CSCE 5300 Introduction to Big data and Data Science
ICE-3

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 bank data. For Petrol, see the slides for details (Lecture 3 Hive) or you may try your own queries using bank data. Implement following.
   - Creation of Table in Hive and Loading of 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.
   - Creation of Table in Hive and Loading of data

     create table olympic (athelete STRING,age INT,country STRING,year STRING,closing STRING,sport STRING,gold INT,silver INT,bronzeINT,total INT) row format delimited fields terminated by '\t' stored as textfile;

     load data local inpath '/home/acadgild/Downloads/olympic_data.csv' into table olympic;

- Find total number of medals won by each country in athletics.
    1. we use SUM function to define total medal and also, we need only Athletics so use where condition.
    2. we need separate country and their relevant total number of medal then we group by country.
- Find the total number of medals each country won display the name along with total medals.
- Find Total number of medals won by each country in swimming.
- Which country got medals for Shooting, year wise classification?

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:
    - Create a table for 3 for movies, user, Rating
    - Now join the two tables. (movies and rating)
    - Find which day of the week most of ratings are posted
4. Using the same data set as Q3. Perform the following Tasks:
   Movie Recommendation Based on Genres
    - Create exploded view of movie id and genre
    - Find for each user, the rank of genre. Ideally you would like to compute weighted
    - Create a temporary view for user and his total ratings by genre
    - Find top 3 genres for each user and create a temporary table for that.

ICE Submission Guidelines

1. ICE Submission is individual.
2. ICE code must be properly commented.
3. ICE should have proper formatting and headings
4. The documentation should include the screenshots of your code/results with explanation.
5. Provide the explanation of the dataset/exercise as per your understanding.
6. The similarity score for your document should be less than 15%.
7. 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.
8. Submission after the deadline is considered as late submission.