CSP595 - Assignment 12

Name: Jagruti Vichare
Email ID: jvichare@hawk.iit.edu
CWID: A20378092

Exercise 1

Apache Cassandra is a leading transactional, scalable, and highly-available distributed database. It is known to manage some of the world's largest datasets on clusters with many thousands of nodes deployed across multiple data centres. 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 [3] in big data applications is attributed to, among other things, its scalable and fault-tolerant peer-to-peer architecture [4], versatile and flexible data model that evolved from the BigTable data model [5], 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.

This paper introduces a rigorous query-driven data modelling methodology for Apache Cassandra. The methodology was shown to be drastically different from the traditional relational data modelling approach in a number of ways, such as query-driven schema design, data nesting and data duplication. This paper elaborates the fundamental data modelling principles for Cassandra and defines mapping rules and mapping patterns to transition from technology-independent conceptual data models to Cassandra-specific logical data models. It also explains the role of physical data modelling and proposed a novel visualization technique, called Chebotko Diagrams, which can be used to capture complex logical and physical data models. Finally, it presents a powerful data modelling tool, called KDM, which automates some of the most complex, error-prone, and time-consuming data modelling tasks, including conceptual-to-logical mapping, logical-to-physical mapping, and CQL generation.

Exercise 2

```
andra@cqlsh:a20378092> DESCRIBE TABLE Music

TE TABLE a20378092.music (
artistname text,
albumname text,
cost int,
numbersold int,
PRIMARY KEY (artistname, albumname)
TH 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_thresho
'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

Ex3.cql File Content

```
SERT INTO Music (artistName, albumName, numberSold, cost) VALUES ('Taylor Swift', 'Fearless', 2300000, 15)
SERT INTO Music (artistName, albumName, numberSold, cost) VALUES ('Black Sabbath', 'Paranoid', 534000, 12)
```

'SELECT * FROM Music;' output

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

rows)
```

Exercise 4

File Content

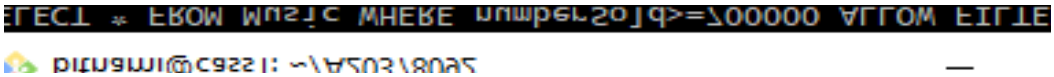
Output

```
tistname | albumname      | cost | numbersold
---------+----------------+------+------------
ty Perry |         Prism |   16 |     800000
ty Perry | Teenage Dream |   14 |     750000

rows)
```

Exercise 5

File Content



Output