# Homework 1 – Hadoop (Fortnite with Hive)

*This document contains information regarding the homework given by the teachers in the PDP (Parallel Distributed Processing) course. For my homework, I was given the Fortnite dataset for which I had to use Hive to perform various queries to achieve specific results.*

## General information

Name: Alexander Arkhipov

Student number: 647833

GitHub URL: <https://github.com/paxer2k/pdp-assignments>

## Explanation of steps

1. Firstly, did a bit of research on Hive to get a rough idea of how it works and what it offers.
2. I then downloaded the Fortnite Statistics dataset, unzipped it and uploaded it as a table in Hive.
3. I gave all the columns in the fields an appropriate name, such as (fn\_x) -> where fn stands for Fortnite and x is the name of the column in the CSV. Furthermore, I made sure to disregards the first row as it contains the names of the columns (header)
4. After uploading the table, I used the SQL window in the Hive view to write the queries that correspond to the assignment given.
5. Lastly, I spent time on the actual assignment by writing the appropriate queries needed to complete the three assignments given to me.

## Explanation of the source code

### Hive query #1: What time of the day is most popular? Make a top 3.

**Explanation**

The following query displays two columns (time\_of\_day & popularity). It looks at the count of each (eg. 2PM, 4PM, etc.) time of day and displays the highest count at the top and goes down the lower it gets. Furthermore, this query only shows the top 3 time of days played by limiting the dataset.

**Code snippet**

****

### Hive query #2: What is the total damage done by players who have an accuracy between 20 and 40%?

**Explanation**

The following query displays two columns (accuracy & total\_damage) It looks for players who have the same accuracy (between 20% and 40%) and adds their damage dealt to other players (total\_damage). In other words, if 3 players have damage of 300 each and an accuracy of 25%, it would be (25% | 900). On top of this, it filters the data by getting rid of the % sign (by replacing it with an empty string) and converting the number string to a float so that the WHERE and AND portions can be queried on numerical data. Furthermore, it groups everything by accuracy and orders by both the accuracy and total\_damage in a descending ordering from highest accuracy to the lowest.

**Code snippet**

****

### Hive query #3: What is the distance travelled per “Mental state” group.

**Explanation**

The following query displays two columns (mental state & total distance traveled). It looks at all of the distance traveled for each mental state and rounds it off to the nearest whole number. The rounding off was done because the expected output in the assignment was shown as a whole number. Furthermore, this query puts the least distance traveled at the top and the most distance traveled at the bottom. (They are all max distances for each mental state, just some are less than others)

**Code snippet**

****

## Screenshots with steps and results:

When uploading the Fortnite Statistics dataset, the file type settings should be set to the following:

A screenshot of a computer

Description automatically generated with medium confidence

You can adhere to the following table structure which is just fn\_nameofcolumn:

A screenshot of a computer

Description automatically generated

Now go to “Hive view”, enter one of the queries in the SQL window and press “Execute” to test it:

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

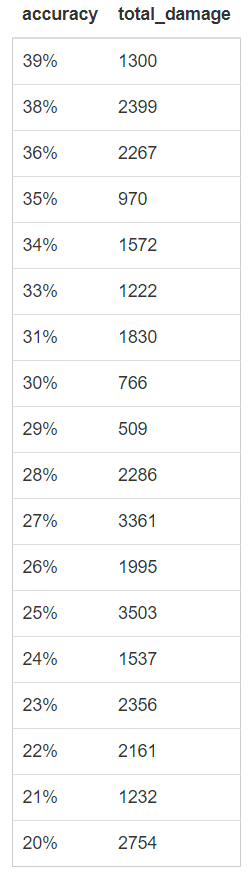
Description automatically generated with medium confidence

### Result #1

A screenshot of a computer

Description automatically generated with medium confidence

### Result #2



### Result #3

A screenshot of a computer

Description automatically generated with medium confidence