CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
EET478	BIG DATA ANALYTICS	PEC	2	1	0	3

Preamble: This course is offered to introduce fundamental algorithmic ideas in processing data. The preliminary concepts of Hadoop and Map Reduce are included as part of this course.

Prerequisite: Nil

Course Outcomes: After the completion of the course the student will be able to

CO 1	Explain the key concepts of data science.
CO 2	Describe big data and use cases from selected business domains
CO 3	Perform big data analytics using Hadoop and related tools like Pig and Hive.
CO 4	Perform preliminary analytics using R language on simple data sets.
CO 5	Differentiate various learning approaches in machine learning to process data, and to interpret the concepts of supervised and unsupervised learning

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO 1	3											2
CO 2	3											2
CO 3	3	2	2		3							2
CO 4	3	2			3							2
CO 5	3	2			3							2

Assessment Pattern

Bloom's Category	Continuous Assessment Tests		End Semester Examination
	1	2	
Remember	15	15	30
Understand	25	25	50
Apply	10	10	20
Analyse			
Evaluate			
Create			

Mark distribution

Total Marks	CIE	ESE	ESE Duration
150	50	100	3 hours

Continuous Internal Evaluation Pattern:

Attendance	: 10 marks
Continuous Assessment Test (2 numbers)	: 25 marks
Assignment/Quiz/Course project	: 15 marks

End Semester Examination Pattern: There will be two parts; Part A and Part B. Part A contain 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer all questions. Part B contains 2 questions from each module of which student should answer any one. Each question can have maximum 2 sub-divisions and carry 14 marks.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Explain the main categories of data that we come across in data science. (K1)
2. Summarize distributed file system with examples. (K1)
3. List the significance of data science. (K2)

Course Outcome 2 (CO2)

1. What are the three characteristics of Big Data, and what are the main considerations in processing Big Data?(K1)
2. Explain Big Data Analytics Lifecycle. (K1)
3. Explain Apache Hadoop ecosystem. (K1)

Course Outcome 3(CO3):

1. Demonstrate the map reduce execution flow to perform word count on data set.(K3)
2. Explain the stages of Map Reduce. (K2)
3. Write short notes on Pig and Hive. (K1)

Course Outcome 4 (CO4):

1. How do you list the preloaded datasets in R? (K2)
2. Use R to find the highest common factor of two numbers. (K3)
3. Why is R useful for data science? (K2)

Course Outcome 5 (CO5):

1. Mention the difference between Data Mining and Machine learning? (K2)
2. What are the different Algorithm techniques in Machine Learning? (K2)
3. Give a popular application of machine learning that you see on day-to-day basis? (K2)

Model Question Paper
QP CODE: **Reg No:_____**
PAGES:3 **Name :_____**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY EIGHTH SEMESTER
B.TECH DEGREE EXAMINATION, MONTH & YEAR
Course Code: EET478
Course Name: BIG DATA ANALYTICS
Max. Marks: 100 Duration: 3 Hours
(2019-Scheme)

PART A
(Answer all questions, each question carries 3 marks)

1. List any six Data Science applications.
2. Briefly explain the data transformation step in the process of Data Science.
3. Explain the important characteristics of Bigdata.
4. List the functions of Namenode in HDFS.
5. Identify the need of MapReduce Partitioner in Hadoop.
6. Differentiate between Hadoop MapReduce and Pig.
7. In R how missing values are represented.
8. How you can import Data in R.
9. Discuss any four examples of machine learning applications.
10. Describe the applications of clustering in various domains.

(10x3 = 30 marks)

PART B
(Answer one full question from each module, each question carries 14 marks)

MODULE I

11.a) Illustrate with an example different stages of data science project.
b. Categorise the different roles associated with a data analysis project. (10+4 =14 marks)

Or

12. a) Explain the data cleansing subprocess of data science process.
b) Discuss in detail about Exploratory Data analysis. (8+6 =14 marks)

MODULE II

13.a) Explain the core components of Apache Hadoop.
b) Write short note on YARN. (8+6 = 14 marks)

Or

14. a) Explain read and write operations in HDFS.
b) What are Blocks in HDFS Architecture. (10+4 = 14 marks)

MODULE III

15.a) With a neat diagram, explain MapReduce architecture?
b) Describe the stages of MapReduce with an example. (5+9 = 14 marks)

Or

16. a) Write short note on Pig and HIVE.
b) Compare NoSQL & RDBMS (10+4 = 14 marks)

MODULE IV

17.a) Explain data frames in R. Illustrate attach (), detach () and search () functions in R.

b) Explain any three functions in R to visualize a single variable. (8+6 = 14 marks)

Or

18. a) What are the data structures in R that is used to perform statistical analyses and create graphs?

b) Mention how you can produce co-relations and covariances with example?

(9+5 = 14 marks)

MODULE V

19.a) Distinguish between classification and regression with an example.

b) Describe in detail with examples (i) Supervised Learning(ii) Unsupervised Learning (iii) Reinforcement Learning. (5+9 = 14 marks)

Or

20. a) Is regression a supervised learning technique? Justify your answer. Compare regression with classification with examples.?

b) Illustrate K means clustering algorithm with an example.? (8+6 = 14 marks)

Syllabus

Module I-Data science in a big data world: Benefits and uses of data science and big data-Facets of data-the big data ecosystem and data science-Data science process-roles-stages in data science project- Defining research goals-Retrieving data-Cleansing, integrating, and transforming data- Data Exploration-Data modelling - Presentation and automation.

(6 hours)

Module II-Big Data Overview–the five V's of big data-State of the Practice in Analytics-Examples of Big Data Analytics-Apache Hadoop and the Hadoop Ecosystem-HDFS-Design of HDFS, HDFS Concepts-Daemons-Reading and Writing Data-Managing File system Metadata- Map Reduce-The Stages of Map Reduce -Introducing Hadoop Map Reduce-Daemons-YARN (8 hours)

Module III-Analysing the Data with Hadoop using Map and Reduce-Developing a Map Reduce Application-Anatomy of a Map Reduce Job- Scheduling-Shuffle and Sort - Task execution.

Big data Management Tools: PIG- : Introduction to PIG, Execution Modes of Pig,Pig Latin, HIVE: Hive Architecture, HIVEQL, Introduction to NoSQL. (Introduction only)

(7 hours)

Module IV -Review of Basic Analytic methods using R- Introduction to R -Data Import and Export -Attribute and Data Types - ordered and unordered factors-arrays and matrices-lists and data frames -Descriptive Statistics-Exploratory Data Analysis-Dirty Data-Visualizing a Single Variable-Examining Multiple Variables-statistical models in R-Graphical Procedures-High-level plotting commands-Low-level plotting commands.

(7 hours)

Module V -Machine learning -Introduction to Machine Learning, Examples of Machine Learning applications-Supervised Learning- Regression – Single variable, Multi variable-Classification – Logistic Regression- Unsupervised Learning - Clustering: K-means-Reinforcement Learning-Model Selection and validation-k-Fold Cross Validation-Measuring classifier performance- Precision, recall

(7 hours)

Text/ Reference Books

1. Davy Cielen, Arno D. B. Meysman, and Mohamed Ali ,"Introducing Data Science - Big data, machine learning, and more, using Python tools" , Dreamtech Press 2016
2. Michael Minelli, Michelle Chambers, and AmbigaDhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses",Wiley,2013
3. EMC Education Services, "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley ,January 2015
4. Tom White,"Hadoop: The Definitive Guide", Third Edition, O'Reilley,2012.
5. Eric Sammer,"Hadoop Operations",O'Reilly Media, Inc ,2012
6. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
7. "Programming Pig", Alan Gates, O'Reilley,2011.

8. Ethem Alpaydın, "Introduction to Machine Learning (Adaptive Computation and Machine Learning)", MIT Press, 2004.
9. Shai Shalev-Shwartz, Shai Ben-David, "Understanding Machine Learning: From Theory to Algorithms", Cambridge University Press, 2014
10. Christopher Bishop, "Pattern Recognition and Machine Learning", Springer, 2007.
11. Matloff, Norman," The art of R programming: A tour of statistical software design". No Starch Press, 2011.
12. Crawley, Michael J. The R book. John Wiley & Sons, 2012.
13. Sourabh Mukherjee, Amit Kumar Das and Sayan Goswami, " Big Data Simplified", Pearson, 1st edition, 2019.
14. Murtaza Haider, "Getting Started with Data Science", Fist Edition, Kindle Edition, IBM Press, 2015.
15. Thomas Erl, Wajid Khattak and Paul Buhler " Big Data Fundamentals:Concepts, Drivers and Techniques", Prentice Hall, Pearson Service, 2016.

Course Contents and Lecture Schedule

No	Topic	No. of Lectures
1	**Module I Data science in a big data world**	**6 hours**
1.1	Data science in a big data world, Benefits and uses of data science and big data-Facets of data	1
1.2	the big data ecosystem and data science-Data science process-roles	1
1.3	Defining research goals-Retrieving data	1
1.4	Cleansing, integrating, and transforming data	1
1.5	Data Exploration	1
1.6	Data modelling - Presentation and automation.	1
2	**Module II -Big Data Overview**	**8 hours**
2.1	the five V's of big data-State of the Practice in Analytics-Examples of Big Data Analytics	1
2.2	Apache Hadoop and the Hadoop Ecosystem- HDFS	2
2.3	Design of HDFS- HDFS Concepts-Daemons-Reading and Writing Data - Managing Filesystem Metadata	2
2.4	Map Reduce-The Stages of MapReduce -Introducing Hadoop MapReduce-Daemons	2
2.5	YARN	1
3	**Module III - Analysing the Data with Hadoop**	**7 hours**
3.1	**Analysing the Data with Hadoop using Map and Reduce**-Developing a Map Reduce Application	1
3.2	Anatomy of a Map Reduce Job- Scheduling-Shuffle and Sort - Task execution	2
3.3	Bigdata Management Tools: PIG- : Introduction to PIG, Execution Modes of Pig,Pig Latin	2
3.4	HIVE: Hive Architecture, HIVEQL,	1

3.5	Introduction to NoSQL	1
4	**Module IV -Review of Basic Analytic methods using R**	**7 hours**
4.1	Introduction to R -Data Import and Export -Attribute and Data Types - ordered and unordered factors-arrays and matrices	2
4.2	lists and data frames -Descriptive Statistics	1
4.3	Exploratory Data Analysis -Dirty Data	1
4.4	Visualizing a Single Variable-Examining Multiple Variables	1
4.5	statistical models in R-.	1
4.6	Graphical Procedures-High-level plotting commands-Low-level plotting commands	1
5	**Module V - Machine learning**	**7 hours**
5.1	Introduction to Machine Learning, Examples of Machine Learning applications	1
5.2	Supervised Learning- Regression – Single variable, Multi variable	2
5.3	Classification – Logistic Regression	1
5.4	Unsupervised Learning - Clustering: K-means	1
5.5	Model Selection and validation-k-Fold Cross Validation	1
5.6	Measuring classifier performance- Precision, recall	1