**Department of Information Technology**

7IT4-21: Big Data Analytics Lab

Odd Sem 2022-23

**Zero Lab**

Faculty: Amol Saxena Asst. Prof-IT

**Lab Objectives:**

The Big Data Analytics lab provides an introduction to Big Data, and the ways Hadoop can be used for Big Data implementations. In this lab the students will learn how Hadoop solves Big Data challenges, and which core Hadoop components can work together to create a rich Hadoop ecosystem applicable for solving many real-world problems. Hadoop provides distributed data storage and the HDFS and HBase are the two main Hadoop data storage mechanisms. Hadoop’s main execution framework is MapReduce, a programming model for distributed, parallel data processing, breaking jobs into mapping phases and reduce phases (thus the name). Because of the nature of how MapReduce works, Hadoop brings the processing to the data in a parallel fashion, resulting in fast implementation. This lab also enables students to perform data analytics related experiments using Pig and Hive platforms on top of Apache Hadoop.

**Lab Outcomes:**

1. **Apply** the installation procedure of the Hadoop framework to successfully install the Hadoop ecosystem in three different operating modes i.e. Standalone, Pseudodistributed, Fully distributed.
2. **Implement** file management tasks in Hadoop and some data structures like linked lists, stacks, queues in Java.
3. **Run** MapReduce programs like word count, matrix multiplication and the programs that mine Weather dataset.
4. **Analyze** datasets by executing Pig latin scripts and Hive queries in Hadoop environment.
5. **Design** solutions of some real life big data problems using Hadoop ecosystem.
6. **Write** lab records in professional manner and submit timely without copying from other sources.
7. Orally **present** the lab work in unambiguous manner with proper linkage with required concepts.

**LO PO & PSO Mapping**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LO | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
| LO1 | - | - | - | - | 3 | - | - | - | - | - | - | - | - | - | 2 |
| LO2 | 3 | - | - | - | - | - | - | - | - | - | - | - | 2 |  | 3 |
| LO3 | - | 3 |  | - | - | - | - | - | - | - | - | - | 2 | 2 |  |
| LO4 | - | 3 | - | - |  | - | - | - | - | - | - | - | 3 |  |  |
| LO5 | - | - | 3 | - | - | - | - |  | - | - | - | - | - | 3 | 2 |
| LO6 | - | - | - | - | - | - | - | 3 | - | - | - | - | - | - | - |
| LO7 | - | - | - | - | - | - | - | - | - | 3 | - | - | - | - | - |

**RTU Syllabus**

7IT4-21: Big Data Analytics Lab

Credit: 2 Max. Marks: 100(IA:60, ETE:40)

0L+0T+4P End Term Exam: 2 Hours

|  |  |
| --- | --- |
| **S No.** | **List of Experiments** |
| 1 | Implement the following Data structures in Java  i) Linked Lists ii) Stacks iii) Queues iv) Set v) Map |
| 2 | Perform setting up and Installing Hadoop in its three operating modes: Standalone, Pseudodistributed, Fully distributed. |
| 3 | Implement the following file management tasks in Hadoop:   * Adding files and directories * Retrieving files * Deleting files   Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities |
| 4 | Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm. |
| 5 | Write a Map Reduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. |
| 6 | Implement Matrix Multiplication with Hadoop Map Reduce |
| 7 | Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data. |
| 8 | Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes. |
| 9 | Solve some real life big data problems. |

**Lab Plan/ List of Experiments**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exp.No.** | **Description of Experiment** | **Batch** | **Scheduled date of Exp.** | **Actual date of Exp.** |
| 1 | Zero Lab | A1 | 23-08-2022 |  |
| A2 | 22-08-2022 |  |
| A3 | 17-08-2000 |  |
| 2 | Implement the following Data structures in Java  i) Linked Lists ii) Stacks | A1 | 30-08-2022 |  |
| A2 | 29-08-2022 |  |
| A3 | 24-08-2022 |  |
| 3 | Implement the following Data structures in Java  iii) Queues iv) Set v) Map | A1 | 06-09-2022 |  |
| A2 | 05-09-2022 |  |
| A3 | 31-08-2022 |  |
| 4 | Perform setting up and Installing Hadoop in its three operating modes: Standalone, Pseudodistributed, Fully distributed. | A1 | 13-09-2022 |  |
| A2 | 12-09-2022 |  |
| A3 | 07-09-2022 |  |
| 5 | Implement the following file management tasks in Hadoop:   * Adding files and directories * Retrieving files * Deleting files   Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities | A1 | 20-09-2022 |  |
| A2 | 19-09-2022 |  |
| A3 | 14-09-2022 |  |
| 6 | Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm. | A1 | 27-09-2022 |  |
| A2 | 26-09-2022 |  |
| A3 | 21-09-2022 |  |
| 7 | Write a Map Reduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. | A1 | 04-10-2022 |  |
| A2 | 03-10-2022 |  |
| A3 | 28-09-2022 |  |
| 8 | Implement Matrix Multiplication with Hadoop Map Reduce | A1 | 11-10-2022 |  |
| A2 | 10-10-2022 |  |
| A3 | 12-10-2022 |  |
| 9 | Install and Run Pig then write Pig Latin scripts to sort, group your data. | A1 | 01-11-2022 |  |
| A2 | 31-10-2022 |  |
| A3 | 02-11-2022 |  |
| 10 | Write Pig Latin scripts to filter join, project, and filter your data. | A1 | 15-11-2022 |  |
| A2 | 07-11-2022 |  |
| A3 | 09-11-2022 |  |
| 11 | Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes. | A1 | 22-11-2022 |  |
| A2 | 14-11-2022 |  |
| A3 | 16-11-2022 |  |
| 12 | Solve some real life big data problems. | A1 | 29-11-2022 |  |
| A2 | 21-11-2022 |  |
| A3 | 23-11-2022 |  |

**Marking Scheme**

**RTU Marks Scheme**

|  |  |  |
| --- | --- | --- |
| Maximum Marks Allocation | | |
| Sessional | End-Term | Total |
| 60 | 40 | 100 |

**Marks Division**

|  |  |  |
| --- | --- | --- |
| **1st / 2nd Midterm and End-term** | | |
| Experiment | Viva | Total |
| 30 | 10 | 40 |
| **Attendance Performance (Internal)** | | |
| Attendance | Performance | Total |
| 10 | 30 | 40 |

**Assessment of an Experiment**

Total Marks – 10

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attendance | Discipline | Performance | Record | Viva | Total |
| 2 | 2 | **3** | 1 | 2 | 10 |

**Software required**

* For experiments 1 and 2, JAVA (JAVA 8 or above) is required to be installed.
* For rest of the experiments this lab required Hadoop environment as given below.

This section describes how to set up and configure a single-node Hadoop installation so that you can quickly perform simple operations using Hadoop MapReduce and the Hadoop Distributed File System (HDFS).

**Required Software for Big Data Analytics Lab**

**Supported Platforms**

1. GNU/Linux is supported as a development and production platform. Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.

Link is given below for installation instructions

<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html>

1. Windows is also a supported platform but the followings steps are for Linux only. To set up Hadoop on Windows, see [wiki page](http://wiki.apache.org/hadoop/Hadoop2OnWindows) (link is given below for installation instructions).

<https://cwiki.apache.org/confluence/display/HADOOP2/Hadoop2OnWindows>

OR

<https://gist.github.com/vorpal56/5e2b67b6be3a827b85ac82a63a5b3b2e>

(Installation instructions enclosed as a separate document)

**Download**

To get a Hadoop distribution, download a recent stable release from one of the [Apache Download Mirrors](http://www.apache.org/dyn/closer.cgi/hadoop/common/) (link is given below).

<https://www.apache.org/dyn/closer.cgi/hadoop/common/>