Posted on 10/11/2020, Due on 10/25/2020

corresponding TA: Mehdi Ataei ataei@mie.utoronto.ca

Login to Queens cluster > click on Hadoop > Ambari > Hive 2.0. It is advisable to create your own database and upload the necessary tables.

General Theoretical Questions

Q1. (15 marks)

1) What are two main functions and the components of HDFS?

2) What are main functions and the components of YARN?

Give short explanations.

Q2. (8 marks) Data is replicated at least thrice on HDFS. Does it imply that any alterations or calculations done on one copy of the data will be reflected in the other two copies also?

Q3. (12 marks) Write the commands for the following HDFS tasks:

1) **Which command is used to copy a file from HDFS to local file system?**
2) **Which command is used to move files within HDFS?**
3) **Which command is used to print the contents of the file?**
4) **Which command is used to move file from local to HDFS?**
5) **Create a new directory in your personal directory and title it Lab1_2020_Fall. What command did you use? Prints-screen your folder and its contents to show that the new folder has been created**
6) **Move a file /tmp/flights.csv to the newly created folder Lab1_2020_Fall. What command did you use and please provide a print screen of your personal folder/directory to prove that flights.csv has been moved.**

Hive @Ambari:

Let us use Hive View 2.0 offered in Ambari:

Note. In order to work through this set of questions, please ensure you download locally the NHL Game Datasets: https://www.kaggle.com/martinellis/nhl-game-data

Specifically, the following datasets should be loaded using Ambari:

- game.csv
- team_info.csv
- player_info.csv
- game_team_stats.csv

- game_skaters_states.csv

For each step, please submit a copy of your HQL query and a screenshot of resulting table. If you have issues importing tables in Ambari, you may also use **nhl_mataei database**, which has all the tables included.

Practice CASE Statement: (15 marks)

Q4. Display the table with season, home team name, away team name, total_goals (home_goals + away_goals), and using CASE statement, categorize the total_goals by following:

- **total_goals >= 7, 'High scoring game'**
- **total_goals < 7 and total_goals > 4 , 'Mid scoring game'**
- **elsewhere, 'Low scoring game'**

only where the home team wins ('home win OT')

Practice OVER and GOUPBY Function: (20 marks)

Q5. Display the table showing player_id, sum, max, and average number of goals and assists of each player using the GROUP BY function ordered by the sum of the goals + sum of assists (descending). The table consists of seven columns.

Try getting the same results using OVER() function with partition window. Are the results same?

Practice JOINS: (15 marks)

Q6. Create a table that lists last name, first name, and nationality of all the players in the team with team_id=2 who play defense (primariPosition is D)

Practice Subquery: (15 marks)

Q7. Calculate the sum, maximum, average of all the shots in each game that the home team scored more than 3 goals.

Q8. Show the average hits per player for all the home teams with odd team ids.