FACULTY OF COMPUTER APPLICATIONS
B.Sc. (DS)

- **Sem.** : 5
- **Subject Code** : 05DS0503
- **Subject** : Big Data Frameworks
- **Course Objectives** :
  1. Students will be able to understand concept of big data and Hadoop.
  2. Students will be able to know the implementation of HDFS and Map Reduce.
  3. Students will be able to develop proficiency in creating hive queries.
  4. Students will be able to understand how to write and execute Pig Script.
  5. Students will be able to understand concept of HBase.

- **Prerequisites** : Knowledge of Java Programming, Python Programming and Database Management System

Unit No	Topics Covered	No of lectures required
1	**Introduction to Big Data and Hadoop:** • Definition of Big data, characteristics of Big Data, • Understanding Big data, • Types of Big Data, • Traditional Versus Big Data Approach, Technologies Available for Big Data, • Infrastructure for Big Data, Big Data Challenges. • Introducing Hadoop, • Why Hadoop?, Why not RDBMS?, • RDBMS versus Hadoop, • Distributed Computing Challenges, • History of Hadoop, • Hadoop Overview, • Use Case of Hadoop, • Hadoop Distributors, • HDFS (Hadoop Distributed File System),	**10**

	• Processing Data with Hadoop, • Managing Resources and Applications with Hadoop YARN (Yet another Resource Negotiator), • Interacting with Hadoop Ecosystem	
2	**MapReduce** • Map Reduce: MapReduce and New Software stack (Distributed File Systems, Physical Organization of Compute Nodes), • The Map Tasks, • Grouping by Key, • The Reduce Tasks, Combiners, • Details of MapReduce Execution, • Coping With Node Failures. • MapReduce, Grouping and Aggregation by MapReduce, Matrix Multiplication, • Matrix Multiplication with One MapReduce Step.	**10**
3	**Hive** • Introduction • What is HIVE?, • HIVE Architecture, • HIVE data Types, • HIVE File Formats, • HIVE query Language, • RCFile implementation, • Sharding, • User-Defined Functions (UDF)	**08**
4	**Pig :** • Introduction • What is Pig? • The anatomy of Pig, • Pig on Hadoop, • Use Case for Pig- ETL Processing, • Pig Latin overview, • Datatypes in Pig, • running Pig, Execution modes of Pig, • HDFS commands, • Relational operators, • Eval function, • complex Data Types, • Piggy Bank,	**10**

	• User-Define Functions, • Parameter substitution, • Diagnostic Operator, • Word Count Example using Pig, • When to use and not use Pig, • Pig at Yahoo, • Pig vs HIVE	
5	**HBase :** • Introduction, • Architecture, Installation, shell, • General commands, • Admin API, • Commands related to table (create, list, enable, disable, describe, alter, drop), • Commands related to data (create, update, read, delete), • Scan, Count, truncate, security	**07**

Course Outcomes:

1. Define the concept of Big Data and basic Big Data terminologies.
2. Implement HDFS and Map Reduce in Hadoop.
3. Implement Hive queries.
4. Apply Pig concepts to create Pig Scripts.
5. Implement the concepts of HBase.

Course Outcomes – Program Outcomes Mapping Table:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PSO1	PSO2	PSO3
CO1	H	L	L		M			L		L	
CO2	H	H	L	M	L			L	H	H	
CO3	H	H	L	M	M			L	M		
CO4	H	H	L		M			L	M		
CO5	H	H	L		M			L		H	

Text Book :

1. **Big Data and Analytics by Seema Acharya, Subhashini Chhellappan Willey, Second edition**
2. **“Big Data Analytics”, Radha Shankarmani, M Vijayalakshmi, 2nd Edition, Wiley**
3. **HBase: The Definitive Guide by Lars George**

Marwadi University

FACULTY OF COMPUTER APPLICATIONS
B.Sc. (DS)

Reference Books :

1. **Professional NoSQL, Shashank Tiwari, WROX, First Edition**
2. **Big Data Fundamentals: Concepts, Drivers and techniques, Thomas Erl,Wajid Khattak,and Paul Buhler: Pearson, First Edition**
3. **Hadoop Real World Solutions, Jonathan R. Owens, Brian Femiano, Jon Lentz, Packt Publication, First Edition**

Web References:

1. **http://www.bigdatauniversity.com**
2. **https://www.tutorialspoint.com/hbase**

App References:

1. **https://play.google.com/store/apps/details?id=com.iitsysco.learn_big_data_hadoop**
2. **https://play.google.com/store/apps/details?id=com.vrpmeone.LearnHadoop**

Syllabus Coverage from text /reference book & web/app reference:

Unit	Book	Chapter Numbers
1	1	5
2	2	4
3	1	9
4	1	10
5	3	2,3

PRACTICALS

Unit No	List of Practicals
1	**Hadoop :** 1. Hadoop installation steps. 2. Which path you need to set in hadoop-env? 3. Properties setting in : a. Core-site.xml b. Hdfs-site.xml c. Mapred.xml d. Yarn-site.xml

	4. How to start and stop hadoop server? 5. Which command is used to list the server names? Start all the servers and list the server names.
2	**Map Reduce :** 1. Create a hello.txt file in local system and copy it into HDFS. 2. Word count program – MapReduce (List all the steps with code) 3. Vowel count program using map reduce. 4. Matrix multiplication using map reduce. 5. Prepare an "input" folder containing multiple text files. Create a program using MapReduce that would accept the path to the "input" folder and generate an "output" folder having a text file containing the total number of occurrences of each single word present in text document. For example, if the text containing in input files is as follows: "We thank you for your visit to Ahmedabad. We hope that you would visit us again." 6. Write a program to perform Union, Intersection and Difference operation using MapReduce on following files. Input files: a) Content of file 1 (apple, orange, mango, apple, banana) b) Content of file 2 (apple, apple, plum, kiwi, kiwi, mango, mango) c) Content of file 3 (orange, orange, plum, grapes, kiwi, mango, apple
3	**Hive :** **Question : Write steps to start Hive.** **Ex – 1 :** a. Create database Payroll b. Create Table Employee, Department c. Run DDL and DML commands d. Create table via loading data from files e. Create table form existing schema f. Run Data Retrieval queries , Joins, HIVEQL **Ex – 2 :** word count using Hive **Ex – 3 :** Create a partition table for customer schema to reward the customers based on their life time values. Custid Customers Life time values 1001 Jack 25000 1002 Smith 8000

	1003 David 12000 1004 John 15000 1005 Scott 12000 1006 Joshi 28000 1007 Ajay 12000 1008 Vinay 30000 1009 Joseph 21000 a) Create partition table if life time value is 12000. b) Create partition table for all life time values
4	**Pig :** **Question: Write steps to start Pig.** **Ex – 1 : Create Movies.csv file** a. Running Pig program in Local and MapReduce Mode b. Working with Pig Operators (FOREACH, ASSERT, FILTER, GROUP, ORDER BY, DISTINCT, JOIN, LIMIT, SAMPLE, SPLIT) c. Working with Pig functions d. Error handling in Pig e. Debugging in Pig **Ex – 2 :** Pig script for word count **Ex-3** Working with Pig Operators/Functions (LOAD, DUMP, FOREACH, GROUP, DISTINCT, LIMIT, ORDER BY, JOIN, UNION, SPLIT, SAMPLE, AVG, MAX, COUNT, TUPLE, MAP,PIGGY BANK, PARAMETER SUBSTITUTION, DESCRIBE, **Ex-4** Simple Problems like 1) Write a pig script to load and store “Student data”.(Student file contain Roll no, Name, Marks and GPA). a) Filter all the students who are having GPA>5. b) Display the name of all Students in Uppercase. c) Group tuples of students based on their GPA. d) Remove duplicates tuple of Student list. e) Display first three tuples from “student” relation. f) Display the names of students in ascending order. g) Join two relation namely Student and department (Rno,DeptNo,DeptName) based on the values contain in the roll no column. h) Merge content of two relation Student and department. i) Partition a relation based on the GPA’s acquired by students. j) To calculate the average marks for each student. k) Calculate maximum marks of each student. l) Count the number of tuples in a bag. **EX-5**

<table>
<tr><td></td><td>Load the file menu.csv (Category,Name, Price) and write one Pig script
a) Which meals cost more than 30.00?
b) Which meals contain the word “Panner ”?
c) Which are the 10 most expensive meals?
d) For every day, what’s the average price for a meal?
e) For every day, what’s the most expensive meal?
EX-6
Write a program to count Word on Pig.
Ex-7
Write a pig script to spilt customers for reward program based on their life time values. If Life time values is >1000 and < =2000 then Silver Program If Life time values is >20000 then Gold Program
Input :
<table><tr><th>Customers</th><th>Lifetime value</th></tr><tr><td>Jack</td><td>25000</td></tr><tr><td>Smith</td><td>8000</td></tr><tr><td>David</td><td>12000</td></tr><tr><td>John</td><td>15000</td></tr><tr><td>Scott</td><td>12000</td></tr><tr><td>Lucy</td><td>28000</td></tr><tr><td>Ajay</td><td>12000</td></tr><tr><td>Vinay</td><td>30000</td></tr><tr><td>Joseph</td><td>21000</td></tr><tr><td>Joshi</td><td>25000</td></tr></table>
Ex-8
Create a data file for below schemas: Order: CustomerId, ItemId, ItemName, OrderDate, DelivaryDate
Customer: CustomerId, CustomerName, Address, City, State, Country
a) Load Order and Customer Data.
b) Write a pig latin Script to determine number of items bought by each customer.
EX-9
Create a UDF to convert name into uppercase</td></tr>
<tr><td>5</td><td>HBase
Write Java program to perform CRUD operations of HBase by creating Employee table</td></tr>
</table>