Hadoop assignment

Name: Nick Glas

Student number: 670516

Assignment/dataset: Fortnite

Date: 12/05/2023

Git URL: https://github.com/nickglas/fortnite\_assignment

Contents

[Taken steps 3](#_Toc134797229)

[Explanation source code 4](#_Toc134797230)

[Results 8](#_Toc134797231)

# Taken steps

To run the Pig script, I tried installing different versions of the Hadoop environment. The instructions provided in the slides didn't work, so I thought I needed to install specific packages. I first tried version 2.5, but it required an older version of Linux, Python, and Pip, which caused some problems with outdated repositories. Then I tried version 3.0, which worked well until I reached the Ambari dashboard. Unfortunately, I couldn't connect the dashboard to the sandbox environment.

I went back to version 2.5 and realized that I didn't need to install anything extra because the PIG tool was already there. I started learning about Pig and ran some scripts successfully in the Shell. However, I faced difficulties when importing a CSV file because I had to use the Hadoop file system. Eventually, I discovered that I could import data and run Pig scripts through Ambari's web portal, which made things easier.

# Explanation source code

In this chapter, we will take a closer look at the source code and explain how it works. We will go through the code step by step to understand how it functions. By doing this, we will learn how the code achieves its goals and what each line of code does. Our goal is to make the code easier to understand and help you gain insights into how the code works.

For this assignment in Fortnite, I chose to work with a provided CSV file that was given to us by our lecturer. This file contains various columns with data related to Fortnite gameplay. You can find this file in the GitHub repository as well. In order to complete the assignment, I focused on several key columns:

* Date: This column indicates the date when the Fortnite game was played.
* Mental State: This column describes the mental state of the player during the game, which can be either "sober" or "high."
* Headshots: The "Headshots" column represents the number of headshots the player made during the game.
* Damage to Players: This column records the amount of damage the player inflicted on other players.

The assignment was further divided into three smaller tasks:

* Determining the most popular day: I had to identify the top 10 days based on game participation.
* Calculating the total damage done by players with an accuracy between 90% and 100%: I needed to find the cumulative damage inflicted by players who had an accuracy rating within this range.
* Calculating the total number of headshots per "Mental State" group: I was tasked with calculating the overall count of headshots for each "Mental State" group.
* In the following code, I will explain how I tackled each of these assignments and provide a detailed overview of my approach.

In order to load the data frame I used the piggybank jar plugin. The extension in Apache Pig is like a toolbox that contains extra tools and features created by users. It's not included with the basic version of Pig, so that is why I needed to download and register it.

The next step is to import the csv file and assign in to the variable d\_with\_headers. As you can see I tried to assign the good data types for most fields, but I was required to change some fields afterwards because I wasn’t able to do it right away.

A screenshot of a computer code

Description automatically generated with medium confidence

I quickly found out that my code was breaking because the headers are still present in the csv file. I tried to remove those headers by using the piggybank library, but that didn’t work out of the box. What I did was a dirty but quick way to remove the first header row. I filtered out the row that had the date\_str not equal to ‘Date’ and assigned the value to the variable called ‘d’

A picture containing text, font, line

Description automatically generated

The data that I needed was still in a format which I could not use. The accuracy\_str column still has a percentage after the numbers. I replaced the percentage sign with an empty whitespace and created a new field called accuracy. Of course I appended all the other fields as well. This is done by using the asterisk.



The first assignment was a bit tricky because I didn’t find a good way to convert the date fields to a usable format. When I open the file in excel it appended the year to the field, but when I open the same file in a text editor, it showed another value. I have sent an email to our lecturer asking for guidance, but I didn’t hear from him. So for the first assignment I used the group operator to group the dates together. I then get the group lengths based on the group. The next step is to assign the length to the ordered groups variable where I also provide the sorting value.

A screenshot of a computer program

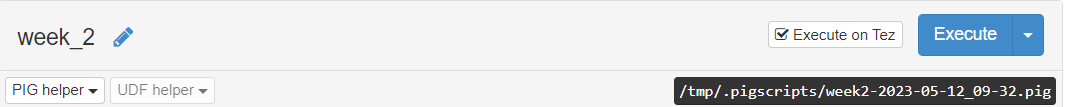
Description automatically generated with low confidence

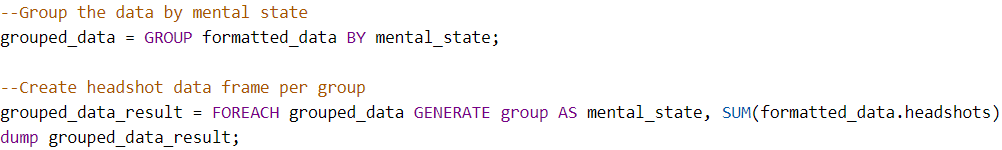
The next assignment was to get the total amount of damage done to players where the accuracy of the player is between 90% and 100% This is done by filtering the accuracy field. I then group all the result and sum up the damage done to players, resulting in a new variable called total\_damage.

A picture containing text, screenshot, font

Description automatically generated

The last assignment is to calculate all the headshots per mental state. As stated before, there are two mental states that the player can be in. This is a sober mental state and a high mental state. I grouped both mental states and then calculated the headshots per group. In order for this to work I needed to enable the ‘Execute on Tez’ mode.





# Results

For assignment one I had to print every day in order to make a list. It was supposed to be a top 10, but the data didn’t include more than 7 days. The results are ordered meaning that the top date is the most played that and the last that is the least played date.

A picture containing text, font, white, typography

Description automatically generated

Assignment two was calculating the total amount of damage done to players where the accuracy of the player is between 90% and 100% once again the data didn’t include many filtered rows (only 1) so the results came back with a single row with a value of 313.



The last assignment is to calculate all the headshots per mental state. The results came back as follows.

A picture containing text, tool, design

Description automatically generated

The total log of this script is as follows.

A screenshot of a computer

Description automatically generated with low confidence