# 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 query selects the ‘fn\_timeofday’ column from the ‘fn\_statistics’ table and counts the number of occurrences of each unique value in that column, aliasing it as ‘popularity’. The rows are then grouped based on the unique values in the ‘fn\_timeofday’ column. The result set is ordered by the popularity column in descending order, representing the most popular time of day. Finally, the query limits the output to the top 3 rows with the highest popularity.

**Code snippet**

****

**Result**

A screenshot of a computer

Description automatically generated with medium confidence

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

**Explanation**

The query selects the ‘fn\_accuracy’ column from the ‘fn\_statistics’ table and calculates the sum of the ‘fn\_damagetoplayers’ column and aliases the columns as ‘accuracy’ and ‘total\_damage’. It filters the rows based on the ‘fn\_accuracy’ column falling within the range of 20 to 40 (inclusive), after removing the percentage sign. The percentage sign gets replaced with an empty string. The rows are then grouped based on the unique values in the ‘fn\_accuracy column’. The result set is ordered first by the ‘fn\_accuracy’ column in descending order, after removing the percentage sign and casting it as a float, and then by the ’total\_damage’ column in descending order which puts the highest accuracy at the top and the lowest at the bottom.

**Code snippet**

****

**Result**

A screenshot of a cell phone

Description automatically generated with low confidence

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

**Explanation**

The query selects the ‘fn\_mentalstate’ column from the ‘fn\_statistics’ table and calculates the rounded sum of the ‘fn\_distancetraveled’ column and aliases the columns as ‘mental\_state’ and ‘total\_distance’. It groups the rows based on the unique values in the ‘fn\_mentalstate’ column. The result set provides the mental state and the total distance traveled for each unique mental state. The least distance traveled it put at the top meanwhile the most distance traveled is put at the bottom.

**Code snippet**

****

**Result**

A screenshot of a computer

Description automatically generated with medium confidence

## Steps to execute queries

Firstly, click on the 9 squares in the top right menu and head over to the “Hive View”.

A screenshot of a computer

Description automatically generated with medium confidence

Now click on “Upload Table” in the navigation bar and select the gear icon next to the “File type”. Make sure this stays on CSV:



Make sure to copy the file type settings over. Don’t forget to put first row as a header, otherwise the last query will not show proper results. If you forget this, add “WHERE fn\_mentalstate <> ‘Mental State’” to the LAST query, will which ignore the header in the query.

A screenshot of a computer

Description automatically generated with medium confidence

Upload of the CSV file is possible either by checking the “Upload from Local” button which will allow you to upload locally or from the Hortonworks HDFS directly, but you’ll have to provide the path. After doing this, feel free to adhere to the following table structure, which is (fn\_nameofcolumn). Also, do not forget to change the data types of the columns.

A screenshot of a computer

Description automatically generated

Now click on “Query” in the navigation bar and paste in your query into the “Query Editor” and press the “Execute” button. This will run your query and provide you with the corresponding result(s).

A screenshot of a computer

Description automatically generated with medium confidence