# **7007 Lab Test (G1)**

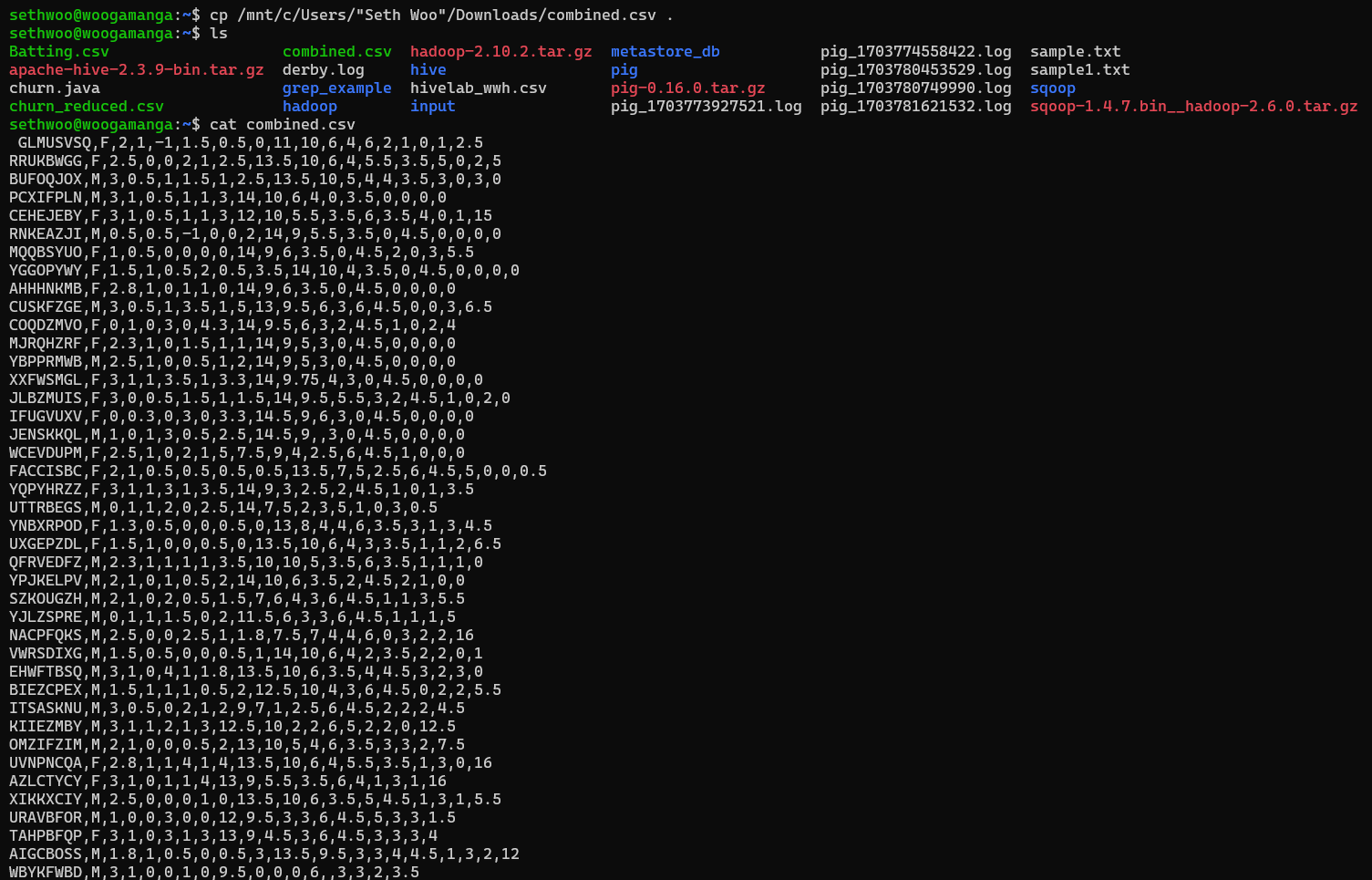
|  |  |
| --- | --- |
| **Name** | **Matrix ID** |
| Woo Wai Hong | 22065374 |
| Chua Xing Zhao | S2188563 |

**Part 1**

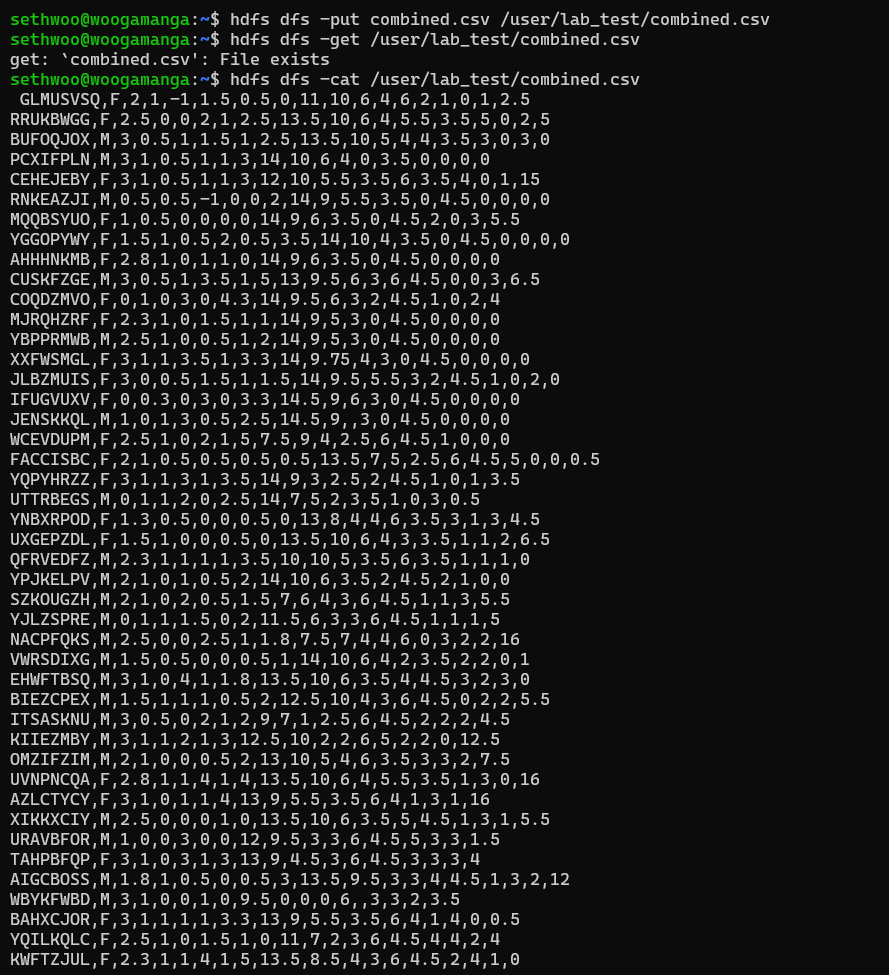
**Question 1**

Dataset used: Set01 & Set02

Copied combined dataset from Downloads folder to Linux file system



Imported combined dataset to HDFS



Problems with dataset:

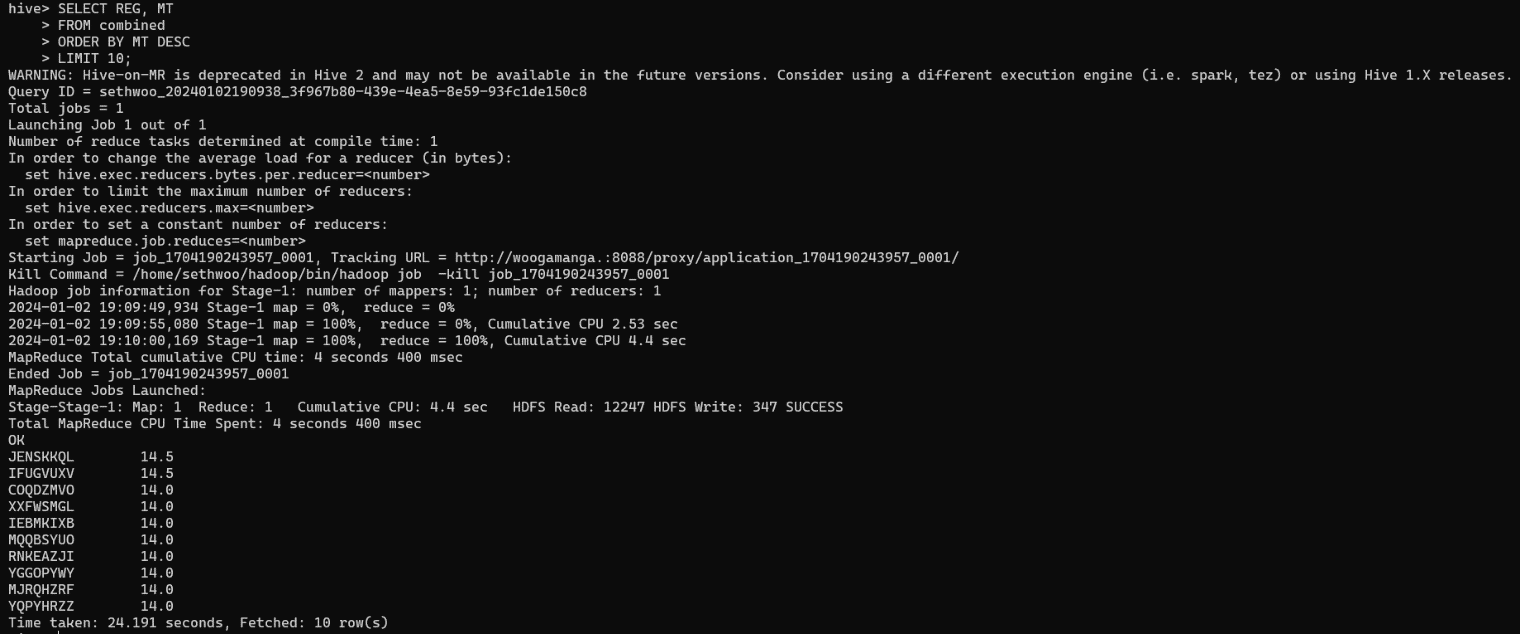
Some of the columns name started with numeric eg: 1FE, 2FE, 3FE, 4FE, 5FE, 6FE. The naming conventions for hive such as table names or column names must start with an alphabetical character.

The column “REG. NO.” has “.” which is not recommended to use in the column name as it follows the standard SQL practice.

**Question 2**

**(a)**

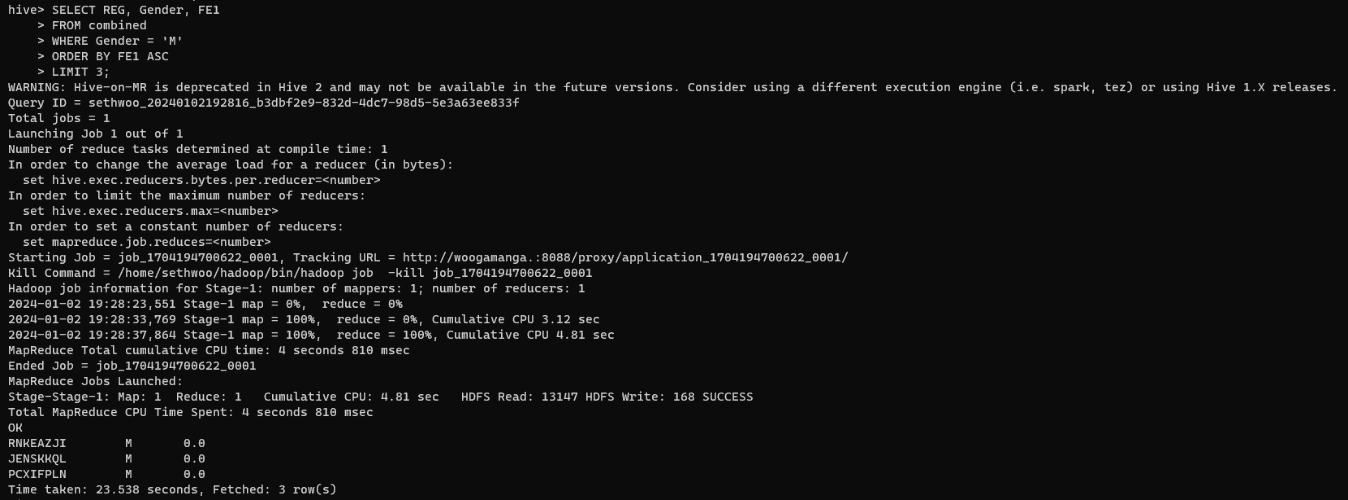
SELECT REG,MT FROM combined ORDER BY MT DESC LIMIT 10



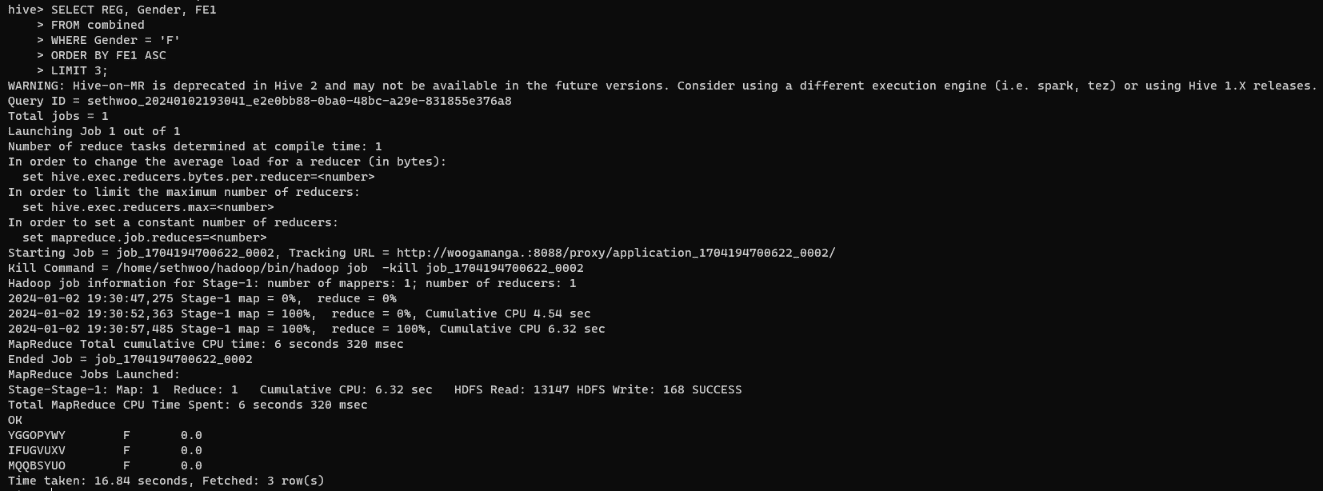
These are the top 10 students with the highest MT score.

**(b)**

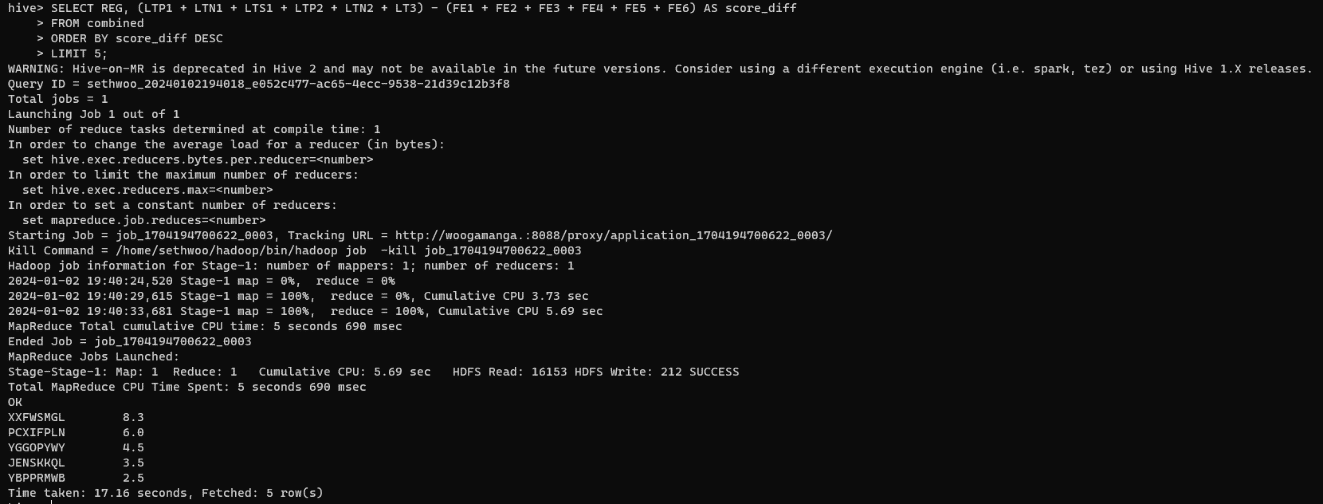
SELECT REG, Gender, FE1 FROM combined WHERE Gender = 'M' ORDER BY FE1 ASC LIMIT 3



SELECT REG, Gender, FE1 FROM combined WHERE Gender = 'F' ORDER BY FE1 ASC LIMIT 3



**(c)**

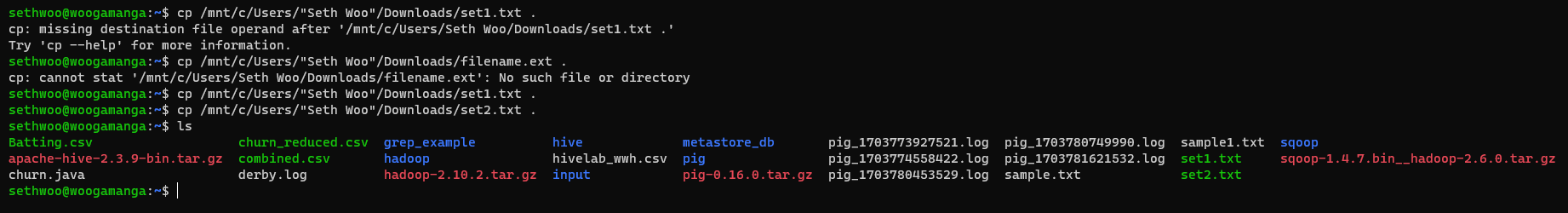


**Part 2**

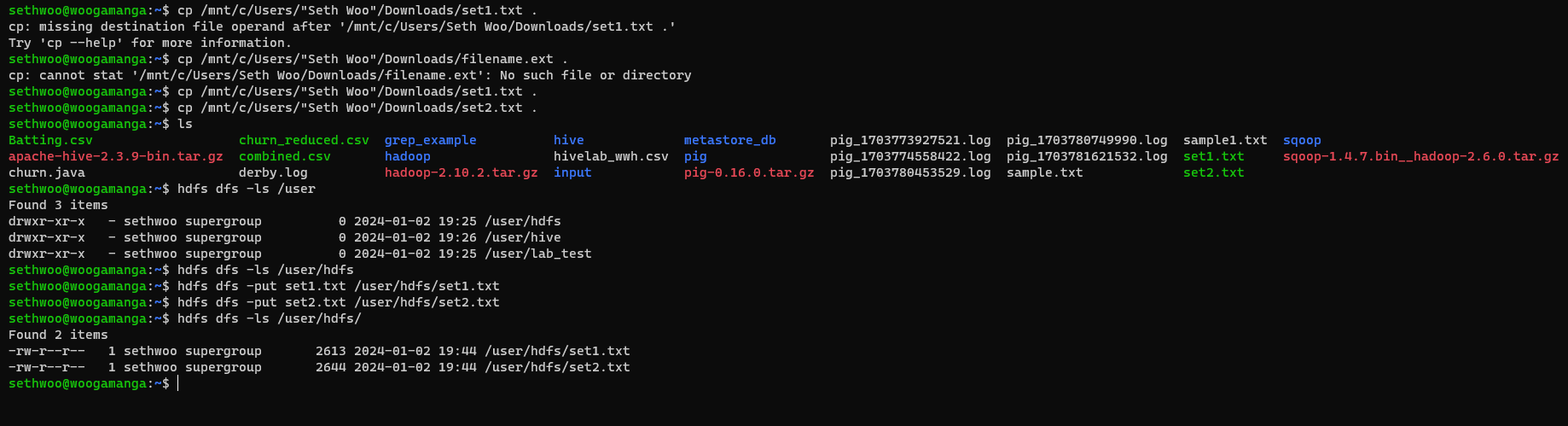
**Question 1**

Text file used: Set01.txt & Set02.txt

Copied two text files from Downloads folder to Linux file system:

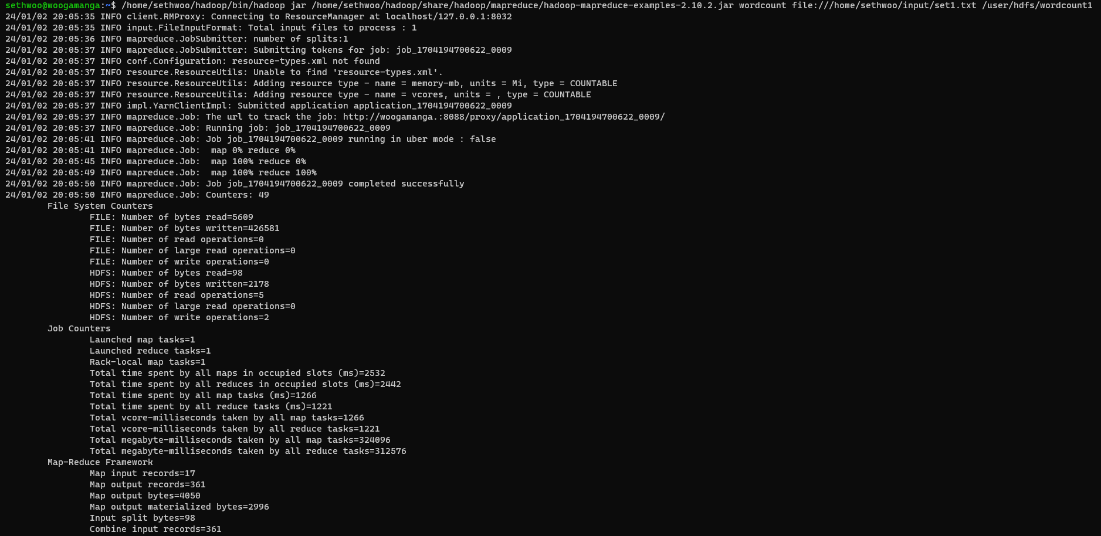


Imported two text file to HDFS

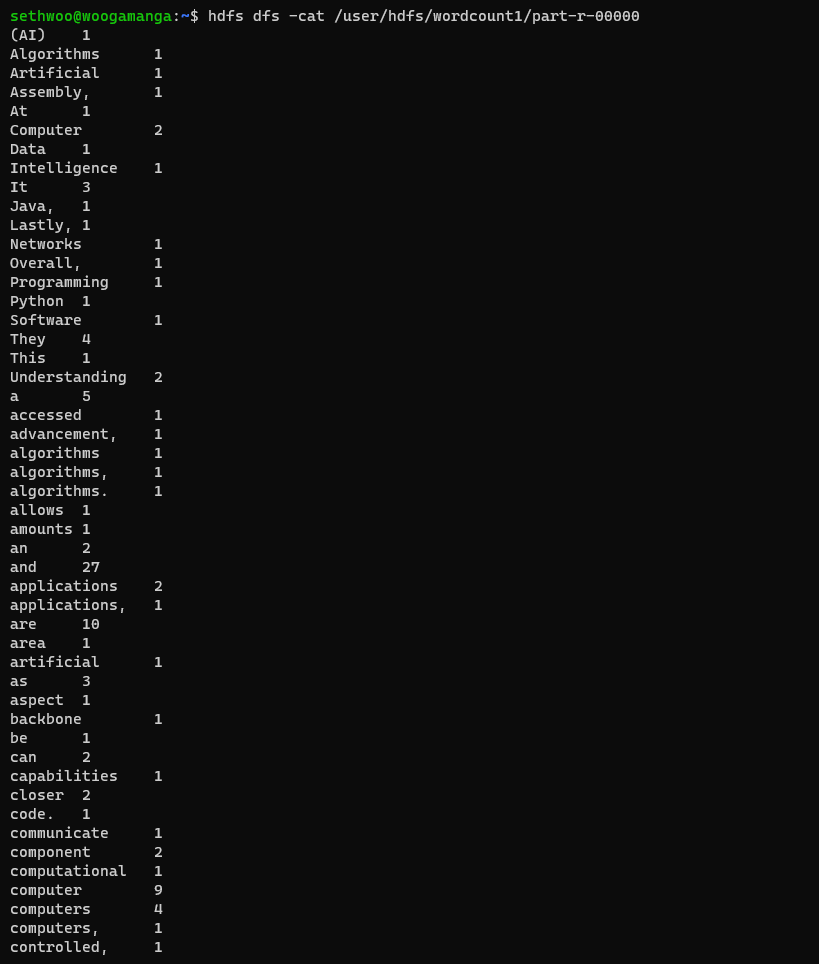


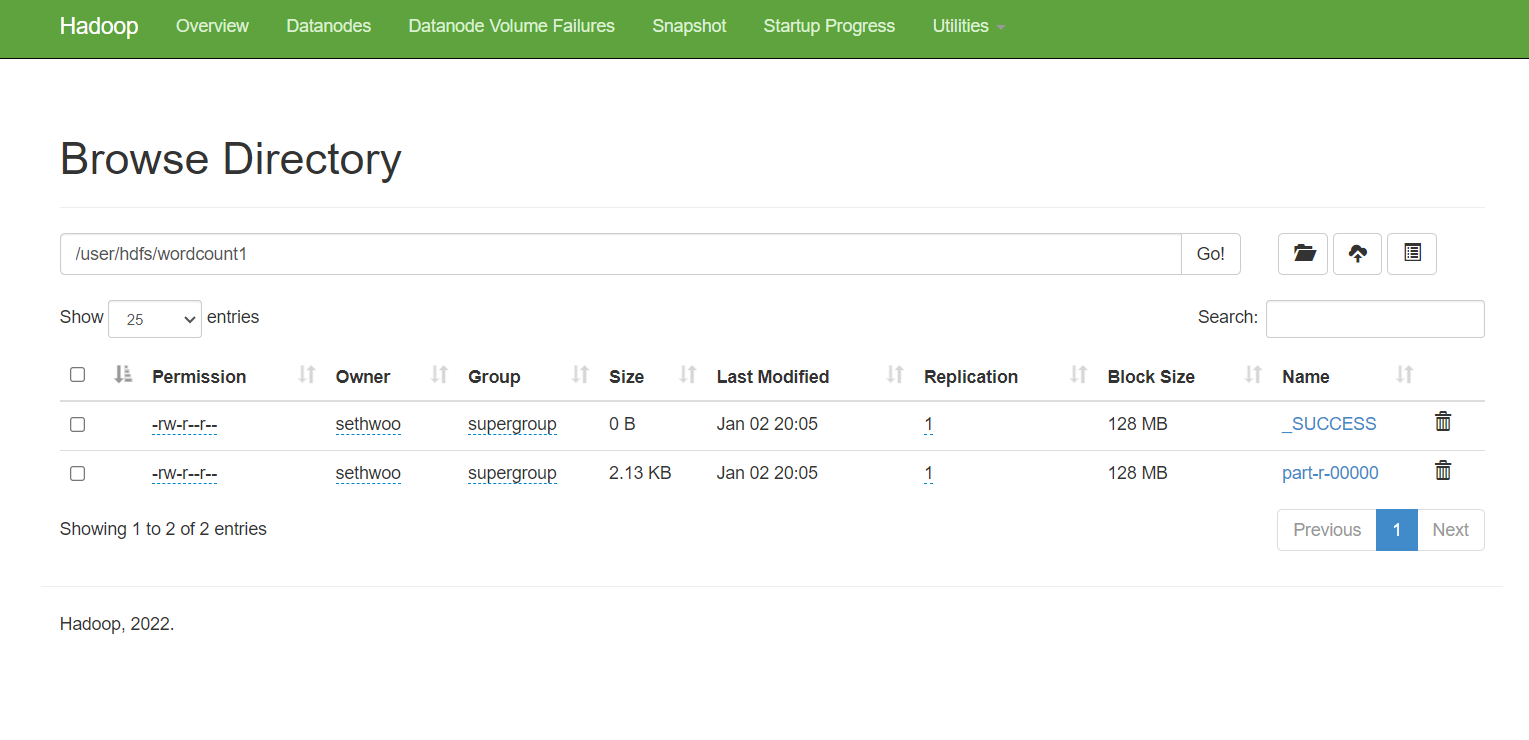
**Question 2**

Run wordcount on set1.txt

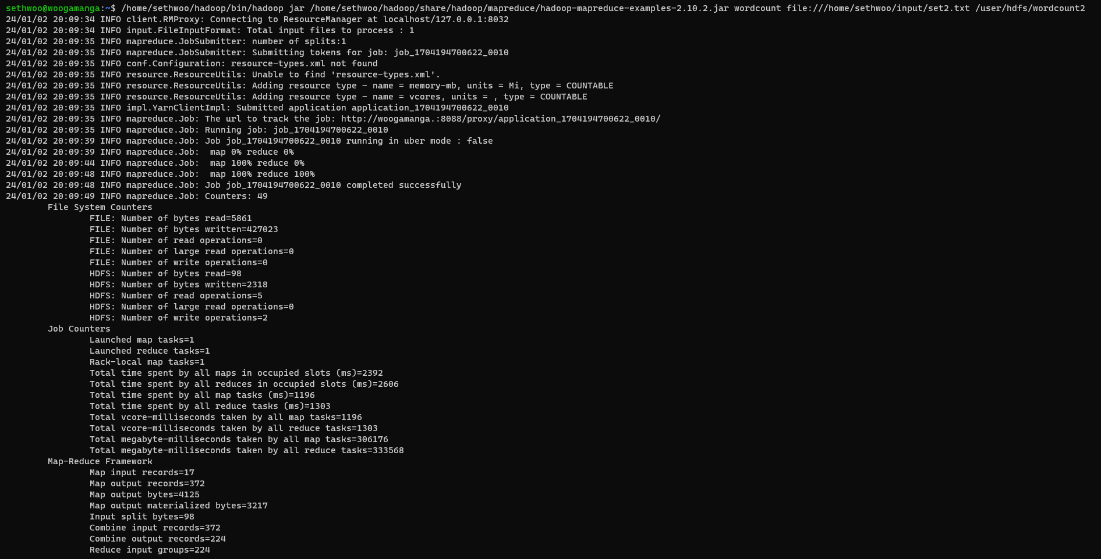


Output of wordcount1 from set1.txt

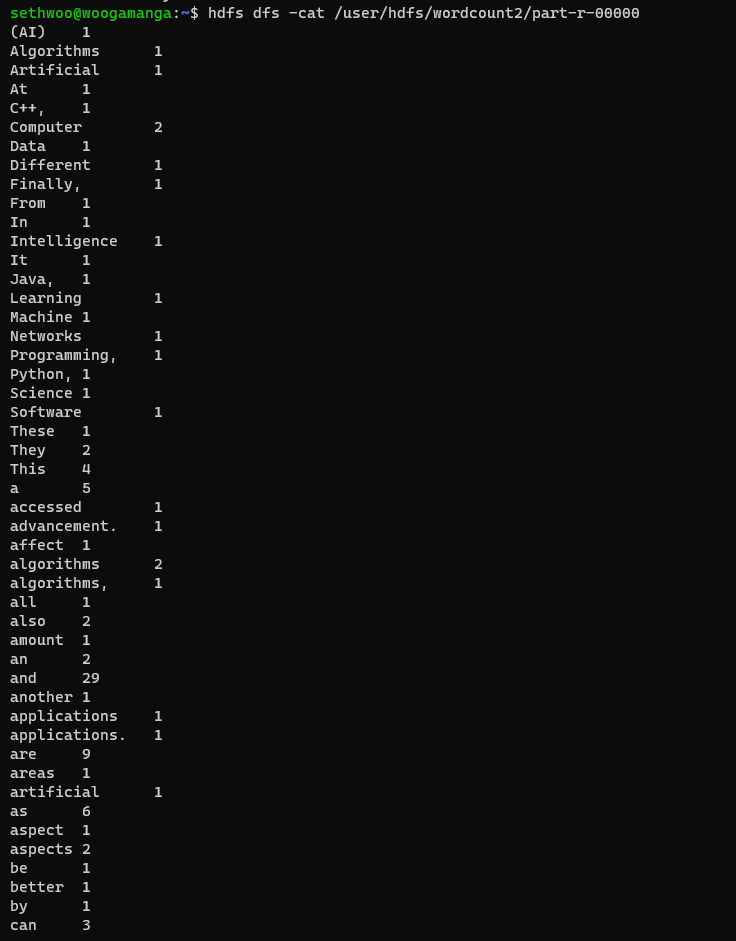


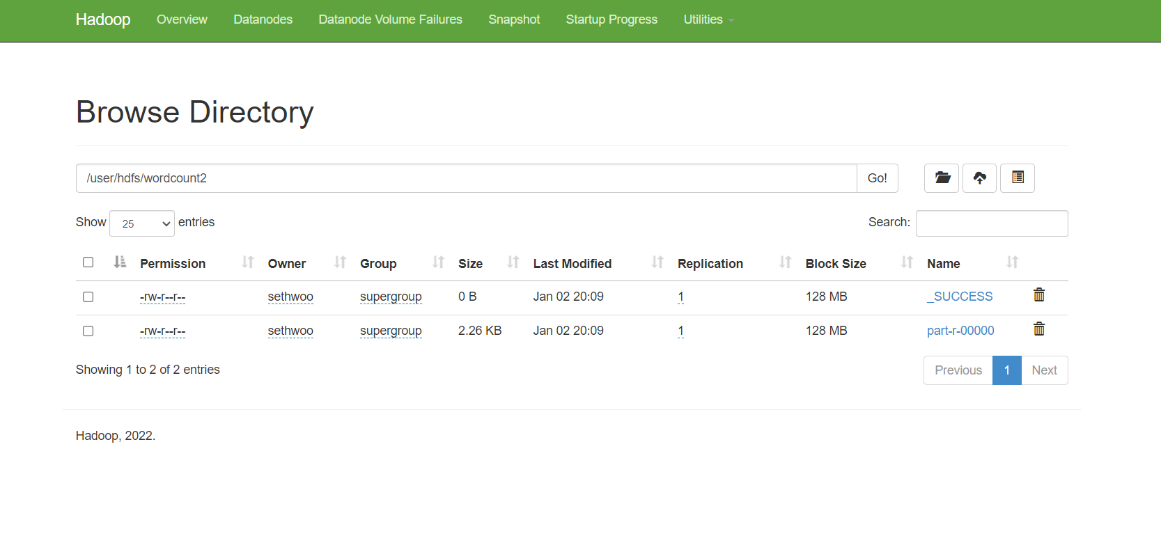


Run wordcount for set02.txt



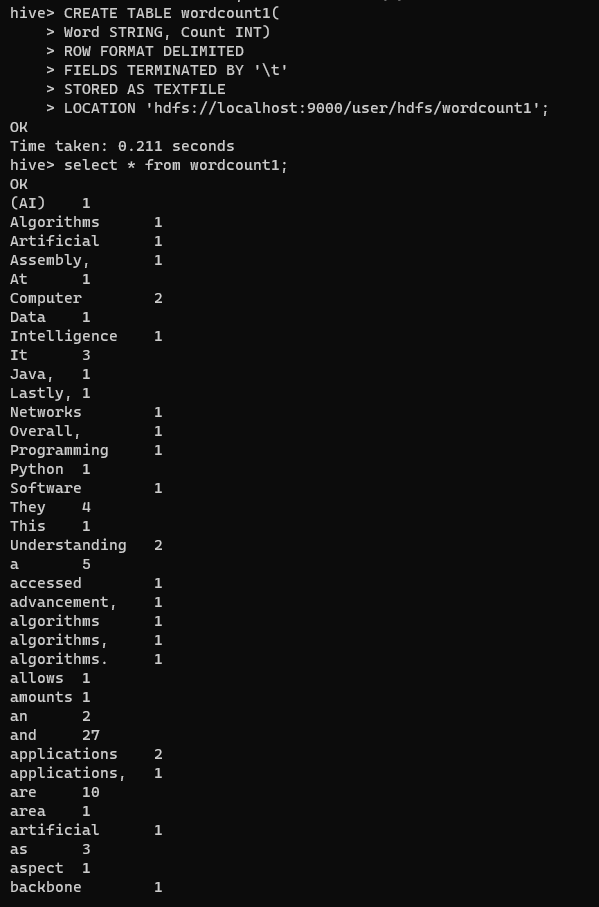
Output of wordcount2 for set02.txt



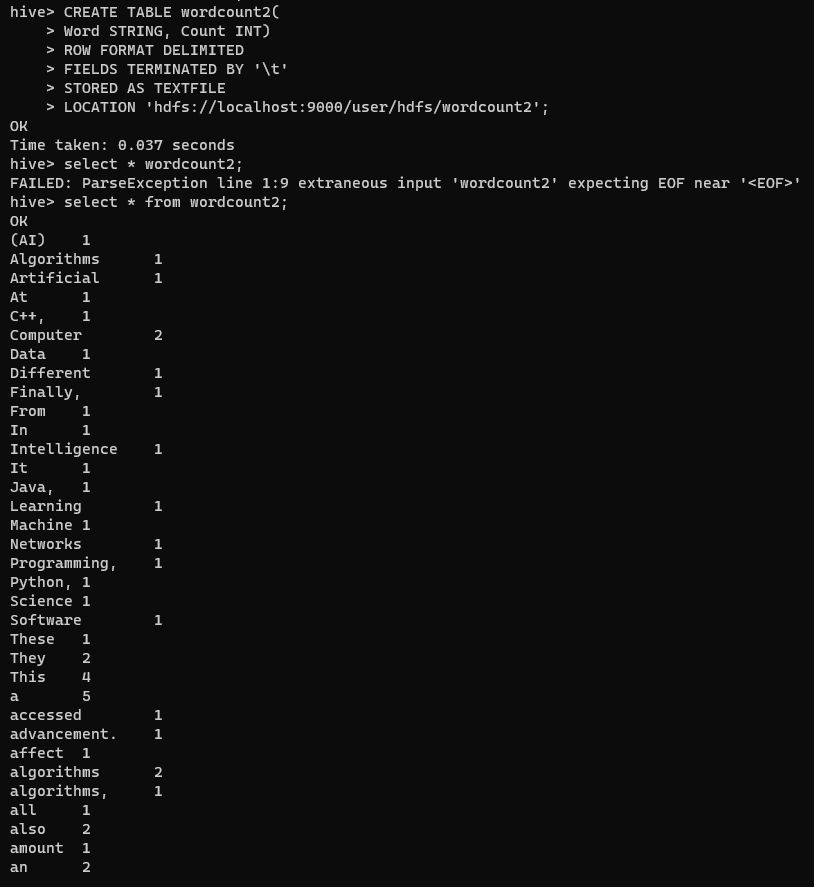


**Question 3**

Imported result of wordcount1 for set1 into hive table wordcount1.

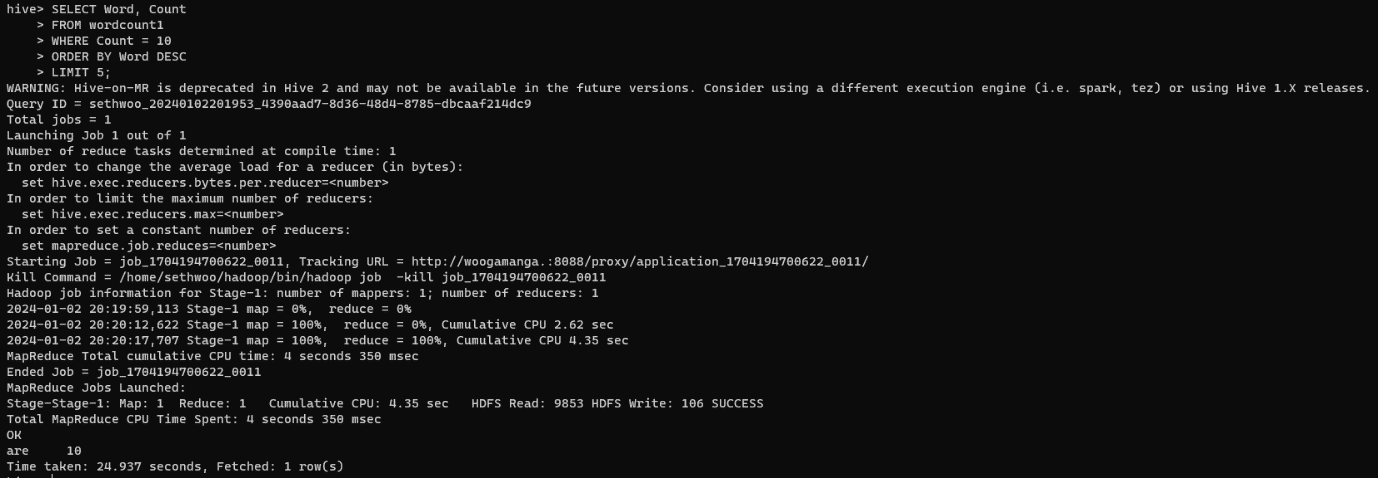


Imported result of wordcount2 for set2 into hive table wordcount2.



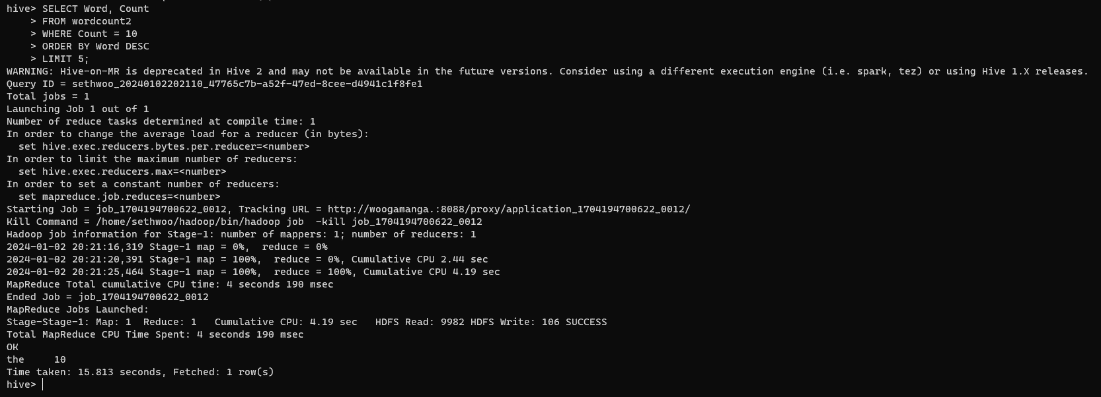
(a)

Set 1



There are only 1 word that has 10 counts in set1 which is “are”.

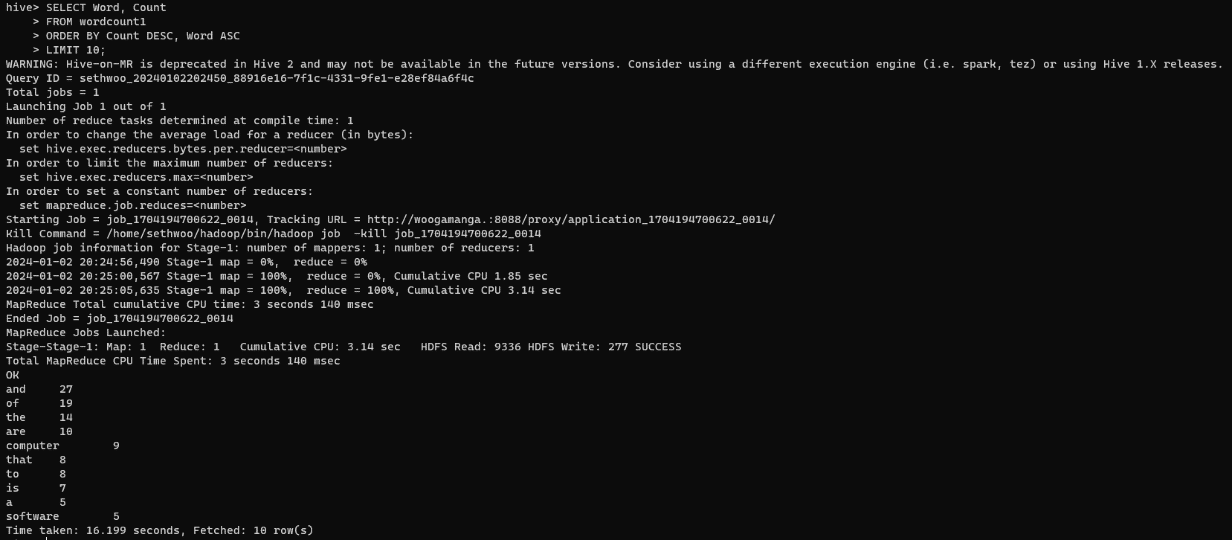
Set 2



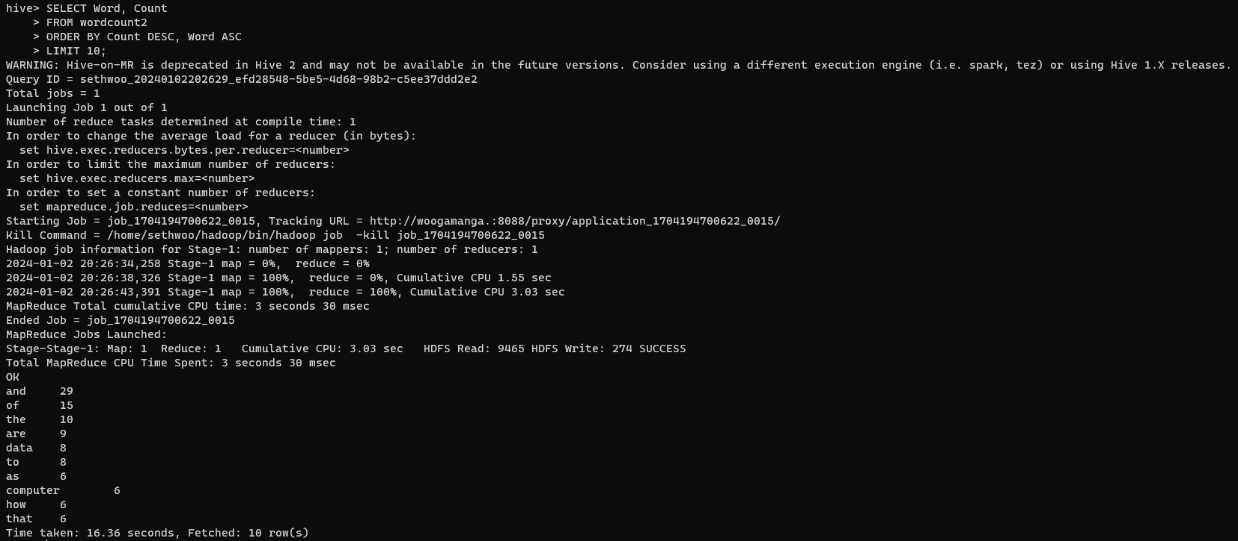
There are only 1 word that has 10 counts in set2 which is “the”.

(b)

Set 1

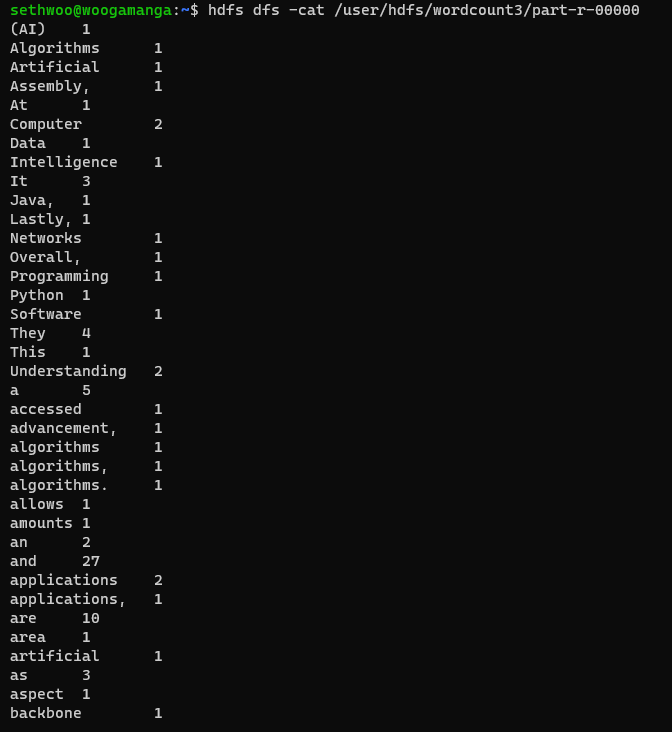


Set 2



**Question 4**

We clean the dataset by removing all the commas and full stops. This will improve the result accuracy as the wordcount in the previous dataset considered the font and comma as a new font.



U,{c8668e8c-965a-4d76-a0a0-80122fd9b778}{60},3.125,3.125