Ministry of Higher Education and Scientific Research
University Center of Mila
Institute of Mathematics and Computer Science
Department of Computer Science
Master 2 I2A – Big Data
2025/2026

Practical Works2 – Introduction to HDFS and Hadoop Commands

Section A – Exploring the Hadoop Environment and Commands

Objective:

Get familiar with the Hadoop environment, verify that all necessary tools are working, and learn to interact with HDFS using essential commands.

1. Verify Services:

- Use ssh: **http://localhost:4200**
- Use login : root
- Use password: hadoop
- Check that services like HDFS, YARN, MapReduce2, Hive, and Spark are running.
- Or verify using the terminal:

```
In BASH:
jps
```

You should see daemons like **NameNode**, **DataNode**, **ResourceManager**, **NodeManager**, etc.

2. Explore HDFS

```
In BASH:
hdfs dfs -ls /
```

3. Create a Personal Directory

```
In BASH:
sudo -u hdfs hdfs dfs -mkdir /user/<your_name>

sudo -u hdfs hdfs dfs -chown  /<your_name>:/<your_name>  /user/<your_name>

hdfs dfs -ls /user
```

4. Basic File Operations

```
In BASH:
echo "Hadoop lab test" > test.txt

Sudo useradd <yourname>

Sudo passwd <yourname>

hdfs dfs -put test.txt /user/<your_name>/

hdfs dfs -cat /user/<your_name>/test.txt
```

- Try copying back and deleting:

```
In BASH:
hdfs dfs -get /user/<your_name>/test.txt .

hdfs dfs -rm /user/<your_name>/test.txt
```

5. Check File Details

```
In BASH:
hdfs fsck /user/<your_name>/test.txt -files -blocks -locations
```

6. Mini Challenge

- Create /user/<your_name>/practice/ and upload multiple files.
- Display them recursively using:

```
In BASH:
hdfs dfs -ls -R /user/<your_name>/
```

- Use -du -h to check total size.

Section 2 – Use Case: Managing Sales Data for a Supermarket

Objective:

Use HDFS to organize and manage a large dataset of supermarket sales.

Students simulate a small data engineering task: storing, viewing, and combining sales records.

1. Create the Folder Structure:

```
In BASH:
hdfs dfs -mkdir -p /data/sales/{2025-01,2025-02,2025-03}
```

You should see daemons like **NameNode**, **DataNode**, **ResourceManager**, **NodeManager**, etc.

2. Generate Example Sales Files

Locally create three small CSV files:

```
In yaml:
date,store,product,quantity,price

2025-01-05,Algiers,Milk,30,120

2025-01-05,Constantine,Bread,100,20

2025-01-06,Oran,Butter,15,250
```

Save as:

- sales_2025-01.csv
- sales_2025-02.csv
- sales_2025-03.csv

3. Upload to HDFS

```
In BASH:
hdfs dfs -put sales_2025-01.csv /data/sales/2025-01/

hdfs dfs -put sales_2025-02.csv /data/sales/2025-02/

hdfs dfs -put sales_2025-03.csv /data/sales/2025-03/
```

4. Check Distribution & Replication

```
In BASH:
hdfs dfs -du -h /data/sales

hdfs fsck /data/sales/2025-01/sales_2025-01.csv -blocks -locations
```

5. Read All Data Together

Combine all CSVs

```
In BASH:
hdfs dfs -cat /data/sales/*/*.csv | head
```

6. Mini Data Query Challenge

Using command-line tools, answer:

- How many total transactions occurred?

```
hdfs dfs -cat /data/sales/*/*.csv | wc -l
```

- How many times was “Milk” sold?

```
hdfs dfs -cat /data/sales/*/*.csv | grep -c "Milk"
```

7. Deliverables

- Screenshot of folder hierarchy (hdfs dfs -ls -R /data/sales)
- File sizes and replication results
- Command outputs from the mini queries