Contents

MASTER OF SCIENCEIN DATA SCIENCE (M.Sc. DS)

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset.

Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

NEHRU MEMORIAL COLLEGE (AUTONOMOUS),
PUTHANAMPATTI, TIRUCHIRAPPALLI – 621 007.

M.Sc. Data Science – Course Structure under CBCS
(For the candidates admitted from the academic year 2017-2018 onwards)

Sem	Course	Course Code	Subjects	Ins. Hrs/ Week	Crs	Exam Hrs	Marks		
							Int	Ext	Total
I	CC-I	17PDS101	Mathematics for Data Science	6	4	3	25	75	100
	CC-II	17PDS102	Probability and Statistics	6	4	3	25	75	100
	CC-III	17PDS103	Data Base Systems	6	4	3	25	75	100
	CC-IV	17PDS104	Data Science for Business	6	4	3	25	75	100
	CP-I	17PDS105P	Data Base Systems – Lab	6	4	3	40	60	100
	TOTAL			**30**	**20**				**500**
II	CC-V	17PDS106	Machine Learning	6	5	3	25	75	100
	CC-VI	17PDS107	Python and R Programming	6	5	3	25	75	100
	CP-II	17PDS108P	Python and R Programming – Lab	6	4	3	40	60	100
	CEC-I	17PDS109E	Any one from the list	6	5	3	25	75	100
	CEC-II	17PDS110E	Any one from the list	6	5	3	25	75	100
	TOTAL			**30**	**24**				**500**
III	CC-VII	17PDS111	NoSQL and Big Data Query Languages	6	5	3	25	75	100
	CC-VII	17PDS112	Multivariate Techniques for Data Analysis	6	5	3	25	75	100
	CP-III	17PDS113P	Hadoop–Lab I	6	4	3	40	60	100
	CEC-III	17PDS114E	Any one from the list	6	5	3	25	75	100
	CEC-IV	17PDS115E	Any one from the list	6	5	3	25	75	100
	TOTAL			**30**	**24**				**500**
IV	CC-VIII	17PDS116	Exploratory and Descriptive Data Analytics	6	5	3	25	75	100
	CC-IX	17PDS117	Cloud Computing	6	5	3	25	75	100
	CP-IV	17PDS118P	Hadoop–Lab II	6	4	3	40	60	100
	CEC-V	17PDS119E	Any one from the list	6	4	3	25	75	100
	Project	17PDS120P	Project Work	6	4	-	25	75	100
	TOTAL			**30**	**22**				**500**
GRAND TOTAL				**120**	**90**				**2000**

Note:

Course:

CC - Core Course
CP - Core Practical
CEC - Core Elective Course

Course Code:

17 denotes the **Academic Year**, PDS denotes **Post graduate Data Science**, 101 refers the **numerical code of subjects**, P refers **Practical** and E refers **Elective**.

LIST OF ELECTIVE COURSES (For 2017-2018)

Elective Course – I		Elective Course – II	
1	Data Mining and Data Warehousing	1	Big Data Analytics
2	Text Mining	2	Applied Statistics
Elective Course – III		**Elective Course – IV**	
1	Hadoop Eco Systems	1	Social Networks
2	Web Mining	2	Artificial Intelligence
Elective Course – V			
1	Data Analytics for Internet of Things		
2	Mobile computing		

Note:

Project : 100 Marks
Dissertation : 80 Marks
Viva Voice : 20 Marks

Core Papers – 10
Core Practical – 4
Elective Papers – 5
Project – 1

Sl. No	Subject	Internal	External
1	Theory	25	75
2	Practical	40	60

Note:

1. Separate passing minimum is prescribed for Internal and External
   a) The passing minimum for CIA shall be 40% out of 25marks (i.e. 10 marks)
   b) The passing minimum for University Examinations shall be 40% out of 75marks (i.e. 30 marks)
   c) The passing minimum not less than 50% in the aggregate.

CORE COURSE - I
MATHEMATICS FOR DATA SCIENCE

Objectives:

- To understand the concepts and operations of matrix algebra needed for computing graphics modeling
- To understand and apply the class of functions which transform a finite set into another finite set which relates to input output functions in computer science.
- To impart discrete knowledge in computer engineering through finite automata and Context free grammars

UNIT I

Matrix Algebra: Matrices, Rank of Matrix, Solving System of Equations - Eigen Values and Eigen Vectors-Inverse of a Matrix - Cayley Hamilton Theorem

UNIT II

Basic Set Theory: Basic Definitions - Venn Diagrams and set operations - Laws of set theory - Principle of inclusion and exclusion - partitions- Permutation and Combination - Relations- Properties of relations - Matrices of relations - Closure operations on relations - Functions - injective, subjective and objective functions.

UNIT III

Mathematical Logic: Propositions and logical operators - Truth table - Propositions generated by a set, Equivalence and implication - Basic laws- Some more connectives - Functionally complete set of connectives- Normal forms - Proofs in Propositional calculus - Predicate calculus.

UNIT IV

Formal Languages: Languages and Grammars-Phrase Structure Grammar- classification of Grammars-Pumping Lemma For Regular Languages-Context Free Languages.

UNIT V

Finite State Automata: Finite State Automata-Deterministic Finite State Automata(DFA), Non Deterministic Finite State Automata (NFA)-Equivalence of DFA and NFA-Equivalence of NFA and Regular Languages

Text Books:

1. Kenneth H.Rosen, “Discrete Mathematics and Its Applications”, Tata McGraw Hill, Fourth Edition, 2002 (Unit 1,2 & 3).
2. Hopcroft and Ullman, “Introduction to Automata Theory, Languages and Computation”, Narosa Publishing House, Delhi, 2002. (Unit 4,5).

CORE COURSE - II
PROBABILITYAND STATISTICS

Objective:

- To equip the students with a working knowledge of probability, statistics and modeling in the presence of uncertainties.

Unit – I

Measures of Central Tendency & Measures of Dispersion: Frequency Distribution - Histogram - Stem and leaf diagram - Frequency Polygon - Mean - Median - Mode - Range - Quartile Deviation - Mean Deviation - Box whisker plot -Standard Deviation - Coefficient of Variation

Unit – II

Skewness - Correlation & Regression: Karl Pearson's coefficient of Skewness - Bowley's coefficient of Skewness - Scatter Diagram - Karl Pearson's coefficient of correlation - Spearman's rank correlation coefficient - Linear Regression and Estimation - Coefficients of regression

Unit – III

Theory of Attributes & Hypothesis: Classes and Class Frequencies - Consistency of Data - Independence of Attributes - Association of Attributes. Hypothesis Type I and Type II errors. Tests of significance – Student's t-test:Single Mean - Difference of means - paired t-test - Chi-Square test:Test of Goodness of Fit - Independence Test

Unit – IV

Introduction to Probability & Conditional Probability: Random experiment - Sample space - Events - Axiomatic Probability - Algebra of events. Conditional Probability - Multiplication theorem of Probability - Independent events - Bayes' Theorem

Unit – V

Random variables & Mathematical Expectation: Discrete random variable - Continuous random variable - Two-dimensional random variable - Joint probability distribution. Expected value of a random variable - Expected value of a function of a random variable - Properties of Expectation and Variance - Covariance.

Text Book:

1. Fundamentals of Mathematical Statistics – 1st Edition S.C.Gupta, V.K.Kapoor , S Chand.
2. Introduction to Probability & Statistics – 4th Edition J.Susan Milton, Jesse C. Arnold Tata McGraw Hill.
3. Fundamentals of Statistics: 7th edition S C Gupta, Himalaya Publishing house.
4. Probability and Statistics with Reliability, Queuing, And Computer Science.
5. Applications (English) 1st Edition: Kishore Trivedi, PHI.
6. Schaum's Outlines Probability, Random Variables & Random Process 3rd Edition Tata McGraw Hill.
7. Probability & Statistics for Engineers: Dr J Ravichandran, Wiley.
8. Statistics for Business and Economics: DrSeema Sharma, Wiley.
9. Applied Business Statistics 7th Edition Ken Black, Wiley.

CORE COURSE - III
DATA BASE SYSTEMS

COURSE OBJECTIVES

- To understand the fundamentals of data models and conceptualize and depict a database system using ER diagram
- To make a study of SQL and relational database design.
- To know about data storage techniques and query processing.
- To impart knowledge in transaction processing, concurrency control techniques and recovery procedures.

Unit I

Introduction – purpose of database systems – Data Abstraction – Data models –Instances and schemes – Data independence – DDL – DML – Database users –ER model – Entity sets – Keys – ER diagram – relational model – Structure –Relations Algebra – Views.

Unit II

SQL – QBE – QUEL – Basic structure – various Operations – Relational database design issues – Normalization – normalization using functional and Multi value dependencies.

Unit III

File and system structure – overall system structure – file Organization – data dictionary – Indexing and hashing – basic concept B and B+ tree indices –Static and Dynamic hash functions.

Unit IV

Recovery and atomicity – failures classification and types – Transaction model and Log based recovery, schedules – serial and non-serial types – Serialization of schedules and views – testing for seriability – lock based protocols – time based protocols – validation techniques – multiple Granularity – multiversion schemes – insert and delete Operations.

Unit V

Distributed data bases – structure of distributed databases – Trade offs in Distributing the database – Transparency and autonomy – distributed query processing – recovery in distributed systems – commit protocols – security and integrity violations – authorization and views – security specification –encryption – Statistical databases.

Text Book(s):

1. Henry F.Korth, and Abraham Silberschatz,,Sudarshan “Database SystemConcepts”, McGraw Hill, 4th Edition, 2002

References:

1. Pipin C. Desai, “An Introduction to data base systems”, Galgotia PublicationsPrivate Limited, 1991.
2. C.J. Date, “An Introduction to Database Systems”, 3rd Edition, AddisonWesley 1983.

CORE COURSE – IV
DATA SCIENCE FOR BUSINESS

Objective:

- Able to apply fundamental algorithmic ideas to process data.
- Document and transfer the results and effectively communicate the findings using visualization techniques.

Unit – I

Introduction: Data-Analytic Thinking - The Ubiquity of Data Opportunities -Data Science, Engineering, and Data-Driven Decision Making - Data Processing and "Big Data" - Data and Data Science Capability as a Strategic Asset.

Business Problems and Data Science Solutions - From Business Problems to Data Mining Tasks -Supervised Versus Unsupervised Methods - The Data Mining Process.

Unit – II

Introduction to Predictive Modeling: From Correlation to Supervised Segmentation - Models, Induction, and Prediction - Supervised Segmentation -Visualizing Segmentations - Trees as Sets of Rules.

Unit – III

Overfitting and Its Avoidance – Generalization – Overfitting– Overfitting Examined -From Holdout Evaluation to Cross-Validation - Learning Curves – Overfitting Avoidance and Complexity Control.

Unit – IV

Similarity, Neighbors, and Clusters - Similarity and Distance - Nearest-Neighbor Reasoning - Some Important Technical Details Relating to Similaritiesand Neighbors – Clustering - Stepping Back: Solving a Business Problem Versus Data Exploration.

Unit – V

Decision Analytic Thinking: What Is a Good Model? - Visualizing Model Performance - Representing and Mining Text - Other Data Science Tasks and Techniques.

Text Book:

1. Foster Provost and Tom Fawcett, "*Data Science for Business*",Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2013, ISBN: 978-1-449-36132-7.

CORE PRACTICAL - I
DATA BASE SYSTEM - LAB

1. To implement Data Definition language
1.1. Create, alter, drop, truncate

1.2. To implement Constraints.
1.2.1. (a). Primary key, (b).Foreign Key, (c). Check, (d). Unique, (e).Null,(f). Not null , (g) . Default, (h). Enable Constraints, (i). Disable Constraints(j). Drop Constraints

2. To implement DML, TCL and DRL
2.1. (a).Insert, (b).Select, (c).Update, (d).Delete, (e).commit, (f).rollback,(g).save point, (i). Like'%', (j).Relational Operator.

3. To implement Nested Queries & Join Queries
3.1.(a). To implementation of Nested Queries
3.2.(b). (a) Inner join, (b).Left join, (c).Right join (d).Full join

4. To implement Views
4.1. (a). View, (b).joint view, (c).force view, (d). View with check option

CORE COURSE - V
MACHINE LEARNING

Objective:

- Understanding a very broad collection of machine learning algorithms and problems.
- To learn algorithmic topics of machine learning and mathematically deep enough to introduce the required theory.

UNIT I

Introduction: Learning Problems – Perspectives and Issues – Concept Learning – Version Spaces and Candidate Eliminations – Inductive bias – Decision Tree learning – Representation – Algorithm – Heuristic Space Search.

UNIT II

Neural Networks and Genetic Algorithms: Neural Network Representation – Problems – Perceptrons – Multilayer Networks and Back Propagation Algorithms – Advanced Topics – Genetic Algorithms – Hypothesis Space Search – Genetic Programming – Models of Evalution and Learning.

UNIT III

Bayesian and Computational Learning: Bayes Theorem – Concept Learning – Maximum Likelihood – Minimum Description Length Principle – Bayes Optimal Classifier – Gibbs Algorithm – Naïve Bayes Classifier – Bayesian Belief Network – EM Algorithm – Probability Learning – Sample Complexity – Finite and Infinite Hypothesis Spaces – Mistake Bound Model.

UNIT IV

Instant Based Learning: K- Nearest Neighbor Learning – Locally weighted Regression – Radial Bases Functions – Case Based Learning.

UNIT V

Advanced Learning: Learning Sets of Rules – Sequential Covering Algorithm – Learning Rule Set – First Order Rules – Sets of First Order Rules – Induction on Inverted Deduction – Inverting Resolution – Analytical Learning – Perfect Domain Theories – Explanation Base Learning – FOCL Algorithm – Reinforcement Learning – Task – Q-Learning – Temporal Difference Learning

Text Books:

1. MehryarMohri, AfshinRostamizadeh, AmeetTalwalkar, ―Foundations of Machine Learning (Adaptive Computation and Machine Learning Series)‖, MIT Press, 2012.
2. Tom M. Mitchell, ―Machine Learning‖, McGraw-Hill, 1 edition, 1997.

CORE COURSE – VI
PYTHON AND R PROGRAMMING

Objective:

- Understanding the basic concepts of Python and R.
- Visualizing the results of analytics effectively.

Unit I

Python Concepts, Data Structures, Classes: Interpreter – Program Execution – Statements – Expressions – Flow Controls – Functions - Numeric Types – Sequences - Strings, Tuples, Lists and - Class Definition – Constructors – Inheritance – Overloading – Text & Binary Files - Reading and Writing.

Unit II

Data Wrangling: Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String Manipulation, Regular Expressions.

Unit III

Visualization in Python:

Unit – IV

Data manipulation with R: Modes and classes of R objects - R object structure and mode conversion - Vector - Factor and its types - Missing values in R - Basic Data Manipulation - Acquiring data - Factor manipulation

Unit - V

R and Databases - R and different databases - R and Excel - R and MS Access - Relational databases in R - The file hash package - The ff package - R and sqldf - Data manipulation using sqldf

Text Books:

1. Mark Lutz, *"Programming Python"*, O'Reilly Media, 4th edition, 2010.
2. Mark Lutz, "Learning Python", O'Reilly Media, 5th Edition, 2013.
3. JaynalAbedin, "*Data Manipulation with R*", Published by Packt Publishing Ltd, January 2014, ISBN 978-1-78328-109-1.

CORE PRACTICAL – II
PYTHON AND R PROGRAMMING – LAB

Objective:

- Preparing and pre-processing data using python and R
- Visualizing the results of analytics effectively

List of Python Programs:

1. Write Python applications using variables, data types, strings and functions.
2. Write Python applications using loops, arrays, sorting and hashes.
3. Write Python applications using dictionaries, lists and tuples.
4. Twitter API Integration for tweet Analysis

List of R Programs:

1. Getting started with R
   a) Data types and data structures
   b) Flow control and looping
   c) Writing and calling functions
   d) Debugging and functions as objects
2. Pre-processing and Preparing data
   a) Raw data
   b) Clean data
3. Exploratory data analytics / Statistical and Machine learning methods
   a) Descriptive Statistics
   b) Hypothesis testing
   c) Linear Regression
   d) Logistic Regression
4. Data Visualization
   a) ggplot2

Case Study using R or Python

Sample Data sets may include any one from the following

1. tax data,
2. automotive data,
3. social media data,
4. stock market data,
5. employment data,
6. sports data, etc.

CORE COURSE - VII
NOSQL AND BIG DATA QUERY LANGUAGES

Objective:

- To learn the difference between conventional SQL query language and NoSQL basic concepts.
- To design and build big data query language.

Unit – I

Technology Evolution: Emerging Database Landscape: The Database Evolution – The Scale-Out Architecture – Database Workloads – Database Technologies for Managing the Workloads – Requirements for the Next Generation Data Warehouses – The Next Generation Database Architecture.

Unit – II

An Overview of NoSQL - Defining NoSQL - What NoSQL is and what it is not - List of NoSQL Databases - Characteristics of NoSQL -RDBMS approach – Challenges -NoSQL approach.

Unit – III

NoSQL Storage Types - Storage types - Column-oriented databases - Document store - Key-value store -Multi-storage type databases - Advantages and Drawbacks - Transactional application - Computational application - Web-scale application.

Unit IV

Introduction – installation and execution – PIG Data Model – PIG Latin – Input, Output-Relational Operators – User Defined Functions – Join Implementations – Integrating Pig with Legacy Code and Map Reduce –Developing and Testing Pig Latin Scripts – Embedding Pig Latin in Python – Evaluation Function in Java- Load Functions – Store Functions.

Unit V

Introduction – Data Types and File Formats – Databases in Hive – HiveQL: Data Definition – Data Manipulation – Queries – Views – Indexes – Schema Design.

Text Books:

1. SoumendraMohanty, MadhuJagadeesh, and HarshaSrivatsa, “Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics”, Published by Apress Media, 2013.
2. GauravVaish, “Getting Started with NoSQL”, Published by Packt Publishing Ltd., 2013,ISBN 978-1-84969-4-988.
3. Gates, *“A. Programming Pig”*, O'Reilly Media, Inc., 2011.
4. Capriolo, E., Wampler, D., &Rutherglen, J., *“Programming hive”,* O'Reilly Media, Inc., 2012.

CORE COURSE - VIII
MULTIVARIATE TECHNIQUES FOR DATA ANALYSIS

Objective:

- Data characteristics and form of Distribution of the Data Structures.
- Understanding the usage of multivariate techniques for the problem under the consideration.
- For drawing valid inferences and to plan for future investigations.

Unit - I

Meaning of Multivariate Analysis, Measurements Scales - Metric measurement scales and Non-metric measurement scales, Classification of multivariate techniques (Dependence Techniques and Inter-dependence Techniques), Applications of Multivariate Techniques in different disciplines.

Unit - II

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

Unit - III

Cluster Analysis: Objectives and Assumptions, Research design in cluster analysis, Deriving clusters and assessing overall fit (Hierarchical methods, Non Hierarchical Methods and Combinations), Interpretation of clusters and validation of profiling of the clusters.

Unit - IV

Discriminant Analysis- concept, objective and applications.Procedure for conducting discriminant analysis. Stepwise discriminate analysis and Mahalanobis procedure. Logic model.

Unit - V

Linear Programming problem - Formulation, graphical method, simplex method.Integer Programming.Transportation and Assignment problem.

Books for Study:

1. Joseph F Hair, William C Black, “Multivariate Data Analysis” , Pearson Education, 7th edition, 2013.
2. T. W. Anderson , “An Introduction to Multivariate Statistical Analysis, 3rd Edition”, Wiley, 2003.
3. William r Dillon, John Wiley & sons, “Multivariate Analysis methods and applications”, Wiley, 1984.
4. Naresh K Malhotra, Satyabhusan Dash, “Marketing Research Anapplied Orientation”, Pearson, 2011.
5. Hamdy A Taha, “Operations Research”, Pearson, 2012.
6. S R Yaday, A K Malik, “Operations Research”, Oxford, 2014.

CORE PRACTICAL – III
HADOOP LAB - I

Objective:

- To understand setting up of Hadoop Cluster
- To solve problems using Map Reduce Technique
- To solve Big Data problems

LIST OF EXERCISES

Hadoop Ecosystem:

1. Set up a pseudo-distributed, single-node Hadoop cluster backed by the Hadoop Distributed File System, running on Ubuntu Linux. After successful installation on one node, configuration of a multi-node Hadoopcluster(one master and multiple slaves).
2. Combiner Using MapReduce
3. Multiple Reducer using MapReduce
4. MapReduce application for word counting on Hadoop cluster
5. Installation of Hadoop Ecosystems
    a. Pig
    b. Hive
    c. Hbase
    d. Sqoop
    e. Zookeeper
    f. Flume

MongoDB

6. Installation of MongoDB
7. Reading CSV file and loading into MongoDB
8. Reading JSON file and Loading into MongoDB
9. Reading MongoDB and writing into MySQL

CORE COURSE - IX
EXPLORATORY AND DESCRIPTIVE DATA ANALYTICS

Objectives:

- To understand exploratory analytics and descriptive analytics

Unit I

Introduction – Data Analytics – EDA – Need for EDA – EDA – Objectives – Google Trend analysis – Explore trends – R Visualization – Packages – Lattice – ggplot2 – understanding plots – aesthetics - Statistical function – Histogram – Box Plot – Density Plot – Scatter Plots – Summarizing Data in R.

Unit II

Variable Analysis – One variable – Understanding outliers through – histogram, boxplot, density plot – dataset – pseudo dataset of Facebook Exploring two variables – Understanding Variables and relationships – scatter plots – correlations – condition means – Explore multivariate variables – Visualization of variables using aesthetics in R – Case study – Explore Diamond dataset for prize prediction.

Unit III

Data types – Categorical – Binary – ordinal – Nominal – Continuous – Discrete – Data dimensions – Univariate – bivariate – multivariate – Numerical Measures – Central Tendency – Mean – Median – Mode - Understanding data using central tendency – plotting histogram density plots and inference of plot - Variability Measure – Variance -Range - IQC -and standard Deviation – Sum of squares – Squared Deviations – Absolute Deviations – Identify outlier using Inter Quartile Range – Visualization using boxplot.

Unit IV

Data standardizing –Z Score – Negative Z Score – Continuous Distributions – Compute proportions – Relative Frequency histogram -Normalized Distribution using Ztable – Probability Distributions -Probability of mean – location of mean distribution - Sampling distributions — K lout Sampling Distribution –Understanding Shape of Distribution – Standard Error -Standard Deviation of sampling distribution –Ratio of Sampling Distribution -Central Limit Theorem R –Mean of sample means Advanced Analytics.

Unit V

Case Study –EDA analytics on dataset Movies –Social network using R –Prediction of Movie ratings –Descriptive Analytics on Movie Dataset.

Text Books:

1. Hadley Wickham, ggplot2: Elegant graphics for data analysis, Springer (2009) http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-98140-6/contents/.
2. Phil Spector, Data Manipulation with R, Springer, (2008) http://www.springer.com/statistics/computanional+statistics/book/978-0-387-74730-9
3. Leland Wilkinson, The Grammar of Graphics, Springer (2005), http://www.springerlink.com.proxy.lib.umich.edu/content/978-0-387-24544-7/contents/
4. Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, C A.

CORE COURSE – X
CLOUD COMPUTING

Objectives:

- To understand the concepts of cloud and utility computing
- To understand the various issues in cloud computing
- To appreciate the emergence of cloud as the next generation computing paradigm

Unit I

Evolution of Cloud Computing - System Models for Distributed and Cloud Computing - NIST Cloud Computing Reference Architecture - IaaS – On-demand Provisioning – Elasticity in Cloud - Examples of IaaS Providers - PaaS - Examples of PaaS Providers – SaaS - Examples of SaaS Providers - Public, Private and Hybrid Clouds – Google App Engine, Amazon AWS - Cloud Software Environments - Eucalyptus, Open Nebula, Open Stack, Nimbus

Unit II

Basics of Virtual Machines - Process Virtual Machines – System Virtual Machines – Emulation – Interpretation – Binary Translation - Taxonomy of Virtual Machines. Virtualization – Management Virtualization —Hardware Maximization –Architectures – Virtualization Management –Storage Virtualization –Network Virtualization

Unit III

Comprehensive Analysis –Resource Pool –Testing Environment –Server Virtualization – Virtual Workloads –Provision Virtual Machines –Desktop Virtualization –Application Virtualization –Work with AppV –Mobile OS for smart phones –Mobile Platform Virtualization –Collaborative Applications for Mobile platforms

Unit IV

Architectural Design of Compute and Storage Clouds -Inter Cloud Resource Management - Resource Provisioning and Platform Deployment -Global Exchange of Cloud Resources

Unit V

Security Overview –Cloud Security Challenges –Software as a Service Security –Security Governance –Risk Management –Security Monitoring –Security Architecture Design –Data Security –Application Security –Virtual Machine Security.

References:

1. Kai Hwang, Geoffrey C Fox, Jack G Dongarra, "Distributed and Cloud Computing, From Parallel Processing to the Internet of Things", Morgan Kaufmann Publishers, 2012.
2. John W.Rittinghouse and James F.Ransome, "Cloud Computing: Implementation, Management, and Security", CRC Press, 2010.
3. Toby Velte, Anthony Velte, Robert Elsenpeter, "Cloud Computing, A Practical Approach", McGraw-Hill Osborne Media, 2009.
4. Jim Smith, Ravi Nair, "Virtual Machines: Versatile Platforms for Systems and Processes", Elsevier/Morgan Kaufmann, 2005.
5. Danielle Ruest, Nelson Ruest, “Virtualization: A Beginner‟s Guide”, McGraw-Hill Osborne Media, 2009.
6. RajkumarBuyya, Christian Vecchiola, and ThamaraiSelvi, “Mastering Cloud Computing”, Tata McGraw Hill, 2013.

CORE PRACTICAL – IV
HADOOP LAB -II

Objective:

- To solve big data problems using Pig and Hive

LIST OF EXERCISES

Pig Programming Language

1. Components of Pig
2. Pig Data Model
3. Pig vs SQL
4. Filtering and Transformation of Data
5. Grouping and Sorting
6. Combining and Splitting
7. Processing Logs in Pig

Hive

8. Hive Query Language
9. Hive Data Models
10. Hive Functions
11. Process tweets in Hive

PROJECT WORK

Objective:

The student can get the knowledge to prepare the document, to implement tools for the specific problem and learn the industrial need programs for their placement.

PROJECTWORK

SL	Area of Work	Maximum Marks
1.	**PROJECT WORK:** (i) Plan of the Project	20
	(ii) Execution of the plan / Collection of data / Organization of materials / Fabrication Experimental study / Hypothesis, Testing etc., and Presentation of the report.	45
	(iii) Individual Initiative	15
2.	VIVA VOCE EXAMINATION	20
	TOTAL	**100**

Note: PASSING MINIMUM – 50 MARKS

ELECTIVE COURSE I
1.1 DATA MINING AND DATA WAREHOUSING

Objective:

- On successful completion of the course the students should have: Understood data mining techniques- Concepts and design of data warehousing.

Unit I

Introduction – What is Data mining – Data Warehouses – Data Mining Functionalities – Basic Data mining tasks – Data Mining Issues – Social Implications of Data Mining– Applications and Trends in Data Mining.

Unit II

Data Preprocessing: Why preprocess the Data? –Data Cleaning - Data Integration and Transformation – Data Reduction – Data cube Aggregation – Attribute Subset Selection Classification: Introduction – statistical based algorithms – Bayesian Classification. Distance based algorithms – decision tree based algorithms – ID3.

Unit III

Clustering: Introduction - Hierarchical algorithms – Partitional algorithms – Minimumspanning tree – K-Means Clustering - Nearest Neighbour algorithm.Association Rules:What is an association rule? – Methods to discover an association rule–APRIORI algorithm – Partitioning algorithm.

Unit IV

Data Warehousing: An introduction – characteristics of a data warehouse – Data marts – other aspects of data mart. Online analytical processing: OLTP & OLAP systems.

Unit V

Developing a data warehouse : Why and how to build a data warehouse – Data warehouse architectural strategies and organizational issues – Design consideration – Data content –meta data – distribution of data – tools for data warehousing – Performance Considerations

Text Books

1. Jiawei Han and Miceline Kamber, “Data Mining Concepts and Techniques”,Morgan Kaulmann Publishers, 2006.
2. Margaret H Dunham, “Data mining Introductory & Advanced Topics”, Pearson Education , 2003.
3. C.S.R.Prabhu, “Data Warehousing Concepts, Techniques, Products &Applications”,PHI, Second Edition.

References:

1. Pieter Adriaans, DolfZantinge, “Data Mining” Pearson Education, 1998.
2. Arun K Pujari, “Data Mining Techniques”,Universities Press(India) Pvt, 2003.
3. S.Rajashekharan, G A VijaylakshmiBhai,”Neural Networks,FuzzyLogic,andGenetic Algorithms synthesis and Application”, PHI

ELECTIVE COURSE - I
1.2 TEXT MINING

Objective:

- To understand the basic issues and types of text mining
- To appreciate the different aspects of text categorization and clustering

Unit I

Overview of text mining- Definition- General Architecture– Algorithms– Core Operations – Pre-processing– Types of Problems- basics of document classification- information retrieval-clustering and organizing documents- information extraction- prediction and evaluation-Textual information to numerical vectors -Collecting documents- document standardization-tokenization- lemmatization- vector generation for prediction- sentence boundary determination -evaluation performance.

Unit II

Text Categorization – Definition – Document Representation –Feature Selection - Decision Tree Classifiers - Rule-based Classifiers - Probabilistic and Naive Bayes Classifiers - Linear Classifiers- Classification of Linked and Web Data - Meta-Algorithms– Clustering – Definition- Vector Space Models - Distance-based Algorithms- Word and Phrase-based Clustering -Semi-Supervised Clustering

Unit III

Information retrieval and text mining- keyword search- nearest-neighbor methods- similarity-web-based document search- matching- inverted lists- evaluation. Information extraction-Text Summarization Techniques - Topic Representation - Influence of Context - Indicator Representations - Pattern Extraction - Apriori Algorithm – FP Tree algorithm

Unit IV

Probabilistic Models for Text Mining -Mixture Models - Stochastic Processes in Bayesian Nonparametric Models - Graphical Models - Relationship between Clustering, Dimension Reduction and Topic Modeling - Latent Semantic Indexing - Probabilistic Latent Semantic Indexing -Latent Dirichlet Allocation.

Unit V

Visualization Approaches - Architectural Considerations - Visualization Techniques in Link Analysis - Example- Mining Text Streams - Text Mining in Multimedia - Text Analytics in Social Media - Text Mining Applications and Case studies

Books for Study:

1. Sholom Weiss, NitinIndurkhya, Tong Zhang, Fred Damerau *"The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data"*, Springer, paperback 2010
2. Ronen Feldman, James Sanger -*"The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data"*-Cambridge University press, 2006.
3. Charu C. Aggarwal,ChengXiangZhai**,"***Mining Text Data"*, Springer; 2012.

ELECTIVE COURSE II
2.1 BIG DATA ANALYTICS

Objective:

- To give an insight and trends in Big Data and Analytics.
- To write Map-Reduce based Applications.

Unit – I

Introduction to Big Data – distributed file system–Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce.

Unit – II

Introduction to Hadoop and Hadoop Architecture - Big Data – Apache Hadoop&HadoopEcoSystem, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.

Unit – III

Hdfs, Hive and Hiveql, HbaseHdfs-Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting And Aggregating, Map Reduce Scripts, Joins & Sub queries, HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIG, Zookeeper , how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.

Unit – IV

Map Reduce - Introduction – distributed file system – algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce – Hadoop - Understanding the Map Reduce architecture - Writing HadoopMapReduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution.

Unit – V

Challenges in Big Data Analytics Methodology: Challenges in Big Data Analysis – Big Data Analytics Methodology – Analyse and Evaluate Business Use Case – Develop Business Hypotheses. Extract value from Big Data. – Data Scientist: The New Skill – The Big Data Workflow.

Books for Study:

1. Soumendra Mohanty, Madhu Jagadeesh, and Harsha Srivatsa, “Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics”, Published by Apress Media, 2013.
2. Anand Rajaraman, Jure Leskovec, Jeffery D. Ullman, “Mining of Massive Datasets”, Springer, July 2013.
3. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, “Practical Data Science Cookbook”, Packt Publishing Ltd., 2014.
4. Nathan Yau, “Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics”, Wiley, 2011.
5. Boris lublinsky, Kevin T. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, ISBN: 9788126551071, 2015.

ELECTIVE COURSE – II
2.2 APPLIED STATISTICS

Objectives

- To acquire more knowledge in applied statistics.

Unit I

Introduction: Meaning, Definition, Statistics as a data, Statistics as a Method. Importance, Functions, and Limitations of Statistics in Data Science. Finite and Infinite population, Hypothetical and existent population, census method, sample method, Random sampling (Non-Random sampling, simple Random Sampling, Restricted Random Sampling), Statistical Sampling, Systematic Sampling, Clustering Sampling, Judgment Sampling, Quota Sampling, Convenience or Churk Sampling, Statistical Errors, Absolute Error, Relative error, Reducing Sample Error, Test of Reliability Error.

Unit II

Classification and Tabulation: Overview of Classification, Statistical Series, Types of Series, Frequency Distribution, Continuous or Grouped Frequency Distribution. Magnitude of C lass intervals, Cumulative Frequency Distribution, Two Way Frequency Distribution. Measures of Central Tendency: Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode. Dispersion: Overview, Mean Deviation, Standard Deviation, Combined Standard Deviation.

Unit III

Correlation: Overview, Types of Correlation, Karl Pearson's Coefficient Correlation, Correlation and Probable Error, Rank Coefficient Correlation. Regression: Overview, Correlation and Regression, Graphical Method, Algebraic Method, Regression Line, Regression Equation, Mathematical Properties, Standard Error of Estimate. Association of attributes: Introduction, Classification, Correlation and Association, Types of Association, Comparison of Observed and Expected Frequencies, Yule's Coefficient of Association, Yule's Coefficient of Colligation, Pearsons' Coefficient of Contingency Partial Association.

Unit IV

Probability: Introduction, Mathematical Properties, Permutation, Combination, Trail, Sample Events, Sample Space, Mutually Exclusive Cases, Exhaustive Events, Independent Events, Dependent Events, Simple and Compound Events, Classical, Relative Frequency, Theory of Probability, Personalistic view of Probability, Addition and Multiplication Theorem, odds. Theoretical Distribution: Binominal Distribution, Obtaining Coefficient, Poison Distribution, Normal Distribution.

Unit V

Sampling Theory and test of significance: Introduction Estimation, Hypothesis, Standard Error, Test of Significance for Attributes, Test of Significance for Large Samples. Test of Significance for Small Samples. Chie Square Test: Introduction, Assumption, Uses of X2 Test of Goodness of fit, X2 Test of Independence, Yale's Correction, X2 test of Homogeneity, Additive Property.

Books for Study:

1. R.S.N. Pillai, Bagavathi, "Statistics Theory and Practice, S.Chand& Company, 2013
2. Douglas C. Montgomery, George C. Runger., "Applied Statistics for Engineers", John Wiley & Sons. Inc, 2003

ELECTIVE COURSE III
3.1 HADOOP ECO SYSTEMS

Objective:

UNIT I

Hadoop Clusters and the Hadoop Ecosystem: Topics - What is Hadoop Cluster? Pseudo Distributed mode, Type of clusters, Hadoop Ecosystem,Pig, Hive, Oozie, Flume, SQOOP.

UNIT II

Hive and HiveQL: Topics - What is Hive?, Hive vsMapReduce, Hive DDL – Create/Show/Drop Tables,Internal and External Tables, Hive DML – Load Files & Insert Data, Hive Architecture &Components, Difference between Hive and RDBMS, Partitions in Hive.

UNIT III

Apache SQOOP, Flume: Topics - Why and what is SQOOP? SQOOP Architecture, Benefits of SQOOP, ImportingData Using SQOOP, Apache Flume Introduction, Flume Model and Goals, Features ofFlume, Flume Use Case.

UNIT IV

NoSQL Databases: Topics - What is HBase? HBase Architecture, HBase Components, Storage Model of HBase,HBasevs RDBMS, Introduction to Mongo DB, CRUD, Advantages of MongoDB overRDBMS, Use case.

UNIT V

Oozie and Zookeeper:

Topics - Oozie – Simple/Complex Flow, Oozie Workflow, Oozie Components, Demo onOozie Workflow in XML, What is Zookeeper? Features of Zookeeper, Zookeeper DataModel.

TEXT BOOKS:

1. Tom White**, "**Hadoop: The Definitive Guide", Third Edition, O'Rielly**.**
2. Deepak Vohra, "Practical hadoop Ecosystem", Apress publications.
3. EelcoPlugge, Peter Membrey and Tim Hawkins, "The Definitive Guide toMongoDB", Apress publications.

ELECTIVE COURSE – III
3.2 WEB MINING

Objective:

- To understand the characteristics of the Internet and data mining
- To know about the web crawling algorithm implementation
- To study the web data collection and analysis of web data for new patterns

UNIT I

Introduction: World Wide Web, History of the Web and the Internet, What is DataMining? What is Web Mining? Introduction to Association Rule Mining, Supervised Learning & Unsupervised Learning. Information Retrieval and WebSearch: Basic Concepts of Information Retrieval, Information Retrieval Models, Relevance Feedback, Evaluation Measures, Text and Web Page Pre-Processing,Inverted Index and Its Compression, Latent Semantic Indexing, Web Search, Meta-Search: Combining Multiple Rankings, Web Spamming.

UNIT II

Social Network Analysis: Introduction, Co-Citation and Bibliographic Coupling,PageRank, HITS Algorithm, Community Discovery. Web Crawling:A Basic crawler Algorithm, Implementation Issues, Universal Crawlers, Focused crawlers,Topical Crawlers, Evaluation, Crawler Ethics and Conflicts.

UNIT III

Structured Data Extraction: Wrapper Generation, Preliminaries, Wrapper Induction, Instance-Based Wrapper Learning, Automatic Wrapper Generation:Problems, String Matching and Tree Matching, Building DOM Trees, Extractionbased on a Single List Page and Multiple Pages.

UNIT IV

Information Integration: Introduction to Schema Matching, Pre-Processing forSchema Matching, Schema -Level Matching, Domain and Instance-LevelMatching, Combining Similarities, 1: m Match, Integration of Web QueryInterfaces, Constructing a Unified Global Query Interface.

UNIT V

Web Usage Mining: Data Collection and Pre-Processing, Data Modelingfor WebUsage Mining, Discovery and Analysis of Web Usage Patterns, Recommender Systems and Collaborative Filtering, Query Log Mining, Computational Advertising.

TEXT BOOK

1. Wilbert Liu, Bing," Web Data Mining", 2nd Edition,Elseiver, 2011.

REFERENCE

1.SoumenChakrabarti, "Mining the Web", Morgan-Kaufmann Publishers, Elseiver, 2002.

ELECTIVE COURSE IV
4.1 SOCIAL NETWORKS

Objectives:

- To provide an all-round enrichment of knowledge in the area of social collaboration, mobility platform, business intelligence and analytics with cloud computing.

UNIT I

SMAC – What does it means? The convergence of 4 Disruptive Technologies - The Old and the New - SMAC as a game changer for the Indian IT industry

UNIT II

Social Collaboration in IT - Why is it important for businesses today? - What is driving social media in India? - How businesses are connecting with their customers on social networks - How businesses are using social networks to increase sales - Benefits of social media for businesses - Social Media Analytics

UNIT III

Importance of Mobility - Growth of Devices & Data Traffic - Applications: Types of Applications - Use of Mobile Applications - Global Enterprise Mobility Market Opportunity - Rise of India as an APP Superpower - How Indian IT Players can make the most out of Enterprise Mobility

UNIT IV

Analytics and its importance in IT space - What is the need for it? - Drivers for Big Data - How it all fits together? - Big Data Analytics - The value of Big Data analytics: Cases and return estimates - Big Data Market Opportunity.

UNIT V

Cloud Computing - Deployment Models - Service Types - Player Roles - The New Gold Rush : The Cloud Advantage - Migration to Clouds - Cloud in use - Market Size & Growth Estimates

Text Books:

1. Feroz Khan, "SMAC: Digital Discipline Building Enterprise", McGraw-Hill Education.

ELECTIVE COURSE – IV
4.2 ARTIFICIAL INTELLIGENCE & EXPERT SYSTEMS

Objective:

- To gain basic knowledge about AI, knowledge representation techniques and Expert system development to enable students to pursue research.

UNIT – I

Artificial Intelligence:AI problem – AI technique – level of the model – defining the problem – production systems – production system characteristics – Heuristic search techniques.

UNIT – II

Knowledge Representation: Representations and Mappings – issues in knowledge representation – predicate logic – representing knowledge using rules – symbolic reasoning under uncertainity.

UNIT – III

Natural language processing: Syntactic processing – semantic analysis – parallel and distributed AI – learning – learning in problem solving – explanation – based learning – discovery – analogy – formal learning theory.

UNIT – IV

Expert Systems: Introduction – architecture of expert systems – knowledge representation – decomposition / Hierarchy of knowledge – augmented transition networks – semantic analysis of knowledge.

UNIT – V

Knowledge Base and chaining functions: Modeling of uncertain reasoning – coherence of knowledge base – reductions of sets of rules – syntactic semantic analysis discursive grammar – the semiotic square – analyseachantieue narrative grammar – applications of semiotic theory of artificial intelligence.

Text Books:

1. Elaine Rich and Kevin Knight, ―Artificial Intelligence‖, Tata Mc-Graw Hill Edition, 2nd Edition,1995.
2. Eugene charniat and Drew McDermot, ―Introduction to Artificial Intelligence‖, Addison Wesley, 1985.
3. Jean-Louis Ermine, ―Expert Systems Theory and Practice‖, Prentice-Hall of India Pvt. Ltd., 2001.

ELECTIVE COURSE V
5.1 DATA ANALYTICS FOR INTERNET OF THINGS

Objective:

- To learn the concepts about Internet of things
- To understand and implement smart systems

UNIT I

Big Data Platforms for the Internet of Things: network protocol- data dissemination –current state of art- Improving Data and Service Interoperability with Structure, Compliance, Conformance and Context Awareness: interoperability problem in the IoT context- Big Data Management Systems for the Exploitation of Pervasive Environments - Big Data challenges and requirements coming from different Smart City applications

UNIT II

On RFID False Authentications: YA TRAP – Necessary and sufficient condition for false authentication prevention - Adaptive Pipelined Neural Network Structure in Self-aware Internet of Things: self-healing systems- Role of adaptive neural network- Spatial Dimensions of Big Data: Application of Geographical Concepts and Spatial Technology to the Internet of Things- Applying spatial relationships, functions, and models

UNIT III

Fog Computing: A Platform for Internet of Things and Analytics: a massively distributed number of sources - Big Data Metadata Management in Smart Grids: semantic inconsistencies – role of metadata

UNIT IV

Toward Web Enhanced Building Automation Systems: heterogeneity between existing installations and native IP devices - loosely-coupled Web protocol stack –energy saving in smart building- Intelligent Transportation Systems and Wireless Access in Vehicular Environment Technology for Developing Smart Cities: advantages and achievements- Emerging Technologies in Health Information Systems: Genomics Driven Wellness Tracking and Management System (GO-WELL) – predictive care – personalized medicine

UNIT V

Sustainability Data and Analytics in Cloud-Based M2M Systems - potential stakeholders and their complex relationships to data and analytics applications - Social Networking Analysis - Building a useful understanding of a social network - Leveraging Social Media and IoTto Bootstrap Smart Environments: lightweight Cyber Physical Social Systems - citizen actuation

Text Books:

1. Stackowiak, R., Licht, A., Mantha, V., Nagode, L.," Big Data and The Internet of Things Enterprise Information Architecture for A New Age", Apress, 2015.
2. Dr. John Bates , "Thingalytics - Smart Big Data Analytics for the Internet of Things", John Bates, 2015.

ELECTIVE COURSE – V
5.2 MOBILE COMPUTING

Objective:

- To understand the concepts of pervasive computing and learn the technologies for developing applications on mobile platforms.

UNIT – I

Wireless networks- emerging technologies- Blue tooth, WiFi, WiMAX, 3G ,WATM.-Mobile IP protocols -WAP push architecture-WML scripts and applications.

UNIT – II

Mobile computing environment—functions-architecture -design considerations ,content architecture -CC/PP exchange protocol ,context manager. Data management in WAECodafile system- caching schemes- Mobility QOS. Security in mobile computing.

UNIT – III

Handoff in wireless mobile networks-reference model-handoff schemes. Location management in cellular networks - Mobility models- location and tracking management schemes- time, movement ,profile and distance based update strategies. ALI technologies.

UNIT – IV

Pervasive Computing- Principles, Characteristics- interaction transparency, context aware, automated experience capture. Architecture for pervasive computing- Pervasive devices- embedded controls.- smart sensors and actuators -Context communication and access services

UNIT – V

Open protocols- Service discovery technologies- SDP, Jini, SLP, UpnP protocols–data synchronization- SyncML framework - Context aware mobile services -Context aware sensor networks, addressing and communications - Context aware security.

Text Books:

1. Ivan Stojmenovic ,“Handbook of Wireless Networks and Mobile Computing”, John Wiley & sons Inc, Canada, 2002.
2. Asoke K Taukder,Roopa R Yavagal, “Mobile Computing”, Tata McGraw Hill Pub Co., New Delhi, 2005.
3. SengLoke, “Context-Aware Computing Pervasive Systems”, Auerbach Pub., New York, 2007.
4. UweHansmannetl, “Pervasive Computing”, Springer, New York,2001.