Compte Rendu
Bases de données NoSQL et Big Data

TP2
Le traitement Batch avec Hadoop Streaming

Réalisé par
SAIDANE Yassine

Partie I : Comprendre Hadoop Streaming

Hadoop Streaming est un API qui nous permet d'utiliser n'importe quel langage de programmation qui peut lire et écrire à l'entrée/sortie standard.

Partie II: Exécution d'un job avec Hadoop Streaming

Mapper : mapperWC.py

mapperWC.py > ...

```python
#!/usr/bin/env python3

"""mapper.py"""
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        # tab-delimited; the trivial word count is 1
        print(f"{word}\t1")
```

Reducer : reducerWC.py

reducerWC.py > ...

```python
#!/usr/bin/env python3
import sys

# Initializer variables
current_word = None
current_count = 0
word = None

# Iterate through input lines, which are sorted by key (word) in ascending order
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 to an integer
    try:
        count = int(count)
    except ValueError:
        # If the conversion fails, skip this line
        continue
    # If the 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
if current_word == word:
    print('{}\t{}'.format(current_word, current_count))
```

Input : input.txt

input.txt

```
hello how are you
are you here
```

Exécution du programme en local

```
PS E:\Documents\Academic\ISIMM\Semestre 7\Big Data\TP2> cat input.txt | python mapperWC.py | sort | python reducerWC.py
are     2
hello   1
here    1
how     1
you     2
PS E:\Documents\Academic\ISIMM\Semestre 7\Big Data\TP2>
```

Exécution du programme sur le cluster

1. *Transferer les scripts dans le cluster Hadoop*

```
docker cp mapperWC.py namenode:/mapperWC.py
docker cp reducerWC.py namenode:/reducerWC.py
```

2. *Entrer dans le container du **namenode***

```
PS C:\Users\ASUS> docker exec -it namenode bash
root@d307457e10fd:/#
```

3. *Créer des fichiers textes et les transférer dans HDFS*

```
root@d307457e10fd:/# cd root
root@d307457e10fd:~# mkdir input
root@d307457e10fd:~# echo "Hello World" > input/f1.txt
root@d307457e10fd:~# echo "Hello Docker" > input/f2.txt
root@d307457e10fd:~# echo "Hello Hadoop" > input/f3.txt
root@d307457e10fd:~# echo "Hello MapReduce" > input/f4.txt
root@d307457e10fd:~# hdfs dfs -mkdir -p /input
root@d307457e10fd:~# hdfs dfs -put ./input/* /input/
2023-10-17 16:16:57,640 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2023-10-17 16:16:58,215 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2023-10-17 16:16:58,264 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2023-10-17 16:16:58,300 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
root@d307457e10fd:~#
```

4. *Exécuter le programme MapReduce*

```
> -files mapperWC.py,reducerWC.py \
> -input /input \
> -output /output \
> -mapper mapperWC.py \
> -reducer reducerWC.py
packageJobJar: [/tmp/hadoop-unjar2871292629335823059/] [] /tmp/streamjob1381590660291170974.jar tmpDir=null
2023-10-17 17:29:21,643 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.2:8032
2023-10-17 17:29:21,938 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.6:10200
2023-10-17 17:29:21,982 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.2:8032
2023-10-17 17:29:21,983 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.6:10200
2023-10-17 17:29:22,225 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1697556801243_0009
2023-10-17 17:29:22,380 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:22,505 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:22,546 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:22,663 INFO mapred.FileInputFormat: Total input files to process : 4
2023-10-17 17:29:22,702 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:23,153 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:23,173 INFO mapreduce.JobSubmitter: number of splits:4
2023-10-17 17:29:23,368 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 17:29:23,391 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1697556801243_0009
2023-10-17 17:29:23,391 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-17 17:29:23,673 INFO conf.Configuration: resource-types.xml not found
2023-10-17 17:29:23,674 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-17 17:29:23,985 INFO impl.YarnClientImpl: Submitted application application_1697556801243_0009
2023-10-17 17:29:24,099 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1697556801243_0009/
2023-10-17 17:29:24,104 INFO mapreduce.Job: Running job: job_1697556801243_0009
2023-10-17 17:29:32,409 INFO mapreduce.Job: Job job_1697556801243_0009 running in uber mode : false
2023-10-17 17:29:32,411 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-17 17:29:41,639 INFO mapreduce.Job:  map 25% reduce 0%
2023-10-17 17:29:42,649 INFO mapreduce.Job:  map 50% reduce 0%
2023-10-17 17:29:43,658 INFO mapreduce.Job:  map 75% reduce 0%
2023-10-17 17:29:48,750 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-17 17:29:50,774 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-17 17:29:50,795 INFO mapreduce.Job: Job job_1697556801243_0009 completed successfully
2023-10-17 17:29:50,901 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=75
                FILE: Number of bytes written=1164152
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=394
                HDFS: Number of bytes written=46
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
```

```
root@d307457e10fd:/# hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \
> -files mapperWC.py,reducerWC.py \
> -input /input \
> -output /output \
> -mapper mapperWC.py \
> -reducer reducerWC.py
packageJobJar: [/tmp/hadoop-unjar4259469173695533595/] [] /tmp/streamjob6569406924102417900.jar tmpDir=null
2023-10-17 16:34:52,600 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.2:8032
2023-10-17 16:34:52,823 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.6:10200
2023-10-17 16:34:52,874 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.2:8032
2023-10-17 16:34:52,877 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.6:10200
2023-10-17 16:34:53,221 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1697556801
243_0002
2023-10-17 16:34:53,393 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:53,561 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:53,600 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:54,193 INFO mapred.FileInputFormat: Total input files to process : 4
2023-10-17 16:34:54,252 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:54,712 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:54,732 INFO mapreduce.JobSubmitter: number of splits:4
2023-10-17 16:34:54,962 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-17 16:34:55,012 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1697556801243_0002
2023-10-17 16:34:55,013 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-17 16:34:55,347 INFO conf.Configuration: resource-types.xml not found
2023-10-17 16:34:55,348 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-17 16:34:56,326 INFO impl.YarnClientImpl: Submitted application application_1697556801243_0002
2023-10-17 16:34:56,467 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1697556801243_0002/
2023-10-17 16:34:56,479 INFO mapreduce.Job: Running job: job_1697556801243_0002
2023-10-17 16:35:13,416 INFO mapreduce.Job: Job job_1697556801243_0002 running in uber mode : false
2023-10-17 16:35:13,418 INFO mapreduce.Job:  map 0% reduce 0%
```

```
2023-10-17 17:29:50,795 INFO mapreduce.Job: Job job_1697556801243_0009 completed successfully
2023-10-17 17:29:50,901 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=75
                FILE: Number of bytes written=1164152
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=394
                HDFS: Number of bytes written=46
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Rack-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=93760
                Total time spent by all reduces in occupied slots (ms)=32976
                Total time spent by all map tasks (ms)=23440
                Total time spent by all reduce tasks (ms)=4122
                Total vcore-milliseconds taken by all map tasks=23440
                Total vcore-milliseconds taken by all reduce tasks=4122
                Total megabyte-milliseconds taken by all map tasks=96010240
                Total megabyte-milliseconds taken by all reduce tasks=33767424
        Map-Reduce Framework
                Map input records=4
                Map output records=8
                Map output bytes=70
                Map output materialized bytes=142
                Input split bytes=340
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=142
                Reduce input records=8
                Reduce output records=5
                Spilled Records=16
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=564
                CPU time spent (ms)=4870
                Physical memory (bytes) snapshot=1395916800
```

```
                Total vcore-milliseconds taken by all reduce tasks=4122
                Total megabyte-milliseconds taken by all map tasks=96010240
                Total megabyte-milliseconds taken by all reduce tasks=33767424
        Map-Reduce Framework
                Map input records=4
                Map output records=8
                Map output bytes=70
                Map output materialized bytes=142
                Input split bytes=340
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=142
                Reduce input records=8
                Reduce output records=5
                Spilled Records=16
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=564
                CPU time spent (ms)=4870
                Physical memory (bytes) snapshot=1395916800
                Virtual memory (bytes) snapshot=28664881152
                Total committed heap usage (bytes)=1290797056
                Peak Map Physical memory (bytes)=310616064
                Peak Map Virtual memory (bytes)=5064470528
                Peak Reduce Physical memory (bytes)=190902272
                Peak Reduce Virtual memory (bytes)=8409866240
        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=54
        File Output Format Counters
                Bytes Written=46
2023-10-17 17:29:50,901 INFO streaming.StreamJob: Output directory: /output
root@d307457e10fd:/#
```

5. *Voir les résultats*

```
root@d307457e10fd:/# hdfs dfs -ls /output
Found 2 items
-rw-r--r--   3 root supergroup          0 2023-10-17 17:29 /output/_SUCCESS
-rw-r--r--   3 root supergroup         46 2023-10-17 17:29 /output/part-00000
root@d307457e10fd:/# hdfs dfs -cat /output/part-00000
2023-10-17 17:31:03,063 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Docker  1
Hadoop  1
Hello   4
MapReduce       1
World   1
root@d307457e10fd:/#
```

6. *Améliorer les codes*

mapperWC.py > ...

```python
#!/usr/bin/python3

"""mapper.py"""
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        i = 0
        while i < len(word):
            if not word[i].isalpha():
                word = word.replace(word[i], "")
            else:
                i += 1
        word = word.lower()
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        # tab-delimited; the trivial word count is 1
        print(f"{word}\t1")
```