Task 01. Hadoop. HDFS.

Industrial Machine Learning on Hadoop and Spark, Fall 2025

Task start date: September 21, 2025, 23:59CET.
Hard Deadline: **September 28, 2025, 23:59CET.**

Assignment Statement

This assignment is aimed at introducing you to the Apache Hadoop infrastructure.
The task consists of the single part:

1. Performing basic actions in the HDFS filesystem

1 Basic actions in HDFS (5% for each point + 15% for questions)

In this section, you need to write a script `hdfs.sh` that performs the following sequence of actions from the root directory of the Hadoop cluster's `namenode`:

1. Create a local file `test.txt` with a size of 100Mb
2. Create HDFS directories `temp` and `logs`
3. Upload the file `test.txt` into the `temp` directory
4. View the properties of the uploaded file
5. Move the file `test.txt` into the `logs` directory
6. Set the replication factor for the file to 1
7. Copy `test.txt` to `test2.txt`
8. Copy the directory `logs` into `logs2` using `hadoop distcp`
9. Set file permissions to read and write only for the owner for `test2.txt` in the `logs2` directory
10. Display the properties of all files in `logs2`
11. View the size of all directories in `/`
12. Delete the `logs` directory
13. Run `fsck` on the `logs2` directory
14. View the HDFS report via `dfsadmin`
15. Move `/logs2/test2.txt` to the local folder `/`
16. Append the contents of the local file `test2.txt` to the end of the file `/logs2/test.txt` in HDFS
17. Output the size of each file in `/logs2` in Mb

Each step must correspond to a single command. The execution result (stdout, stderr) of `hdfs.sh` must be saved in the file `hdfs.output.txt`.

Also, answer the following questions in the file `hdfs.answers.md`:

1. How many block replicas are missing after running the `fsck` command? Explain the reason for the missing replicas. **(7%)**
2. What is the size of the HDFS filesystem? **(8%)**

As your submission, you must provide three files named exactly as follows: `hdfs.sh`, `hdfs.output.txt`, `hdfs.answers.md`.