Graded Project

Impala

Week 9

Table of Content

[Business Requirement 3](#__RefHeading___Toc395_4031633032)

[Data Set Description 3](#__RefHeading___Toc309_2316712879)

[Learning Outcomes 3](#__RefHeading___Toc311_2316712879)

[Grading Criteria 4](#__RefHeading___Toc397_4031633032)

[Load Data Into HDFS 4](#__RefHeading___Toc315_2316712879)

[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. 4](#__RefHeading___Toc399_4031633032)

[2. Find out the average HP (Hit points) of all the Pokémon 6](#__RefHeading___Toc246_2316712879)

[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’ 7](#__RefHeading___Toc401_4031633032)

[4. Find out the number of powerful and moderate HP Pokémons present 8](#__RefHeading___Toc1084_28677486501)

[5. Find out the top 10 Pokémons according to their HP’s 8](#__RefHeading___Toc317_2316712879)

# 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

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

# [**Load Data Into**](https://www.geeksforgeeks.org/hive-load-data-into-table/) **HDFS**

The first step is to create a folder and upload data into HDFS

**On the CloudX Lab web console:**

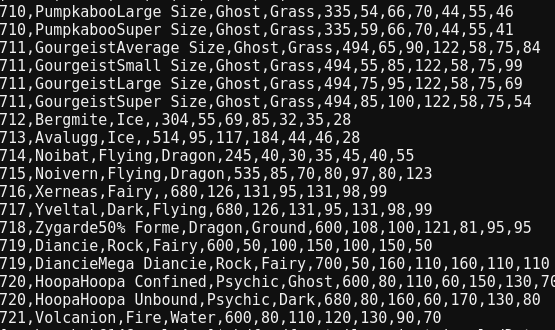
ls

hdfs dfs -ls

hdfs dfs -mkdir project-impala

hdfs dfs -put Dataset-Impala-Project.csv project-impala

hdfs dfs -tail project-impala/Dataset-Impala-Project.csv



## 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.

**Create database:**

create database project2;

use project2;

**Create table pokemon:**

create table if not exists pokemon (pokemonid\_number int, name string, type1 string, type2 string, total int, hp int, attack int, defense int, sp\_atk int, sp\_def int, speed int)

row format delimited

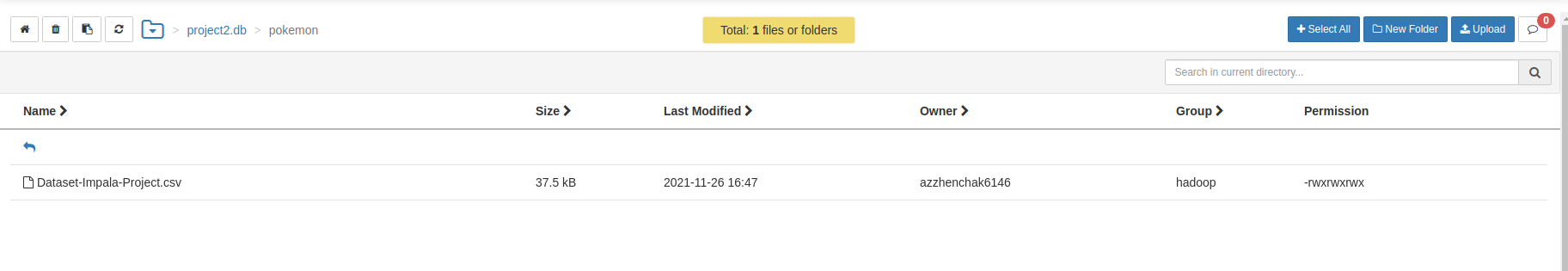
fields terminated by ","

stored as textfile;

**Load the data into the table:**

load data inpath 'project-impala/Dataset-Impala-Project.csv' overwrite into table pokemon;

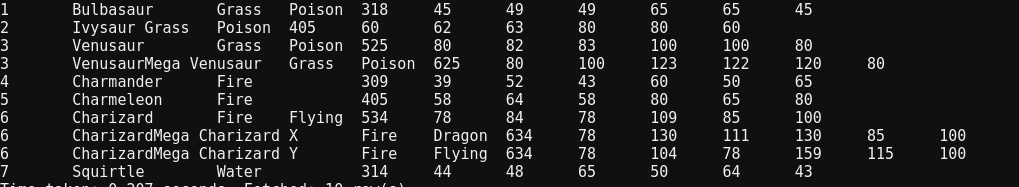
**Check data in Ambari cloudxlab:**

**Verify that the data has been loaded:**

**Show first 10 rows:**

select \* from pokemon limit 10;

**Output**

**Show number of rows in pokemon table:**

select count(\*) from pokemon;

**Output:**

800

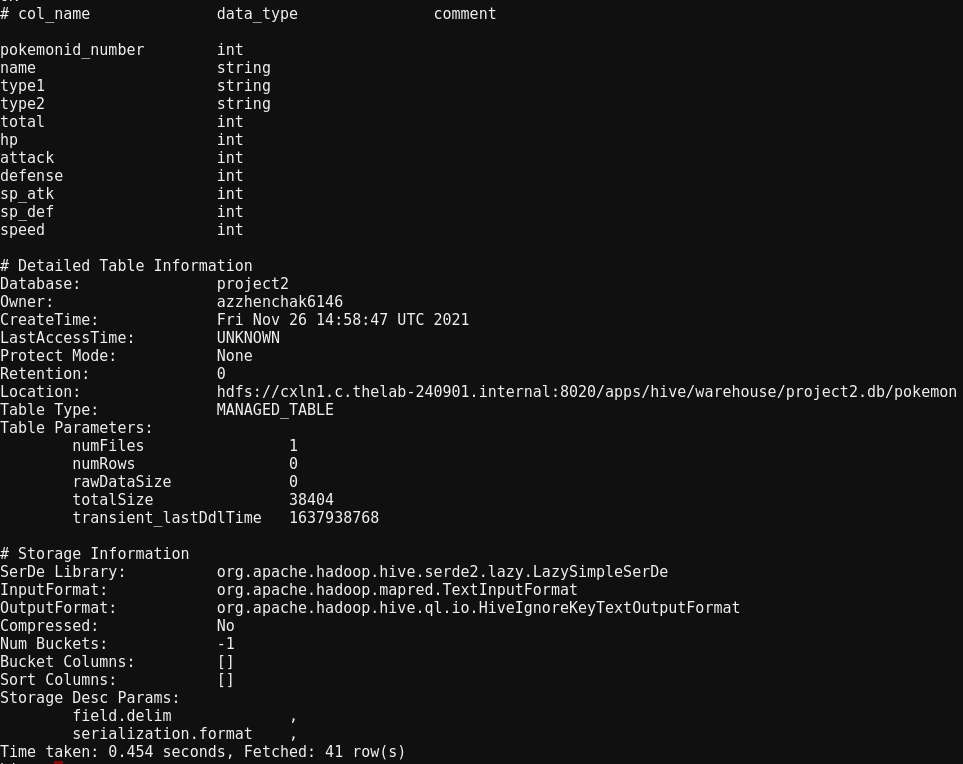
select count(distinct name) from pokemon;

**Output:**

800

There are 800 different pokemon in the dataset

describe formatted pokemon;



## Find out the average HP (Hit points) of all the Pokémon

**Average HP of all the pokemon**

select avg(hp) from pokemon;

**Output:**

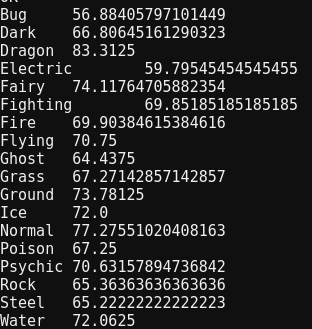
69.25875

Average Hit point of the Pokémon is 69.25875

Let`s see what is the average hp in each group with same type1

**Average HP of pokemons grouped by type1**

select type1, avg(hp) from pokemon group by type1

**Output:**

Dragon Pokemons have the highest average score among other types (type1).

## 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’

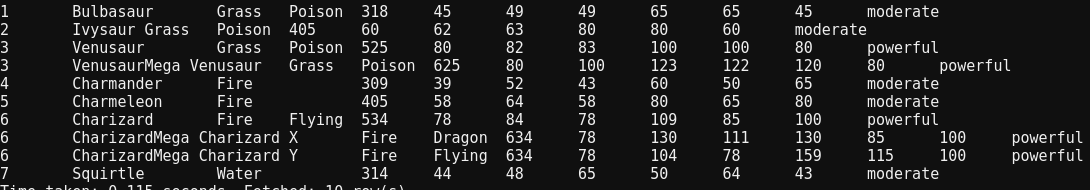
We will create additional column pokemon\_rate baced on average hp. So, pokemons which have hp greater than average are considered as powerful and other are moderate.

create table if not exists pokemon1 as select \*, IF(hp>=69.25875,'powerful',IF(hp<69.25875,'moderate', '')) AS power\_rate from pokemon;

**First 10 rows in pokemon1 table**

select \* from pokemon1 limit 10;

**Output:**

****

## Find out the number of powerful and moderate HP Pokémons present

select power\_rate, count(\*) from pokemon1 group by power\_rate;

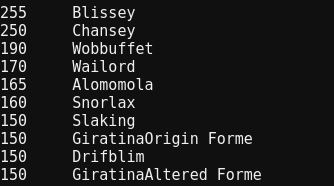
**Output:**



As a result we have 422 pokemons with moderate rate and 378 pokemons with powerful rate

## Find out the top 10 Pokémons according to their HP’s

select hp, name from pokemon order by hp desc limit 10;

**Output:**

In the above screenshot there is printed top 10 Pokemons according to their HP’s