DR. BABASAHEB AMBEDKAR TECHNOLOGICAL UNIVERSITY, LONERE-RAIGAD
Department of Computer Engineering

B.Tech. Final Year A.Y. 2022-2023
Semester: - VII (ODD) Subject: -Big Data Analytics Laboratory
Subject Code: - BTCOL707 (A)
Subject Co-ordinator: - Ms. Harsha Gaikwad
Laboratory Name: - B3

List of Experiments:

1. Perform setting up and Installing Hadoop in its two operating modes:
   a) Pseudo distributed,
   b) Fully distributed.
2. Implement the following file management tasks in Hadoop: -
   a) Adding files and directories
   b) Retrieving files
   c) Deleting files
3. To understand the overall programming architecture using Map Reduce API
4. Store the basic information about students such as roll no, name, date of birth and address of student using various collection types such as List, Set and Map
5. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.
   a) Find the number of occurrence of each word appearing in the input file(s)
   b) Performing a Map Reduce Job for word search count (look for specific keywords in a file)
6. Install and Run Hbase then use Hbase DDL and DML commands
7. Install, Deploy & configure Apache Spark Cluster. Run apache spark applications using Scala.
8. Basic CRUD operations in MongoDB
9. Retrieve various types of documents from students collection
10. Data analytics using Apache Spark on Amazon food dataset, find all the pairs of items frequently reviewed together.

a) Write a single Spark application that:

i. Transposes the original Amazon food dataset, obtaining a Pair RDD of the type: → <user_id> → <list of the product_ids reviewed by user_id>

ii. Counts the frequencies of all the pairs of products reviewed together;

iii. Writes on the output folder all the pairs of products that appear more than once and their frequencies. The pairs of products must be sorted by frequency.