**Pyspark Exercise -2**

**(a) To use the “friends\_test” dataset. Col1 is ID, Col2 is name, Col 3 is Age, Col 4 is num of friends. Understand mapvalues function of RDD in spark and find the average number of friends for each unique age present in the dataset.**

**(b) Use the “temp.csv” dataset. Column headers are present in the dataset. Understand filter operations and filter out only the “TMIN” values from the “desc” column. With the resultant data (RDD) find the following:**

**a. Minimum temperature (overall)**

**b. Minimum temperature for every ItemID**

**c. Minimum temperature for every StationID.**

**(c) Use the same dataset, filter only “TMAX” column and find the maximum temperatures just like the ones mentioned above.**

1. The following is the directory structure in my Ubuntu WSL:

/home/hadoop/hadoop/

├── pyspark\_friends/

├── input.txt/

├── WordCount.java/

├── all the wordcount classes/

├── part-r-00000/

A screenshot of a computer

Description automatically generated

1. Setting up the directory structure



1. Inside the directory, add the datasets

A computer screen shot of a black screen

Description automatically generated



1. Write the Pyspark Scripts

For (a)

A computer screen with white text and green text

Description automatically generated

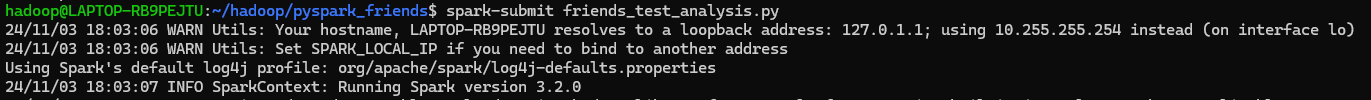
For(b) and (c)

A computer screen shot of a computer program

Description automatically generated

1. Run the programs

For friends\_test.py



A screenshot of a computer

Description automatically generated

A computer screen with white text

Description automatically generated

For temp\_analysis.py

A screen shot of a computer

Description automatically generated





A computer screen shot of numbers and symbols

Description automatically generated





A computer screen with white text

Description automatically generated

1. Stop all Daemons

A screen shot of a computer program

Description automatically generated