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

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

I YEAR I SEMESTER PAPER– I MATHS FOR DATA SCIENCE

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

1. Gilbert Strang, *Linear Algebra and its Applications*. Thomson /Brooks Cole (Available in a Greek Translation).
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. *Calculus,* publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir . and Frank R. Giordano. Thomas's Calculus, Pearson
   12th Edition 2009.
5. David C. Lay, Linear Algebra and Its Applications, 4th Editoin.
6. Yourself saad, Iterative Methods for spare Linear Systems.

Student Activity:

1. Find the Eigenvectors of A = { 1 1 1 1,2 3 4 5,3 4 5 6}
2. Find orthogonal S =Spam{(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.

I YEAR II SEMESTER PAPER– II

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

Understanding data: Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and Un-Structured - Example Applications. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution.

Unit-2:

Introduction to R- Features of R - Environment - R Studio. Basics of R-Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - R help functions - R Data Structures - Control Structures. Vectors: Definition- Declaration - Generating - Indexing - Naming - Adding & Removing elements - Operations on Vectors - Recycling - Special Operators - Vectorized if- then else-Vector Equality – Functions for vectors - Missing values - NULL values - Filtering & Subsetting.

Unit-3:

Matrices - Creating Matrices - Adding or Removing rows/columns - Reshaping - Operations - Special functions on Matrices. Lists - Creating List – General List Operations - Special Functions - Recursive Lists. Data Frames - Creating Data Frames - Naming - Accessing -

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

Unit- 4:

Input / Output – Reading and Writing datasets in various formats - Functions - Creating User-defined functions - Functions on Function Object - Scope of Variables - Accessing Global, Environment - Closures - Recursion. Exploratory Data Analysis - Data Preprocessing - Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary - Handling Missing values and Outliers - Normalization
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

1.Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014. 2.Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, "Mining of Massive Datasets",
Cambridge University Press, 2014.
3.Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.
4.W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013. 5.Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, "Practical Data
Science Cookbook", Packt Publishing Ltd., 2014.
6. Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", Wiley, 2011.
7. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.

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

6. Individual and group project reports like "COVID-19 Analysis", "Estimated Quanrantain Period for Covid-19 Contacts", 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

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**"

	C1	*C2*	*C3*	*C4*	*C5*
C1	*0*	*12*	*13*	*8*	*20*
C2	*12*	*0*	*15*	*28*	*88*
C3	*13*	*15*	*0*	*6*	*9*
C4	*8*	*28*	*6*	*0*	*33*
C5	*20*	*88*	*9*	*33*	*0*

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

Section	Student no	M1	M2	M3
A	1	45	54	45
A	2	34	55	55
A	3	56	66	64
B	1	43	44	45
B	2	67	76	78
B	3	76	68	37

- 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.

	price	
	S1	S2
Roll	1.5	1
Bun	2	2.5
Cake	5	4.5
Bread	16	17

	demand.quantity			
	Roll	**Bun**	**Cake**	**Bread**
P1	6	5	3	1
P2	3	6	2	2
P3	3	4	3	1

- 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:

employee details as follows		
	emp_no:1	
	name: Ram	
	salary	
		basic: 10000
		hra: 2500
		da: 4000
	deductions	
		pf: 1100
		tax: 200
	total salary	
		gs(Gross Salary):
		ns(Net Salary)

- 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.

Hadoop Cluster Setup – SSH & Hadoop Configuration –HDFS Administering – Monitoring & Maintenance.

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 Pogram (WordCount program): The Driver Code - The Mapper class - The Reducer class.

Unit-V:

Writing first MapReduce Program - Hadoop's Streaming API - Using Eclipse for Rapid Development – YARN Vs MapReduce Advanced MapReduce Concepts: Partitioner – Combiner – Joins – Map-side Join – Reduce-side Join - Case Study: Weblog Analysis done using Mapper, Reducer, Combiner, Partitioner, etc.

References

1. Boris lublinsky, Kevin t. Smith Alexey Yakubovich, "Professional Hadoop Solutions".
   Wiley, ISBN : 9788126551071, 2015.
2. Chris Eaton, Dirk Deroos et al., "Understanding Big Data", McGraw Hill , 2010.
3. Tom White, "HADOOP" : The definitive Guide", O Reilly 2012.
4. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", PACKT publishing, 2013.

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:
   - Adding files and directories, List the files and directories
   - Retrieving files
   - Deleting files
   - Copying files from one folder to another in HDFS
   - 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 developping 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

Classification and Prediction - Basic Concepts of Classification and Prediction, General Approach to solving a classification problem- Logistic Regression - LDA - Decision Trees: Tree Construction Principle – Feature Selection measure – Tree Pruning - Decision Tree construction Algorithm, Random Forest, Bayesian Classification-Accuracy and Error Measures- Evaluating the Accuracy of the classifier / predictor- Ensemble methods and Model selection.

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

Cluster Analysis: Basic concepts and Methods – Cluster Analysis – Partitioning methods – Hierarchical methods – Density Based Methods – Grid Based Methods – Evaluation of Clustering – Advanced Cluster Analysis: Probabilistic model based clustering – Clustering High – Dimensional Data – Clustering Graph and Network Data – Clustering with Constraints- Outlier 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

1. Adelchi Azzalini, Bruno Scapa, "Data Analysis and Data mining" , 2nd Ediiton, Oxford Univeristy Press Inc., 2012.
2. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2011.
3. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th Edition, TataMc Graw Hill Edition , 2007.
4. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, Easter Economy Edition, PHI, 2006.
5. Joseph F Hair, William C Black etal, "Multivariate Data Analysis", Pearson Education, 7th edition, 2013.

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 MINIG 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

Introduction To Big Data Acquisition: Big data framework – fundamental concepts of Big Data Management and analytics – Current challenges and trends in Big Data Acquisition. Map Reduce Algorithm- Hadoop Storage [HDFS], Common Hadoop Shell commands - Anatomy of File Write and Read, NameNode, Secondary NameNode, and DataNode - Hadoop Configuration – Pig Configuration – Hive Configuration - HBase Configuration.

Unit-II

Data Collection And Transmission: Big data collection – Strategies – Types of Data Sources – Structured Vs Unstructured data – ELT vs ETL – storage infrastructure requirements – Collection methods – Log files – sensors – Methods for acquiring network data (Libcap-based and zero-copy packet capture technology) – Specialized network monitoring softwares (Wireshark, Smartsniff and Winnetcap) – Mobile equipments, Transmission methods, Issues.

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

1. Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications', John Wiley & Sons, 2014.
2. Tom White " Hadoop: The Definitive Guide" Third Edit on, O'reily Media, 2012.
3. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
4. Min Chen. Shiwen Mao, Yin Zhang. Victor CM Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.
5. Michael Minelli, Michele Chambers Ambiga Dhiraj, "Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends", John Wiley & Sons, 2013.
6. Raj. Pethuru " Handbook of Research on Cloud Infrastructures for Big Data Analytics", IGI Global.

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:

(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, 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