Big Data

Compte Rendu TP2

I- Exécution d'un job avec hadoop streaming

- Exécution du programme en local :
  - Le code de mapper :

```python
import sys


#input comes from STDIN
for line in sys.stdin:
    #remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    for word in words:
        print(f"{word}\t1")

```

  - Le code de reducer:

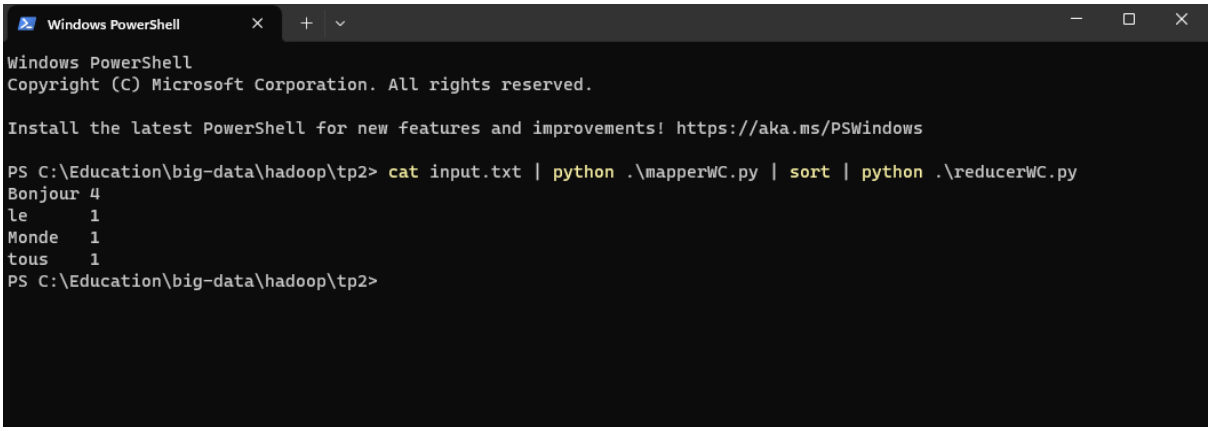
```python
import sys
# Initialise variables
current_word = None
current_count = 0

word = None
# Iterate Through input lines, which are sorted by key (word) in ascending or
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()
    #split the key (word) and value (count) by a tab character
    word, count = line.split("\t",1)
    # convert the count into integer
    try:
        count = int(count)
    except ValueError:
        # if conversin fails, skip this line
        continue
    # If current word is the same as the previous word, increment the count
    if current_word == word:
        current_count += count
    else:
        # If the word changes, print the result for the previous word
        if current_word:
            print('{}\t{}'.format(current_word,current_count))
        # reset the variables for the new word
        current_word = word
        current_count = count
# print the result for the last word
print('{}\t{}'.format(current_word,current_count))

```

  - le fichier input.txt

    Contenu : Bonjour Bonjour Bonjour tous le Monde Bonjour

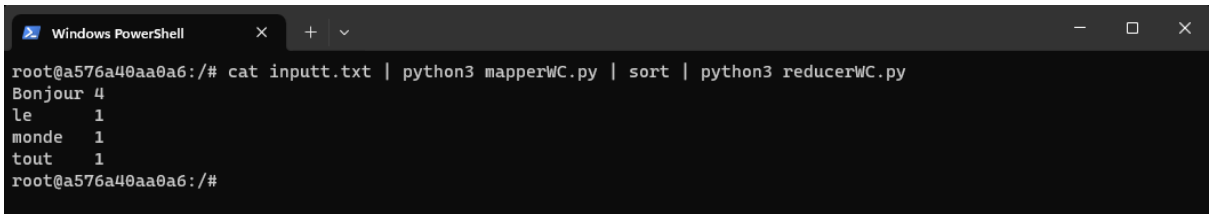
- Exécution du programme :

  > cat input.txt | python .\mapperWC.py | sort | python .\reducerWC.py



- Exécution du programme dans le cluster :
  - Exécution du programme :

    > cat input.txt | python .\mapperWC.py | sort | python .\reducerWC.py

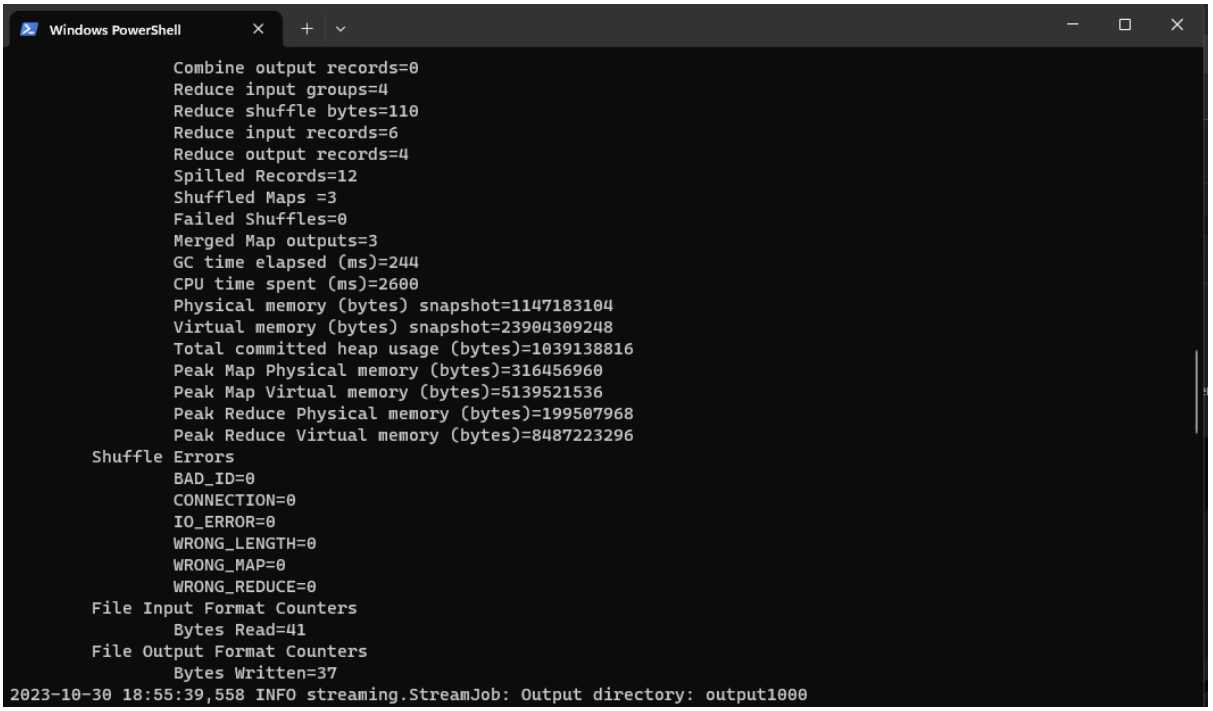


  - Les fichiers input :

  - Exécution de job map reduce :

La commande utilisé :

> hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -file ./mapperWC.py -mapper 'python3 mapperWC.py' -file ./reducerWC.py -reducer 'python3 reducerWC.py' -input input -output output500 -mapper mapperWC.py



- Affichage de resultat :

La commande utilisé :

> hdfs dfs -cat output1000/part-00000