# CSP554—Big Data Technologies

## Assignment #11 – Cassandra

**Exercise 1)**

**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 paper describes about the Cassandra database and how the traditional methodology differs from the new proposed data modelling methodology the associated mapping rules, patterns and then demonstrating a tool that would automate the data modelling process. Cassandra, a highly available database is used in many big data systems as it is scalable and fault tolerant system and can handle million transaction per second. The paper states that the traditional approaches focus on organization of data and understanding relations and minimizing the data redundancy. The paper says that in Cassandra, unlike traditional model’s application queries are given an important priority over others and as same data can be stored in different tables there is data redundancy also present. It adds on by stating that complex queries and data aggregation is not supported by Cassandra Query Language (CQL) unlike SQL. The authors have given an importance on application workflow and access patterns that helps to map the conceptual data model to logical data model so that they can support application queries according to the workflow. It eliminates normalization hence data duplication is introduced and hence developing a query driven approach instead of data driven approach. The author further explains about Cassandra Query language its working along with an example and how is it different from SQL. The conceptual data model is designed through two stages one is through Entity Relationship model that helps in managing data and the application workflow diagram that helps to see how a data driven application needs to access data and the author supports it by using the library database use case. The data modelling principles that help in building the model are to know the data, know the queries, data nesting and data duplication. And based on these rules, 5 mapping rules are also established that have entity relationship attributes, equality search attributes, inequality attributes, ordering attributes and key attributes. Further these mapping rules help in providing mapping patterns that do the automation of Cassandra database design. It concludes in the final step by doing query optimization of logical data model that provides a physical data model. The Chebotko Diagrams help in visualization for complex logical and physical data model. KDM tool helped in CQL generation, automating data models and the mappings.

**Exercise 2)**

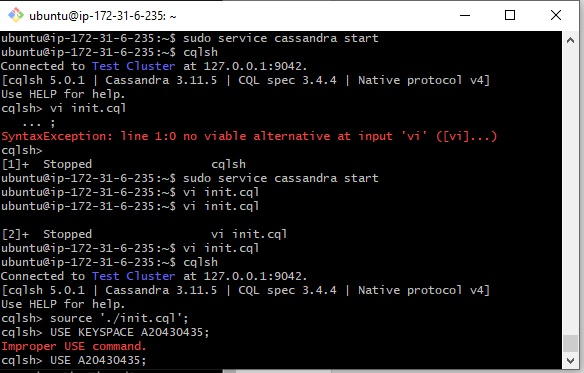
**Follow the instructions “Installing and Starting Cassandra” included on the backboard with this assignment.**

1. **You might want to open two terminal windows and connect them (via ssh) to the Cassandra VM you set up in step “a”. This will allow you to edit files in one window and start cqlsh in the other to execute them.**
2. **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 };**

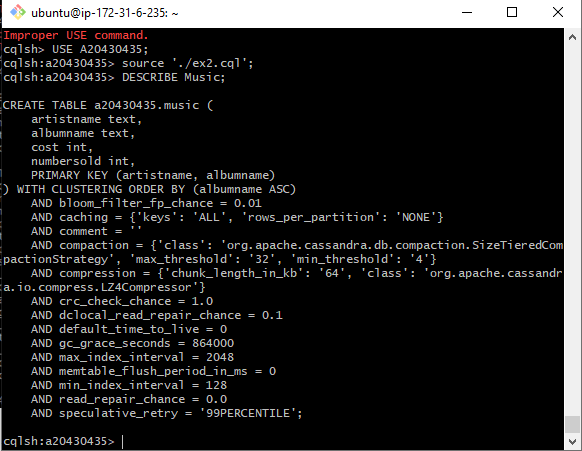
1. **Then execute this file in the CQL shell as follows…**

**source ‘./init.cql’;**



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

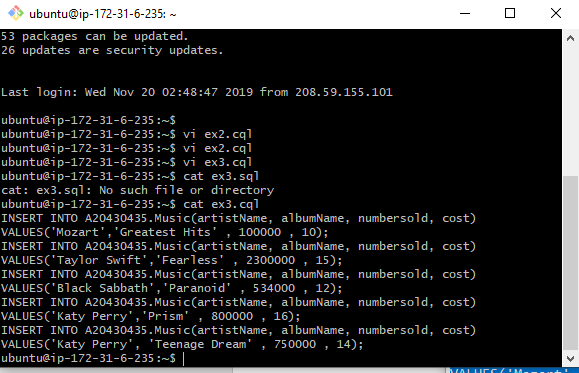
**USE KEYSPACE <IIT id>;**



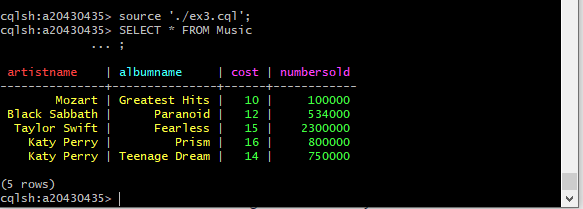
**Exercise 3)**

**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’…**

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

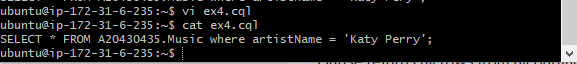


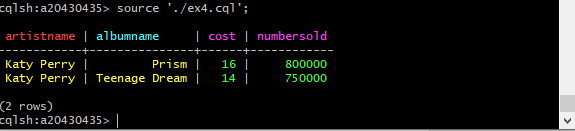
1. **Execute the command ‘SELECT \* FROM Music;’ and provide the output of this command as another result of the exercise.**



**Exercise 4)**

**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.**





**Exercise 5)**

**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**.

