CSEE5590-0001/490-0003: Big Data Programming

Lesson Plan # 4

ICP Feedback and Submission Link:

https://forms.gle/xMAmr3zATrtMG5cX7

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

Dataset: (In your GitHub datasets folder)

https://umkc.box.com/s/hx82th050ysf557mewozonatmusmjn7b

https://umkc.box.com/s/8fzos5jl3cjsnklh3yl67c14b4jjvo2p

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

Dataset:

https://umkc.box.com/s/f918eea7k6mw6h7qiwj4b8im97c6hy84

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

   Dataset:

   https://umkc.box.com/s/m3i7oabkj00boxuiskv5d4aoklh85w3x

   Perform following tasks:

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

   2. For movies table:
      - List all movies with genre of movie is “Action” and “Drama”

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

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

      You can refer following document for reference:

      https://umkc.box.com/s/1dcugk08caqzitgqvrthiqe5n6sgznd5

ICP Submission Guidelines:

1. ICP Submission is in pairs of three/four students.
2. Once completed, must be presented to TA or Instructor before the completion of the class
3. Submission after the deadline is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation should be pushed to GitHub.
5. Submit your screenshots as well to GitHub and documentation. The screenshot should have both the code and the output.
6. Submit a demo video 2-3 min showing your assignment with a voice over explaining your work if you are unable to complete ICP within the deadline due to genuine reason.
7. Provide the video submission link through the GitHub and submission form https://forms.gle/xMAmr3zATrtMG5cX7

Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: *https://catalog.umkc.edu/special-notices/academic-honesty/*