

Big Data

Compte Rendu TP2

I- Exécution d'un job avec hadoop streaming

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

```
1 import sys
2
3
4 #input comes from STDIN
5 for line in sys.stdin:
6     #remove leading and trailing whitespace
7     line = line.strip()
8     # split the line into words
9     words = line.split()
10    for word in words:
11        print(f"{word}\t1")
12
```

- Le code de reducer:

```
1 import sys
2 # Initialise variables
3 current_word = None
4 current_count = 0
5
6 word = None
7 # Iterate Through input lines, which are sorted by key (word) in ascending or
8 for line in sys.stdin:
9     # Remove leading and trailing whitespace
10    line = line.strip()
11    #split the key (word) and value (count) by a tab character
12    word, count = line.split("\t",1)
13    # convert the count into integer
14    try:
15        count = int(count)
16    except ValueError:
17        # if conversion fails, skip this line
18        continue
19    # If current word is the same as the previous word, increment the count
20    if current_word == word:
21        current_count += count
22    else:
23        # If the word changes, print the result for the previous word
24        if current_word:
25            print('{}\t{}'.format(current_word,current_count))
26        # reset the variables for the new word
27        current_word = word
28        current_count = count
29 # print the result for the last word
30 print('{}\t{}'.format(current_word,current_count))
31
```

- 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

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Education\big-data\hadoop\tp2> cat input.txt | python .\mapperWC.py | sort | python .\reducerWC.py
Bonjour 4
le      1
Monde   1
tous    1
PS C:\Education\big-data\hadoop\tp2>
```

- Exécution du programme dans le cluster :

- Exécution du programme :

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

```
Windows PowerShell
root@a576a40aa0a6:/# cat inputt.txt | python3 mapperWC.py | sort | python3 reducerWC.py
Bonjour 4
le      1
monde   1
tout    1
root@a576a40aa0a6:/#
```

- Les fichiers input :

```
Windows PowerShell
root@a576a40aa0a6:/# ls input
f1.txt f2.txt f3.txt f4.txt
root@a576a40aa0a6:/#
```

- 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**

```
Windows PowerShell
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=110
Reduce input records=6
Reduce output records=4
Spilled Records=12
Shuffled Maps =3
Failed Shuffles=0
Merged Map outputs=3
GC time elapsed (ms)=244
CPU time spent (ms)=2600
Physical memory (bytes) snapshot=1147183104
Virtual memory (bytes) snapshot=23904309248
Total committed heap usage (bytes)=1039138816
Peak Map Physical memory (bytes)=316456960
Peak Map Virtual memory (bytes)=5139521536
Peak Reduce Physical memory (bytes)=199507968
Peak Reduce Virtual memory (bytes)=8487223296
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=41
File Output Format Counters
  Bytes Written=37
2023-10-30 18:55:39,558 INFO streaming.StreamJob: Output directory: output1000
```

- Affichage de resultat :

La commande utilisé :

> hdfs dfs -cat output1000/part-00000

```
root@a576a40aa0a6:/# hdfs dfs -cat output1000/part-00000
2023-10-30 18:57:28,073 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, r
emoteHostTrusted = false
Docker 1
Hello 3
MapReduce 1
World 1
```