Naam : Murat Celem  
Studentnummer : 642213

URL GitHhub:

https://github.com/DatIsHetLeven/MapReduce

Er zullen vast meerdere manieren zijn om de code te laten runnen, maar dit is hoe ik het gedaan heb.  
  
Inloggen via putty.

Switchen naar root.

Naar de juiste directory gaan.(of zelf 1 aanmaken)

In deze directory de codes plaatsen en dataset.

Vervolgens de volgende code uitvoeren :

python avgHeight.py Soccer2019.csv > output.txt  
python voetOpdracht.py Soccer2019.csv > outputVoet.txt

in dezelfde directory dient de dataset te zitten (Soccer2019.csv).

avgHeight.py staat in dit document onder ‘opdracht 1’ op de volgende pagina.

voetOpdracht.py staat in document onder het kopje ‘opdracht 2’.

**Code Opdracht 1**

# coding=utf-8

from mrjob.job import MRJob

class AvgValueByFoot(MRJob):

def mapper(self, \_, line):

# split the CSV line into an array of individual columns

columns = line.split(',')

# extract the value column and remove any currency symbols

value\_str = columns[11].replace('€', '').replace('M', '000000').replace('K',$

# check if the value is 0 and skip the record

if value\_str == '0':

return

# extract the preferred foot column and remove any single quotes

foot\_str = columns[14].replace("'", "")

try:

# convert value to integer

value = int(value\_str)

yield foot\_str, (value, 1)

except ValueError:

# if the value cannot be converted to an integer, skip the record

pass

def combiner(self, foot, values):

total\_value = 0

num\_players = 0

for value, count in values:

total\_value += value

num\_players += count

yield foot, (total\_value, num\_players)

def reducer(self, foot, values):

total\_value = 0

num\_players = 0

for value, count in values:

total\_value += value

num\_players += count

avg\_value = total\_value / num\_players

yield foot, avg\_value

if \_\_name\_\_ == '\_\_main\_\_':

AvgValueByFoot.run()

Uitleg van de code hierboven:  
De functie mapper splitst de gegvens van elke rij in kolommen en haalt ook weer de waardes op van de nationaliteit en de lengtes van de spelers.

De funcite reducer wordt gebruikt om de uitvoer uit de mapper te gebruiken en daar verder mee te rekenen. Hier wordt de gemiddelde hoogte per land berkeent door alle hoogtes bij elkaar op te tellen en te delen door het aantal spelers van een bepaalde nationaliteit.

**Code Opdracht 2**

# coding=utf-8

from mrjob.job import MRJob

class AvgValueByFoot(MRJob):

def mapper(self, \_, line):

# split the CSV line into an array of individual columns

columns = line.split(',')

# extract the value column and remove any currency symbols

value\_str = columns[11].replace('€', '').replace('M', '000000').replace('K',$

# check if the value is 0 and skip the record

if value\_str == '0':

return

# extract the preferred foot column and remove any single quotes

foot\_str = columns[14].replace("'", "")

try:

# convert value to integer

value = int(value\_str)

yield foot\_str, (value, 1)

except ValueError:

# if the value cannot be converted to an integer, skip the record

pass

def combiner(self, foot, values):

total\_value = 0

num\_players = 0

for value, count in values:

total\_value += value

num\_players += count

yield foot, (total\_value, num\_players)

def reducer(self, foot, values):

total\_value = 0

num\_players = 0

for value, count in values:

total\_value += value

num\_players += count

avg\_value = total\_value / num\_players

yield foot, avg\_value

if \_\_name\_\_ == '\_\_main\_\_':

AvgValueByFoot.run()

Uitleg van de code hierboven :  
Er wordt een methode genaamd mapper die elke regel in de dataset verwerkt

Line.split zorgt dat de regels gesplits worden met een komma.

Value\_str = columns.replace zorgt ervoor dat de euro teken verwijderd worden, dat was 1 van de problmen waar tegen aangelopen werd, omdat hij dat niet als geldige int zag.

Vervolgens heb je de combiner die ervoor zorgt dat alle waardes met dezelfde voorkeursvoet die optelt(dus de aantallen en de waardes)

Als laatst heb je de reducer die combineert dit en berekent hierdoor de gemiddelde waardes.

**Output**Aangezien Hadoop niet op mijn eigen machine werkte, heb ik gebruik gemaakt van een leenlaptop van Inholland. Vandaar dat ik foto’s gemaakt heb i.p.v. screenshots.

Ik heb gebruik gemaakt van putty en via de terminal om de code laten uitvoeren. De output van de twee opdrachten zie je hieronder

Opdracht 1. De gemiddelde lengte van een voetballer berekenen per land.

Afbeelding met tekst, computer, schermopname, Elektronisch apparaat

Automatisch gegenereerde beschrijving

Opdracht 2 : De gemiddelde waarde berekenen voor een voorkeursvoet van een voetballer.

Afbeelding met tekst, computer, Uitvoerapparaat, Elektronisch apparaat

Automatisch gegenereerde beschrijving

Je ziet dat de output als volgt is:  
“” 354375 oftewel 354k

“Left” 1795238 oftewel 1.8M

“Right” 1603704 oftewel 1.6M

De eerste waarde is het gemiddelde van de spelers zonder voorkeursvoet.