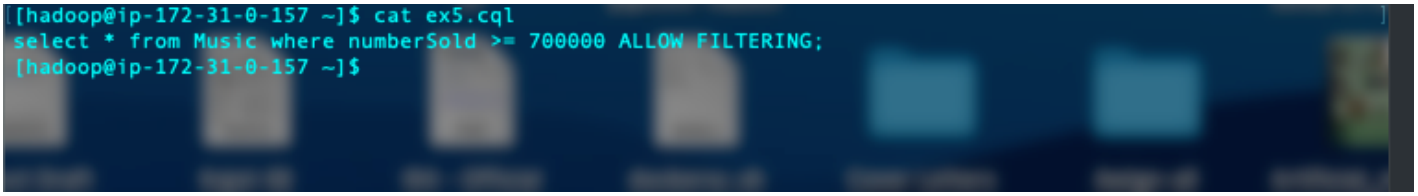
rishabhjain — hadoop@ip-172-31-0-157:~ — ssh -i ~/Desktop/new-key-pair-emr.pem hadoop@ec2-3-21...

```
[hadoop@ip-172-31-0-157 ~]$ vi ex4.cql
[hadoop@ip-172-31-0-157 ~]$ cat ex4.cql
select * from Music where artistName = 'Katy Perry';
[hadoop@ip-172-31-0-157 ~]$
```

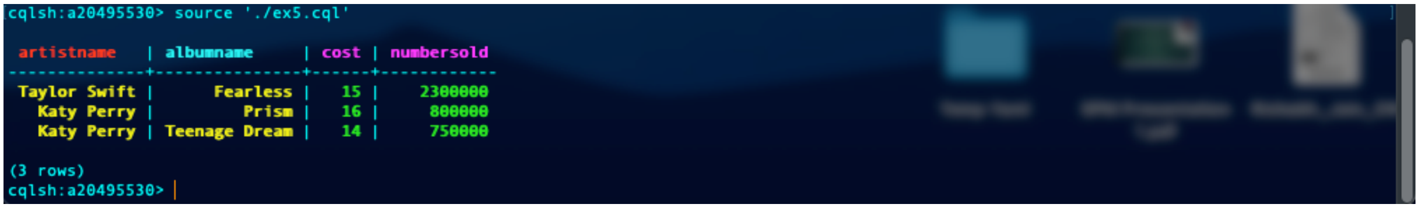


```
cqlsh:a20495530> source './ex4.cql'

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

(2 rows)
cqlsh:a20495530>
```

Exercise 5) (2 points)



```
[hadoop@ip-172-31-0-157 ~]$ cat ex5.cql
select * from Music where numberSold >= 700000 ALLOW FILTERING;
[hadoop@ip-172-31-0-157 ~]$
```



```
cqlsh:a20495530> 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)
cqlsh:a20495530>
```