greatlearning

Project: Impala

Business Requirement:

Imagine that you are working with one of the largest gaming companies in the world. Your manager asks you to analyses the data from the game to get some more insights. The game that we are talking about is Pokemon Go. Pokémon Go is a free-to-play, location-based augmented reality game developed by Niantic for iOS and Android devices. It was released only in July 2016 and only in selected countries. You can download Pokémon for free of cost and start playing. You can also use PokéCoins to purchase Pokéballs, the in-game item you need to be able to catch Pokémon.

Data Set Description:

The dataset consists of 11 columns and their respective description is as follows:

Pokemonid_Number: This column represents id of each Pokémon.

Name: This column represents the name of the Pokémon.

Type 1: This column represents the property of a Pokémon.

Type 2: This column represents the extended property of the same Pokémon.

A Pokémon may be one or both the types. For instance, Charmander is a Fire type, while Bulbasaur is both a Grass type as well as a Poison type. With the current 18-type system, there are 324 possible ways to assign these types to Pokémon, along with 171 unique combinations. As of Generation VI, 133 different type combinations have been used.

Total: This column represents the sum of all character points of a Pokémon (HP, Attack, Defense, Sp. Atk, Sp. Def, and Speed).

HP (Hit Points): This column represents Pokémon Hit Points, which is a value that determines how much damage a Pokémon can receive. When a Pokémon's HP is down to '0', the Pokémon will faint. HP is the most frequently affected stat of them all, as a depleting HP is a key factor in winning a battle.

Attack: This column represents the Attack stat.

Defense: This column represents the Defense stat.

Sp. Atk: This column represents a Pokémon's Special Attack stat.

Sp. Def: This column represents a Pokémon's Special Defense stat.

Speed: This column represents the speed stat of a Pokémon.

Learning Outcomes:

After successfully completing the project, the participants will be able to

- Use Impala as a SQL tool for analysing Big Data
- Get understanding about writing queries using Impala
- Approach a business problem and model the solution

Grading Criteria: 50 Marks

Participants can use hive shell to explore the problem and find the solution, since the queries of Hive and Impala are the same. Connect with a hive shell and perform the following analysis

1. Create a Database and use the same for analysis. Create a Table named pokemon and Load the data to table. Verify that the data has been loaded. (10 Marks)

2. Find out the average HP (Hit points) of all the Pokémon (10 Marks)

3. Create and insert values of existing table 'pokemon' into a new table 'pokemon1', with an additional column 'power_rate' to find the count of 'powerful' and 'moderate' from the table 'pokemon1' (10 Marks)

4. Find out the number of powerful and moderate HP Pokémons present (10 Marks)

5. Find out the top 10 Pokémons according to their HP's (10 Marks)