Rapport de TP – 2

Réalise par : LAABID ABDESSAMAD

Github : https://github.com/aplusInDev/hadoop_tps/tree/main/tp2

Introduction

Ce TP vise à implémenter deux jobs MapReduce sous Hadoop pour analyser des données météorologiques. Les objectifs sont :

1. **Calculer la température maximale par année** à partir d'un fichier structuré (`jour:mois:année:température:ville`).
2. **Compter le nombre de mois distincts** ayant enregistré une température supérieure à un seuil donné.

Les scripts Python (`mapper_1.py`, `reducer_1.py`, `mapper_2.py`, `reducer_2.py`) et les scripts shell (`apply_1.sh`, `apply_2.sh`) sont conçus pour fonctionner avec Hadoop Streaming.

Objectifs

- Maîtriser l'écriture de mappers et reducers en Python pour Hadoop.
- Manipuler des données structurées avec MapReduce.
- Exploiter Hadoop Streaming pour exécuter des jobs distribués.

Méthodologie

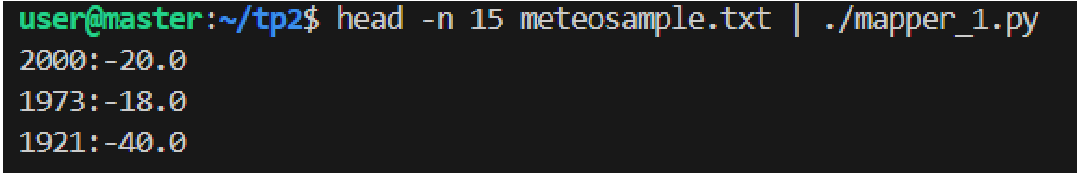
1. Température Maximale par Année

Fonctionnement :

- **Mapper (`mapper_1.py`)** :
    - Lit chaque ligne d'entrée.
    - Extrait l'année et la température.
    - Émet des paires `<année>:<température>`.

```python
#!/usr/bin/env python3
""" Mapper module for processing weather data and finding the maximum
temperature per year."""
import sys


def mapper():
    """ Mapper function to read input from stdin and output year and max
temperature """
    for line in sys.stdin:

        line = line.strip()
        fields = line.split(':')
        if len(fields) < 4:
            continue
        year = fields[2]
        temperature = fields[3]
        try:
            year = int(year)
            temperature = float(temperature)
            print(f"{year}\t{temperature}")
        except ValueError:
            continue


if __name__ == "__main__":
    mapper()
```

`# Exemple de sortie du mapper_1.py`



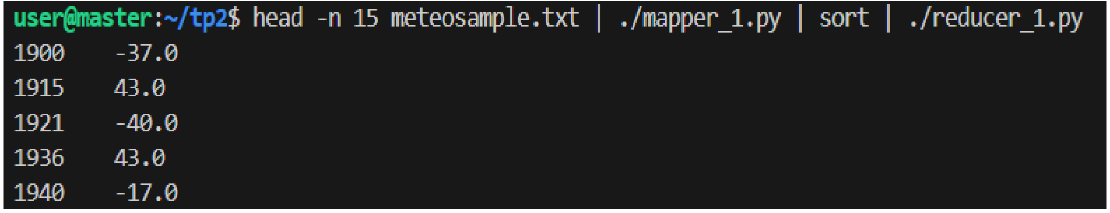
```
user@master:~/tp2$ head -n 15 meteosample.txt | ./mapper_1.py
2000:-20.0
1973:-18.0
1921:-40.0
```

- **Reducer (`reducer_1.py`)** :

- Agrège les températures par année.
- Garde la valeur maximale pour chaque année.

```
#!/usr/bin/env python3
""" Reducer module for processing weather data and finding the maximum
temperature per year."""
import sys


def reducer():
    """ Reducer function to read input from stdin and output year and max
temperature """
    current_year = None
    max_temp = -float('inf')  # Initialize to negative infinity

    for line in sys.stdin:
        line = line.strip()
        key, value = line.split('\t', 1)
        year = int(key)
        temperature = float(value)

        if year == current_year:
            if temperature > max_temp:
                max_temp = temperature
        else:
            if current_year is not None:
                print(f"{current_year}\t{max_temp}")
            current_year = year
            max_temp = temperature

    if current_year is not None:
        print(f"{current_year}\t{max_temp}")

if __name__ == "__main__":
    reducer()
```

```
# Exemple de sortie du reducer_1.py
```



```
user@master:~/tp2$ head -n 15 meteosample.txt | ./mapper_1.py | sort | ./reducer_1.py
1900    -37.0
1915    43.0
1921    -40.0
1936    43.0
1940    -17.0
```

Commande d’exécution (`apply_1.sh`):

```
hdfs dfs -mkdir -p /data/tp2/input_1
hdfs dfs -copyFromLocal hadoop_tps/tp2/meteosample.txt /data/tp2/input_1/

cleanup() {
    echo "Cleaning up HDFS directories..."
    hdfs dfs -rm -r /data/tp2/ /output_1 2>/dev/null
}

trap cleanup ERR INT TERM EXIT

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
  -files hadoop_tps/tp2/mapper_1.py,hadoop_tps/tp2/reducer_1.py \
  -mapper "python3 mapper_1.py" \
  -reducer "python3 reducer_1.py" \
  -input /data/tp2/input_1/* \
  -output /output_1

echo "Results: ------------------";
hdfs dfs -cat /output_1/part-*
```

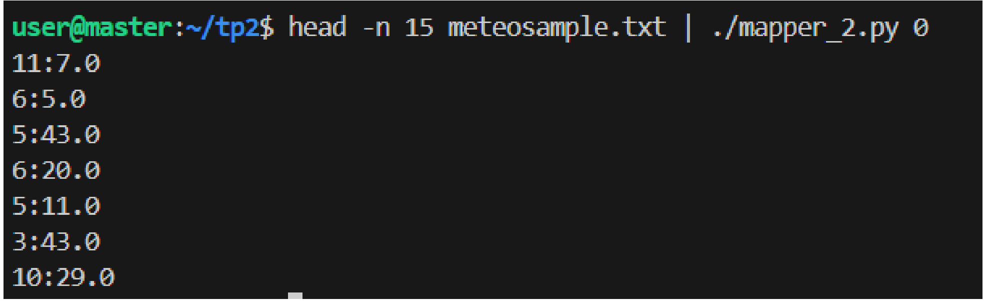
2. Comptage des Mois avec Température > Seuil

Fonctionnement :

- **Mapper (`mapper_2.py`)** :
    - Prend un seuil en argument (ex : `0`).
    - Filtre les lignes où `température > seuil`.
    - Émet des paires `<mois>:<température>`.

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


# Check if the argument is a valid integer
if len(sys.argv) != 2:
    print("Usage: python mapper_2.py <temperature>")
    sys.exit(1)

try:
    temperature_argument = int(sys.argv[1])
except ValueError:
    print("Error: Argument must be an integer.")
    sys.exit(1)


def mapper():
    for line in sys.stdin:
        line = line.strip()
        fields = line.split(':')
        if len(fields) != 5:
            continue
        _, month, _, temperature_str, _ = fields
        try:
            month = int(month)
            temperature = float(temperature_str)
            if temperature > temperature_argument:
                print(f"{month}\t{temperature}")
        except ValueError:
            continue

if __name__ == "__main__":
    mapper()
```

```
# Exemple de sortie du mapper_2.py (seuil=0)
```

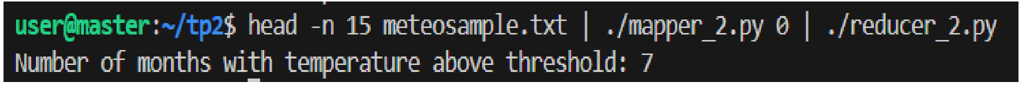


```
user@master:~/tp2$ head -n 15 meteosample.txt | ./mapper_2.py 0
11:7.0
6:5.0
5:43.0
6:20.0
5:11.0
3:43.0
10:29.0
```

- **Reducer (`reducer_2.py`)** :
    - Compte le nombre de mois **distincts** ayant dépassé le seuil.

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

def reducer():
    unique_months = set()

    for line in sys.stdin:
        line = line.strip()
        parts = line.split('\t')
        if len(parts) != 2:
            continue
        month_str, _ = parts
        try:
            month = int(month_str)
            unique_months.add(month)
        except ValueError:
            continue


    print(f"Number of months with temperature above threshold: {len(unique_months)}")

if __name__ == "__main__":
    reducer()
```

```
# Exemple de sortie du reducer_2.py
```



```
user@master:~/tp2$ head -n 15 meteosample.txt | ./mapper_2.py 0 | ./reducer_2.py
Number of months with temperature above threshold: 7
```

Commande d’exécution (`apply_2.sh`**):**

```
if [$# -eq 0]; then
    echo "Usage: $0 <threshold>"
    exit 1
fi

THRESHOLD=$1

hdfs dfs -mkdir -p /data/tp2/input_2
hdfs dfs -copyFromLocal hadoop_tps/tp2/meteosample.txt /data/tp2/input_2/

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
    -files hadoop_tps/tp2/mapper_2.py,hadoop_tps/tp2/reducer_2.py \
    -mapper "python3 mapper_2.py $THRESHOLD" \
    -reducer "python3 reducer_2.py" \
    -input /data/tp2/input_2/* \
    -output /output_2

echo "Results for Months with Temperature > $THRESHOLD: ------------------";
hdfs dfs -cat /output_2/part-*

hdfs dfs -rm -r /output_2
hdfs dfs -rm -r /data/tp2/
```

Résultats et Analyse

Résultats Attendus

1. **Job 1** : Un fichier listant chaque année avec sa température maximale.
2. **Job 2** : Un nombre indiquant combien de mois ont dépassé le seuil.

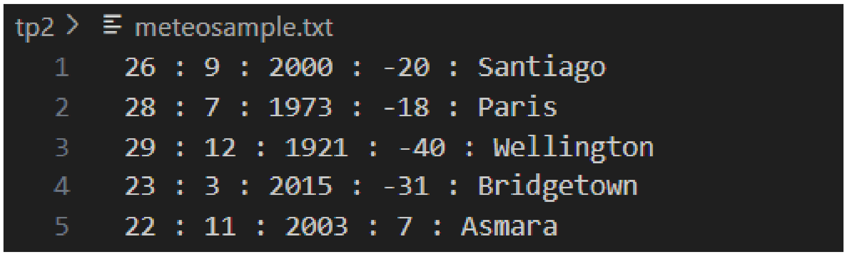
Conclusion

Ce TP a permis de :

- Pratiquer l’écriture de mappers et reducers en Python pour Hadoop.
- Manipuler des données structurées avec MapReduce.
- Identifier des erreurs courantes (ex : gestion des types de données, logique de comptage).

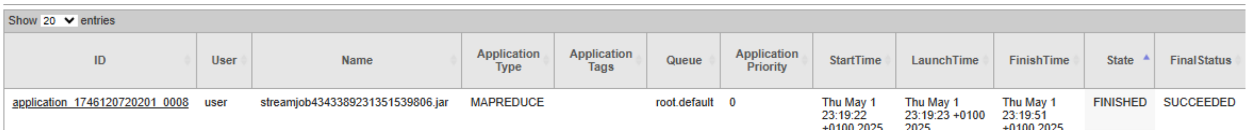
Annexes

Exemple de Données d’Entrée (`meteosample.txt`)



```
26 : 9 : 2000 : -20 : Santiago
28 : 7 : 1973 : -18 : Paris
29 : 12 : 1921 : -40 : Wellington
23 : 3 : 2015 : -31 : Bridgetown
22 : 11 : 2003 : 7 : Asmara
```

Sortie du Job 1



Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
application_1746120720201_0008	user	streamjob4343389231351539806.jar	MAPREDUCE		root.default	0	Thu May 1 23:19:22 +0100 2025	Thu May 1 23:19:23 +0100 2025	Thu May 1 23:19:51 +0100 2025	FINISHED	SUCCEEDED

```
user@master:~/tp2$ bash apply_1.sh

2025-05-01 22:23:13,742 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/user/tp2/mapper_1.py, /home/user/tp2/reducer_1.py, /tmp/hadoop-unjar6035087189545399534/] [] /tmp/streamjob5497562387453498258.jar tmpDir=null

2025-05-01 22:23:15,150 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-05-01 22:23:15,407 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-05-01 22:23:15,899 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/user/.staging/job_1746120720201_0009

2025-05-01 22:23:16,663 INFO mapred.FileInputFormat: Total input files to process : 1

2025-05-01 22:23:16,811 INFO mapreduce.JobSubmitter: number of splits:2

2025-05-01 22:23:17,258 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746120720201_0009

2025-05-01 22:23:17,259 INFO mapreduce.JobSubmitter: Executing with tokens: []

2025-05-01 22:23:17,515 INFO conf.Configuration: resource-types.xml not found

2025-05-01 22:23:17,517 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2025-05-01 22:23:17,655 INFO impl.YarnClientImpl: Submitted application application_1746120720201_0009

2025-05-01 22:23:17,704 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1746120720201_0009/

2025-05-01 22:23:17,708 INFO mapreduce.Job: Running job: job_1746120720201_0009

2025-05-01 22:23:27,419 INFO mapreduce.Job: Job job_1746120720201_0009 running in uber mode : false

2025-05-01 22:23:27,421 INFO mapreduce.Job:  map 0% reduce 0%
```

```
2025-05-01 22:23:37,950 INFO mapreduce.Job:  map 100% reduce 0%

2025-05-01 22:23:44,090 INFO mapreduce.Job:  map 100% reduce 100%

2025-05-01 22:23:46,234 INFO mapreduce.Job: Job job_1746120720201_0009 completed successfully

2025-05-01 22:23:46,412 INFO mapreduce.Job: Counters: 54

        File System Counters

                FILE: Number of bytes read=1337

                FILE: Number of bytes written=938841

                FILE: Number of read operations=0

                FILE: Number of large read operations=0

                FILE: Number of write operations=0

                HDFS: Number of bytes read=4722

                HDFS: Number of bytes written=634

                HDFS: Number of read operations=11

                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=2

                Launched reduce tasks=1

                Data-local map tasks=2

                Total time spent by all maps in occupied slots (ms)=15472

                Total time spent by all reduces in occupied slots (ms)=4298

                Total time spent by all map tasks (ms)=15472

                Total time spent by all reduce tasks (ms)=4298
        Total vcore-milliseconds taken by all map tasks=15472
        Total vcore-milliseconds taken by all reduce tasks=4298
        Total megabyte-milliseconds taken by all map tasks=15843328
        Total megabyte-milliseconds taken by all reduce tasks=4401152
    Map-Reduce Framework
        Map input records=100
        Map output records=100
        Map output bytes=1131
        Map output materialized bytes=1343
        Input split bytes=208
        Combine input records=0
        Combine output records=0
        Reduce input groups=99
        Reduce shuffle bytes=1343
        Reduce input records=100
        Reduce output records=62
        Spilled Records=200
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=272
        CPU time spent (ms)=2890
        Physical memory (bytes) snapshot=725188608
        Virtual memory (bytes) snapshot=8150073344
```

```
		Total committed heap usage (bytes)=476053504

		Peak Map Physical memory (bytes)=265273344

		Peak Map Virtual memory (bytes)=2714906624

		Peak Reduce Physical memory (bytes)=205529088

		Peak Reduce Virtual memory (bytes)=2720739328

	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=4514

	File Output Format Counters

		Bytes Written=634

2025-05-01 22:23:46,414 INFO streaming.StreamJob: Output directory: /output_1

1900	-2.0

1903	-41.0

1907	-42.0
```

Sortie du Job 2 (seuil=0)

```
user@master:~/tp2$ bash apply_2.sh

2025-05-01 22:27:37,387 WARN streaming.StreamJob: -file option is deprecated, please use
generic option -files instead.
```

```
packageJobJar: [/home/user/tp2/mapper_2.py, /home/user/tp2/reducer_2.py, /tmp/hadoop-unjar14856059120974465139/] [] /tmp/streamjob7114440649114425558.jar tmpDir=null

2025-05-01 22:27:38,801 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-05-01 22:27:39,013 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

2025-05-01 22:27:39,484 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/user/.staging/job_1746120720201_0010

2025-05-01 22:27:40,202 INFO mapred.FileInputFormat: Total input files to process : 1

2025-05-01 22:27:40,369 INFO mapreduce.JobSubmitter: number of splits:2

2025-05-01 22:27:41,192 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746120720201_0010

2025-05-01 22:27:41,193 INFO mapreduce.JobSubmitter: Executing with tokens: []

2025-05-01 22:27:41,547 INFO conf.Configuration: resource-types.xml not found

2025-05-01 22:27:41,549 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2025-05-01 22:27:41,675 INFO impl.YarnClientImpl: Submitted application application_1746120720201_0010

2025-05-01 22:27:41,745 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1746120720201_0010/

2025-05-01 22:27:41,751 INFO mapreduce.Job: Running job: job_1746120720201_0010

2025-05-01 22:27:51,249 INFO mapreduce.Job: Job job_1746120720201_0010 running in uber mode : false

2025-05-01 22:27:51,260 INFO mapreduce.Job:  map 0% reduce 0%

2025-05-01 22:28:01,780 INFO mapreduce.Job:  map 100% reduce 0%

2025-05-01 22:28:08,986 INFO mapreduce.Job:  map 100% reduce 100%

2025-05-01 22:28:11,091 INFO mapreduce.Job: Job job_1746120720201_0010 completed successfully

2025-05-01 22:28:11,323 INFO mapreduce.Job: Counters: 54

    File System Counters

        FILE: Number of bytes read=482

        FILE: Number of bytes written=937137

        FILE: Number of read operations=0

        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=4722
        HDFS: Number of bytes written=55
        HDFS: Number of read operations=11
        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=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=17163
        Total time spent by all reduces in occupied slots (ms)=4897
        Total time spent by all map tasks (ms)=17163
        Total time spent by all reduce tasks (ms)=4897
        Total vcore-milliseconds taken by all map tasks=17163
        Total vcore-milliseconds taken by all reduce tasks=4897
        Total megabyte-milliseconds taken by all map tasks=17574912
        Total megabyte-milliseconds taken by all reduce tasks=5014528
    Map-Reduce Framework
        Map input records=100
        Map output records=48
        Map output bytes=380
        Map output materialized bytes=488
        Input split bytes=208
        Combine input records=0
        Combine output records=0
        Reduce input groups=46
        Reduce shuffle bytes=488
        Reduce input records=48
        Reduce output records=1
        Spilled Records=96
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=343
        CPU time spent (ms)=3570
        Physical memory (bytes) snapshot=795410432
        Virtual memory (bytes) snapshot=8167067648
        Total committed heap usage (bytes)=665845760
        Peak Map Physical memory (bytes)=319426560
        Peak Map Virtual memory (bytes)=2722783232
        Peak Reduce Physical memory (bytes)=211435520
        Peak Reduce Virtual memory (bytes)=2726576128
    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=4514
    File Output Format Counters
        Bytes Written=55
2025-05-01 22:28:11,325 INFO streaming.StreamJob: Output directory: /output_2
Number of months with temperature above threshold: 12
Deleted /output_2
Deleted /data/tp2
```