

```
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> docker-compose up -d
[+] Running 2/2
 ✓ Container postgres_traffic Started 0.5s
 ✓ Container hadoop_cluster Started 0.6s
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project>
```

## 2-Download Data:

```
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> .\scripts\download-data.ps1
Downloading Paris Traffic Data...
Source: Open Data Paris - Comptages routiers permanents

Downloading from Paris Open Data API...

Download successful!
File: .\data\paris_traffic_data.csv
Size: 31.63 MB

Next step: Load data into PostgreSQL
Run: python .\data-loader\load_paris_traffic.py .\data\paris_traffic_data.csv
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project>
```

## 3- Load Data into PostgreSQL:

```
Next step: Load data into PostgreSQL
cd data-loader
>> pip install -r requirements.txt
>> python load_paris_traffic.py ../data/paris_traffic_data.csv
WARNING: Ignoring invalid distribution -orch (c:\users\ramib\appdata\local\programs\python\python310\lib\site-packages)
Collecting pandas==2.1.4 (from -r requirements.txt (line 1))
  100000 rows - Done! Loaded 100000 rows
```

## 4-Install Sqoop

```
docker exec -it hadoop_cluster bash
/sqoop-scripts/install-sqoop.sh
```

## 5-Transfer to hdfs

```

PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> docker exec -it hadoop_cluster bash /sqoop-scripts/import-traffic-to-hdfs.sh
>>
Starting namenodes on [hadoop]
hadoop: WARNING: HADOOP_NAMENODE_OPTS has been replaced by HDFS_NAMENODE_OPTS. Using value of HADOOP_NAMENODE_OPTS.
hadoop: namenode is running as process 6478. Stop it first and ensure /tmp/hadoop-root-namenode.pid file is empty before retry.
Starting datanodes
WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop: WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop: WARNING: HADOOP_DATANODE_OPTS has been replaced by HDFS_DATANODE_OPTS. Using value of HADOOP_DATANODE_OPTS.
hadoop: datanode is running as process 6603. Stop it first and ensure /tmp/hadoop-root-datanode.pid file is empty before retry.
Starting secondary namenodes [hadoop]
hadoop: WARNING: HADOOP_SECONDARYNAMENODE_OPTS has been replaced by HDFS_SECONDARYNAMENODE_OPTS. Using value of HADOOP_SECONDARYNAMENODE_OPTS.
hadoop: secondarynamenode is running as process 6834. Stop it first and ensure /tmp/hadoop-root-secondarynamenode.pid file is empty before retry.
WARNING: YARN_CONF_DIR has been replaced by HADOOP_CONF_DIR. Using value of YARN_CONF_DIR.
Starting resourcemanager
Starting nodemanagers
hadoop: nodemanager is running as process 7207. Stop it first and ensure /tmp/hadoop-root-nodemanager.pid file is empty before retry.
8577 Jps
6834 SecondaryNameNode
8210 ResourceManager
7207 NodeManager
6603 DataNode
6478 NameNode
Warning: /opt/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/../accumulo does not exist! Accumulo imports will fail.

```

```

Total megabyte-milliseconds taken by all map tasks=4548608
Map-Reduce Framework
  Map input records=100000
  Map output records=100000
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=65
  CPU time spent (ms)=4730
  Physical memory (bytes) snapshot=403914752
  Virtual memory (bytes) snapshot=2586849280
  Total committed heap usage (bytes)=329777152
  Peak Map Physical memory (bytes)=403914752
  Peak Map Virtual memory (bytes)=2586849280
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=19885548
2025-11-15 20:05:45,856 INFO mapreduce.ImportJobBase: Transferred 18.9643 MB in 25.3309 seconds (766.6309 KB/sec)
2025-11-15 20:05:45,860 INFO mapreduce.ImportJobBase: Retrieved 100000 records.
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-11-15 20:05 /user/bigdata/traffic/_SUCCESS
-rw-r--r-- 1 root supergroup 19885548 2025-11-15 20:05 /user/bigdata/traffic/part-m-00000

```

## Verification:

```

PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> docker exec -it hadoop_cluster hdfs dfs -ls /user/bigdata/traffic/
>>
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-11-15 20:05 /user/bigdata/traffic/part-m-00000
0 | head -10 1 root supergroup 19885548 2025-11-15 20:05 /user/bigdata/traffic/part-m-00000
>> C:\Users\ramib\OneDrive\Bureau\Big-Data-Project>

```

## Puis:

```

docker exec -it hadoop_cluster hdfs dfs -cat
/user/bigdata/traffic/part-m-00000 | head -10

```

```
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> docker exec -it hadoop_cluster bash -c "hdfs dfs -cat /user/bigdata/traffic/pa
rt-m-00000 | head -10"
>>
1,5325,PE Reuilly,2024-10-04 19:00:00.0,5955.00,18.70000,Pré-saturé,528,SE Charenton,2818,PE Daumesnil,Ouvert,1996-10-28,2023-01-0
1,48.83194119,2.40515737,2025-11-15 19:08:22.194938
2,829,Bd Gal de Gaulle (le Kremlin),2024-11-03 13:00:00.0,null,24.60056,Pré-saturé,467,A6b_Elisee_Reclus,475,Pte_d'Italie-PE_RN7,I
nvalide,1996-10-03,2023-01-01,48.81588802,2.35872760,2025-11-15 19:08:22.194938
3,4010,Av_de_Clichy,2024-12-04 13:00:00.0,223.00,13.72667,Fluide,2037,Av_de_Clichy-Moquet,2029,Av_de_Clichy-Bodin,Invalide,1996-10
1,48.83194119,2.40515737,2025-11-15 19:08:22.194938
2,829,Bd Gal de Gaulle (le Kremlin),2024-11-03 13:00:00.0,null,24.60056,Pré-saturé,467,A6b_Elisee_Reclus,475,Pte_d'Italie-PE_RN7,I
nvalide,1996-10-03,2023-01-01,48.81588802,2.35872760,2025-11-15 19:08:22.194938
3,4010,Av_de_Clichy,2024-12-04 13:00:00.0,223.00,13.72667,Fluide,2037,Av_de_Clichy-Moquet,2029,Av_de_Clichy-Bodin,Invalide,1996-10
3,4010,Av_de_Clichy,2024-12-04 13:00:00.0,223.00,13.72667,Fluide,2037,Av_de_Clichy-Moquet,2029,Av_de_Clichy-Bodin,Invalide,1996-10
-03,2023-01-01,48.89143425,2.31880792,2025-11-15 19:08:22.194938
4,1067,Quai d'Issy,2025-01-11 18:00:00.0,769.00,4.43000,Fluide,560,Quai Andre Citroen-am_Victor,593,Quai_Issy_Y-Echangeur_Gariglia
no,Invalide,1996-10-03,2023-01-01,48.83972716,2.27020334,2025-11-15 19:08:22.194938
5,5460,SI_Passy,2025-02-18 21:00:00.0,null,null,Inconnu,2856,SI_Passy,2864,Lacs_a_Passy-Fortifications-BP,Ouvert,1996-11-06,2023-0
4,1067,Quai d'Issy,2025-01-11 18:00:00.0,769.00,4.43000,Fluide,560,Quai Andre Citroen-am_Victor,593,Quai_Issy_Y-Echangeur_Gariglia
no,Invalide,1996-10-03,2023-01-01,48.83972716,2.27020334,2025-11-15 19:08:22.194938
5,5460,SI_Passy,2025-02-18 21:00:00.0,null,null,Inconnu,2856,SI_Passy,2864,Lacs_a_Passy-Fortifications-BP,Ouvert,1996-11-06,2023-0
no,Invalide,1996-10-03,2023-01-01,48.83972716,2.27020334,2025-11-15 19:08:22.194938
5,5460,SI_Passy,2025-02-18 21:00:00.0,null,null,Inconnu,2856,SI_Passy,2864,Lacs_a_Passy-Fortifications-BP,Ouvert,1996-11-06,2023-0
5,5460,SI_Passy,2025-02-18 21:00:00.0,null,null,Inconnu,2856,SI_Passy,2864,Lacs_a_Passy-Fortifications-BP,Ouvert,1996-11-06,2023-0
1-01,48.85791852,2.26121221,2025-11-15 19:08:22.194938
6,1505,Bd Serurier,2025-03-01 15:00:00.0,null,13.19611,Fluide,801,Bd Serurier-Zenith,800,Pte_Pantin-Av_Jean_Jaures,Ouvert,1996-10-
03,2023-01-01,48.89014230,2.39535847,2025-11-15 19:08:22.194938
7,5935,Bd_de_Clichy,2025-04-01 10:00:00.0,null,19.32444,Pré-saturé,3083,Clichy-Fromentin,3082,Place_Blanche,Ouvert,1997-09-23,2023
-01-01,48.88355608,2.33384206,2025-11-15 19:08:22.194938
8,4621,Souterrain Invalides_VGP,2025-05-06 17:00:00.0,1895.00,7.09278,Fluide,2467,Cours_Albert_Ier-VGP,163,Souterrain_Cours_la_Rei
ne,Ouvert,1996-10-16,2023-01-01,48.86459641,2.30529562,2025-11-15 19:08:22.194938
9,1,Quai du Louvre,2025-06-04 21:00:00.0,null,null,Inconnu,1,Pt_Carrousel_Droite,0,Quai du Louvre (Musee),Invalide,1996-10-03,2023
-01-01,48.85983846,2.33424182,2025-11-15 19:08:22.194938
10,5167,Georges_Lafont,2025-07-03 17:00:00.0,null,null,Inconnu,2729,Pte_St_Cloud-Av_Georges_Lafont,2496,Pte_St_Cloud,Invalide,2005
-01-01,2019-06-01,48.83662295,2.25604286,2025-11-15 19:08:22.194938
cat: Unable to write to output stream.
PS C:\Users\ramib\OneDrive\Bureau\Big-Data-Project> █
```

## Overview 'hadoop:9000' (active)

← ↻ 🔍 localhost:9870/dfshealth.html#tab-overview

Non Heap Memory used 56.31 MB of 57.75 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1006.85 GB
Configured Remote Capacity:	0 B
DFS Used:	19.45 MB (0%)
Non DFS Used:	25.01 GB
DFS Remaining:	930.61 GB (92.43%)
Block Pool Used:	19.45 MB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Sat Nov 15 21:01:22 +0100 2025
Last Checkpoint Time	Sat Nov 15 21:01:17 +0100 2025
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 336

Journal Manager	State
FileJournalManager(root=/root/hdfs/namenode)	EditLogFileOutputStream(/root/hdfs/namenode/current/edits_inprogress_000000000000000336)

## NameNode Storage

Storage Directory	Type	State
/root/hdfs/namenode	IMAGE_AND_EDITS	Active

## DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	1006.85 GB	19.45 MB (0%)	930.61 GB (92.43%)	19.45 MB	1