Big Data Technologies Lab Tasks

Task 1

Prepare a Linux (preferably Ubuntu) based environment if you have not already. For those having Mac OS do not require to configure for the Linux environment.

You are required to prepare a Linux OS platform in order to install, configure and later solve example problems using Hadoop framework.

Minimum specifications of the system:

- 2 core CPU
- 4 GB RAM
- 25 GB Storage

If you are preparing the VM, be careful of the specifications of the host system (your main system) while configuring.

You need to submit a short report including the specifications information of the prepared VM with real recordings of the terminal highlighting the username and hostname along with cpu info (for e.g. using *lscpu* command) and memory info (for e.g. using *free* command).

Current Mac/Linux owners can submit the report based on their existing systems configuration.

Video report Submission URL: 078BCT
(In this location, you need to create a folder with your CRN in 'THA078BCT001' format - case sensitive)
Report Format: .mp4
File Name: BDT_Lab.mp4 (Inside your roll number folder)
[All tasks should have a combined video report]

Note: For those, who are planning not to fill the board exam form of 'Big Data Technologies' (regular absentees), are not obliged to submit the assignment.

Task 2
Install hadoop-3.4.0 in your OS environment you prepared in your ***Task 1*** section.

The purpose of this assignment is to familiarize students with the installation process of Apache Hadoop 3.4.0, a distributed data processing framework, and to document the process as a professional report. This will provide hands-on experience in configuring a big data platform and showcasing your ability to produce technical documentation.

Prerequisites:

- JDK 8 or later
- set environment variables such as ***JAVA_HOME*** and ***HADOOP_HOME*** which will be used during the installation process
- It is recommended to create a separate user to use the hadoop system.

Post Download Actions:
After downloading the hadoop-3.4.0 binaries, configure following files:

- core-site.xml
- hdfs-site.xml
- mapred-site.xml
- yarn-site.xml

After formatting the Hadoop filesystem using ***hdfs namenode -format*** and starting the hadoop services using ***start-all.sh*** or similar command, following services should be in the running state.

- NameNode and DataNode
- ResourceManager and NodeManager

Verify the Correct Installation by:

- Accessing the Hadoop web interface at `http://localhost:9870` (for NameNode).
- Running basic Hadoop commands such as:
    - `hadoop version`
    - `hdfs dfs -ls /`

Prepare the Demo Video:

Your video report should contain the installation steps including the following information in your report:

1. **System Specifications:**
    - Hardware and software prerequisites (You have configured in ***Task 1***)
2. **Installation Proof:**
    - Proof of successful installation, including:
        - Hadoop version output (should contain the hostname information of the OS environment you prepared in the ***Task1)***
        - Command outputs (should contain the hostname information of the OS environment you prepared in the ***Task1)***
        - Recording of the web interface

Task 3

Solve the following problems using the provided resources. Separate helping instructions are provided for each problem. You can follow the instructions if you want.

Problem 1:

Count the frequency of words in a given plain text file using hadoop framework.

Problem 2:

Find the average/mean age of male and female died in titanic disaster.

Prepare the Report:

Your video report should contain the following information:

1. **Problem Definition**
2. **Input Data (Table for CSV format)**
3. **Output**

You can find the resources for both problem here: Lab_Problems

Submission Deadline: 22nd Feb, 2026