CSCE 5300 Introduction to Big data and Data Science
ICE-4

Lesson Title: *Hive*

Lesson Description: *Hadoop Dependent Query Based NoSQL Database Hive*

Lesson Overview:

Hive is a data warehousing system to store structured data on Hadoop file system and provides an easy query these data by execution Hadoop MapReduce plans. In this exercise we will learn basics of Hive QL.

In Class Exercise:

1. Create Hive Tables and Perform Queries for Use Case based on Petrol or hotel_bookings data. For Petrol, see the slides for details or you may try your own queries using hotel_bookings data.

Queries should include applying below type of queries. (If not possible, provide justification why it is not possible in your environment)
   - Order by query
   - Group by query
   - Sort by
   - Cluster By
   - Distribute By

2. Create Hive Tables and Perform Queries for Use Case based on Olympics Data. See the Slides for details.

3. Create Hive Tables and Perform Queries for Use Case based on Movielens dataset which has 3 datasets as movies, users and ratings.

Perform following tasks:

A. Create 3 tables called movies, ratings and users. Load the data into tables.

B. For movies table:
   - List all movies with genre of movie is "Action" and "Drama"

C. For Ratings table:
- List movie ids of all movies with rating equal to 5.

D. Find top 11 average rated "Action" movies with descending order of rating. (Hint: Need to perform join operation on Movies and Ratings table)

ICE Submission Guidelines

1. ICE Submission is individual.
2. ICE code has to be properly commented.
3. The documentation should include the screenshots of your code/results with explanation.
4. Provide the explanation of the dataset/exercise as per your understanding.
5. The similarity score for your document should be less than 15%.
6. All you need to do is submit the source code (properly commented) and documentation (.pdf/.doc) with explanation and screenshot of source code having input logic and output results.
7. Submission after the deadline is considered as late submission.