PDP assignment 3 – Chris Lips

Name: Christiaan Lips

Student number: 623434

Github: https://github.com/Lipsinator/Hadoop-fundamentals-Chris-Lips

**Steps:**

1. For this assignment I couldn’t get spark to work on Hadoop so I found a workaround online that lets you install spark on windows which happened to workout for me. Link: <https://phoenixnap.com/kb/install-spark-on-windows-10>
2. Run spark with command: “C:\Spark\spark-2.4.5-bin-hadoop2.7\bin\spark-shell”
3. Execute the file from the terminal when spark is running.

**Results:**

**Assignment 3A:**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

**Assignment 3B:**

****

**Assignment 3C:**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

**Code explanation:**

1. First we load the data from the csv into a data frame.
2. For assignment A I select the fields survived, passenger class and Sex to group by sex and passenger class. Then the average survived is calculated with the result as show above.
3. For assignment B I select the age class and survived from the data frame. Then I filter on age under or equals 10 to then count the children that survived. Finally we calculate the probability of surviving on the titanic as a child on the third class. The results is 41,5 percent.
4. For assignment C the fare and passenger class is selected and then grouped by the passengers and class to calculate the average fare. This result is shown above aswell.