SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

III YEAR VI SEMESTER

(Cluster 1) Paper-VIII: Elective

A-1 Foundations of Data Science

Course Code: BSCS68A1T

Course Objectives

Modern scientific, engineering, and business applications are increasingly dependent on data, existing traditional data analysis technologies were not designed for the complexity of the modern world. Data Science has emerged as a new, exciting, and fast-paced discipline that explores novel statistical, algorithmic, and implementation challenges that emerge in processing, storing, and extracting knowledge from Big Data.

Course Outcomes

1. Able to apply fundamental algorithmic ideas to process data.
2. Learn to apply hypotheses and data into actionable predictions.
3. Document and transfer the results and effectively communicate the findings using visualization techniques.

UNIT I

INTRODUCTION TO DATA SCIENCE : Data science process – roles, stages in data science project – working with data from files – working with relational databases – exploring data – managing data – cleaning and sampling for modelling and validation – introduction to NoSQL.

UNIT II

MODELING METHODS : Choosing and evaluating models – mapping problems to machine learning, evaluating clustering models, validating models – cluster analysis – K- means algorithm, Naïve Bayes – Memorization Methods – Linear and logistic regression – unsupervised methods.

UNIT III

INTRODUCTION TO R Language: Reading and getting data into R – ordered and unordered factors – arrays and matrices – lists and data frames – reading data from files.

UNIT IV

PROBABILITY DISTRIBUTIONS in R - Binomial, Poisson, Normal distributions. - Manipulating objects - data distribution.

UNIT V

DELIVERING RESULTS : Documentation and deployment – producing effective presentations– Introduction to graphical analysis – plot() function – displaying multivariate data – matrix plots – multiple plots in one window - exporting graph - using graphics parameters in R Language. Case studies.

Reference Books

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.

SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

(Cluster 1) Paper-VIII: Elective –A-1
Foundations of Data Science
Lab Course Code: BSCS68A1P

Objectives:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis. **Outcomes**:

1) At end student will learn to handle the data through R.
2) Student will familiar with loading and unloading of packages.

I. Installing R and R studio
II. Basic Operations in r
1. Arthematic Operations
2. Comments and spacing
3. Logical Operators - <, <=, >, >=, = , !=, &&, 1

III.
1. Getting data into R, Basic data manipulation
2. Vectors, Materials, operation on vectors and matrices.

IV. IV.

1. Basic Plotting
2. Quantitative data
3. Frequency plots
4. Box plots
5. Scatter plot
6.Categorial data
7. Bar charts
8. Pie charts

V. Loops and functions

1. if, if else, while, for break, next, repeat.

2. Basic functions- Print(), exp( ), Log( ), sqrt( ), abs( ), sin( ), Cos( ), tan( ), factorial( ), rand ().

SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

III B.Sc Computer Science VI-Semester

MODEL QUESTION PAPER

Paper - VIII: Elective – II: (Cluster A) A1. Foundations of Data Science

Time : 3 Hours Max.Marks : 75

SECTION – A

Answer any **FIVE** of the following questions. **5 x 5 M = 25 M**

1. What is sampling for modelling and validation?
2. Explain evaluating clustering model.
3. What is linear regression?
4. Write about k-means algorithm.
5. Differentiate supervised and unsupervised learning.
6. What are matrix plots?
7. What is poisson distribution?
8. Write about Lists in 'R' language.

SECTION - B

Answer **ALL** the following questions. **5 x 10 M = 50 M**

9. a) What is Data Science? Explain its roles and stages in Data Science. (or)

b) Explain different properties and characteristics of Relational Databases.

10. a) What is Machine Learning? What is its role in Data Science?

(or)

b) What is Cluster Analysis? Explain K-means algorithm.

11. a) Explain the characteristics of ‘R’ language? How do we read data into ‘R’?

(or)

b) Explain Arrays and Matrices in ‘R’ language.

12. a) Briefly explain Binomial Distribution.

(or)

b) What is normal distribution? Explain its representation in ‘R’ language with an example.

13. a) Explain plot() function in ‘R’ language.

(or)

b) Explain about Graph Exploration in ‘R’ language.

SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

III YEAR VI SEMESTER
(Cluster 1) Paper-VIII: Elective –A-2
BIG DATA TECHNOLOGY
Course Code: BSCS68A2T

Course Objective

The Objective of this course is to provide 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.

Course Outcome

1. Learn tips and tricks for Big Data use cases and solutions.

2. Learn to build and maintain reliable, scalable, distributed systems with Apache Hadoop.

3. Able to apply Hadoop ecosystem components.

UNIT I

INTRODUCTION TO BIG DATA: Introduction – distributed file system – Big Data and its importance, Four V's in bigdata, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce.

UNIT II

INTRODUCTION HADOOP : Big Data – Apache Hadoop & Hadoop EcoSystem – Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.

UNIT- III

HADOOP ARCHITECTURE: Hadoop Architecture, Hadoop Storage: HDFS, Common Hadoop Shell commands , Anatomy of File Write and Read., NameNode, Secondary NameNode, and DataNode, Hadoop MapReduce paradigm, Map and Reduce tasks, Job, Tasktrackers - Cluster Setup – SSH & Hadoop Configuration – HDFS Administering – Monitoring & Maintenance.

UNIT-IV

HIVE AND HIVEQL, HBASE:-Hive Architecture and Installation, Comparison with Traditional Database, HiveQL - Querying Data - Sorting And Aggregating, Map Reduce Scripts, Joins & Subqueries,

UNIT-V

HBase concepts- Advanced Usage, Schema Design, Advance Indexing - Zookeeper - how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.

Reference Books

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.

2 .Big Data Black Book(Covers Hadoop 2, Map Reduce, Hive, Yarn, Pig & Data Visualization)- Dream Tech Publications

3. Chris Eaton, Dirk deroos et al. , "Understanding Big data ", McGraw Hill, 2012.
4. Tom White, "HADOOP: The definitive Guide" , O Reilly 2012.
5. Vignesh Prajapati, "Big Data Analytics with R and Haoop", Packet Publishing 2013.
6. Tom Plunkett, Brian Macdonald et al, "Oracle Big Data Handbook", Oracle Press, 2014.
7. Jy Liebowitz, "Big Data and Business analytics",CRC press, 2013.

SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

Paper-VIII: Elective –A-2

BIG DATA TECHNOLOGY LAB

Course Code: BSCS68A2P

Objectives:

- Understand what Hadoop is
- Understand what Big Data is
- Learn about other open source software related to Hadoop

Outcomes:

i) Get help on the various Hadoop commands

ii) Observe a Map-Reduce job in action

1. Implement the following Data Structures in Java

a) Linked Lists

b) Stacks

c) Queues

d) Set

e) Map

2. (i) Perform setting up and Installing Hadoop in its three operating modes:
Standalone Pseudo distributed

Fully distributed

(ii) Use the web based tools to monitor your Hadoop setup.

3. Implement the following file management tasks
in Haddop. Adding files and directories

Retrieving
files Deleting
files

SRI A S N M GOVERNMENT COLLEGE, PALAKOL, W.G. DT
(Affiliated to Adikavi Nannaya University, Rajahmundry)
(Accredited with NAAC "B" Grade with 2.61 CGPA points)

III B.Sc Computer Science VI-Semester

MODEL QUESTION PAPER

Paper - VIII: Elective – II: (Cluster A) A2. BIG DATA TECHNOLOGY

Time : 3 Hours Max.Marks : 75

SECTION - A

Answer any **FIVE** of the following questions. **5 x 5 M = 25 M**

1. Write the importance of Big Data.
2. Write some applications of Map Reduce.
3. What is Data Serialization?
4. Write about inputs and outputs of Map Reduce.
5. What is HDFS?
6. Write about Map Reduce Paradigm.
7. Write about joins in HiveQL
8. Write about Zookeeper.

SECTION - B

Answer **ALL** the following questions. **5 x 10 M = 50 M**

9. a) What is distributed file system? Explain the significance of four V's in Big Data.
   (or)

   b) Explain briefly about Big Data applications.

10. a) What is Big Data? Explain the characteristics of APACHE Hadoop.

    (or)

b) Explain how do we move data in and out of Hadoop.

11. a) Explain Hadoop architecture.
(or)

b) Explain Hadoop shell commands.

13. a) Explain Hive architecture and installation. (or)

b) Compare Traditional data file with Hive.

14. a) Explain the concepts of HBase and write its uses. (or)

b) Explain how a schema design is done in HBase.
