Big Data Technologies Assignments

Assignment 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 screenshots 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.

Report Submission URL: 077BEI
(In this location, you need to create a folder with your CRN in 'THA077BEI001' format - case sensitive)
Report Format: PDF
File Name: Assignment1.pdf (Inside your roll number folder)
Submission Deadline: 3rd Jan, 2025

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.

Assignment 2
Install hadoop-3.4.0 in your OS environment you prepared in your ***Assignment 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 Report:

Your PDF report should contain the installation steps including the following sections in your report:

1. **Introduction:**
    - Brief overview of Apache Hadoop and its significance
2. **System Requirements:**
    - Hardware and software prerequisites (You have configured in ***Assignment 1***)
3. **Installation Steps:**
    - Detailed steps with screenshots and explanations
4. **Validation:**
    - Proof of successful installation, including:
        - Hadoop version output (should contain the hostname information of the OS environment you prepared in the ***Assignment1)***
        - Command outputs (should contain the hostname information of the OS environment you prepared in the ***Assignment1)***
        - Screenshots of the web interface
5. **Challenges and Solutions:**
    - Describe any issues faced during the installation and how they were resolved
6. **Conclusion:**
    - Summarize the learning outcomes and the importance of Hadoop in big data applications

Submit the following (Inside a folder named Assignment2 (case sensitive) of your roll number folder):

- A PDF document of the report (Assignment2.pdf)
- Configuration files (`core-site.xml`, `hdfs-site.xml`, `mapred-site.xml`,`yarn-site.xml`)

Submission Deadline: 17th Jan, 2025