KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
Deemed to be University
BHUBANESWAR-751024

School of Computer Engineering
Autumn Semester 2021-22

Course Handout

1. **Course code** : CS 3032
2. **Course Title** : Big Data
3. **LTP Structure** :

L	T	P	Total	Credit
3	0	0	3	3

4. **Course Coordinator** : Dr. Sarita Tripathy
5. **Class start date** : 05/07/2021
6. **Course offered to the School** : Computer Engineering
7. **Course Faculty** : School of Computer Engineering

Sr#	Faculty
1	Dr. Sarita Tripathy(C)
2	Dr. Mainak Bandhopadhyaya
3	Prof.Abhaya Kumar Sahoo
4	Prof.Manas Ranjan Biswal
5	Prof. Ipsita Paul

Class Timing:

CSE G1 Class Timing	IT-G1 Class Timing
Monday (12:00-1:00PM)	Monday-(11:00-12:00Noon)
Wednesday (8:00-9:00AM)	Thursday-(1:00-2:00PM)
Friday (3:00-4:00PM)	Friday(10:00-11AM)

8. **Course Objective:**

- To understand the concepts and principles of big data.
- To explore the big data stacks and the technologies associated with it.
- To evaluate the different NOSQL databases and frameworks required to handle the big data.
- To apply the techniques for analysis of big data using R tool.
- To formulate the concepts, principles and techniques focussing on the applications to industry and real world experience.

9. **Course Outcome:**

CO #	Detail
CO1	Identity the basic characteristics of big data and deploy a structured life cycle approach.
CO2	Classify and examine the data under big data stack and associated technologies.
CO3	Evaluate big data technologies to analyze big data and create models.
CO4	Compose efficient data analysis techniques using R tools.
CO5	Understand and interpret the big data through various visualization techniques
CO6	Contextually integrate and correlate large amounts of information to gain faster insights for real time scenarios.

10. Course Contents

The course focuses on basic and essential topics in Big Data.

Unit #	Unit	Detailed Area
1	Introduction to Big Data	Importance of Data, Characteristics of Data Analysis of Unstructured Data, Combining Structured and Unstructured Sources. Introduction to Big Data Platform – Challenges of conventional systems – Web data – Evolution of Analytic scalability, analytic processes and tools, Analysis vs reporting – Modern data analytic tools, Types of Data, Elements of Big Data, Big Data Analytics, Data Analytics Lifecycle.
2	Big Data Technology Foundations	Exploring the Big Data Stack, Data Sources Layer, Ingestion Layer, Storage Layer, Physical Infrastructure Layer, Platform Management Layer, Security Layer, Monitoring Layer, Analytics Engine, Visualization Layer, Big Data Applications, Virtualization. Introduction to Streams Concepts – Stream data model and architecture – Stream Computing, Sampling data in a stream – Filtering streams, Counting distinct elements in a stream.
3	Big Data Tools	NOSQL, MapReduce – Hadoop, HDFS, Hive, MapR – Hadoop - YARN - Pig and PigLatin, Jaql - Zookeeper - HBase, Cassandra-Oozie, Lucene- Avro, Mahout. Hadoop Distributed file systems.
4	Data Analysis Through R	Exploring R: Exploring Basic Features of R, Programming Features, Packages, Exploring RStudio, Handling Basic Expressions in R, Basic Arithmetic in R, Mathematical Operators, Calling Functions in R, Working with Vectors, Creating and Using Objects, Handling Data in R Workspace, Creating Plots, Using Built-in Datasets in R, Reading Datasets and Exporting Data from R, Manipulating and Processing Data in R.
5	Frameworks And Visualization	Distributed and Parallel Computing for Big Data, Visualizations – Visual data analysis techniques, interaction techniques; Systems and applications. Exploring the Use of Big Data in Business Context, Use of Big Data in Social Networking, Business Intelligence, Product Design and Development.

11. Text Book:

Big Data Analytics, G. Sudha Sadasivam & R. Thirumahal, Oxford University Press 2020

12. Reference Books:

RB1. Big Data and Analytics, Seema Acharya, Subhashini Chellappan, Infosys Limited, Publication: Wiley India Private Limited,1st Edition 2015

RB2. Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (Editor), Wiley, 2014

RB3. Stephan Kudyba, Thomas H. Davenport, Big Data, Mining, and Analytics, Components of Strategic Decision Making, CRC Press, Taylor & Francis Group. 2014

RB4. Norman Matloff , THE ART OF R PROGRAMMING, No Starch Press, Inc.2011

RB5. Big Data For Dummies, Judith Hurwitz et al. Wiley 2013.

RB6. Glenn J. Myatt, Making Sense of Data, John Wiley & Sons, 2007 Pete Warden,Big Data Glossary, O'Reilly, 2011.

13. Pre-requisites:

- DBMS

14. Lesson Plan:

Lecture No.	Unit	Topics
1-2	Introduction	• Importance of Data • Characteristics of Data Analysis of Unstructured Data

Lecture No.	Unit	Topics
		• Combining Structured and Unstructured Sources
3-6	Introduction to Big Data Platform	• Challenges of conventional systems • Web data
		• Evolution of Analytic scalability • Analytic processes and tools
		• Analysis vs reporting • Modern data analytic tools
		• Types of Data • Elements of Big Data
		• Data Analytics Lifecycle • Discussion
7-11	Big Data Technology Foundations	• Exploring the Big Data Stack • Data Sources Layer • Ingestion Layer
		• Storage Layer • Physical Infrastructure Layer • Platform Management Layer
		• Security Layer • Monitoring Layer
		• Analytics Engine • Visualization Layer
		• Big Data Applications, Virtualization.
12-14	Introduction to Streams Concepts	• Stream data model and architecture
		• Stream Computing • Sampling data in a stream
		• Filtering streams • Counting distinct elements in a stream.
15-22	Big Data Tools:	• NOSQL, MapReduce – Hadoop
		• HDFS
		• Hive
		• MapR – Hadoop -YARN - Pig and PigLatin
		• Jaql - Zookeeper – Hbase
		• Cassandra- Oozie
		• Avro
		• Mahout
23-30	Data Analysis through R:	• Exploring Basic Features of R, Programming Features
		• Packages
		• Exploring R Studio, Handling Basic Expressions in R,
		• Basic Arithmetic in R, Mathematical Operators, Calling Functions in R,
		• Working with Vectors, Creating and Using Objects.
		• Handling Data in R Workspace
		• Creating Plots, Using Built-in Datasets in R
		• Reading Datasets and Exporting Data from R, Manipulating and Processing Data in R
31-36		• Distributed and Parallel Computing for Big Data

Lecture No.	Unit	Topics
	Framework & visualization	• Visualizations – Visual data analysis techniques
		• Interaction techniques; Systems and applications.
		• Exploring the Use of Big Data in Business Context
		• Use of Big Data in Social Networking, Business Intelligence
		• Product Design and Development

15. Assessment Components:

Sr #	Assessment Component	Time	Weightage/ Marks	Date	Course Lecture No.		Mode
					From	To	
1	Mid-Semester Examination	1.5 Hrs	20	13/09/2021 - 17/09/2021	1	18	Closed Book
2	Activity based Teaching and Learning	Through out semester	30	Throughout semester	1	36	Open Book, Closed Book and Presentation
3	End-Semester Examination	3 Hrs	50	29/11/2021 – 11/12/2021	1	36	Closed Book

16. Assessment plan for activity based learning:

Considering the guidelines circulated and after discussing with the faculty members, following activity based teaching and learning is proposed and Component wise distributions of the activities are listed below.

Problem Solving	Critical Thinking	Interactivity	Quiz
Assignment = 5	Research Paper Presentation = 10	Viva= 5	10

17. Activity List – The list of activities and learning practices are further detailed below.

17.1 Problem Solving (Assignment – 5 marks) – Expectation from this activity is to solve set of problems individually and submit before due date. The assignments should be hand-written and the soft copies to be submitted to the subject faculty member before due date. The 1st assignment to be submitted before mid semester and the other after mid-semester and the exact date will be initiated in the class by respective subject faculty member.

17.2 Critical Thinking (Research Paper Presentation = 10 marks) - Expectation from this activity is to go through research papers in order to identify a pertinent research need. Few pages short compiling covering research needs, background, literature summary, methodology and conclusion to be presented to subject faculty. This is group based activity and the formation of group will be decided in the class.

17.3 Quiz (10 marks): Expectation from this activity is to answer subject questions online. Two quizzes with easy, moderate and difficulty level will be conducted before the mid and end of semester. Faculties are free to give their own questions in the quiz. Evaluation is to be done by respective subject teacher.

17.4 Interactivity (Viva – 5 marks): - Expectation from this activity is to know the level of understanding of the student and participation in the class.Randomly any student will be asked questions related to the topics covered in the class.

18. **Course Materials:** Course Material will be provided for all topics which can be used as reference. The material consists of –

- Lecture Notes
- Home Work
- Research Papers
- Class Work
- Supplementary Reading

19. **Attendance:** Every student is expected to be regular (in attendance) in all lecture classes, tutorials, labs, tests, quizzes, seminars etc and in fulfilling all tasks assigned to him / her. Attendance will be recorded and 75% attendance is compulsory.

20. **Makeup:**
    - No make-up examination will be scheduled for the mid semester examination. However, official permission to take a make-up examination will be given under exceptional circumstances such as admission in a hospital due to illness / injury, calamity in the family at the time of examination.
    - A student who misses a mid-semester examination because of extenuating circumstances such as admission in a hospital due to illness / injury, calamity in the family may apply in writing via an application form with supporting document(s) and medical certificate to the Dean of the School for a make-up examination.
    - Applications should be made within five working days after the missed examination.