From: THE REGISTRAR

To
The Principal,
Visakha Govt. Degree & P.G. College,
College for Women,
Visakhapatnam.

Sir/Madam,

Sub: Approval of Syllabus and Model Question Paper of
Data Science & Internet of Things, w.e.f. 2020-21-Requesting – Reg.

Ref: 1. Letter dt. 31-07-2021, received from the Principal, Visakha Govt.
Degree & P.G. College for Women, Visakhapatnam, enclosing the
Syllabus and Model Question Paper of Data Science & Internet of
Things, w.e.f. 2020-21.
2. No.C-II(4)Govt/New Market Oriented Courses/Affin/2020-21,
College Development Council, Date 02-08-2021.

* * *

With reference to the above, I am by direction to inform that the Revised Choice
Based Credit System, U.G. Courses (w.e.f. 2020-2021) Syllabus and Model
Question Paper of Data Science & Internet of Things has been approved.

Hence, I request to arrange to circulate the same among the Teaching Staff and
Students concerned and placed in A.U. website.

Thanking you,

Yours faithfully,

(B. RAMACHANDER)
ASSISTNAT REGISTRAR (ACADEMIC)

Copies to:
1. The Dean of Academic Affairs, A.U., VSP.
3. The Dean, CDC, A.U., Vsp.
4. The Dean, Confidential, A.U., Vsp.
5. The Controller of Examinations, A.U., Vsp.
6. The Secretary to V.C., Rector Table, P.A. to Registrar, A.U., Vsp.
7. The Director, Computer Centre, A.U., Vsp.
8. O.C. & O.O.F.
REVISED SYLLABUS OF B.Sc. (Data Science) UNDER CBCS FRAMEWORK WITH EFFECT FROM 2020-2021

PROGRAMME: THREE-YEAR B.Sc. (Computers – Statistics – Data science)

Market oriented course in Computer Science

Data Science

(With Learning Outcomes, Unit-wise Syllabus, References, Co-curricular Activities)

For 1, 2, 3 & 4 Semesters)

(To be Implemented from 2020-21 Academic Year)
## Data Science

<table>
<thead>
<tr>
<th>Semester</th>
<th>Paper</th>
<th>Subject</th>
<th>Hrs</th>
<th>Credits</th>
<th>IA</th>
<th>ES</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRST YEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMESTER I</td>
<td>I</td>
<td>Maths for Data science</td>
<td>4</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maths for Data science tutorial</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>SEMESTER II</td>
<td>II</td>
<td>Introduction to Data science With R</td>
<td>4</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R Programming Lab</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>SECOND YEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMESTER III</td>
<td>III</td>
<td>Big Data Technology</td>
<td>4</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Big Data Technology through Hadoop Lab</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>SEMESTER IV</td>
<td>IV</td>
<td>Data Mining and Data Analysis</td>
<td>4</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data Mining and Data Analysis lab</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>Big data Acquisition and Analysis.</td>
<td>4</td>
<td>3</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Big data Acquisition and Analysis lab</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>
Objective

The course is a brief overview of the basic tools from Linear Algebra and Multivariable Calculus that will be needed in subsequent course of the program.

Outcome

By completing the course the students will have been reminded of the basic tools of Linear Algebra and Multivariable Calculus needed in subsequent courses in the program notably:

- Fundamental properties of matrices, their norms, and their applications.
- Differentiating/Integrating multiple variable functions and the role of the gradient and the hessian matrix.
- Basic properties of optimization problems involving matrices and functions of multiple variables.

Unit-I

Matrices and Basic Operations, Special structures Matrices and Basic Operations, Interpretation of matrices as linear mappings and some examples.

Square Matrices, Determinants, Properties of determinants, singular and non-singular matrices, examples, finding an inverse matrix.

Unit-II

Eigen values and Eigenvectors Characteristic Polynomial, Definition of Left/Right Eigen values and Eigenvectors, Caley – Hamilton theorem, singular value Decomposition, Interpretation of Eigen values/vectors.

Unit-III

Linear Systems Definition, applications, solving linear systems, linear inequalities, linear programming.

Unit-IV

Real-valued functions of two or more variables. Definition, examples, simple demos, applications.

Unit-V

Analysis elements Distance, Limits, Continuity, Differentiability, the gradient and the Gaussian.
Optimization problems Simple examples, motivation, the role of the Hessian maxima and minima and related extreme conditions.

Integration Double integrals, Fubini’s theorem, properties, applications.

References


Student Activity:

1. Find the Eigenvectors of $A = \{ 1 1 1 1, 2 3 4 5, 3 4 5 6 \}$
2. Find orthogonal $S = \text{Span}\{(1 1 1 1), (1 4 4 0), (-1 4 4 0), (-4 2 2 0)\}$

I YEAR I SEMESTER MATHS FOR DATA SCIENCE

Tutorial

1. Study various applications of Matrices.
2. Study different polynomial functions and their uses.
3. Take one real world example and apply the Linear System solution.
4. Study some real valued functions and its applications.
5. Study and solve one optimization problem.
INTRODUCTION TO DATA SCIENCE WITH R

Objective

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection, preparation, analysis, modeling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Outcomes

1. Recognize various disciplines that contribute to a successful data science effort.
2. Understand the processes of data science - identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
3. Be aware of the challenges that arise in data sciences.
4. Develop and appreciate various techniques for data modeling and mining.
5. Be cognizant of ethical issues in many data science tasks.
6. Be comfortable using commercial and open source tools such as the R language and its associated libraries for data analytics and visualization.
7. Learn skills to analyze real time problems using R
8. Able to use basic R data structures in loading, cleaning the data and preprocessing the data.
9. Able to do the exploratory data analysis on real time datasets
10. Able to understand and implement Linear Regression
11. Able to understand and use - lists, vectors, matrices, dataframes, etc.

Unit-1:
Introduction to Data Science- Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team

Unit-2:

Unit-3:
Adding - Removing - Applying Special functions to Data Frames - Merging Data Frames - Factors and Tables.

Unit- 4:


Data Visualization in R : Types of visualizations - packages for visualizations - Basic Visualizations, Advanced Visualizations and Creating 3D plots.

Unit- 5:

Inferential Statistics with R - Types of Learning - Linear Regression- Simple Linear Regression - Implementation in R - functions on lm() - predict() - plotting and fitting regression line. Multiple Linear Regression - Introduction - comparison with simple linear regression - Correlation Matrix - F-Statistic - Target variables Vs Predictors - Identification of significant features - Implementation of Multiple Linear Regression in R.

References


Student Activity

Databases need to undergo pre-processing to be useful for data mining. Dirty data can cause confusion for the data mining procedure, resulting in unreliable output. Data cleaning includes smoothing noisy data, filling in missing values, identifying and removing outliers, and resolving inconsistencies.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)
A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity)

B. General

1. Group Discussion

2. Try to solve MCQ’s available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills


7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work

I YEAR II SEMESTER PAPER– II

R Programming LAB

1) Installing R and R studio

2) Create a folder DS_R and make it a working directory. Display the current working directory

3) installing the "ggplot2", "caTools", "CART" packages
4) load the packages "ggplot2", "caTools".
5) Basic operations in r
6) Working with Vectors:
   • Create a vector v1 with elements 1 to 20.
   • Add 2 to every element of the vector v1.
   • Divide every element in v1 by 5
   • Create a vector v2 with elements from 21 to 30. Now add v1 to v2.
7) Getting data into R, Basic data manipulation
8) Using the data present in the table given below, create a Matrix "M"

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0</td>
<td>12</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>C2</td>
<td>12</td>
<td>0</td>
<td>15</td>
<td>28</td>
</tr>
<tr>
<td>C3</td>
<td>13</td>
<td>15</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>C4</td>
<td>8</td>
<td>28</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>C5</td>
<td>20</td>
<td>88</td>
<td>9</td>
<td>33</td>
</tr>
</tbody>
</table>

   • Find the pairs of cities with shortest distance.
9) Consider the following marks scored by the 6 students

<table>
<thead>
<tr>
<th>Section</th>
<th>Student no</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>45</td>
<td>54</td>
<td>45</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>54</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>66</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>43</td>
<td>44</td>
<td>45</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>67</td>
<td>76</td>
<td>78</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>76</td>
<td>68</td>
<td>37</td>
</tr>
</tbody>
</table>

   • create a data structure for the above data and store in proper positions with proper names
   • display the marks and totals for all students
   • Display the highest total marks in each section.
   • Add a new subject and fill it with marks for 2 sections.

   • Three people denoted by P1, P2, P3 intend to buy some rolls, buns, cakes and bread. Each of them needs these commodities in differing amounts and can buy them in two shops S1, S2. The individual prices and desired quantities of the commodities are given in the following table "demand."

<table>
<thead>
<tr>
<th>Roll</th>
<th>S1</th>
<th>S2</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roll</td>
<td></td>
<td></td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Bun</td>
<td>2</td>
<td>2.5</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Cake</td>
<td>5</td>
<td>4.5</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Bread</td>
<td>16</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

   • Create matrices for above information with row names and col names.
   • Display the demand.quantity and price matrices
   • Find the total amount to be spent by each person for their requirements in each shop
   • Suggest a shop for each person to buy the products which is minimal.
10) Consider the following employee details:

<table>
<thead>
<tr>
<th>Employee Details as follows</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp_no: 1</td>
</tr>
<tr>
<td>name: Ram</td>
</tr>
<tr>
<td>salary</td>
</tr>
<tr>
<td>basic: 30000</td>
</tr>
<tr>
<td>hra: 2500</td>
</tr>
<tr>
<td>dz: 4000</td>
</tr>
<tr>
<td>deductions</td>
</tr>
<tr>
<td>pf: 1100</td>
</tr>
<tr>
<td>tax: 200</td>
</tr>
<tr>
<td>total_salary</td>
</tr>
<tr>
<td>ns(Gross Salary):</td>
</tr>
<tr>
<td>ns(Net Salary):</td>
</tr>
</tbody>
</table>

• Create a list for the employee data and fill gross and net salary.
• Add the address to the above list
• Display the employee name and address
• Remove street from address
• Remove address from the List.

11) Loops and functions - Find the factorial of a given number
12) Implementation of Data Frame and its corresponding operators and functions
13) Implementation of Reading data from the files and writing output back to the specified file
14) Treatment of NAs, outliers, Scaling the data, etc
15) Applying summary() to find the mean, median, standard deviation, etc
16) Implementation of Visualizations - Bar, Histogram, Box, Line, scatter plot, etc.
17) Implementation of Linear and multiple Linear Regression
18) Fitting regression line
II YEAR III SEMESTER PAPER– III
BIG DATA TECHNOLOGY

Objectives:
This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

Outcome
1. Learn tips and tricks for Big Data use cases and solutions.
2. Acquire knowledge of HDFS components, Namenode, Datanode, etc.
3. Acquire knowledge of storing and maintaining data in cluster, reading data from and writing data to Hadoop cluster.
4. Able to maintain files in HDFS
5. Able to write MapReduce applications to access data present on HDFS
6. Able to read different formats of files into map-reduce application.
7. Able to develop MapReduce applications to analyze Big Data related to the real world use cases.
8. Able to write MapReduce applications that can take data from multiple datasets and join them
9. Able to optimize the performance of Map-Reduce application

Unit-I: Introduction to Big Data
Introduction – Distributed File System – Big Data and its importance, Characteristics of Big Data, Limitation of Conventional Data Processing Approaches, Need of big data frameworks, Big data analytics, Limitations of Big Data and Challenges, Big data applications

Unit-II
Hadoop: Basic Concepts of Hadoop and its features - The Hadoop Distributed File System (HDFS)- Anatomy of a Hadoop Cluster - Hadoop cluster modes - Hadoop Architecture, Hadoop Storage - Hadoop daemons (Name node-Secondary name node-Job tracker-Task tracker-Data node, etc) - Anatomy of Read & Write operations – Interacting HDFS using command-line (HDFS Shell and FS shell commands) - Interacting HDFS using Java APIs – Dataflow – Blocks – Replica - YARN.

Unit-III
Hadoop Ecosystem Components – Schedulers- Fair and Capacity, Hadoop 2.0 Vs Hadoop 3.0 and its new features.


Unit-IV
Hadoop MapReduce - Introduction - Phases in MapReduce Framework - Anatomy of MapReduce Job run - Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce
Types and Formats, Map Reduce Features. Understanding Basic MapReduce Program (WordCount program): The Driver Code - The Mapper class - The Reducer class.

Unit-V:


References


Student Activity:

Case Study I: Centers for Medicare & Medicaid Services: The Integrity of Healthcare Data and Secure Payment Processing.

Case Study II: Movie Lens Data set Analysis

Case Study III: Web Server Log Analysis using MapReduce.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)
2. Student seminars (on topics of the syllabus and related aspects (individual activity))
3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))
4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity)
B. General

1. Group Discussion
2. Try to solve MCQ’s available online.
3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)
2. Closed-book and open-book tests
3. Problem-solving exercises
4. Practical assignments and laboratory reports
5. Observation of practical skills
6. Individual and group project reports like “Movie Lens Data Analysis”, "Youtube Click stream Data Analysis", etc.
7. Efficient delivery using seminar presentations,
8. Viva voce interviews.
9. Computerized adaptive testing, literature surveys and evaluations,
10. Peers and self-assessment, outputs form individual and collaborative work
II YEAR III SEMESTER PAPER– III

BIG DATA TECHNOLOGY Through Hadoop LAB

1. Implement the following Data Structures in Java
   a) Linked Lists
   b) Stacks
   c) Queues
   d) Set
   e) Map

2. Hadoop Cluster Setup
   (i) Perform setting up and Installing Hadoop in its three operating modes: Standalone
   Pseudo distribute
d Fully distribute
d (ii) Use web based tools to monitor your Hadoop setup.

3. Implement the following file management tasks in Hadoop:
   o Adding files and directories, List the files and directories
   o Retrieving files
   o Deleting files
   o Copying files from one folder to another in HDFS
   o Copying files from Local File System to HDFS

4. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm

5. Write a Map Reduce program that mines weather data (NCDC). 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. Data available at: ftp://ftp.ncdc.noaa.gov/pub/data/noaa/.
   • Find average, max and min temperature for each year in NCDC data set
   • Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.

6. Implement Matrix Multiplication program with Hadoop Map Reduce.

7. Stop word elimination problem:
   Input:
   • A large textual file containing one sentence per line
   • A small file containing a set of stop words (One stop word per line)
Output:
- A textual file containing the same sentences of the large input file without the words appearing in the small file.

8. Write a MapReduce Application to implement Combiners
9. Write a MapReduce Application to implement Reduce-side Join
10. Write a MapReduce Application to implement Map-side Join

Outcome:
- Able to develop MapReduce applications to analyze Big Data related to the real world use cases.
- Able to setup, configure and manage Hadoop cluster on single node
- Able to access the Hadoop cluster through Web UI.
- Able to track the execution of MapReduce jobs through Web UI
- Able use Joins, partitioner, combiners as and when needed while developing MapReduce application to analyze the Big Data.

II YEAR IV SEMESTER PAPER- IV
DATA MINING AND DATA ANALYSIS

Objective
- To learn data analysis techniques.
- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Outcome
Students who complete this course will be able to
1. To understand and demonstrate data mining
2. Compare various conceptions of data mining as evidenced in both research and application.
3. Characterize various kinds of patterns that can be discovered by association rule mining.
4. Evaluate mathematical methods underlying the effective application of data mining.
5. To Analyze the data using statistical methods
6. Gain hands-on skills and experience on data mining tools.

Unit-I
Data mining - KDD Vs Data Mining, Stages of the Data Mining Process-Task Primitives, Data Mining Techniques – Data Mining Knowledge Representation. Major Issues in Data Mining – Measurement and Data – Data Preprocessing – Data Cleaning - Data transformation- Feature Selection - Dimensionality reduction

Unit-II: Predictive Analytics

Unit-III : Classification and Descriptive Analytics

Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Prediction. Descriptive Analytics - Mining Frequent Itemsets - Market based model – Association and Sequential Rule Mining

Unit - IV : Cluster Analysis

Unit-V: Factor Analysis

Factor Analysis: Meaning, objectives and Assumptions, Designing a factor analysis, Deriving factors and assessing overall factors, Interpreting the factors and validation of factor analysis.

References

Student Activity

Case Study I: Analysis and Forecasting of House Price Indices
Case Study II: Customer Response Prediction and Profit Optimization
Case Study III: Iris Species Prediction
RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity)

B. General

1. Group Discussion

2. Try to solve MCQ’s available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills

6. Individual and group project reports like “Movie Lens Data Analysis”, "COVID-19 Analysis", etc.

7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work

II YEAR IV SEMESTER PAPER– IV
DATA MINING AND DATA ANALYSIS LAB
1. Data Analysis – Getting to know the Data (Using ORANGE WEKA or R Programming)
   - Parametric – Means, T-Test, Correlation
   - Prediction for numerical outcomes – Linear regression, Multiple Linear Regression
   - Correlation analysis
   - Preparing data for analysis
     - Pre-Processing techniques

2. Data Mining (Using ORANGE WEKA or R Programming)
   - Implement clustering algorithm
   - Implement Association Rule mining
   - Implement classification using
     - Decision tree
     - Back Propagation
     - Logistic Regression
     - Decision Tree
     - Random Forest
     - Naive Bayes
     - Support Vector Machines
   - Visualization methods
II YEAR IV SEMESTER PAPER– V

BIG DATA ACQUISITION AND ANALYSIS

Objective

Learn to develop Hadoop applications for storing processing and analyzing data stored in Hadoop cluster. The course is mainly covering Big Data tools for Data Transformation (Apache PIG), Data Analysis (HIVE) and for handling unstructured data HBase. To Understand the complexity and volume of Big Data and their challenges. To analyse the various methods of data collection. To comprehend the necessity for pre-processing Big Data and their issues.

Outcome

1. Identify the various sources of Big Data
2. Able to collect and store Big Data from various sources
3. Able to write Pig Scripts- Extract, Transform and Load the data on HDFS
4. Able to write Hive Scripts- Extract, Transform, Load and Analyse the data present in HDFS
5. Able to write scripts to extract data from structured and un-structured data for analytics
6. Able to extract and process semi and un-structured data using HBase

Unit- I


Unit-II


Unit-III

Apache Pig - Introduction - Pig features - Pig Architecture - Pig Execution modes, Pig Grunt shell and Shell commands. Pig Latin Basics: Data model, Data Types, Operators - Pig Latin Commands - Load & Store , Diagnostic Operators, Grouping, Cogroup, Joining, Filtering, Sorting, Splitting - Built-In Functions, User define functions. Pig Execution Modes: Batch Mode – Embedded Mode – Pig Execution in Batch Mode –Use cases - Map Reduce programs with Pig – Pig Vs SQL

Unit-IV

Hive: Introduction - Hive Features - Hive architecture -Hive Meta store - Hive data types -
Hive Tables - Table types - Creating database, Altering database, Create table, alter table, Drop table, Built-In Functions - Built-In Operators, User defined functions(UDFs), View, Pig Vs Hive.

**HiveQL** – Introduction, HiveQL Select, HiveQL – MapReduce using HiveQL OrderBy, Group By Joins, LIMIT, Distribute By , Cluster By - Sorting And Aggregation – Partitioning: Static & Dynamic partitioning – Index Creation - Bucketing – Analysis of MapReduce execution – Hive Optimization – Setting Hiivng Parameters. Comparison between MapReduce, Hive QL and SQL. UseCase: Implementation of MapReduce programs with HiveQL.

**Unit-V**

**Hbase** : HBasics, Features of HBase, Concepts, Clients, Example, Hbase Versus RDBMS, Limitations of HBase

**Big Data Privacy And Applications**: Data Masking – Privately identified Information (PII) – Privacy preservation in Big Data – Popular Big Data Techniques and tools – Applications-Social Media Analytics – Fraud Detection.

**References**

**Student Activity:**

**Case study I**: “BankAmeriDeals” provides cash-back offers to credit and debit-card customers based upon analyses of their prior purchases.

**Case Study II: GOOGLE**: Working with the U.S. Centers for Disease Control, tracks when users are inputting search terms related to flu topics, to help predict which regions may experience outbreaks.

**Case Study III**: Twitter data Analysis

**RECOMMENDED CO-CURRICULAR ACTIVITIES:**

(Therefore, curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)
A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity)

B. General

1. Group Discussion

2. Try to solve MCQ’s available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills

6. Individual and group project reports like “Movie Lens Data Analysis”, ”Youtube Click stream Data Analysis, Twitter Data Analysis, etc

7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work
II YEAR IV SEMESTER Paper-V

Data Acquisition and Analysis Lab

1. Hadoop Cluster Setup
   - Perform setting up and installing Hadoop in its three operating modes:
     - standalone
     - Pseudo distributed
     - Fully distributed
   - Use web-based tools to monitor your Hadoop setup.

2. Install and run Pig and also use Pig Shell commands to display the list of files in HDFS

3. Install and run Hive and also use Hive Shell commands to display the list of files in HDFS

4. Install and run HBase and also use HBase Shell commands to display the version and user of HBase

5. Use Hive to create, alter, and drop databases, tables, views, functions, and indexes

6. Write and execute Pig Script to Load data into a Pig relation without a schema

7. Write and execute Pig Script Load data into a Pig relation with a schema

8. Write a Pig script to find the word count in a text file

9. Write a Pig Script that mines weather data (NCDC). 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. Data available at: ftp://ftp.ncdc.noaa.gov/pub/data/noaa/.
   - Find average, max and min temperature for each year in NCDC data set
   - Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.

10. Write HiveQL command to create Weather table and to find the year-wise maximum temperature

11. Write a Pig Script to remove null and duplicate values from the given input file.

12. Write Pig scripts to implement filter, project, sort, group by, joins

13. Write Hive Query to create database, managed table, external table, join, index, view, etc

14. Create a table in HBase and insert the data into with Shell

15. Display the data present in a HBase table using Shell