Hadoop et Map Reduce

Anthony Moisan

December 22, 2019

Contents

Préambule : ce projet est associé au cours Hadoop et Map_Reduce et se compose de deux projets pour implémenter Map Reduce et pour utiliser via Python et la librairie hdfs3 l'utilisation d'hadoop. Ce Notebook est autoportant avec des chemins en relatif avec l'arborescence des fichiers et dossiers associés.

1 Pré-requis

[1]:
```
!pip freeze > requirements.txt
```

Permet de créer un environnement avec la même configuration de dépendances : pip install -r requirements.txt

2 Projet Map Reduce

2.1 Mapper et Reducer

2.1.1 Mapper

Le mapper va charger uniquement : * le premier champs et extraire le mois (deux premiers caractères) * le second champs pour ensuite filter les départs * le troisième champs pour connaitre la ville * le 11ème champs pour éviter le double comptage * le 13ème champs pour le maximum de seats

Les clés, valeurs exposés correspondent au mois et nombre de seats en filtrant sur la ville Sydney dans le cas d'un départ de cette ville et dans le cas où le champs stop est à 0.

2.1.2 Reducer

Le réducer utilise un dictionnaire permettant de stocker par mois ou un dictionnaire de dictionnaire pour agréger par année/mois les résultats en les sommant. Un test est réalisé pour s'assurer que le nombre de seats est un entier. On trie le dictionnaire de résultats pour l'affficher par ordre croissant.

Deux mapper et reducer ont été effectués : * un pour faire des agrégations uniquement par mois * un autre pour faire des agrégations par mois et par an

2.2 Test en local du mapper et reducer

2.2.1 Mapper et reducer avec agrégation par mois

[2]:
```
!cat Reservation/data.csv | python Reservation/Reservation_mapper.py | python␣
↪Reservation/Reservation_reduce.py
```

```
Month : 01      Number of passengers from Sydney : 6190421
Month : 02      Number of passengers from Sydney : 4766537
Month : 03      Number of passengers from Sydney : 6261788
Month : 04      Number of passengers from Sydney : 4960807
Month : 05      Number of passengers from Sydney : 4928827
Month : 06      Number of passengers from Sydney : 5935277
```

```
Month : 07       Number of passengers from Sydney : 5213361
Month : 08       Number of passengers from Sydney : 5112502
Month : 09       Number of passengers from Sydney : 6590224
Month : 10       Number of passengers from Sydney : 5187097
Month : 11       Number of passengers from Sydney : 5051415
Month : 12       Number of passengers from Sydney : 7250241
```

2.2.2 Mapper et reducter avec agrégation par an et par mois

[3]:
```
!cat Reservation/data.csv | python Reservation/Reservation_mapperMonthYear.py |␣
↪python Reservation/Reservation_reduceMonthYear.py
```

```
Year : 2003 - Month : 09        Number of passengers from Sydney : 487489
Year : 2003 - Month : 12        Number of passengers from Sydney : 557005
Year : 2004 - Month : 03        Number of passengers from Sydney : 554227
Year : 2004 - Month : 06        Number of passengers from Sydney : 539183
Year : 2004 - Month : 09        Number of passengers from Sydney : 550141
Year : 2004 - Month : 12        Number of passengers from Sydney : 606719
Year : 2005 - Month : 03        Number of passengers from Sydney : 596534
Year : 2005 - Month : 06        Number of passengers from Sydney : 569057
Year : 2005 - Month : 09        Number of passengers from Sydney : 579150
Year : 2005 - Month : 12        Number of passengers from Sydney : 618713
Year : 2006 - Month : 01        Number of passengers from Sydney : 628357
Year : 2006 - Month : 02        Number of passengers from Sydney : 552457
Year : 2006 - Month : 03        Number of passengers from Sydney : 600364
Year : 2006 - Month : 04        Number of passengers from Sydney : 581205
Year : 2006 - Month : 05        Number of passengers from Sydney : 576843
Year : 2006 - Month : 06        Number of passengers from Sydney : 567687
Year : 2006 - Month : 07        Number of passengers from Sydney : 600386
Year : 2006 - Month : 08        Number of passengers from Sydney : 588643
Year : 2006 - Month : 09        Number of passengers from Sydney : 571883
Year : 2006 - Month : 10        Number of passengers from Sydney : 593800
Year : 2006 - Month : 11        Number of passengers from Sydney : 573123
Year : 2006 - Month : 12        Number of passengers from Sydney : 619897
Year : 2007 - Month : 01        Number of passengers from Sydney : 623281
Year : 2007 - Month : 02        Number of passengers from Sydney : 561693
Year : 2007 - Month : 03        Number of passengers from Sydney : 601283
Year : 2007 - Month : 04        Number of passengers from Sydney : 563075
Year : 2007 - Month : 05        Number of passengers from Sydney : 562521
Year : 2007 - Month : 06        Number of passengers from Sydney : 553939
Year : 2007 - Month : 07        Number of passengers from Sydney : 583057
Year : 2007 - Month : 08        Number of passengers from Sydney : 575999
Year : 2007 - Month : 09        Number of passengers from Sydney : 564396
Year : 2007 - Month : 10        Number of passengers from Sydney : 585538
Year : 2007 - Month : 11        Number of passengers from Sydney : 576085
Year : 2007 - Month : 12        Number of passengers from Sydney : 612876
Year : 2008 - Month : 01        Number of passengers from Sydney : 622525
```

```
Year : 2008 - Month : 02        Number of passengers from Sydney : 580886
Year : 2008 - Month : 03        Number of passengers from Sydney : 610098
Year : 2008 - Month : 04        Number of passengers from Sydney : 579828
Year : 2008 - Month : 05        Number of passengers from Sydney : 593481
Year : 2008 - Month : 06        Number of passengers from Sydney : 571365
Year : 2008 - Month : 07        Number of passengers from Sydney : 614763
Year : 2008 - Month : 08        Number of passengers from Sydney : 596635
Year : 2008 - Month : 09        Number of passengers from Sydney : 562295
Year : 2008 - Month : 10        Number of passengers from Sydney : 585773
Year : 2008 - Month : 11        Number of passengers from Sydney : 585932
Year : 2008 - Month : 12        Number of passengers from Sydney : 618514
Year : 2009 - Month : 01        Number of passengers from Sydney : 626406
Year : 2009 - Month : 02        Number of passengers from Sydney : 546582
Year : 2009 - Month : 03        Number of passengers from Sydney : 596356
Year : 2009 - Month : 04        Number of passengers from Sydney : 582837
Year : 2009 - Month : 05        Number of passengers from Sydney : 586048
Year : 2009 - Month : 06        Number of passengers from Sydney : 564161
Year : 2009 - Month : 07        Number of passengers from Sydney : 612112
Year : 2009 - Month : 08        Number of passengers from Sydney : 604721
Year : 2009 - Month : 09        Number of passengers from Sydney : 584070
Year : 2009 - Month : 10        Number of passengers from Sydney : 619346
Year : 2009 - Month : 11        Number of passengers from Sydney : 599650
Year : 2009 - Month : 12        Number of passengers from Sydney : 667769
Year : 2010 - Month : 01        Number of passengers from Sydney : 682777
Year : 2010 - Month : 02        Number of passengers from Sydney : 593296
Year : 2010 - Month : 03        Number of passengers from Sydney : 634380
Year : 2010 - Month : 04        Number of passengers from Sydney : 608030
Year : 2010 - Month : 05        Number of passengers from Sydney : 603295
Year : 2010 - Month : 06        Number of passengers from Sydney : 594696
Year : 2010 - Month : 07        Number of passengers from Sydney : 642945
Year : 2010 - Month : 08        Number of passengers from Sydney : 638477
Year : 2010 - Month : 09        Number of passengers from Sydney : 616638
Year : 2010 - Month : 10        Number of passengers from Sydney : 643887
Year : 2010 - Month : 11        Number of passengers from Sydney : 630428
Year : 2010 - Month : 12        Number of passengers from Sydney : 689319
Year : 2011 - Month : 01        Number of passengers from Sydney : 708382
Year : 2011 - Month : 02        Number of passengers from Sydney : 622079
Year : 2011 - Month : 03        Number of passengers from Sydney : 676351
Year : 2011 - Month : 04        Number of passengers from Sydney : 663109
Year : 2011 - Month : 05        Number of passengers from Sydney : 650324
Year : 2011 - Month : 06        Number of passengers from Sydney : 614870
Year : 2011 - Month : 07        Number of passengers from Sydney : 689314
Year : 2011 - Month : 08        Number of passengers from Sydney : 677257
Year : 2011 - Month : 09        Number of passengers from Sydney : 661884
Year : 2011 - Month : 10        Number of passengers from Sydney : 696907
Year : 2011 - Month : 11        Number of passengers from Sydney : 662271
Year : 2011 - Month : 12        Number of passengers from Sydney : 711800
Year : 2012 - Month : 01        Number of passengers from Sydney : 729013
```

```
Year : 2012 - Month : 02        Number of passengers from Sydney : 654120
Year : 2012 - Month : 03        Number of passengers from Sydney : 675644
Year : 2012 - Month : 04        Number of passengers from Sydney : 674681
Year : 2012 - Month : 05        Number of passengers from Sydney : 663790
Year : 2012 - Month : 06        Number of passengers from Sydney : 669329
Year : 2012 - Month : 07        Number of passengers from Sydney : 720682
Year : 2012 - Month : 08        Number of passengers from Sydney : 703989
Year : 2012 - Month : 09        Number of passengers from Sydney : 690536
Year : 2012 - Month : 10        Number of passengers from Sydney : 712018
Year : 2012 - Month : 11        Number of passengers from Sydney : 694402
Year : 2012 - Month : 12        Number of passengers from Sydney : 750584
Year : 2013 - Month : 01        Number of passengers from Sydney : 757168
Year : 2013 - Month : 02        Number of passengers from Sydney : 655424
Year : 2013 - Month : 03        Number of passengers from Sydney : 716551
Year : 2013 - Month : 04        Number of passengers from Sydney : 708042
Year : 2013 - Month : 05        Number of passengers from Sydney : 692525
Year : 2013 - Month : 06        Number of passengers from Sydney : 690990
Year : 2013 - Month : 07        Number of passengers from Sydney : 750102
Year : 2013 - Month : 08        Number of passengers from Sydney : 726781
Year : 2013 - Month : 09        Number of passengers from Sydney : 721742
Year : 2013 - Month : 10        Number of passengers from Sydney : 749828
Year : 2013 - Month : 11        Number of passengers from Sydney : 729524
Year : 2013 - Month : 12        Number of passengers from Sydney : 797045
Year : 2014 - Month : 01        Number of passengers from Sydney : 812512
```

2.3 Execution HDFS

2.3.1 Creation répertoire HDFS

[4]:
```
!hadoop fs -mkdir /Reservation
```

2.3.2 Copie du fichier data.csv sur HDFS

[5]:
```
!hadoop fs -put Reservation/data.csv /Reservation/
```

[6]:
```
!hadoop fs -ls /Reservation/
```

```
Found 1 items
-rw-r--r--   1 cloudera supergroup    6126033 2019-12-22 03:01
/Reservation/data.csv
```

2.3.3 Lancement MAP Reduce par mois

[7]:
```
!sudo hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/
 ↪hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar \
-mapper "python $(pwd)/Reservation/Reservation_mapper.py" \
-reducer "python $(pwd)/Reservation/Reservation_reduce.py" \
```

```
-input /Reservation/data.csv \
-output /Reservation/OutMonth
```

```
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-
streaming-2.6.0-cdh5.13.0.jar] /tmp/streamjob5819363448156649822.jar tmpDir=null
19/12/22 03:01:44 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
19/12/22 03:01:44 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
19/12/22 03:01:46 INFO mapred.FileInputFormat: Total input paths to process : 1
19/12/22 03:01:46 INFO mapreduce.JobSubmitter: number of splits:2
19/12/22 03:01:46 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1576777860458_0017
19/12/22 03:01:47 INFO impl.YarnClientImpl: Submitted application
application_1576777860458_0017
19/12/22 03:01:47 INFO mapreduce.Job: The url to track the job:
http://quickstart.cloudera:8088/proxy/application_1576777860458_0017/
19/12/22 03:01:47 INFO mapreduce.Job: Running job: job_1576777860458_0017
19/12/22 03:01:59 INFO mapreduce.Job: Job job_1576777860458_0017 running in uber
mode : false
19/12/22 03:01:59 INFO mapreduce.Job:  map 0% reduce 0%
19/12/22 03:02:16 INFO mapreduce.Job:  map 50% reduce 0%
19/12/22 03:02:17 INFO mapreduce.Job:  map 100% reduce 0%
19/12/22 03:02:27 INFO mapreduce.Job:  map 100% reduce 100%
19/12/22 03:02:27 INFO mapreduce.Job: Job job_1576777860458_0017 completed
successfully
19/12/22 03:02:27 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=88424
                FILE: Number of bytes written=614166
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6130337
                HDFS: Number of bytes written=660
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=28017
                Total time spent by all reduces in occupied slots (ms)=9181
                Total time spent by all map tasks (ms)=28017
                Total time spent by all reduce tasks (ms)=9181
```

```
                Total vcore-milliseconds taken by all map tasks=28017
                Total vcore-milliseconds taken by all reduce tasks=9181
                Total megabyte-milliseconds taken by all map tasks=28689408
                Total megabyte-milliseconds taken by all reduce tasks=9401344
        Map-Reduce Framework
                Map input records=61065
                Map output records=8655
                Map output bytes=71108
                Map output materialized bytes=88430
                Input split bytes=208
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=88430
                Reduce input records=8655
                Reduce output records=12
                Spilled Records=17310
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1766
                CPU time spent (ms)=8750
                Physical memory (bytes) snapshot=905039872
                Virtual memory (bytes) snapshot=4711632896
                Total committed heap usage (bytes)=696778752
        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=6130129
        File Output Format Counters
                Bytes Written=660
19/12/22 03:02:27 INFO streaming.StreamJob: Output directory:
/Reservation/OutMonth
```

2.3.4 Lancement MAP Reduce par an et par mois

[8]:
```
!sudo hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/
 ↪hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar \
-mapper "python $(pwd)/Reservation/Reservation_mapperMonthYear.py" \
-reducer "python $(pwd)/Reservation/Reservation_reduceMonthYear.py" \
-input /Reservation/data.csv \
-output /Reservation/OutMonthYear
```

```
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-
streaming-2.6.0-cdh5.13.0.jar] /tmp/streamjob8502980822403719350.jar tmpDir=null
19/12/22 03:02:33 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
19/12/22 03:02:34 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
19/12/22 03:02:35 INFO mapred.FileInputFormat: Total input paths to process : 1
19/12/22 03:02:35 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DF
SOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutpu
tStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStre
am.java:894)
19/12/22 03:02:35 INFO mapreduce.JobSubmitter: number of splits:2
19/12/22 03:02:36 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1576777860458_0018
19/12/22 03:02:36 INFO impl.YarnClientImpl: Submitted application
application_1576777860458_0018
19/12/22 03:02:36 INFO mapreduce.Job: The url to track the job:
http://quickstart.cloudera:8088/proxy/application_1576777860458_0018/
19/12/22 03:02:36 INFO mapreduce.Job: Running job: job_1576777860458_0018
19/12/22 03:02:50 INFO mapreduce.Job: Job job_1576777860458_0018 running in uber
mode : false
19/12/22 03:02:50 INFO mapreduce.Job:  map 0% reduce 0%
19/12/22 03:03:05 INFO mapreduce.Job:  map 50% reduce 0%
19/12/22 03:03:06 INFO mapreduce.Job:  map 100% reduce 0%
19/12/22 03:03:16 INFO mapreduce.Job:  map 100% reduce 100%
19/12/22 03:03:17 INFO mapreduce.Job: Job job_1576777860458_0018 completed
successfully
19/12/22 03:03:18 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=131699
                FILE: Number of bytes written=700782
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6130337
                HDFS: Number of bytes written=7383
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
```

```
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=27122
                Total time spent by all reduces in occupied slots (ms)=9330
                Total time spent by all map tasks (ms)=27122
                Total time spent by all reduce tasks (ms)=9330
                Total vcore-milliseconds taken by all map tasks=27122
                Total vcore-milliseconds taken by all reduce tasks=9330
                Total megabyte-milliseconds taken by all map tasks=27772928
                Total megabyte-milliseconds taken by all reduce tasks=9553920
        Map-Reduce Framework
                Map input records=61065
                Map output records=8655
                Map output bytes=114383
                Map output materialized bytes=131705
                Input split bytes=208
                Combine input records=0
                Combine output records=0
                Reduce input groups=107
                Reduce shuffle bytes=131705
                Reduce input records=8655
                Reduce output records=107
                Spilled Records=17310
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1007
                CPU time spent (ms)=8870
                Physical memory (bytes) snapshot=955269120
                Virtual memory (bytes) snapshot=4695879680
                Total committed heap usage (bytes)=835715072
        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=6130129
        File Output Format Counters
                Bytes Written=7383
19/12/22 03:03:18 INFO streaming.StreamJob: Output directory:
/Reservation/OutMonthYear
```

2.3.5 Traitements en cas d'exécutions répétées

On supprime les résultats des traces des traitements sauvegardés Map/Reduce d'une exécution précédente.

[9]:
```
!sudo rm $(pwd)/Reservation/OutMonthYear/*
!sudo rm $(pwd)/Reservation/OutMonth/*
```

2.3.6 Recupération du fichier de résultats en local ou visualisation

[10]:
```
!hadoop fs -copyToLocal /Reservation/OutMonth/part-00000 $(pwd)/Reservation/
↪OutMonth/
```

[11]:
```
!hadoop fs -cat /Reservation/OutMonth/part-00000
```

```
Month : 01      Number of passengers from Sydney : 6190421
Month : 02      Number of passengers from Sydney : 4766537
Month : 03      Number of passengers from Sydney : 6261788
Month : 04      Number of passengers from Sydney : 4960807
Month : 05      Number of passengers from Sydney : 4928827
Month : 06      Number of passengers from Sydney : 5935277
Month : 07      Number of passengers from Sydney : 5213361
Month : 08      Number of passengers from Sydney : 5112502
Month : 09      Number of passengers from Sydney : 6590224
Month : 10      Number of passengers from Sydney : 5187097
Month : 11      Number of passengers from Sydney : 5051415
Month : 12      Number of passengers from Sydney : 7250241
```

[12]:
```
!hadoop fs -copyToLocal /Reservation/OutMonthYear/part-00000 $(pwd)/Reservation/
↪OutMonthYear/
```

[13]:
```
!hadoop fs -cat /Reservation/OutMonthYear/part-00000
```

```
Year : 2003 - Month : 09        Number of passengers from Sydney : 487489
Year : 2003 - Month : 12        Number of passengers from Sydney : 557005
Year : 2004 - Month : 03        Number of passengers from Sydney : 554227
Year : 2004 - Month : 06        Number of passengers from Sydney : 539183
Year : 2004 - Month : 09        Number of passengers from Sydney : 550141
Year : 2004 - Month : 12        Number of passengers from Sydney : 606719
Year : 2005 - Month : 03        Number of passengers from Sydney : 596534
Year : 2005 - Month : 06        Number of passengers from Sydney : 569057
Year : 2005 - Month : 09        Number of passengers from Sydney : 579150
Year : 2005 - Month : 12        Number of passengers from Sydney : 618713
Year : 2006 - Month : 01        Number of passengers from Sydney : 628357
Year : 2006 - Month : 02        Number of passengers from Sydney : 552457
Year : 2006 - Month : 03        Number of passengers from Sydney : 600364
Year : 2006 - Month : 04        Number of passengers from Sydney : 581205
Year : 2006 - Month : 05        Number of passengers from Sydney : 576843
```

```
Year : 2006 - Month : 06        Number of passengers from Sydney : 567687
Year : 2006 - Month : 07        Number of passengers from Sydney : 600386
Year : 2006 - Month : 08        Number of passengers from Sydney : 588643
Year : 2006 - Month : 09        Number of passengers from Sydney : 571883
Year : 2006 - Month : 10        Number of passengers from Sydney : 593800
Year : 2006 - Month : 11        Number of passengers from Sydney : 573123
Year : 2006 - Month : 12        Number of passengers from Sydney : 619897
Year : 2007 - Month : 01        Number of passengers from Sydney : 623281
Year : 2007 - Month : 02        Number of passengers from Sydney : 561693
Year : 2007 - Month : 03        Number of passengers from Sydney : 601283
Year : 2007 - Month : 04        Number of passengers from Sydney : 563075
Year : 2007 - Month : 05        Number of passengers from Sydney : 562521
Year : 2007 - Month : 06        Number of passengers from Sydney : 553939
Year : 2007 - Month : 07        Number of passengers from Sydney : 583057
Year : 2007 - Month : 08        Number of passengers from Sydney : 575999
Year : 2007 - Month : 09        Number of passengers from Sydney : 564396
Year : 2007 - Month : 10        Number of passengers from Sydney : 585538
Year : 2007 - Month : 11        Number of passengers from Sydney : 576085
Year : 2007 - Month : 12        Number of passengers from Sydney : 612876
Year : 2008 - Month : 01        Number of passengers from Sydney : 622525
Year : 2008 - Month : 02        Number of passengers from Sydney : 580886
Year : 2008 - Month : 03        Number of passengers from Sydney : 610098
Year : 2008 - Month : 04        Number of passengers from Sydney : 579828
Year : 2008 - Month : 05        Number of passengers from Sydney : 593481
Year : 2008 - Month : 06        Number of passengers from Sydney : 571365
Year : 2008 - Month : 07        Number of passengers from Sydney : 614763
Year : 2008 - Month : 08        Number of passengers from Sydney : 596635
Year : 2008 - Month : 09        Number of passengers from Sydney : 562295
Year : 2008 - Month : 10        Number of passengers from Sydney : 585773
Year : 2008 - Month : 11        Number of passengers from Sydney : 585932
Year : 2008 - Month : 12        Number of passengers from Sydney : 618514
Year : 2009 - Month : 01        Number of passengers from Sydney : 626406
Year : 2009 - Month : 02        Number of passengers from Sydney : 546582
Year : 2009 - Month : 03        Number of passengers from Sydney : 596356
Year : 2009 - Month : 04        Number of passengers from Sydney : 582837
Year : 2009 - Month : 05        Number of passengers from Sydney : 586048
Year : 2009 - Month : 06        Number of passengers from Sydney : 564161
Year : 2009 - Month : 07        Number of passengers from Sydney : 612112
Year : 2009 - Month : 08        Number of passengers from Sydney : 604721
Year : 2009 - Month : 09        Number of passengers from Sydney : 584070
Year : 2009 - Month : 10        Number of passengers from Sydney : 619346
Year : 2009 - Month : 11        Number of passengers from Sydney : 599650
Year : 2009 - Month : 12        Number of passengers from Sydney : 667769
Year : 2010 - Month : 01        Number of passengers from Sydney : 682777
Year : 2010 - Month : 02        Number of passengers from Sydney : 593296
Year : 2010 - Month : 03        Number of passengers from Sydney : 634380
Year : 2010 - Month : 04        Number of passengers from Sydney : 608030
Year : 2010 - Month : 05        Number of passengers from Sydney : 603295
```

```
Year : 2010 - Month : 06        Number of passengers from Sydney : 594696
Year : 2010 - Month : 07        Number of passengers from Sydney : 642945
Year : 2010 - Month : 08        Number of passengers from Sydney : 638477
Year : 2010 - Month : 09        Number of passengers from Sydney : 616638
Year : 2010 - Month : 10        Number of passengers from Sydney : 643887
Year : 2010 - Month : 11        Number of passengers from Sydney : 630428
Year : 2010 - Month : 12        Number of passengers from Sydney : 689319
Year : 2011 - Month : 01        Number of passengers from Sydney : 708382
Year : 2011 - Month : 02        Number of passengers from Sydney : 622079
Year : 2011 - Month : 03        Number of passengers from Sydney : 676351
Year : 2011 - Month : 04        Number of passengers from Sydney : 663109
Year : 2011 - Month : 05        Number of passengers from Sydney : 650324
Year : 2011 - Month : 06        Number of passengers from Sydney : 614870
Year : 2011 - Month : 07        Number of passengers from Sydney : 689314
Year : 2011 - Month : 08        Number of passengers from Sydney : 677257
Year : 2011 - Month : 09        Number of passengers from Sydney : 661884
Year : 2011 - Month : 10        Number of passengers from Sydney : 696907
Year : 2011 - Month : 11        Number of passengers from Sydney : 662271
Year : 2011 - Month : 12        Number of passengers from Sydney : 711800
Year : 2012 - Month : 01        Number of passengers from Sydney : 729013
Year : 2012 - Month : 02        Number of passengers from Sydney : 654120
Year : 2012 - Month : 03        Number of passengers from Sydney : 675644
Year : 2012 - Month : 04        Number of passengers from Sydney : 674681
Year : 2012 - Month : 05        Number of passengers from Sydney : 663790
Year : 2012 - Month : 06        Number of passengers from Sydney : 669329
Year : 2012 - Month : 07        Number of passengers from Sydney : 720682
Year : 2012 - Month : 08        Number of passengers from Sydney : 703989
Year : 2012 - Month : 09        Number of passengers from Sydney : 690536
Year : 2012 - Month : 10        Number of passengers from Sydney : 712018
Year : 2012 - Month : 11        Number of passengers from Sydney : 694402
Year : 2012 - Month : 12        Number of passengers from Sydney : 750584
Year : 2013 - Month : 01        Number of passengers from Sydney : 757168
Year : 2013 - Month : 02        Number of passengers from Sydney : 655424
Year : 2013 - Month : 03        Number of passengers from Sydney : 716551
Year : 2013 - Month : 04        Number of passengers from Sydney : 708042
Year : 2013 - Month : 05        Number of passengers from Sydney : 692525
Year : 2013 - Month : 06        Number of passengers from Sydney : 690990
Year : 2013 - Month : 07        Number of passengers from Sydney : 750102
Year : 2013 - Month : 08        Number of passengers from Sydney : 726781
Year : 2013 - Month : 09        Number of passengers from Sydney : 721742
Year : 2013 - Month : 10        Number of passengers from Sydney : 749828
Year : 2013 - Month : 11        Number of passengers from Sydney : 729524
Year : 2013 - Month : 12        Number of passengers from Sydney : 797045
Year : 2014 - Month : 01        Number of passengers from Sydney : 812512
```

2.3.7 Supression de l'ensemble du répertoire et de son contenu

[14]:
```
!hadoop fs -rm -r /Reservation/
```

```
Deleted /Reservation
```

3 HDFS

Installation de la librairie hdfs3

[15]:
```
conda install -c conda-forge hdfs3
```

```
Collecting package metadata (current_repodata.json): done
Solving environment: done

# All requested packages already installed.


Note: you may need to restart the kernel to use updated packages.
```

3.1 Connexion à HDFS via hdfs3

[16]:
```
from hdfs3 import HDFileSystem
```

[17]:
```
hdfs = HDFileSystem(host="localhost", port=8020)
```

3.2 Création d'un répertoire sur HDFS

[18]:
```
hdfs.mkdir("/HDFS_Data/")
```

[19]:
```
hdfs.ls("/")
```

[19]:
```
['//HDFS_Data',
 '//Repertoire1',
 '//benchmarks',
 '//enterprise-deployment.json',
 '//hbase',
 '//solr',
 '//test',
 '//tmp',
 '//user',
 '//var']
```

3.3 Mettre les fichiers de données sur HDFS

[20]:
```
hdfs.put("HDFS_Data/data1.txt", "/HDFS_Data/data1.txt")
hdfs.put("HDFS_Data/data2.txt", "/HDFS_Data/data2.txt")
hdfs.put("HDFS_Data/data3.txt", "/HDFS_Data/data3.txt")
hdfs.put("HDFS_Data/data4.txt", "/HDFS_Data/data4.txt")
```

[21]:
```
hdfs.ls("/HDFS_Data/")
```

[21]:
```
['/HDFS_Data/data1.txt',
 '/HDFS_Data/data2.txt',
 '/HDFS_Data/data3.txt',
 '/HDFS_Data/data4.txt']
```

3.4 Comptabiliser le nombre de mots

[22]:
```
def count_words(file):
    word_counts = {}
    for line in file:
        for word in line.strip().split():
            if (word in word_counts) :
                word_counts[word] += 1
            else:
                word_counts[word] = 1
    return word_counts
```

Juste pour vérifier le fonctionnement de la fonction prédéfinie

[23]:
```
print(count_words(['Anthony Anthony Toto', 'Toto Titi']))
```

```
{'Anthony': 2, 'Toto': 2, 'Titi': 1}
```

On parcourt l'ensemble des fichiers hdfs, on les ouvre, on ajoute le contenu dans une liste puis on appelle la fonction pour déterminer le nombre de mots.

[24]:
```
import os
listFile = []
for file in hdfs.glob(os.path.join('/HDFS_Data', '*.txt')):
    with hdfs.open(file) as f:
        print(f.info())
        content = f.read()
        content = (str)(content)
        listFile.append(content)

resultWord = count_words(listFile)
```

```
{'kind': 'file', 'name': '/HDFS_Data/data1.txt', 'last_mod': 1577012659, 'size':
67742, 'replication': 3, 'block_size': 67108864, 'owner': 'cloudera', 'group':
```

```
'supergroup', 'permissions': 511, 'last_access': 1577012659, 'encryption_info':
None}
{'kind': 'file', 'name': '/HDFS_Data/data2.txt', 'last_mod': 1577012659, 'size':
68130, 'replication': 3, 'block_size': 67108864, 'owner': 'cloudera', 'group':
'supergroup', 'permissions': 511, 'last_access': 1577012659, 'encryption_info':
None}
{'kind': 'file', 'name': '/HDFS_Data/data3.txt', 'last_mod': 1577012659, 'size':
67933, 'replication': 3, 'block_size': 67108864, 'owner': 'cloudera', 'group':
'supergroup', 'permissions': 511, 'last_access': 1577012659, 'encryption_info':
None}
{'kind': 'file', 'name': '/HDFS_Data/data4.txt', 'last_mod': 1577012659, 'size':
67911, 'replication': 3, 'block_size': 67108864, 'owner': 'cloudera', 'group':
'supergroup', 'permissions': 511, 'last_access': 1577012659, 'encryption_info':
None}
```

[25]:
```
for key in sorted(resultWord) :
    print("%s %s"% (key, resultWord[key]))
```

```
' 4
Aenean 159
Aliquam 172
Aliquam. 1
Class 35
...
vulputate 142
vulputate, 11
vulputate. 28
vulputate.\r\n\r\nSed 1
```

[26]:
```
print("Le nombre de mots qui se repete : " + str(len(resultWord)))
```

```
Le nombre de mots qui se repete : 881
```

[27]:
```
print("Le top 10 des mots les plus repetes\n")
top10 = sorted(resultWord.items(), key=lambda k_v: k_v[1], reverse=True)[:10]
for (mot,count) in top10 :
    print("Le mot " + str(mot) + " est repete " + str(count))
```

```
Le top 10 des mots les plus repetes

Le mot et est repete 544
Le mot sit est repete 514
Le mot ac est repete 504
Le mot in est repete 477
Le mot sed est repete 452
Le mot id est repete 425
Le mot eget est repete 419
Le mot ut est repete 415
```

```
Le mot quis est repete 415
Le mot vel est repete 413
```

3.5 Suppression du répertoire HDFS_Data

[28]:
```
!hadoop fs -rm -r /HDFS_Data/
```

```
Deleted /HDFS_Data
```

[29]:
```
!hadoop fs -ls
```

```
Found 8 items
drwxr-xr-x   - cloudera cloudera          0 2019-12-12 13:39 Magasin
drwxr-xr-x   - cloudera cloudera          0 2019-12-13 02:02 SalaireMinMax
drwxr-xr-x   - cloudera cloudera          0 2019-12-12 14:46 WordCount
drwxr-xr-x   - cloudera cloudera          0 2019-12-12 13:17 countword
-rw-r--r--   1 cloudera cloudera      53655 2019-12-12 05:43 enterprise-
deployment.json
drwxr-xr-x   - cloudera cloudera          0 2019-12-17 01:51 orders
drwxr-xr-x   - cloudera cloudera          0 2019-12-12 13:00 temperature
drwxr-xr-x   - cloudera cloudera          0 2019-12-12 06:03 test
```