Lab 01

HDFS Fundamentals

Hands-On Exploration of Distributed Storage

Course: SE446 - Big Data Engineering

Week: 2 • **Duration:** 60 minutes • **Difficulty:** Beginner

جامعة الفيصل
Alfaisal University

Prof. Anis Koubaa

Prince Sultan University • Alfaisal University

Learning Objectives

By the end of this lab, you will be able to:

LO1. **Connect** to a remote HDFS cluster using SSH
LO2. **Explore** the HDFS cluster architecture (NameNode, DataNodes)
LO3. **Navigate** the HDFS filesystem using basic commands
LO4. **Upload** files to HDFS and observe data distribution
LO5. **Understand** HDFS blocks, replication, and metadata

Prerequisites

- You have received your cluster credentials via email
- Terminal access (Mac/Linux) or PowerShell (Windows)
- Basic knowledge of Linux commands

1 Connecting to the Cluster

1.1 SSH Login

Open your terminal and connect to the cluster using the following command:

```
ssh <your_username>@134.209.172.50
```

Example:

Login Example If your username is `akoubaa`, use:

```
ssh akoubaa@134.209.172.50
```

Enter your password when prompted (from your credentials email).

Success Check

You should see a welcome message and a command prompt like: `akoubaa@master-node:~$`

Warning:

Critical: Two Separate Filesystems! **IMPORTANT:** The cluster has TWO different filesystems:

1. **Linux Filesystem (Local):**
   - Path: `/home/akoubaa`
   - Commands: `cd`, `ls`, `cat`, `pwd`
   - This is your SSH login environment
2. **HDFS Filesystem (Distributed):**
   - Path: `/user/akoubaa`
   - Commands: `hdfs dfs -ls`, `hdfs dfs -cat`, etc.
   - **Cannot use `cd`** - HDFS is NOT a mountable filesystem!

Common Mistake:

```
# WRONG - This will fail!
cd /user/akoubaa

# CORRECT
hdfs dfs -ls /user/akoubaa
```

2 Exploring Cluster Architecture

2.1 Check Cluster Status

View the overall cluster health and capacity:

```
hdfs dfsadmin -report
```

Sample Output:

```
Configured Capacity: 102719209472 (95.66 GB)
Present Capacity: 90407325696 (84.20 GB)
DFS Remaining: 90407235584 (84.20 GB)
DFS Used: 90112 (88 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
Live datanodes (2):
...
```

Key Insight:

What to Look For When you run this command, observe:

- **Configured Capacity:** Total storage available (~95 GB)
- **DFS Used:** Currently occupied space
- **Live datanodes:** Should show 2 worker nodes
- **Replication factor:** Default is 2

Lab Question 1: How much storage is available in the cluster? _____

2.2 Check Cluster Configuration

View the replication factor:

```
hdfs getconf -confKey dfs.replication
```

Expected Output:

```
2
```

This means each file is replicated 2 times for fault tolerance.

Try these additional queries:

```
# Block size (how large each block is)
hdfs getconf -confKey dfs.blocksize
```

Sample Output:

```
134217728
```

This is 128 MB in bytes (134217728 ÷ 1024 ÷ 1024 = 128 MB).

```
# NameNode address
hdfs getconf -confKey fs.defaultFS
```

Sample Output:

```
hdfs://0.0.0.0:9000
```

Lab Question 2: What is the default block size in bytes? Convert to MB: _____

3 HDFS Filesystem Navigation

3.1 List Your Home Directory

Check your personal HDFS directory:

```
hdfs dfs -ls /user/<your_username>
```

Example:

Example for User akoubaa

```
hdfs dfs -ls /user/akoubaa
```

Expected: Empty (no output) if this is your first time, or a list of files/directories you created previously.

Warning:

Remember: HDFS vs Linux Filesystem **Do NOT try:** `cd /user/akoubaa` (This will fail!)
Always use: `hdfs dfs` commands to interact with HDFS.
The `/user` directory exists in HDFS, not in the Linux filesystem at `/home`.

3.2 Create Directories

Practice creating directories in HDFS:

```
# Create a single directory
hdfs dfs -mkdir /user/<your_username>/lab01

# Create nested directories
hdfs dfs -mkdir -p /user/<your_username>/data/input
hdfs dfs -mkdir -p /user/<your_username>/data/output
```

Verify creation:

```
hdfs dfs -ls /user/<your_username>
hdfs dfs -ls -R /user/<your_username>
```

Key Insight:

The -R Flag The `-R` flag stands for *recursive*, meaning it lists all subdirectories and their contents.

3.3 Understanding Permissions

When you list directories, you'll see output like:

```
drwxr-xr-x   - akoubaa hadoop     0 2026-02-02 10:30
   /user/akoubaa/lab01
```

Definition:

Permission Format

- `d`: Indicates this is a directory
- `rwxr-xr-x`: Permissions (owner can read/write/execute)
- `akoubaa`: Owner of the directory
- `hadoop`: Group
- `0`: Current size in bytes
- Remaining fields: Timestamp and path

4 Uploading Files to HDFS

4.1 Create a Test File

Create a small file on the local filesystem:

```
echo "Hello HDFS! This is my first file." > test.txt
echo "Big Data is awesome!" >> test.txt
echo "HDFS stores data in blocks." >> test.txt
```

Verify the local file:

```
cat test.txt
ls -lh test.txt
```

4.2 Upload to HDFS

Copy the file from local storage to HDFS:

```
hdfs dfs -put test.txt /user/<your_username>/lab01/
```

Key Insight:

Alternative Upload Methods HDFS provides multiple commands for uploading:

```
# Same as -put
hdfs dfs -copyFromLocal test.txt
   /user/<your_username>/lab01/test2.txt
```

Verify upload:

```
hdfs dfs -ls /user/<your_username>/lab01/
```

4.3 Read File from HDFS

Display file contents:

```
hdfs dfs -cat /user/<your_username>/lab01/test.txt
```

5 Understanding Blocks and Replication

5.1 Create a Large File

Generate a file large enough to span multiple blocks:

```
# Create a ~150 MB file (larger than default 128 MB block size)
dd if=/dev/urandom of=bigfile.dat bs=1M count=150
```

Check the local file size:

```
ls -lh bigfile.dat
```

Warning:

Upload Time This upload will take approximately 10-30 seconds depending on network speed. Be patient!

5.2 Upload Large File

```
hdfs dfs -put bigfile.dat /user/<your_username>/lab01/
```

5.3 Check File Status

View detailed information about the file:

```
hdfs fsck /user/<your_username>/lab01/bigfile.dat -files -blocks
   -locations
```

Definition:

FSCK Output The `fsck` (filesystem check) command shows:

- **Total size**: Approximately 150 MB
- **Number of blocks**: 2 (one 128 MB block + one smaller block)
- **Block locations**: Which DataNodes store each block
- **Replication**: Each block appears on 2 DataNodes

Lab Question 3: How many blocks does your 150 MB file have? _____

Lab Question 4: On which DataNodes are the blocks stored? _____

5.4 View File Statistics

Get quick file information:

```
hdfs dfs -stat "Size: %b bytes, Replication: %r, Block Size: %o"
   /user/<your_username>/lab01/bigfile.dat
```

6 File Operations

6.1 Copy Files Within HDFS

```
# Copy file to another location in HDFS
hdfs dfs -cp /user/<your_username>/lab01/test.txt
   /user/<your_username>/data/input/
```

6.2 Move/Rename Files

```
# Rename a file
hdfs dfs -mv /user/<your_username>/lab01/test2.txt
   /user/<your_username>/lab01/renamed.txt
```

6.3 Download from HDFS

Copy a file from HDFS back to local storage:

```
hdfs dfs -get /user/<your_username>/lab01/test.txt downloaded_test.txt
```

Verify local download:

```
cat downloaded_test.txt
```

6.4 Delete Files

Warning:

Permanent Deletion HDFS does not have a recycle bin! Files deleted are gone permanently.

```
# Delete a single file
hdfs dfs -rm /user/<your_username>/lab01/renamed.txt

# Delete directory and all contents
hdfs dfs -rm -r /user/<your_username>/lab01/temp
```

7 Cluster Metadata Analysis

7.1 Check Disk Usage

See how much space your files use:

```
# Your directory size
hdfs dfs -du -h /user/<your_username>

# Summary only
hdfs dfs -du -s -h /user/<your_username>
```

Key Insight:

Replication and Storage If you have a 150 MB file with replication factor 2, the total storage consumed is:

$$150 \text{ MB} \times 2 = 300 \text{ MB}$$

7.2 Count Files and Directories

```
hdfs dfs -count /user/<your_username>
```

Output format:

```
DIR_COUNT   FILE_COUNT   CONTENT_SIZE   PATHNAME
```

7.3 View Namespace Quota

Check your storage limit:

```
hdfs dfs -count -q /user/<your_username>
```

Key Insight:

Student Quota Each student has a 300 MB quota. Stay within this limit!

8 Advanced Exploration (Optional)

8.1 Examine Specific Blocks

For a specific block (get block ID from previous `fsck` output):

```
hdfs fsck /user/<your_username>/lab01/bigfile.dat -blockId <block_id>
```

8.2 Check Cluster Health

See if any blocks are under-replicated or corrupted:

```
hdfs fsck / -list-corruptfileblocks
```

Expected Result: No corrupt blocks.

8.3 Explore Root Directory

Warning:

Read-Only Access You can view the root directory but cannot modify it.

```
hdfs dfs -ls /
hdfs dfs -ls /user
```

You can see all students' directories, but you cannot access their contents (permissions prevent this).

Lab Completion Checklist

Mark these tasks as complete:

- ☐ Successfully connected to cluster via SSH
- ☐ Viewed cluster status with `dfsadmin -report`
- ☐ Created directories in your HDFS home
- ☐ Uploaded a small text file
- ☐ Uploaded a large file (>128 MB)
- ☐ Examined file blocks and replication
- ☐ Performed file operations (copy, move, delete)
- ☐ Checked disk usage and metadata

Lab Questions (Submit Answers)

1. What is the total configured capacity of the cluster?

2. How many DataNodes are live in the cluster?

3. What is the default replication factor?

4. How many blocks does your 150 MB file have? Why?

5. If the replication factor is 2, how much space does a 100 MB file consume in total?

6. What happens if one DataNode fails? Will your data be lost? Why or why not?

Key Takeaways

1. **HDFS is a distributed filesystem** that splits files into blocks for parallel processing
2. **Default block size is 128 MB** – files larger than this span multiple blocks
3. **Replication provides fault tolerance** – each block exists on multiple nodes
4. **Metadata is managed by the NameNode** – it tracks where blocks are stored
5. **DataNodes store the actual data** – they report heartbeats to the NameNode

Troubleshooting

Problem: Can't connect via SSH

Solution: Check your username and password. Contact instructor if the issue persists.

Problem: "Permission denied" error
Solution: You can only write to `/user/<your_username>`. Don't try to write elsewhere.

Problem: "No space left" error
Solution: You've exceeded your 300 MB quota. Delete some files with `hdfs dfs -rm`.

Problem: File upload is very slow
Solution: Normal for large files. A 1visit website 50 MB file may take 30-60 seconds.

Additional Resources

- HDFS Commands Cheat Sheet
- Course Slides: Week 2B & 2C
- Textbook: Chapter 2 – Distributed Storage with HDFS

Congratulations!

You've completed your first HDFS lab!

Next Steps: Explore more HDFS commands and prepare for MapReduce in Week 3