



))

Figure 1

Rapport de TP Map Reduce

LASFER Nisrine
Module: BIG DATA
SUJET : Création d'un programme MapReduce .

Contents

0.1	Enoncé	3
0.2	Création d'utilisateur	3
0.3	Chargement des fichiers du ML vers VM	4
0.4	Afficher le contenu du fichier	5
0.5	Le code du mapper et reducer	6
0.6	Execution des fichiers	6

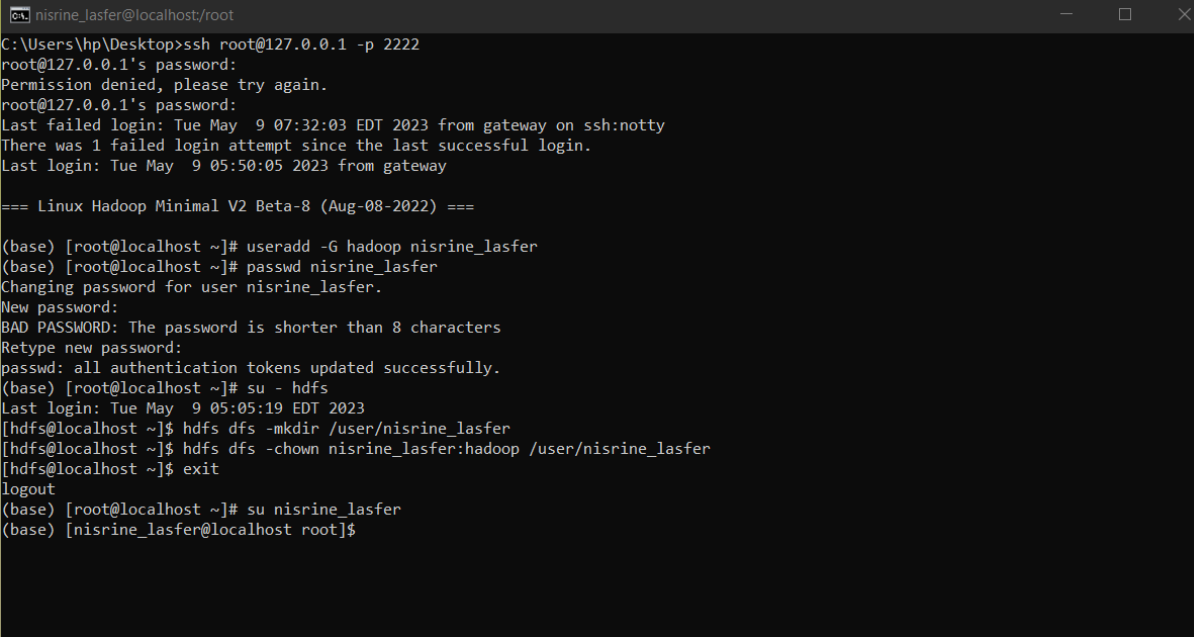
0.1 Enoncé

On considère le fichier de données "africa_covid19_daily_deaths_national.txt" (disponible dans leDrive. Ce fichier contient le nombre de décès par jour et par pays africain. Proposez un programme MapReduce qui calcule le total des décès par pays.

Exemple de sortie : Maroc, x Gabon, y... Où x et y sont le nombre de décès sur toute la période étudiée pour les pays correspondants.

0.2 Création d'utilisateur

On commence par la création d'un utilisateur nommée 'nirine_lasfer':



```
nirine_lasfer@localhost/root
C:\Users\hp\Desktop>ssh root@127.0.0.1 -p 2222
root@127.0.0.1's password:
Permission denied, please try again.
root@127.0.0.1's password:
Last failed login: Tue May  9 07:32:03 EDT 2023 from gateway on ssh:notty
There was 1 failed login attempt since the last successful login.
Last login: Tue May  9 05:50:05 2023 from gateway

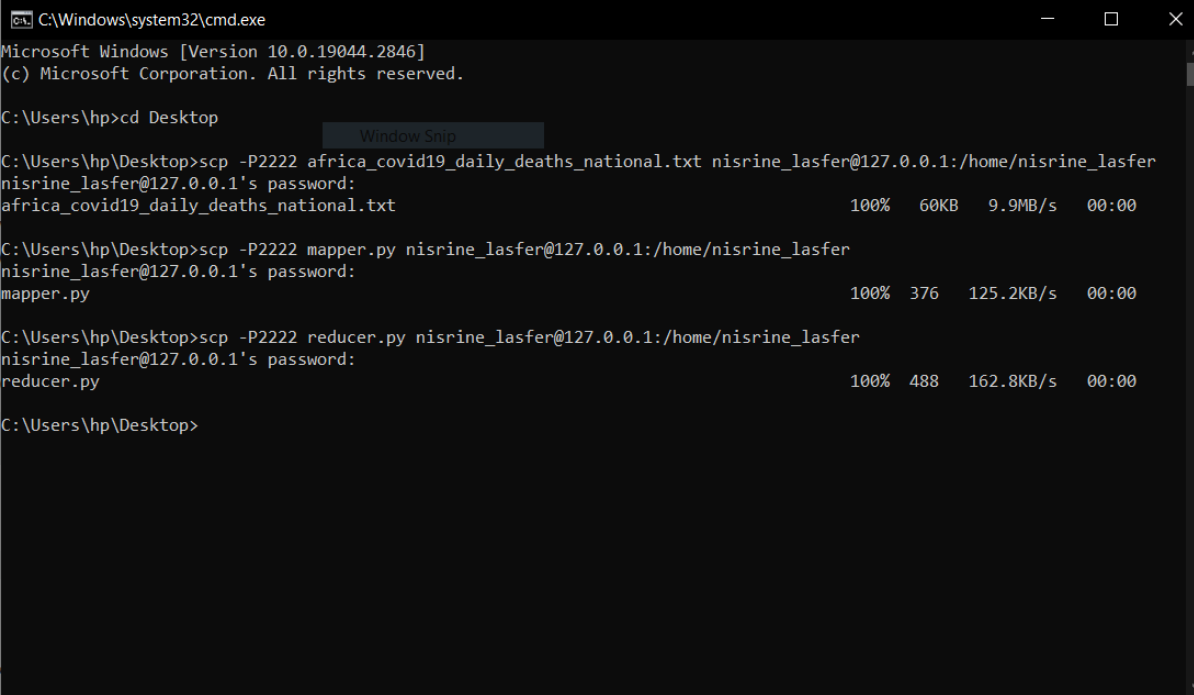
=== Linux Hadoop Minimal V2 Beta-8 (Aug-08-2022) ===

(base) [root@localhost ~]# useradd -G hadoop nirine_lasfer
(base) [root@localhost ~]# passwd nirine_lasfer
Changing password for user nirine_lasfer.
New password:
BAD PASSWORD: The password is shorter than 8 characters
Retype new password:
passwd: all authentication tokens updated successfully.
(base) [root@localhost ~]# su - hdfs
Last login: Tue May  9 05:05:19 EDT 2023
[hdfs@localhost ~]$ hdfs dfs -mkdir /user/nirine_lasfer
[hdfs@localhost ~]$ hdfs dfs -chown nirine_lasfer:hadoop /user/nirine_lasfer
[hdfs@localhost ~]$ exit
logout
(base) [root@localhost ~]# su nirine_lasfer
(base) [nirine_lasfer@localhost root]$
```

Figure 2

0.3 Chargement des fichiers du ML vers VM

On charge les fichiers du machine local vers la machine virtuelle en utilisant la commande scp comme c'est marqué dans la capture ci-dessous :



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 10.0.19044.2846]
(c) Microsoft Corporation. All rights reserved.

C:\Users\hp>cd Desktop

C:\Users\hp\Desktop>scp -P2222 africa_covid19_daily_deaths_national.txt nistrine_lasfer@127.0.0.1:/home/nistrine_lasfer
nistrine_lasfer@127.0.0.1's password:
africa_covid19_daily_deaths_national.txt                                100% 60KB   9.9MB/s   00:00

C:\Users\hp\Desktop>scp -P2222 mapper.py nistrine_lasfer@127.0.0.1:/home/nistrine_lasfer
nistrine_lasfer@127.0.0.1's password:
mapper.py                                                            100% 376    125.2KB/s  00:00

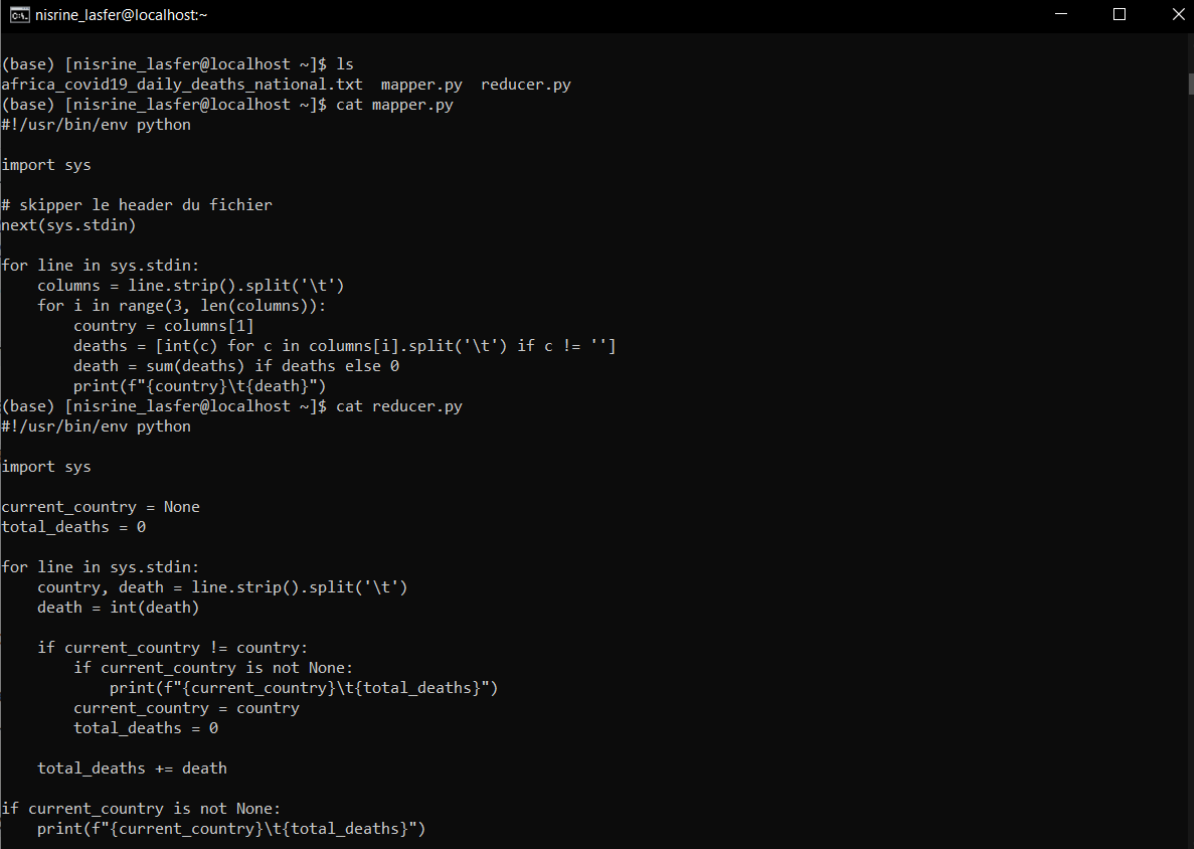
C:\Users\hp\Desktop>scp -P2222 reducer.py nistrine_lasfer@127.0.0.1:/home/nistrine_lasfer
nistrine_lasfer@127.0.0.1's password:
reducer.py                                                            100% 488    162.8KB/s  00:00

C:\Users\hp\Desktop>
```

Figure 3

0.4 Afficher le contenu du fichier

en utilisant `ls` pour afficher le contenu du répertoire et `cat` pour afficher le contenu des fichiers `mapper.py` et `reducer.py` : On note bien que le fichier est séparé par des tabulations et pas des virgules, pour ça on utilise le `split("\t")`. Le fichier `mapper` sert à définir les clé valeur, dans ce cas le clé c'est le `COUNTRY_NAME` la valeur c'est le nombre de décès pour chaque jour de ce pays . Le `reducer` prend comme entrées les valeurs de décès pour chaque pays et la somme de ces valeurs afin de les associer au pays correspondant



```
nisrine_lasfer@localhost:~$ ls
africa_covid19_daily_deaths_national.txt  mapper.py  reducer.py
(nisrine_lasfer@localhost ~)$ cat mapper.py
#!/usr/bin/env python

import sys

# skipper le header du fichier
next(sys.stdin)

for line in sys.stdin:
    columns = line.strip().split('\t')
    for i in range(3, len(columns)):
        country = columns[1]
        deaths = [int(c) for c in columns[i].split('\t') if c != '']
        death = sum(deaths) if deaths else 0
        print(f"{country}\t{death}")
(nisrine_lasfer@localhost ~)$ cat reducer.py
#!/usr/bin/env python

import sys

current_country = None
total_deaths = 0

for line in sys.stdin:
    country, death = line.strip().split('\t')
    death = int(death)

    if current_country != country:
        if current_country is not None:
            print(f"{current_country}\t{total_deaths}")
            current_country = country
            total_deaths = 0

    total_deaths += death

if current_country is not None:
    print(f"{current_country}\t{total_deaths}")
```

Figure 4

0.5 Le code du mapper et reducer

```
Mapper.py -
1 #!/usr/bin/env python
2
3 import sys
4
5 # skipper le header du fichier
6 next(sys.stdin)
7
8 for line in sys.stdin:
9     columns = line.strip().split('\t') #diviser les colonnes
10    for i in range(2, len(columns)): #boucle sur les colonnes de dates
11        country = columns[1] #diviser la colonne du pays
12        deaths = [int(c) for c in columns[i].split('\t') if c != ''] #sautez les valeurs manquantes
13        death = sum(deaths) if deaths else 0
14        print("{}{}{}\t{}\t{}\t{}".format(country, death, country, death, country, death))
15
```

(a) Mapper.py

```
Reducer.py -
1 #!/usr/bin/env python
2
3 import sys
4
5 current_country = None #initialisation du valeur contient le pays
6 total_deaths = 0 #initialisation du valeur contient le total des décès
7
8 for line in sys.stdin:
9     country, death = line.strip().split('\t') #prenant comme entrées le couple clé-valeur de chez le map
10    death = int(death)
11
12    if current_country != country:
13        if current_country is not None:
14            print("{}{}\t{}\t{}\t{}".format(current_country, total_deaths, current_country, total_deaths))
15            current_country = country
16            total_deaths = 0
17
18    total_deaths += death #on calcule la somme des décès
19
20 if current_country is not None:
21     print("{}{}\t{}\t{}\t{}".format(current_country, total_deaths, current_country, total_deaths))
22
```

(b) Reducer.py

0.6 Execution des fichiers

Pour Exécuter les fichier on utilise la commande : `cat africa_covid19_daily_deaths_national.txt |python mapper.py |sort |python reducer.py`

```
nisrine_lasfer@localhost:~$ cat africa_covid19_daily_deaths_national.txt |python mapper.py |sort |python reducer.py
Algeria 3270
Angola 602
Benin 99
Botswana 712
Burkina Faso 157
Burundi 6
Cameroon 1107
Central African Republic 88
Chad 170
Comoros 146
Congo 137
Cote d'Ivoire 287
Democratic Republic of the Congo 769
Djibouti 146
Egypt 13402
Equatorial Guinea 112
Eritrea 10
Eswatini 671
Ethiopia 3725
Gabon 140
Gambia 175
Ghana 779
Guinea 146
Guinea-Bissau 67
Kenya 2781
Lesotho 318
Liberia 85
Libya 3047
Madagascar 666
Malawi 1148
Mali 489
Mauritania 455
Mauritius 17
Mayotte 170
Morocco 9028
Mozambique 815
Namibia 649
Niger 191
Nigeria 2063
Rwanda 335
Sao Tome and Principe 35
Senegal 1111
Sierra Leone 79
Somalia 707
South Africa 54417
South Sudan 115
Sudan 2349
Tanzania 21
Togo 123
Tunisia 10868
Uganda 342
```

Figure 6