CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A)

B.E. (CSE) V Sem

Big Data Analytics Professional Elective-II

Assignment-1, 2024-25

1. Recall the definitions of the 5 Vs of BigData. CO-1 L1
2. Contrast BigData and Data Science CO-1 L2
3. Contrast the features of Hadoop 1 and Hadoop 2. CO -1 L2
4. Explain the working of MapReduce Pipeline. CO-5 L2
5. Discuss the Hadoop ecosystem. Explain how each component of the Hadoop stack (HDFS, MapReduce, YARN) contributes to Big Data processing. CO-1 L2
6. Write the Need of big data for healthcare industry. CO-5 L2
7. Provide examples of commonly used commands in the Hadoop Command-Line Interface (CLI). Explain how these commands are used to manage files in HDFS. CO-1 L2
8. Write about the following Regarding Apache PIG CO2 L3
   - Modes of Pig
   - Different Modes of Pig Execution
   - User-Defined Functions (UDFs) in Pig Latin
9. Write the differences between Hive and HBase and architecture of Hive. Explain how data is stored, managed, and queried in Hive. CO2- L2
10. Consider the Scenario :

create a table that contains details of all the transactions done by the customers of year 2024: **CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’**

Now, after inserting 30,000 tuples in this table, Now just we want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query.

How will you solve this problem and list the steps that will be taking in order to do so?

Prepared By : Dr Raman (CSE1& CSE2)

Dr G Vanitha (CSE2& CSE3)