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

School of Computer Engineering
Autumn Semester 2023-24

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 Faculty** : Dr. Sarita Tripathy
5. **Course offered to the School :** Computer Engineering
6. **Course Objective:**

- To understand the concept 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 formulate the concepts, principles and techniques focusing on the applications to industry and real world experience.
- To contextually integrate and correlate large amounts of information to gain faster insights for real time scenarios.

7. **Course Outcome:**

CO #	Detail
CO1	Understand the concept of big data and its analytics in the real world
CO2	Analyse various big data technology foundations
CO3	Apply filtering technique to stream data
CO4	Apply Hadoop ecosystem paradigm using MapReduce, YARN, Pig, Hive, Scoop, HBase to solve data intensive problems
CO5	Analyse big data framework like Hadoop and NoSQL to efficiently store and process big data to generate analytics
CO6	Present appropriate solutions to big data analytics frameworks and visualization.

8. **Course Contents**

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

Unit #	Unit	Detailed Area
1	Overview of Big Data	Importance of Data, Characteristics of Data, Analysis of unstructured data, Introduction to Big Data, Challenges of conventional systems, Data analytic, Evolution of analytic scalability, Big Data Analytics, Key Big Data terminologies, Big Data analytics lifecycle, Cloud Computing and Big Data.
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.
3	Streaming	Introduction to Streams Concepts – Stream data model and architecture – Stream Computing, Sampling data in a stream –

		Filtering streams, Counting distinct elements in a stream.
4	Hadoop Ecosystem	Introduction to Hadoop, Hadoop Ecosystem, Hadoop Distributed File System, MapReduce, YARN, Pig and PigLatin, Hive, Scoop, HBase
5	Storing Data in Big Data context.	Data Models, RDBMS and Hadoop, Non-Relational Database, Introduction to NoSQL, Types of NoSQL, Polyglot Persistence, Sharding
6	Frameworks And Visualization	Distributed and Parallel Computing for Big Data, Big Data Visualizations – Visual data analysis techniques, interaction techniques, applications

9. Text Book:

TB1. Big Data, Black Book, DT Editorial Services, Dreamtech Press, 2016

10. 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.

11. Pre-requisites:

- DBMS

12. Lesson Plan:

Lecture No.	Unit	Topics	Lesson #
1-6	Overview of Big Data	• Importance of Data • Characteristics of Data, Analysis of Unstructured Data • Combining Structured and Unstructured Sources	1
		• Introduction to Big Data • Challenges of conventional systems	2
		• Data analytic • Evolution of Analytic scalability	3
		• Big Data Analytics • Key Big Data terminologies	4
		• Big Data analytics lifecycle	5
		• Cloud Computing and Big Data • Discussion	6
7-11	Big Data Technology Foundations	• Exploring the Big Data Stack • Data Sources Layer • Ingestion Layer	7
		• Storage Layer • Physical Infrastructure Layer • Platform Management Layer	8
		• Security Layer • Monitoring Layer	9
		• Analytics Engine • Visualization Layer	10
		• Big Data Applications, Virtualization.	11
12-14	Streaming	• Introduction to Streams Concepts • Stream data model and architecture	12

Lecture No.	Unit	Topics	Lesson #
		• Stream Computing • Sampling data in a stream	13
		• Filtering streams • Counting distinct elements in a stream.	14
15-22	Hadoop Ecosystem	• Introduction to Hadoop • Hadoop Ecosystem	15
		• Hadoop Distributed File System • MapReduce	16
		• YARN	17
		• Hive	18
		• Pig and PigLatin	19
		• HBase	20
		• Scoop	21
		• Discussion	22
23-30	Storing Data in Big Data context	• Data Models	23
		• RDBMS and Hadoop	24
		• Non-Relational Database	25
		• Introduction to NoSQL	26
		• Types of NoSQL	27
		• Types of NoSQL cont...	28
		• Polyglot Persistence	29
		• Sharding • Discussion	30
31-36	Framework & visualization	• Distributed and Parallel Computing for Big Data	31
		• Big Data Visualizations – Visual data analysis techniques	32
		• Interaction techniques and applications	33
		• Big Data Visualizations – Visual data analysis techniques cont...	34
		• Big Data Visualizations – Visual data analysis techniques cont...	35
		• Interaction techniques and applications • Discussions	36

13. Assessment Components:

Sr #	Assessment Component	Time	Weightage/ Marks	Course Lecture No.		Mode
				From	To	
1	Mid-Semester Examination	1.5 Hrs	20	1	18	Closed Book
2	Activity based Teaching and Learning	Through out semester	30	1	36	Open Book, Closed Book
3	End-Semester Examination	3 Hrs	50	1	36	Closed Book

Mid-semester question paper comprises of 5 questions and students have to answer any four questions including question no 1, which is compulsory. Weightage for each question is 5. There will be 5 parts in question no 1.

End-semester question paper comprises of 8 questions and students have to answer any six questions including question no 1, which is compulsory. Weightage for 1st question is 10 and 8 for others. There will be 10 parts in question no 1.

14. 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 (2)	Critical Thinking (1)	Interactivity (1)	Quiz (1)	Class Participation (1)
Assignment = 2X5 = 10	Class Test = 1X5 = 5	Class Note = 1X5=5	1X5=5	1X5=5

15. Activity List – The list of activities and learning practices are shown below.

15.1. Assignment – I (Problem Solving)

- It is impossible for a web service to provide following three guarantees at the same time i.e., consistency, availability and partition-tolerance. Justify it with suitable explanation.
- Hotel Booking: are we double-booking the same room? Justify this statement with CAP theorem.
- A mobile health organisation captures patient's physical activities by attaching various sensors on different body parts. These sensors measure the motion of diverse body parts like the rate of turn, the magnetic field orientation etc. A model will be built for effectively deriving information about the motion of different body parts like chest, ankle etc. Indentify the key activities for each stage of Big Data Analytics lifecycle.
- You have to analyze the Adahar card data set against different research queries for example total number of Adahar cards approved by state, rejected by state, total number of Adahar card applicants by gender and total number of Adahar card applicants by age type with visual depiction. How the Big Data and Cloud Computing interlinks in such a case?
- With the emergence of new technologies, new academic trends introduced into Educational system which results in large data which is unregulated and it is also challenge for students to prefer to those academic courses which are helpful in their industrial training and increases their career prospects. Another challenge is to convert the unregulated data into structured and meaningful information. Develop tool that will be helpful in decision making for students to determine courses chosen for industrial trainings. Derive preferable courses for pursuing training for students based on course combinations.
- The reasons to why a system should be built distributed, not just parallel are: Scalability, Reliability, Data sharing, Resources sharing, Heterogeneity and modularity, Geographic construction, and Economic. Explain each of the terms in details.
- Design considerations for distributed systems are: No global clock, Geographical distribution, No shared memory, Independence and heterogeneity, Fail-over mechanism, and Security concerns. Explain each of the terms.
- Explain Hadoop 2.0 HDFS daemons with pictorial depiction. Explain their roles.
- What you understand by Rack awareness and replication. Explain with suitable example.
- Explain the anatomy of file read and file write in Hadoop 2.0 HDFS with pictorial depiction.
- Why was 128 MB chosen as default chunk size in Hadoop 2.X? What was the size in Hadoop 1.x?
- Explain data locality with suitable examples. Explain the difference between moving computation and moving data in a cluster.
- Explain Hadoop 2.0 architecture with pictorial depiction. Explain the concept of blocks in HDFS architecture?
- In what circumstances MapReduce is useful or used? In which cases, it is not suitable.
- Describe each component of Hadoop Ecosystem.
- Design a space-efficient probabilistic data structure that would be used to test False Negative.
- What you understand by data orchestration? How Big Data orchestrations help business in B2B marketing?
- Let, Twitter wants to count “how many unique users visited this month?” What will be the stream elements in such a case?
- Deletion of elements from Bloom filter is not allowed. Why?
- Device an algorithm for uniform sampling from data streams.

- Discuss the use case of bloom filter in authentication and authorization.

Evaluation scheme: Full mark for the correct answer and stepwise mark to be awarded based on the partial correctness of the solution. The students are expected to form a group and solve the problems jointly.

15.2. Assignment – II (Problem Solving)

- Explain data locality with suitable examples. Explain the difference between moving computation and moving data in a cluster.
- What is YARN and what is it used?
- Explain YARN architecture and discuss their roles.
- How an application submitted to YARN is successfully executed?
- What is the need of Apache Pig and how it is different from MapReduce?
- Explain Apache Pig architecture and discuss their roles.
- Discuss Apache Pig execution mode and execution mechanisms.
- Discuss different operators of Apache Pig with suitable examples.
- Discuss Hive DDL (creation of database, tables etc) with suitable examples.
- Discuss different operators of Apache Hive with suitable examples.
- Discuss Hive QL (select-where, select-order by, select – group by, select-join) with suitable example.
- Explain Hive portioning with suitable example.
- How Scoop works and discusses its Import and Export utility.
- What is HBase? Mention the difference between HBase and RDBMS?
- Explain HBase architecture and discuss the concepts of regions. Discuss its storage mechanism.
- What is NoSQL and why it is used?
- Discuss different use cases of NoSQL.
- Explain different types of NoSQL with suitable examples.
- What are the differences between column-oriented and row-oriented database. Explain with pictorial depiction.
- What are the difference between SQL and NoSQL?
- Design considerations for distributed systems are: No global clock, Geographical distribution, No shared memory, Independence and heterogeneity, Fail-over mechanism, and Security concerns. Explain each of the terms.
- The reasons to why a system should be built distributed, not just parallel are: Scalability, Reliability, Data sharing, Resources sharing, Heterogeneity and modularity, Geographic construction, and Economic. Explain each of the terms in details.
- Why data visualization is important in Big Data?
- Explain the techniques used for visualization of data.
- Explain different types of data visualization.
- What are the different visualization interaction techniques?
- Draw the Euler diagram of the sets, X = {1, 2, 5, 8}, Y = {1, 6, 9} and Z={4, 7, 8 , 9}. Then draw the equivalent Venn diagram.
- Your parent has asked you to reduce monthly expenses. So you adopted 2 techniques called as “Buy Less” and “Pay Less”. So, you have started in buying less food and cloths. While buying less food, you have started cooking; reduce eating outside meals, buy nonperishable items in bulk. While paying, you have started sharing the cost, bought items that are on sale, and bought generic items. Draw the Tree/Hierarchical data visualization.
- Let’s assume the score of Big Data has been published for your section. Using 1D data visualization technique, display the “possible grade” against each of you.

Evaluation scheme: Full mark for the correct answer and stepwise mark to be awarded based on the partial correctness of the solution. The students are expected to form a group and solve the problems jointly.

15.3. Class Test

- You are planning the marketing strategy for a new product in your company. Identify and list some limitations of structured data related to this work.
- In what ways does analyzing Big Data help organizations prevent fraud?
- Discuss the techniques of parallel computing.
- Discuss the features of cloud computing that can be used to handle Big Data.
- Discuss similarities and differences between ELT and ETL.
- Draw a step-by-step process in the insertion of element 25, and then 40 into the Bloom Filter of size 10. Then, draw a step-by-step process for lookup/membership test with the elements 10 and 48. The hash functions are: h1(x) = (3x+41) mod 6, and h2(x) = (7x+5). Identify whether any lookup element (i.e. either 10 or 48) is resulting into the case of FALSE POSITIVE?

Evaluation scheme: Full mark for the correct answer and stepwise mark to be awarded based on the partial correctness of the solution.

15.4. Quiz

Q1. Refer to the Fig 1. Draw the equivalent logical, conceptual and physical data model, considering OLAP system.

Logical Structure Design on teaching management system

"Students" Table				"Teachers" Table			
1	StudentNumber	char	10	51	TeacherNumber	char	10
2	StudentName	varchar	20	52	TeacherName	varchar	20
3	AdmissionYear	char	4	53	GraduationSchool	varchar	30
4	CollegeName	varchar	50	54	GraduationDate	datetime	4
5	ProfessionalName	varchar	30	55	CollegeName	varchar	50
6	ClassName	varchar	10	56	ProfessionalName	varchar	30
7	BirthDate	datetime	4	57	TechnicalTitle	varchar	20
8	HomeAddress	varchar	100	58	FinalDegree	varchar	20
9	Postcode	char	6	59	FinalEducation	varchar	20
10	MobilePhone	varchar	20	60	MobilePhone	varchar	20
11	ElectronicMailbox	varchar	50	61	ElectronicMailbox	varchar	50
12	StudentProfile	text	8	62	JobResume	text	8
"Courses" Table				"TestQuestions" Table			
21	CourseNumber	char	10	71	TestQuestionsNumber	char	10
22	CourseName	varchar	40	72	CourseNumber	char	10
23	CourseNature	varchar	40	73	KnowledgePoint	varchar	50
24	OpenSemester	char	1	74	ChapterSection	varchar	20
25	CourseCredit	decimal	4,1	75	TestQuestionsType	varchar	20
26	TheoreticalHours	smallint	2	76	TitleName	varchar	100
27	ExperimentHours	smallint	2	77	TestQuestionsContent	text	8
28	CourseDescription	text	8	78	TestQuestionsAnswer	text	8

Figure 1: Teaching management system

Q2. Draw a diagram illustrating multi-threaded parallel distributed system.
Q3. Calculate the probability that a slot is not hashed with 5 hash functions for 13 bit length bloom filter.
Q4. Develop an algorithm to i) insert an item, and to ii) test the membership (or lookup) in Bloom Filter. Draw a step-by-step process in the insertion of element 25, and then 40 into the Bloom Filter of size 10. Then, draw a step-by-step process for lookup/membership test with the elements 10 and 48. The hash functions are: h1(x) = (3x+4) mod 6, and h2(x) = (7x+5) mod 3. Identify whether any lookup element (i.e. either 10 or 48) is resulting into the case of **FALSE POSITIVE**?

Evaluation scheme: Full mark for the correct answer and stepwise mark to be awarded based on the partial correctness of the solution.

16. **Course Materials:** Course Material will be provided for all topics which can be used as reference. The material consists of –
    - Lecture Notes
    - Home Work
    - Supplementary Reading

- Class Work

17. Activity Planning:

S: Starting class and **C: Class**

Sr #	Activity	Activity Planned Date
1	Assignment 1	S + 11 C
2	Quiz 1	S + 11 C
3	Assignment 2	S + 20 C
4	Class Test	S + 23 C
5	Class Participation	S + 30 C
6	Class Notes	S + 35 C

18. **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.

19. **Make-up:**
    - 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.

20. **Discussion of Mid Semester performance:** Performance of the mid semester examination will be discussed in the class room

21. Pre-end semester total marks: To be announced and discussed in the class.

22. Course Management System: Google Portal - Is a software system designed to facilitate faculties in the management (instructional content, assessment and documentation) of the courses for their students, both faculties and students can monitor the system. Though usually considered as a tool and often used to complement the face-to-face classroom.

23. Chamber consultation hour for doubts clarification:

Sr#	Cabin No	Day & Time
1	Cabin-6, Block-B, Campus - 15.	Will be informed in due time.

24. Academic Dishonesty:
- It may be noted that any kind of copying/plagiarism by any student and/or malpractice in examinations is strictly prohibited.
- In case of the violation of above the Institute will take appropriate and necessary action.

25. **Notices:** All notices regarding the course will be email or communicated via WhatsApp.