B. M. S. INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA, BANGALORE-064
Department of Computer Science & Engineering

COURSE FILE CONTENTS

1. Calendar of Events
2. Time Table
3. Syllabus
4. Lesson Plan
5. Course Outcomes
6. CO – PO /PSO Mapping
7. Gap Analysis, Student activity plan & Articulation
8. List of Students
9. Internal Test Papers
10. Scheme of Evaluation
11. CO – PO Analysis-Excel sheet
12. Articulation for unattained PO`s
13. Three blue books,copy of assignment/Poster/PBL/reports etc

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Avalahalli, Doddaballapur Road, Yelahanka, Bangalore - 560064

Computer Science and Engineering - Calender of Events (CoE) 2019-20 (Even Semeste

VISION OF THE DEPARTMENT	To develop technical professionals acquainted with recent trends and technologies of computer science to serve as valuable resource for the nation/society.
MISSION OF THE DEPARTMENT	Facilitating and exposing the students to various learning opportunities through dedicated academic teaching, guidance and monitoring.

Month	Week	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Working Days	EVENTS
Jan	W-1	26	27	28	29	30	31		6	26-Jan.: 71st Republic Day Celebration; 27-Jan.: Start of VI Sem MCA & IV Sem M.Tech Classes; 31-Jan: PAC Meeting -1
								1		
February	W-2	2	3	4	5	6	7	8	6	3-8 Feb: AICTE Activity for III sem students; 5-Feb.: Start of Jnanavardhan Series - 8
	W-3	9	10	11	12	13	14	15	6	10-Feb.: Start of B.E (II, IV, VI, VIII) Classes; 10-19Feb: Induction Program for I Year; 11,12,14 -Feb: Industry Institute Interaction; 14-Feb: IIC Activity - I; 15-Feb.: Annual Alumni Meet; 15-Feb.: Release of Depts. Magazine/ Publication
	W-4	16	17	18	19	20	21	22	5	17-Feb.: Expert Talk Review; 20- Feb: PAC Meeting -2; 21-Feb.: Mahashivaratri
	W-5	23	24	25	26	27	28	29	6	25-Feb.: PBL Group Formation; 25 -Feb: Industrial visit -1; 29-Feb.: Industrial Visit Review - 1; 29-Feb.: Employers Meet - I; 29 Feb: Alumni Interaction 1
March	W-6	1	2	3	4	5	6	7	6	2-March: PBL Synopsis Submission; 5-Mar.: Start of II & IV Sem MCA and II Sem M.Tech Classes; 7-Mar.: Employers Meet - II; 6-Mar.: IIC Activity - II; 7-Mar: IIC Activity - III
	W-7	8	9	10	11	12	13	14	6	13 Mar : PAC Meeting 3; 14 Mar: Alumni Interaction 2
	W-8	15	16	17	18	19	20	21	6	16-Mar.: Expert Talk Review; 17-Mar.: FIMS Data Entry; 20-Mar.: Visa Workshop - I, Higer Education Dept.
	W-9	22	23	24	25	26	27	28	5	21-26 March IA-1 (B.E: All Semesters); 25-Mar.: Chandramana Ugadi; 27-Mar.: IIC Activity - IV; 28-Mar.: IIC Activity - V; 27-28 Mar: Utsaha 2020
	W-10	29	30	31					6	30-31 Mar.: 8th Sem Project Review - 2 and PBL Review - 1; 31-Mar.: IA - 1 SMS Dispatch; 31-Mar.: Industrial Visit Review - 2; 30-Mar.: IIC Activity - VI
April					1	2	3	4		1-Apr: Student Feedback - 1; 2 -Apr.: Industrial visit -2; 3-Apr.: Work Opportunities for International Students; 4-Apr: Academic Monitoring I; 4-Apr.: Parents-Teachers- Association (PTA) - 1
	W-11	5	6	7	8	9	10	11	4	6-Apr.: Mahaveer Jayanti; 7-9 Apr.: IA - 1 MCA (II & IV Sem) & M.Tech (II Sem); 10-Apr.: Good Friday
	W-12	12	13	14	15	16	17	18	5	14-Apr.: Ambedkar Jayanti; 16-April: Expert Talk Review; 15-Apr.: FIMS Data Entry; 17-Apr.:: Visa Workshop - II, Higer Education Dept; 18 Apr: Alumni Interaction 3
	W-13	19	20	21	22	23	24	25	6	22-25: April: IA - 2 (B.E: All Semesters)
	W-14	26	27	28	29	30			5	26-Apr.: Basava Jayanti; 27-28 April: 8th Sem Project Review - 3; 30-Apr.: Industrial Visit Review - 3
							1	2		1-May: May Day; 01-May: IA - 2 SMS Dispatch; 2-May: Parents- Teachers- Association (PTA) - 2; 2-May: Academic Monitoring II
May	W-15	3	4	5	6	7	8	9	6	4-May: Student Feedback - 2; 7-9 May.: IA - 2 MCA (II & IV Sem) & M.Tech (II Sem); 8-9 May: DJango Workshop by IEEE
	W-16	10	11	12	13	14	15	16	6	14 May : OE/PE Orientation; 15-May: FIMS Data Entry; 15 May : PAC Meeting 4; 16-May: Expert Talk Review
	W-17	17	18	19	20	21	22	23	6	20-May.: End of VI Sem MCA & IV Sem M.Tech Classes; 20-23 May: IA - 3 (B.E: All Semesters)
	W-18	24	25	26	27	28	29	30	5	25-May: Qutub-e-Ramzan; 26-May: BMSIT Open Day; 27-May: Farewell; 30-May: IA - 3 SMS Dispatch; 30-May: HEF Activity: Pre-Departure Orientation; 30-May: Academic Monitoring III
	W-19	31							6	
June			1	2	3	4	5	6		1-June: Last Working Day for B.E (All Semesters); 1-June: Subject Allotment for 2020-21 Odd Sem
	W-20	7	8	9	10	11	12	13	6	8-10 Jun.: IA - 3 MCA (II & IV Sem) &
	W-21	14	15	16	17	18	19	20	6	15-Jun.: FIMS Data Entry; 15-20 Jun: FDP
	W-22	21	22	23	24	25	26	27	6	22-June: Last Working Day of II & IV Sem MCA and II Sem M.Tech; 22 Jun: NBA AUDIT
	W-23	28	29	30					2	
	Total Number of Working Days								127	

CONTINUOUS INTERNAL EVALUATION

COURSE	SEM	START	END
INTERNAL ASSESSMENT - 1			
B.E	IV	21-Mar	26-Mar
B.E	II, VI, VIII	23-Mar	26-Mar
MCA	II & IV	07-Apr	09-Mar
M.Tech	II	07-Apr	09-Mar
M.Tech	IV		
INTERNAL ASSESSMENT - 2			
B.E	IV	22-Apr	25-Apr
B.E	II, VI, VIII	22-Apr	24-Apr
MCA	II & IV	07-May	09-May
M.Tech	II	07-May	09-May
MCA	VI		
INTERNAL ASSESSMENT - 3			
B.E	IV	20-May	23-May
B.E	II, VI, VIII	20-May	22-May
MCA	II & IV	08-Jun	10-Jun
M.Tech	II	08-Jun	10-Jun

SEMESTER END EXAMINATIONS

COURSE	START OF EXAM	END OF EXAM
B.E : II - SEM	15-06-2020	04-07-2020
B.E : IV & VI - SEM	15-06-2020	20-07-2020
B.E : VIII - SEM	03-06-2020	11-06-2020
M.Tech: IV - SEM	03-06-2020	10-06-2020
MCA: IV - SEM	03-07-2020	11-07-2020
M.Tech: II - SEM	01-07-2020	11-07-2020
MCA: II - SEM	03-07-2020	11-07-2020
B.E:VIII - SEM VIVO	15-06-2020	20-06-2020

SUMMER PROJECT/PROFESSIONAL TRAINING

COURSE	SEM	START	END
MCA	VI	22-May	30-May
M.Tech	IV	12-Jun	25-Jun
M.Tech	II	13-Jul	31-Jul

COMMENCEMENT OF ODD SEMESTER (2020-21)

COURSE	SEM	DATE
B.E	II, IV & VI	27-Jul
MCA	II & IV	27-Jul

LIST OF HOLIDAYS

26-Jan	Republic Day
21-Feb	Maha Shivaratri
25-Mar	Chandramana Ugadi
06-Apr	Mahaveer Jayanti
10-Apr	Good Friday
14-Apr	Ambedkar Jayanti
26-Apr	Basava Jayanti
01-May	May Day
25-May	Qutub-e-Ramzan

PARENTS-TEACHERS ASSOCIATION

PTA	DATE
PTA - 1	04-Apr
PTA - 2	02-May

PRACTICAL EXAMINATION

COURSE	SEM	DATES
B.E	II, IV, VI, & VIII	03-Jun to 13-Jun
MCA	II & IV	24-Jun to 29-Jun

	LESSON PLAN	Date: 5/03/2020
	Sub Code & Name : 18SCS21- MANAGING BIG DATA	Page 1 of 6
	Semester : II | Program: M.Tech CSE	
	Academic Year: 2019-20	
	Lesson Plan Author(s) Dr. Anjan Krishnamurthy,	

Prerequisite: The student must be aware of DBMS concepts.

Course Objective:

- Define big data for business intelligence
- Analyze business case studies for big data analytics
- Explain managing of Big data Without SQL
- Develop map-reduce analytics using Hadoop and related tools

Course Outcomes:

After the completion of this course, students will be able to

CO No.	Course Outcome	BT Levels
PCSE.121.1	Summarize the fundamentals and concepts of Big Data.	L2
PCSE.121.2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	L3
PCSE.121.3	Analyze methods and algorithms, to compare them to solve problems.	L4
PCSE.121.4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce	L5

Course Articulation Matrix

CO No.	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6
PCSE.121.1	Summarize the fundamentals and concepts of Big Data.						
PCSE.121.2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	1					1
PCSE.121.3	Analyze methods and algorithms, to compare them to solve problems.				3	1	2
PCSE.121.4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce						2

Course To PO, PSO Mapping

Course	PO1	PO2	PO3	PO4	PO5	PO6
Managing Big Data (18SCS21)	1	0	0	3	1	2

Program Educational Objectives (PEOs)

PEO1 Apply analytical thinking to solve problems through research in the areas of Computer Science and Engineering.

PEO2 Adapt to changing technological trends through life-long learning by exhibiting professional ethics, integrity and career growth.

PEO3 Develop skills to facilitate in providing sustainable solutions by addressing the ever-growing challenges of the society.

Program Outcomes (POs)

The graduates of M. Tech. in Computer Science and Engineering (CSE) Program will be able to:

PO1 Independently carry out research and development work to solve practical problems related to Computer Science and Engineering domain.

PO2 Write and present a substantial technical report/document.

PO3 Demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program.

PO4 Analyse the acquired domain knowledge for providing feasible solution(s).

PO5 Relate the learning outcomes to build requisite competency in professional environment.

PO6 Appraise the need for engaging in lifelong learning.

Course Content

Module -1	Contact Hours
UNDERSTANDING BIG DATA: What is big data – why big data –.Data!, Data Storage and Analysis, Comparison with Other Systems, Rational Database Management System , Grid Computing, Volunteer Computing, convergence of key trends – unstructured data – industry examples of big data – web analytics – big data and marketing – fraud and big data – risk and big data – credit risk management – big data and algorithmic trading – big data and healthcare – big data in medicine – advertising and big data – big data technologies – introduction to Hadoop – open source technologies – cloud and big data – mobile business intelligence – Crowd sourcing analytics – inter and trans firewall analytics. **RBT: L1, L2**	**10 Hours**
Module -2	
NOSQL DATA MANAGEMENT: Introduction to NoSQL – aggregate data models – aggregates – key-value and document data models – relationships – graph databases – schema less databases – materialized views – distribution models – shading — version – map reduce – partitioning and combining – composing map-reduce calculations. **RBT: L1, L2**	**10 Hours**
Module – 3	
BASICS OF HADOOP: Data format – analyzing data with Hadoop – scaling out – Hadoop streaming – Hadoop pipes – design of Hadoop distributed file system (HDFS) – HDFS concepts – Java interface – data flow – Hadoop I/O – data integrity – compression – serialization – Avro – file-based data structures. **RBT: L1, L2, L3**	**10 Hours**
Module-4	
MAPREDUCE APPLICATIONS: MapReduce workflows – unit tests with MRUnit – test data and local tests – anatomy of MapReduce job run – classic Map-reduce – YARN – failures in classic Map-reduce and YARN – job scheduling – shuffle and sort – task execution – MapReduce types – input formats – output formats **RBT: L1, L2, L3**	**10 Hours**
Module-5	
HADOOP RELATED TOOLS: Hbase – data model and implementations – Hbase clients – Hbase examples –praxis. Cassandra – Cassandra data model – Cassandra examples – Cassandra clients –Hadoop integration. Pig – Grunt – pig data model – Pig Latin – developing and testing Pig Latin scripts. Hive – data types and file formats – HiveQL data definition – HiveQL data manipulation – HiveQL queries. **RBT: L1, L2, L3**	**10 Hours**

Course outcomes:
The students shall able to: Describe big data and use cases from selected business domains Explain NoSQL big data management Install, configure, and run Hadoop and HDFS Perform map-reduce analytics using Hadoop Use Hadoop related tools such as HBase, Cassandra, Pig, and Hive for big data Analytics
Question paper pattern: The question paper will have ten questions. There will be 2 questions from each module. Each question will have questions covering all the topics under a module. The students will have to answer 5 full questions, selecting one full question from each module.
Text Books: 1. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012. 2. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.

Reference Books:

1. VigneshPrajapati, Big data analytics with R and Hadoop, SPD 2013.
2. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
3. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.
4. Alan Gates, "Programming Pig", O'Reilley, 2011

Course Schedule – Week wise

WEEK	DAYS	UNIT	MAIN TOPICS	SUB TOPICS	DELIVERY METHOD	BOOKS
1	1	1	Understanding Big Data	Introduction to Big data and its importance	PPT	R1
	2			Convergence of key trends	PPT	R1
	3			A wider variety of data	PPT	R1
	4			Industry examples of Big data - Web analytics, Big data and marketing	PPT	R1
2	1	1	Understanding Big Data	Industry examples of Big data - Fraud and Big data, risk and Big data, credit risk management	PPT	R1
	2			Industry examples of Big data - Big data and algorithmic trading, Big data and healthcare	PPT	R1
	3			Industry examples of Big data - Big data in medicine, advertising and Big data	PPT	R1
	4			Big data technology - Introduction to Hadoop, Open source technologies	PPT	R1
3	1	1	Understanding Big Data	Big data technology - Cloud and Big data, Mobile business intelligence	PPT	R1
	2			Big data technology - Crowd sourcing analytics, Inter and trans firewall analytics	Flipped Mode	R1
	3	2	**NOSQL DATA MANAGEMENT**	Introduction to NoSQL - The Value of Relational Databases, Impedance Mismatch,	PPT	R2

				Application and Integration Databases		
	4			Introduction to NoSQL - Attack of the Clusters, The Emergence of NoSQL	PPT	R2
4	1	2	**NOSQL DATA MANAGEMENT**	Aggregate Data Models - Aggregates, Key-Value and Document Data Models, Column-Family Stores, Summarizing Aggregate-Oriented Databases	Flipped Mode	R2
	2			Data Models - Relationships, Graph Databases, Schemaless Databases, Materialized Views	PPT	R2
	3			Distribution Models - Single Server, Sharding, Master-Slave Replication, Peer-to-Peer Replication, Combining Sharding and Replication	PPT	R2
	4			Consistency - Update Consistency, Relaxing Consistency	PPT	R2
First Internal						
5	1	2	**NOSQL DATA MANAGEMENT**	Version Stamps - Business and System Transactions, Version Stamps on Multiple Nodes	PPT	R2
	2			Map-Reduce - Basic Map-Reduce, Partitioning and Combining, Composing Map-Reduce Calculations	PPT	R2
	3	3	**BASICS OF HADOOP**	History of Hadoop, Data format	PPT	R3
	4			Analyzing data with Hadoop, Scaling out Hadoop streaming,	PPT	R3

				Hadoop pipes		
6	1	3	**BASICS OF HADOOP**	Design of Hadoop distributed file system (HDFS) - HDFS Concepts, Hadoop File systems	Flipped Mode	R3
	2			HDFS - Java interface - Reading Data, Writing Data, Directories	PPT	R3
	3			HDFS - Java interface - Querying the Filesystem, Deleting Data	PPT	R3
	4			Data Flow - Anatomy of a File Read, Anatomy of a File Write, Data Flow - Hadoop Archives Hadoop I/O - Data Integrity	PPT	R3
7	1	3	**BASICS OF HADOOP**	Hadoop I/O - Compression, Hadoop I/O - Serialization	PPT	R3
	2			Avro - data types and schemas Avro - File-Based Data Structures - MapFile, Review	PPT	R3
	3	4	**MAPREDUCE APPLICATIONS**	MapReduce workflows - Decomposing a Problem into MapReduce Jobs, Unit tests with MRUnit, Test data and local tests	PPT	R3
	4			How MapReduce Works - Anatomy of a MapReduce Job Run, Classic Map-reduce	PPT	R3
8	1	4	**MAPREDUCE APPLICATIONS**	YARN, Failures in classic Map-reduce and YARN	PPT	R3
	2			Job Scheduling - The Fair Scheduler, The Capacity Scheduler	PPT	R3
	3			Shuffle and Sort	PPT	R3

	4			Task Execution	PPT	R3
Second Internal						
9	1	4	**MAPREDUCE APPLICATIONS**	MapReduce Types and Formats	Flipped Mode	R3
	2			MapReduce Types - Input Formats, MapReduce Types - Output Formats	PPT	R3
	3	5	**HADOOP RELATED TOOLS**	Hbase - HBasics, Data model Concepts, Implementation	Flipped Mode	R3
	4			Hbase clients, Hbase examples, Hbase - Praxis	PPT	R3
10	1	5	**HADOOP RELATED TOOLS**	Cassandra - Cassandra data model	PPT	R7
	2			Cassandra - Cassandra examples	PPT	R7
	3			Cassandra - Cassandra clients	PPT	R7
	4			Cassandra - Hadoop integration	PPT	R7
11	1	5	**HADOOP RELATED TOOLS**	Pig - Introduction and Grunt - Entering Pig Latin Scripts in Grunt	PPT	R8, R3
	2			Pig - Pig's data model	PPT	R8, R3
	3			Pig - Pig Latin - Input and Output	PPT	R8, R3
	4			Pig - Developing and testing Pig Latin scripts	PPT	R8, R3
12	1	5	**HADOOP RELATED TOOLS**	Hive - Data types and file formats	PPT	R5, R3
	2			HiveQL - Data Definition and Data Manipulation	PPT	R5, R3
	3			HiveQL - Queries, Review	PPT	R5, R3
	4			Hive - Data types and file formats	PPT	R5, R3
Third Internal						

Reference Books:

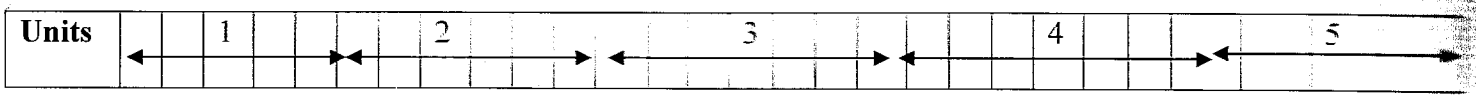
R1. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics:
R2. Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
R3. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
R4. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
R5. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
R6. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
R7. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.
R8. Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilley, 2010.
R9. Alan Gates, "Programming Pig", O'Reilley, 2011.

Details of the Innovative teaching methods:

Topic	Unit	Book	ITM
Hbase - HBasics, Data model Concepts, Implementation Link - https://www.youtube.com/watch?v=bjyH8nlHKHA	5	R3	Flipped Mode
MapReduce Types and Formats Link - https://www.youtube.com/watch?v=dWn9Z19tRMQ	4	R3	Flipped Mode
Design of Hadoop distributed file system (HDFS) - HDFS Concepts, Hadoop File systems Link- https://www.youtube.com/watch?v=1_ly9dZnmWc	3	R3	Flipped Mode
Aggregate Data Models - Aggregates, Key-Value and Document Data Models, Column-Family Stores, Summarizing Aggregate-Oriented Databases Link- https://www.youtube.com/watch?v=6yp4Za9jBxM	2	R2	Flipped Mode
Big data technology - Crowd sourcing analytics, Inter and trans firewall analytics Link- https://www.youtube.com/watch?v=VukxtSfp1tw	1	R1	Flipped Mode
Volunteer Computing Link - https://www.youtube.com/watch?v=AhqNRwyr9II	1		Video Lectures

Course Delivery Plan

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II	I II



Course Unitization for Internals and Semester End Examination

Part	Chapter		Teaching Hours	No. of Questions in		
				Internals I	Internals II	Compensatory Internals
Unit 1	1	UNDERSTANDING BIG DATA	10	4+1*		
Unit 2	2	NOSQL DATA MANAGEMENT	10	2+1*	1	
Unit 3	3	BASICS OF HADOOP	10		3+1*	
Unit 4	4	MAPREDUCE APPLICATIONS	10		2+1*	2+1*
Unit 5	5	HADOOP RELATED TOOLS	10			4+1*

*Represents Innovative and Case Study questions from the units

IA Scheme

Assessment	Weightage in Marks
3 IA test	50
Best two IA average	50 (20)
Assignment	20
Total	**40**

Assignment Rubrics

	Level of Achievement					
	Criteria	Excellent	Good	Average	Poor	Max Score
a	Correlation of Chosen topic with Big Data framework and research potential of the topic	5	4	3	1	5
b	Literature survey with review on latest trends in the topic or prototype submission	8	7	6	3	8
c	New directions of the work in the design and development of latest architecture/system or proposed model	2	1	1	0	2

d	Report submission	5	4	3	1	5
	Total					**20**

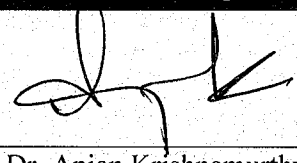
Model Questions

1 Apply a suitable NoSQL data model for crowd source analytics. Outline important features of crowd source analytics that is to be considered for choice of NoSQL DB.
2 Distinguish between Volunteer Computing and Grid Computing. Consider computational architecture viewpoints for the answer.
3 Consider a system which has six I/O channels and each channel can read the data at 150 MB/sec. What is the time taken to read 1 TB and 2 TB of data?
4 Define CAP theorem. Distinguish between CAP concepts with ACID property.
5 List and explain the broad classification of NoSQL data model. Provide appropriate examples with NoSQL DB to explain each data model.
6 Illustrate with examples the need for graph databases in storing Big Data.
7 Analyze the data semantics of a web analytics application and provide illustrious step to perform data analytics. Apply the Map-Reduce technique for web analytics application.
8 Design a columnar family DB for storing the customer and order details of the online e-commerce system using <key,value> pair mechanism.
9 With a neat diagram, explain different types of distribution models. Give the complete categorization and examples for each type of distribution model.
10 With neat diagram, show the anatomy of the file write. State the algorithm used to for file write in the HDFS environment.
11 What is the scale-out in map/reduce environment? Give the roles of the Job-tracker and task-tracker in the overall process.
12 Write code snippet to deal with the compressed data in Hadoop. Explain the need for data compression.
13 With an example, explain the role of network topology in Hadoop architecture. Give the standard representation for topology elements and also provide the distance metrics for various cases.
14 What is clumping? List the factors to be considered for arranging data in the nodes.
15 Design the Hadoop pipe to determine max temperature using the unstructured data provided in weather mining data. Write C++ code for the same.
16 With sample weather mining data, explain the process of the map-reduce for obtaining minimum temperature? Write suitable map and reduce functions.
17 With neat diagram, explain how fair scheduling is performed on map reduce jobs?
18 With an programming example, Illustrate the advantages of the using MRunit test case for testing the logic of the map() and reduce() function designed by the user. Also give the limitation of the MR unit test.
19 Differentiate between HBase and RDBMS
20 List and explain different failures that are possible in Map-Reduce 1 with the help of diagram.
21 With appropriate diagrams, give the anatomy of the job run on Map-Reduce 2 platform
22 Illustrate the process of the shuffle and sort in Map-Reduce platform. Use related diagram wherever required.
23 Port the Student table (Name, RegNo, Semester, CGPA, Address, and Contact No) in RDBMS to a table in Cassandra. Identify the design conditions to port a CA compliant database to AP compliant database.
24 Design a MapReduce function to insert column family into HBase Database.

Course End Survey questions

Managing Big Data (18SCS21)

SNo	Questions	PO
1.	Did the course allow you to independently think to solve problems related to Big Data leading to research work? (Yes/No)	1
2.	Did the course enable you to articulate, present, write reports or documents?	2
3.	Rate the level of your mastery over the course before taking it. (1-Low, 2- Medium, 3 -High)	3
4.	Rate the level of your mastery over the course after taking it. (1-Low, 2- Medium, 3 -High)	3
5.	Are the topics in this course appropriately assisted you in identifying solution?	4
6.	Were you able to do research work in the field of computer science aligned with your course where the work showcases your leadership, integrity and professional ethics?	5
7.	Did make use of the any research tools for the implementation of algorithms?	5
8.	To what extent you grade the quality of contents in this subject?	6
9.	Do you feel topics included in this course will give good background for higher education?	6
10.	Rate the level of the knowledge improvement after the successful completion of this course. (1-Low, 2- Medium, 3 -High)	6

	Prepared by	Approved by
Signature		
Name and Affiliation	Dr. Anjan Krishnamurthy, Associate Professor & PG Coordinator, Dept. of CSE, BMSIT&M	
Date	03/03/2020	

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA – BANGALORE - 64
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Minutes of Meeting with PAC –Batch 2019-20

Date	05 -03-2020	Location	Dept. of CSE
Time	10:30am	Module Coordinator	Dr. Anjan Krishnamurthy
Course Name	Current Course Coordinator	Previous Course Coordinator	
Managing Big Data	Dr. Anjan Krishnamurthy	Dr. Anjan Krishnamurthy	

Sl.No	Discussion	Action By/ Responsible	Action Taken
1	**Agenda:** Course Outcomes, CO – PO Mapping, Gap Identification for Managing Big Data (18SCS21)		
	Defined Cos in the academic year 2018-19	Course Coordinator	

CO No.	Course Outcome	BT Levels
PCSE.121.1	Summarize the fundamentals and concepts of Big Data.	K2
PCSE.121.2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	K3
PCSE.121.3	Analyze methods and algorithms, to compare them to solve problems.	K4
PCSE.121.4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce	K5

CO-PO Mapping

CO No.	Course Outcome	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6
PCSE.121.1	Summarize the fundamentals and concepts of Big Data.						
PCSE.121.2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	1					1
PCSE.121.3	Analyze methods and algorithms, to compare them to solve problems.				3	1	2
PCSE.121.4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce						2

Articulation:

- CO2 is on comparative application of the various forms of data representation

	in NoSQL contribute as an initial topic of research. Hence this is mapped low to PO1 and PO6. • CO3 is analysing various methods and algorithms to solve problems and maps low to PO1 and contributes more to PO3, PO4 and PO6. • CO4 is assignment and practical based and hence maps medium with PO6.		
3	**Action planned to bridge the gap** • Research work related to the design, algorithm portability and system configuration of Hadoop to bridge gap PO1 and PO2. PO3		Research topics /MOOCs

CO-PO Mapping

CO No.	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6
PCSE.121.1	Summarize the fundamentals and concepts of Big Data.						
PCSE.121.2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	1					1
PCSE.121.3	Analyze methods and algorithms, to compare them to solve problems.				3	1	2
PCSE.121.4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce						2

Course Coordinator

www.bmsit.in
hodcse@bmsit.in
Ph: 080-65369470

BMS INSTITUTE OF TECHNOLOGY & MANAGEMENT

(Approved by AICTE, Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka)
Post Box No. 6443, Doddaballapura Main Road, Yelahanka, Bengaluru - 560 064, INDIA
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

II SEM M.TECH (CSE), AC. Year 2019 -20

SL. NO.	USN	NAME OF THE CANDIDATE	Mobile No	Email ID
1	1BY19SCS01	Deepthi M	9741897107	deepthim028@gmail.com
2	1BY19SCS02	Gautham S K	9447566189	skgautham007@gmail.com
3	1BY19SCS03	Madeha Kauser	9482398861	madehakauser2017@gmail.com
4	1BY19SCS04	Meghana Kumar K J	9738826309	meghanakumar98@gmail.com
5	1BY19SCS05	Pujitha J	9663577349	jpujitha1@gmail.com
6	1BY19SCS06	Tejaswini A Kantanavar	7483868815	shreyakantanavar1997@gmail.com
7	1BY19SCS07	Varshini N	9632274110	varshini0505@gmail.com

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

Date: 15th Apr 2020

Course Name: Managing Big Data

Faculty: Dr. Anjan Krishnamurthy

Type of Assignment: MOOCs

Assignment list for MBD 18SCS21 – 2020

Sl. No.	Student Name	Topic Name	RBT	CO	PO1	PO2	PO3	PO4	PO5	PO6
1.	Pujitha	Big Data Modelling and Load Balancing Tools For Hadoop	K4	4						√
2.	Meghana	Intrinsic Security Issues	K4	4						√
3.	Varshini	Big Data Modelling & Hadoop Distributed File System	K4	4						√
4.	Deepthi	Hadoop tools	K4	4						√
5.	Tejaswini A K	Ambari	K4	4						√
6.	Goutham S K	Big Data Modelling and	K4	4						√

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

		Management System								
7.	Madeha	Managing Big Data with MySql	K4	4		√	√			√

Grading policies:

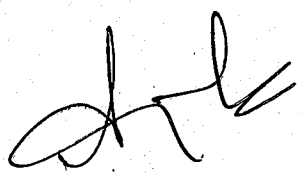
- The last date for the submission of the assignment is on or before 25th May 2020 (hard deadline).
- The assignment must be unique contribution and will undergo rigorous plagiarism process. This similarity index must be less than or equal to 25%.
- The report of the assignment must details out in 2-column IEEE paper format.
- Care to be taken for representation of facts, diagrams, grammar.
- The assignment can be simple prototype implement, deeper exploration of technology, novel thoughts and ideas on the topics.
- A 20 slides ppt must be presented within 5 working days from the submission date.
- Grading will be based on punctual submission of the assignment.

Rubrics:

	5	3	1
Timely completion	Completion of course with Certificate on time	Completion of course after the deadline	Not completed the course only registered

Feed Back and Analysis:

- Students have undergone Online MOOCs courses with respect to Big Data domain. The online certification course helped them to learn new concept and to build their career. Through this assignment students are enabled to attain CO4, PO6. Students will be having additional course certificate also.



Course Coordinator Signature:

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

REFERENCES

(As per IEEE format and must be numbered consecutively in order of first mention)

Example format:

Journal Paper: Name initial, —title‖, Journal name, vol. **(issue), year, pp.11

1. Honig, M.L., Steiglitz, K., and Gopinath, B., —Multichannel signal processing for data communication in the presence of crosstalk‖, *IEEE Trans. Communications.*, vol. 38, (4), 1990 , pp. 551–558.

Conference Proceedings: Name Initial, —title‖, Proceeding of the *, place, year, pp. ***

2. Shin, K.G. and Mckay, N.D. —Open Loop Minimum Time Control of Mechanical Manipulations and its Applications‖ *Proceedings of the Amer. Contr. Conf., San Diego, CA, ,1984,* pp. 1231-1236

Patent: Name initial, —title of patent‖, Patent number, date of patent

3. Bischoff F, —Apparatus for vapor deposition of silicon,‖ U.S. *Patent* 3 335 697, Aug. 15, 1967

Thesis (Masters / Doctoral): Name, initials, —title‖, University, Year

1. Nongpiur, R C, —Near-End Crosstalk Cancellation in xDSL Systems‖ *Doctoral thesis, University of Victoria*, 2005

Annual reports / manual: Name (optional), —title‖, Report number, Agencies, Year

5. The International Technology Roadmap for Semiconductors, Report-7, ITRS, 2011,

Books / Manual / standards data hand books: —Title —, publisher, year

6. —Ferrous Material Testing Procedure — ASTM Standard- vol.3, American Society for Testing Materials, 2003

UC San Diego

COURSE CERTIFICATE

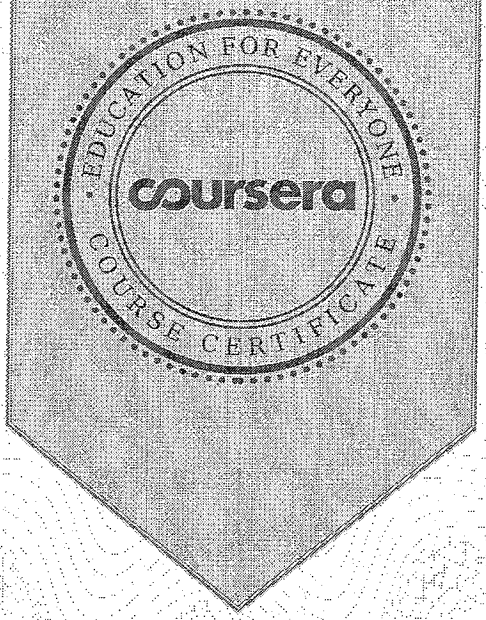
05/23/2020

GAUTHAM SK

has successfully completed

Introduction to Big Data

an online non-credit course authorized by University of California San Diego and offered through Coursera



Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center

Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/6KKFLK89R2HR

Coursera has confirmed the identity of this individual and their participation in the course.

UC San Diego

COURSE CERTIFICATE

05/16/2020

Deepthi M

has successfully completed

Introduction to Big Data

an online non-credit course authorized by University of California San Diego and offered through Coursera

Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center

Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/59WS4GZU5KZM

Coursera has confirmed the identity of this individual and their participation in the course.

Duke
UNIVERSITY

COURSE
CERTIFICATE

05/21/2020

Madeha Kauser

has successfully completed

Managing Big Data with MySQL

an online non-credit course authorized by Duke University and offered through Coursera

Daniel Egger
Executive in Residence and Director,
Center for Quantitative Modeling
Pratt School of Engineering

Jana Schaich Borg
Assistant Research Professor
Social Science Research Institute

Verify at coursera.org/verify/BUGPNA8ZXD54

Coursera has confirmed the identity of this individual and their participation in the course.

UC San Diego

COURSE CERTIFICATE

05/19/2020

Madeha Kauser

has successfully completed

Big Data Modeling and Management Systems

an online non-credit course authorized by University of California San Diego and offered through Coursera

Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center

Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/CMAQNN2DBQAQ

Coursera has confirmed the identity of this individual and their participation in the course.

UC San Diego

COURSE CERTIFICATE

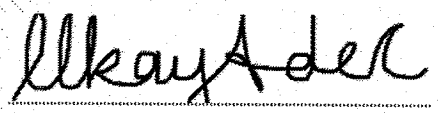
06/05/2020

Pujitha J

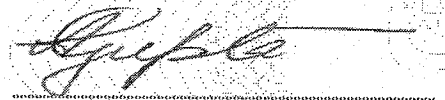
has successfully completed

Big Data Modeling and Management Systems

an online non-credit course authorized by University of California San Diego and offered through Coursera



Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center



Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/WN4P3NE356QK

Coursera has confirmed the identity of this individual and their participation in the course.

UC San Diego

COURSE CERTIFICATE

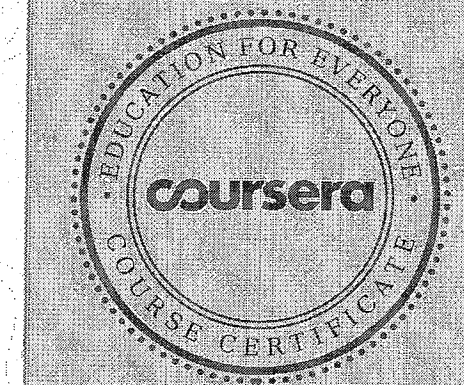
06/05/2020

VARSHINI N

has successfully completed

Big Data Modeling and Management Systems

an online non-credit course authorized by University of California San Diego and offered through Coursera



Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center

Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/2VYRGKTZB8EW

Coursera has confirmed the identity of this individual and their participation in the course.

UC San Diego

COURSE CERTIFICATE

06/06/2020

VARSHINI N

has successfully completed

Introduction to Big Data

an online non-credit course authorized by University of California San Diego and offered through Coursera

Ilkay Altintas
Chief Data Science Officer
San Diego Supercomputer Center

Amarnath Gupta
Research Scientist
San Diego Supercomputer Center

Verify at coursera.org/verify/62BYPFFFGCWL

Coursera has confirmed the identity of this individual and their participation in the course.

Quiz 1

Q:Point out the wrong statement.?
A. Non Relational databases require that schemas be defined before you can add data
B. NoSQL databases are built to allow the insertion of data without a predefined schema
C. NewSQL databases are built to allow the insertion of data without a predefined schema
D. Not required that schemas be defined before you can add data in OODBMS
ANSWER: A

Q:Which of the following are the simplest NoSQL databases?
A. Key-value
B. Wide-column
C. Document
D. Table
ANSWER: A

Q: CAP theorem is applicable for _____ databases?
A. RDBMS
B. NoSQL
C. NewSQL
D. UltraSQL
ANSWER: B

Q:What is clumping?
A. A data replication technique
B. Data division rechnique
C. Clump the data up so that one user mostly gets data from a single server
D. clump the data up so that one user mostly gets data from a different servers
ANSWER: C

Q:Which type of indexes does MongoDB support?
A. Compound Indexes
B. Multikey Indexes
C. Geospatial Indexes
D. All of the above
ANSWER: B

Q:What is crowdsourcing?
A. A collective intellectual gathering of information that comes from the public and then used to complete a business-related task.
B. A collaboration of a crowd of people in one place
C. The science of crowding out marketers.
D. The use of a source in a group people within a crowd.
ANSWER: A

Q: _____ stores are used to store information about networks, such as social connections?
A. Key-value
B. Wide-column
C. Document
D. Graph
ANSWER: D

Q: "Sharding" a database across many server instances can be achieved with _____?
A. LAN
B. SAN
C. MAN
D. PAN
ANSWER: B

Q:NoSQL databases is used mainly for handling large volumes of _____ data?
A. unstructured
B. structured
C. semi-structured
D. joint-structured
ANSWER: A

Q:Which of the following is not a NoSQL database?
A. SQL Server
B. MongoDB
C. Cassandra
D. None of the mentioned
ANSWER: A

Q:What does "Velocity" in Big Data mean?
A. Speed of input data generation
B. Speed of individual machine processors
C. Speed of ONLY storing data
D. Speed of storing and processing data
ANSWER: D

Q:The term Big Data first originated from?
A. Stock Markets Domain
B. Banking and Finance Domain
C. Genomics and Astronomy Domain
D. Social Media Domain
ANSWER: C

Q:Which of the following are examples of Real Time Big Data Processing?
A. Complex Event Processing CEP platforms
B. Stock market data analysis
C. Bank fraud transactions detection
D. both A and C
ANSWER: D

Q:Hadoop is open source?
A. ALWAYS True
B. True only for Apache Hadoop
C. True only for Apache and Cloudera Hadoop
D. ALWAYS False
ANSWER: B

Q:What is the default HDFS block size?
A. 32 MB

B. 64 KB
C. 128 KB
D. 64 MB
ANSWER: D

Q:What is the default HDFS replication factor?
A. 4
B. 1
C. 3
D. 2
ANSWER: C

Q:Read the statement NameNodes are usually high storage machines in the clusters?
A. True
B. False
C. Depends on cluster size
D. True if co-located with Job tracker
ANSWER: B

Q:What is optimal size of a file for distributed cache?
A. <=10 MB
B. >=250 MB
C. <=100 MB
D. <=35 MB
ANSWER: C

Q:A Map reduce job can be written in?
A. Java
B. Ruby
C. Python
D. Any Language which can read from input stream
ANSWER: D

Q:Pig is a?
A. Programming Language
B. Data Flow Language
C. Query Language
D. Database
ANSWER: B

Q:Maximum size allowed for small dataset in replicated join is?
A. 10KB
B. 10 MB
C. 100 MB
D. 500 MB
ANSWER: C

Q:Hadoop is a framework that allows the distributed processing of?

A. Small Data Sets
B. Semi-Large Data Sets
C. Large Data Sets
D. Large and Small Data sets
ANSWER: C

Q:Identify the batch processing scenarios from following?
A. Sliding Window Averages Job
B. Facebook Comments Processing Job
C. Inventory Dynamic Pricing Job
D. Fraudulent Transaction Identification Job
ANSWER: B

Q:Which of the following is not true about Name Node?
A. It is the Master Machine of the Cluster
B. It is Name Node that can store user data
C. Name Node is a storage heavy machine
D. Name Node can be replaced by any Data Node Machine
ANSWER: D

Q:What decides number of Mappers for a MapReduce job?
A. File Location
B. mapreD.map.tasks parameter
C. Input file size
D. Input Splits
ANSWER: D

Q:The source of HDFS architecture in Hadoop originated as which of the following?
A. Azure distributed filesystem
B. Yahoo distributed filesystem
C. Facebook distributed filesystem
D. Google distributed filesystem
ANSWER: B

Q:The default replication factor for HDFS file system in Hadoop is which of the following?
A. 1
B. 3
C. 4
D. 2
ANSWER: D

Q:Which of the following are NOT big data problems?
A. Parsing 5 MB XML file every 5 minutes
B. Processing IPL tweet sentiments
C. Processing online bank transactions
D. both A and C
ANSWER: D

Q:What does "Velocity" in Big Data mean?
A. Speed of input data generation
B. Speed of individual machine processors

C. Speed of ONLY storing data
D. Speed of storing and processing data
ANSWER: D

Q:The term Big Data first originated from?
A. Stock Markets Domain
B. Banking and Finance Domain
C. Genomics and Astronomy Domain
D. Social Media Domain
ANSWER: C

Quiz 2

Q:Hadoop is open source?
A. ALWAYS True
B. True only for Apache Hadoop
C. True only for Apache and Cloudera Hadoop
D. ALWAYS False
ANSWER: B

Q:What is the default HDFS block size?
A. 32 MB
B. 64 KB
C. 128 KB
D. 64 MB
ANSWER: D

Q:What is the default HDFS replication factor?
A. 4
B. 1
C. 3
D. 2
ANSWER: C

Q:The mechanism used to create replica in HDFS is_____ ?
A. Gossip protocol
B. Replicate protocol
C. HDFS protocol
D. Store and Forward protocol
ANSWER: C

Q:Which of the following is not a valid Hadoop config file?
A. mapred-site.xml
B. hadoop-site.xml
C. core-site.xml
D. Masters
ANSWER: B

Q:Read the statement NameNodes are usually high storage machines in the clusters?
A. True
B. False
C. Depends on cluster size
D. True if co-located with Job tracker
ANSWER: B

Q:Which of the following is the correct sequence of MapReduce flow?
A. Combine Reduce Map
B. Map Combine Reduce
C. Reduce Combine Map
D. All of the above
ANSWER: C

Q:Which of the following can be used to control the number of part files?

A. Number of Mappers
B. Number of Reducers
C. Counter
D. Partitioner
ANSWER: B

Q:What is optimal size of a file for distributed cache?
A. <=10 MB
B. >=250 MB
C. <=100 MB
D. <=35 MB
ANSWER: C

Q:Number of mappers is decided by the?
A. Mappers specified by the programmer
B. Available Mapper slots
C. Available heap memory
D. Input Splits
ANSWER: D

Q:Which of the following class is responsible for converting inputs to key-value Pairs of Map Reduce?
A. FileInputFormat
B. InputSplit
C. RecordReader
D. Mapper
ANSWER: C

Q:Maximum size allowed for small dataset in replicated join is?
A. 10KB
B. 10 MB
C. 100 MB
D. 500 MB
ANSWER: C

Q:Who will initiate the mapper?
A. Task tracker
B. Job tracker
C. Combiner
D. Reducer
ANSWER: A

Q:What decides number of Mappers for a MapReduce job?
A. File Location
B. mapreD.map.tasks parameter
C. Input file size
D. Input Splits
ANSWER: D

Q:The default replication factor for HDFS file system in Hadoop is which of the following?
A. 1

B. 3
C. 4
D. 2
ANSWER: D

Q: What happens when a Namenode fails?
A. System fails when the name node fails
B. Secondary Name Node takes over
C. RAM is cleared and all the log files are deleted.
D. None of the above
ANSWER: A

Q: Which of the following is true for Hadoop 1.0?
A. Hadoop does not have active and Passive name nodes
B. Hadoop does have active and Passive name nodes
C. Hadoop does not have Passive name nodes
D. None of the above
ANSWER: A

Q: What is namespace?
A. Node Manager
B. File Manager
C. Manages directories, files and blocks
D. Hadoop Master
ANSWER: C

Q: Which two are a part of the block storage in Hadoop?
A. Logical Data Storage and Schema
B. Physical Data Storage and Meta Files
C. Block Schema and storage
D. Block Management and Physical Storage.
ANSWER: D

Q: What is block pool?
A. pool of RAM buffers
B. Set of RAM and ROM pointers
C. set of blocks that belong to a single namespace
D. None of above
ANSWER: C

Q: What is namespace volume?
A. pool of RAM buffers
B. A Namespace and its block pool
C. set of blocks that belong to a single namespace
D. None of above
ANSWER: B

Q: Which of the following is valid sequence performed during File Read?
A. Read, Get Locations, Read node, close
B. Open, Read, Get Locations, Read node, close
C. Get Locations, Read node, close

ANSWER: B

Q: Which of the following is not an output format in Hadoop?
A. TextoutputFormat
B. ByteoutputFormat
C. SequenceFileOutputFormat
D. DBOutputFormat
ANSWER: B

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Avalahalli, Doddaballapur Main Road, Bengaluru - 560064

THIRD ASSESSMENT TEST, JUN 2019 - 20

Course Name	Managing Big Data	Course Code	18SCS21
Branch & Semester	2nd Sem M.Tech CSE	Date	22-06-2020 10:00 AM to 12:00PM
Name of the Course Coordinator (s)	Dr. Anjan Krishnamurthy	Max. Marks	40 marks

Note: *Answer **FIVE** full questions out of eight.*

Qn. No.		Marks	CO
1.	Illustrate the steps to perform the Gossip protocol in Cassandra.	08 M	**COs:4 K:2**
2.	Explain the different types of failures that are possible in Map-Reduce 1 with the help of diagram.	08 M	**COs:3 K:2**
3.	Differentiate between HBase and RDBMS	08 M	**COs:3 K:4**
4.	Port the Student table (Name, RegNo, Semester, CGPA, Address, and Contact No) in RDBMS to a table in Cassandra. Evaluate the design conditions to port a CA compliant database to AP compliant database.	08 M	**COs:4 K:5**
5.	Write a MapReduce function to insert column family into HBase Database.	08 M	**COs:3 K:2**
6.	With neat diagram, explain how capacity scheduling is performed on map reduce jobs?	08 M	**COs:3 K:2**
7.	Analyze the data model of HBase and bring out important characteristics aspects and suitability to convert traditional RDBMS to NoSQL base.	08 M	**COs:3 K:4**
8.	Illustrate how data integration are performed in Hive and HBase?	08 M	**COs:3 K:2**

Course Outcomes (COs)

CO1: Summarize the fundamentals and concepts of Big Data.

CO2: Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.

CO3: Analyze methods and algorithms, to compare them to solve problems.

CO4: Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce

Bloom's Category

Remembering (K1) **Understanding (K2)** **Applying (K3)** **Analyzing (K4)** **Evaluating (K5)** **Creating (K6)**

Signatures of the Question Paper Scrutiny Committee

Course Coordinator(s)	Module Coordinator(s)	Program Coordinator	Head of the Department
10/6/2020	10/6/2020	10/6/2020	10.6.2020

SCHEME AND SOLUTION for IA 3

SUBJECT CODE: 18SCS21 **SUBJECT:** Managing Big Data

Question No		Marks
1	Gossip Protocol – 4 M Cassandra Architecture – 4 M	08
2	Diagram -03 Types of Failures – Node, Master Node, Client Node, communication, OS (05)	08
3	Any four significant comparison of HBase with RDBMS (4x2=8)	08
4	Steps to convert from RDBMS to NOSQL (4) Design consideration (4)	08

5	Code- (6)
	`// Create a new Job`
	`Job job = new Job(new Configuration());`
	`job.setJarByClass(MyJob.class);`
	`// Specify various job-specific parameters`
	`job.setJobName("myjob");`
	`job.setInputPath(new Path("in"));`
	`job.setOutputPath(new Path("out"));`
	`job.setMapperClass(MyJob.MyMapper.class);`
	`job.setReducerClass(MyJob.MyReducer.class);`
	`// Submit the job, then poll for progress until the job is complete`
	`job.waitForCompletion(true);`
	About HBase (2)
		08

6	Diagram (2) Capacity Scheduling (4) Algorithm (2)	08
7	Steps to convert from RDBMS to HBase (4) Design consideration (4)	08
8	Data Integration in Hive (4) Data Integration in Hbase (4)	08

GUIDELINES FOR III INTERNAL ASSESSMENT TEST (THEORY):

STUDENTS:

1. III Internal Assessment Test is descriptive in nature and will be conducted through Google Class Room (GCR). Hence all the students must be registered for GCR of their respective courses well in advance through official bmsit e-mail id only. For any clarifications regarding registering for GCR, respective faculty members may be contacted.

2. The dates of the test have been notified; 11th Jun to 16th Jun 2020 and the detailed schedule will be released shortly.

3. Internal Assessment Test will be conducted for 30 marks and the total duration is 120 minutes including the uploading of duly signed answer sheets.

4. There will be eight questions (with no sub-sections) in question paper of each course out of which student has to answer any five of his choice.

5. Students to down load the question paper from GCR and write the answers in A-4 sheets which have the following information on header of every sheet.
   - USN, Name, Course Code
   - Course Name ,Branch/Semester/Section
   - Name of the Faculty
   - Student Signature

(It is suggested to have sufficient number of A-4 sheets with the above information on header of each sheet with signature well in advance).

6. After writing all the answers in the A-4 sheets (with above information), scan all the pages, join/merge all the scanned pages into one single PDF. Please follow this link to convert and merge in to one single pdf: https://online2pdf.com/

7. The name of your pdf document must be in this format

   USN_COURSECODE (example: 1BY19SCS02_18SCS21.pdf)

8. Finally upload this pdf in google classroom, click Turn in

9. The link for the video of detailed process https://www.youtube.com/watch?v=gJ4ZYkHJCks

10. Signature of the student at the Footer is mandatory.

Surname	First name	Email addr	State	Started on	Completed	Time taker	Grade/15.(	Q. 1 /0.50	Q. 2 /0.50	Q. 3 /0.50	Q. 4 /0.50
1BY19SCSC	GAUTHAM	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins 34	11.50	0.50	0.50	0.50	0.50
1BY19SCSC	DEEPTHI M	1BY19SCSC	Finished	4 May 202(	4 May 202(	58 mins 32	12.50	0.50	0.00	0.50	0.50
1BY19SCSC	MADEHA K	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins 1 s	10.00	0.50	0.50	0.50	0.00
1BY19SCSC	MEGHANA	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins	8.00	0.50	0.50	0.50	0.00
1BY19SCSC	PUJITHA J	1BY19SCSC	Finished	4 May 202(	4 May 202(	52 mins 44	11.00	0.50	0.50	0.50	0.50
1BY19SCSC	VARSHINI I	1BY19SCSC	Finished	4 May 202(	4 May 202(	54 mins 19	9.50	0.50	0.50	0.50	0.50
1BY19SCSC	TEJASWINI	1BY19SCSC	Finished	4 May 202(	4 May 202(	57 mins 42	11.00	0.50	0.50	0.50	0.50
Overall average							10.50	0.50	0.43	0.50	0.36

Surname	First name	Email addr	State	Started on	Completed	Time taker	Grade/15.(	Q. 1 /0.50	Q. 2 /0.50	Q. 3 /0.50	Q. 4 /0.50
1BY19SCSC	GAUTHAM	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins 34	11.50	0.50	0.50	0.50	0.50
1BY19SCSC	DEEPTHI M	1BY19SCSC	Finished	4 May 202(	4 May 202(	58 mins 32	12.50	0.50	0.00	0.50	0.50
1BY19SCSC	MADEHA K	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins 1 s	10.00	0.50	0.50	0.50	0.00
1BY19SCSC	MEGHANA	1BY19SCSC	Finished	4 May 202(	4 May 202(	59 mins	8.00	0.50	0.50	0.50	0.00
1BY19SCSC	PUJITHA J	1BY19SCSC	Finished	4 May 202(	4 May 202(	52 mins 44	11.00	0.50	0.50	0.50	0.50
1BY19SCSC	VARSHINI I	1BY19SCSC	Finished	4 May 202(	4 May 202(	54 mins 19	9.50	0.50	0.50	0.50	0.50
1BY19SCSC	TEJASWINI	1BY19SCSC	Finished	4 May 202(	4 May 202(	57 mins 42	11.00	0.50	0.50	0.50	0.50
Overall average							10.50	0.50	0.43	0.50	0.36

Q. 17 /0.5(	Q. 18 /0.5(	Q. 19 /0.5(	Q. 20 /0.5(	Q. 21 /0.5(	Q. 22 /0.5(	Q. 23 /0.5(	Q. 24 /0.5(	Q. 25 /0.5(	Q. 26 /0.5(	Q. 27 /0.5(	Q. 28 /0.5(
0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.50	0.00	0.50	0.50	0.50
0.50	0.50	0.50	0.00	0.50	0.00	0.00	0.50	0.50	0.00	0.50	0.00
0.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.00	0.50	0.00	0.50	0.50	0.50	0.00	0.50	0.50
0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.50	0.50	0.00	0.50	0.50
0.43	0.50	0.50	0.07	0.50	0.36	0.43	0.50	0.43	0.29	0.50	0.43

Q. 29 /0.5(	Q. 30 /0.50
0.00	0.00
0.50	0.50
0.00	0.00
0.00	0.00
0.00	0.00
0.00	0.00
0.50	0.00
0.14	0.07

USN : 1BM19SCS03
Name : Madeha Kauser
Course Code : 18SCS21
Course Name : Managing Big Data
Branch / sem : Computer Science, 2nd Sem
Name of faculty : Dr. Anjan Krishnamurthy
Student Signature : Madeha.

III rd Internal Assignment.

Q3	HBase	RDBMS
	• It is distributed, Column-oriented data storage system	It is fixed schema, row-oriented databases with ACID properties & a sophisticated SQL query engine.
	• It has been designed from the ground up with a focus on scale in every direction: tall in no. of rows, wide in no. of columns & horizontally partitioned & replicated.	The emphasis is on a strong consistency, referential integrity, abstraction from the physical layer. & Complex Queries through SQL language.
	• It picks up where Hadoop left off by providing random reads & writes on top of HDFS.	You can easily create secondary indexes, perform complex inner & outer joins & count, sum, sort, group & page your data.
	• It is no sql. • No fixed schema	It requires SQL. It has fixed schema.

22/6/20 Madeha

- The table schemas mirror the physical storage, creating a system for efficient data structures serialization, storage & retrieval.
- The burden is on the application developer
- It is scalable
- Dynamic in nature

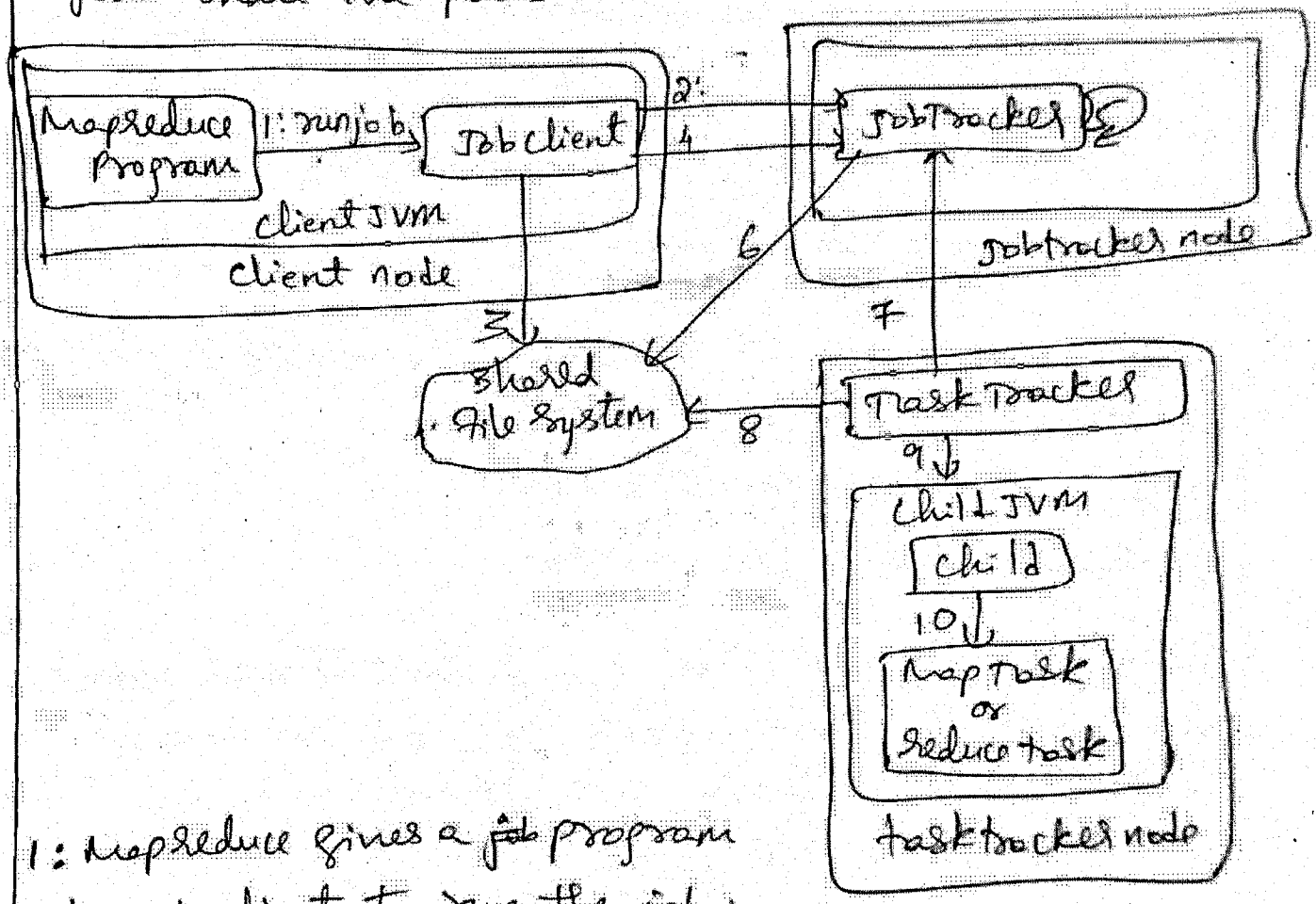
It has a majority of small-to-medium volume applications, there is no substitute for the ease of use, flexibility, maturity, & powerful.
- the scaling involves breaking code rules, loosing ACID restrictions.
- It is not scalable.
- Static in nature.

06 The Capacity scheduler takes a slightly different approach to multiuser scheduling.
- A cluster is made up of a number of queues which may be hierarchical & each queue has an allocated capacity.
- This is like the fair scheduler, except that within each queue, jobs are scheduled using FIFO scheduling.
- It allows users or organizations to simulate a seperate Mapreduce cluster with FIFO scheduling for each user or organization.
- It also supports FIFO job scheduling within

22/6/20 Nadeha

Pools as an option making it like the Capacity Scheduler.

It enforces fair sharing within each pool, so running jobs share the pools resources.



1: Mapreduce gives a job program to job client to run the job.

2: Ask the jobtracker for a new job id.

3: Copies the resources needed to run the job.

4: Tells the jobtracker that it is ready for execution

5: It initializes the job.

22/6/20 Nadeha

6: The job scheduler retrieves the i/p split computed by the client from shared file system.

7: Heartbeats tells the job tracker that a task tracker is alive.

8: It retrieves the job resources from file shared system.

9: Taskrunner launches a new JVM.

10: Then runs each task.

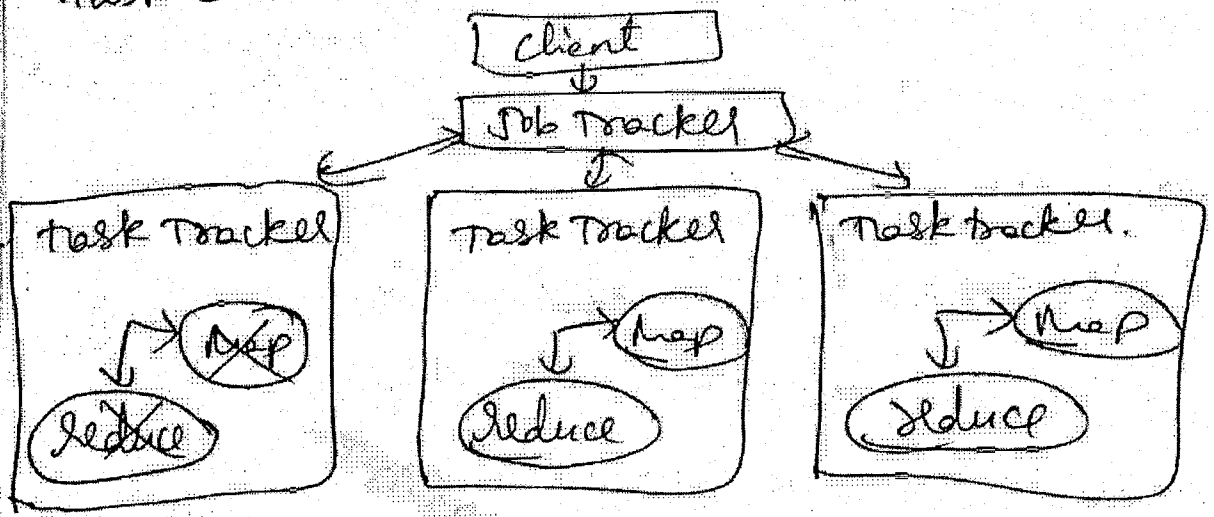
Q2 In the map reduce, there are 3 failures modes.

① Failure of the running task.

② Failure of the tasktracker

③ Failure of the job tracker.

Child task failure: this is a failure caused when there is an exception from the map or reduce task due to error in the code.



22/6/20 Madeha

The child JVM reports the error back to its parent tasktracker before it exists.
- A user log is created for this failure.
- The tasktracker marks the task attempt as failed filling up the client to run another task.
- If a task fails 4 times it will be not be retried further.
- A task attempt may be killed which is different from it failing. It may be killed because of duplicates

(2) Tasktracker failure.
- It is an another failure mode.
- If a tasktracker fails by crashing or running very slowly it will stop sending heartbeats to jobtracker.
- The jobtracker will notice a tasktracker that has stopped sending heartbeats & remove from its pool.
- A task tracker can also be blacklisted by jobtracker, even if the tasktracker has not failed.
- If more than 4 task fails at some time in a tasktracker then jobtracker records this as a fault.
- Blacklisted tasktrackers are not assigned tasks

22/6/20 Madeha

(3) Jobtracker failure.

- It is a most failure mode.
- Hadoop has no mechanism for dealing with failure of jobtracker.
- Machine failing is low.
- After restarting a jobtracker any jobs that were running at time it was stopped will need to be re-submitted.

Q7 It is a data model which is similar to bigdata which are designed to provide quick random access to huge amount of structured data.

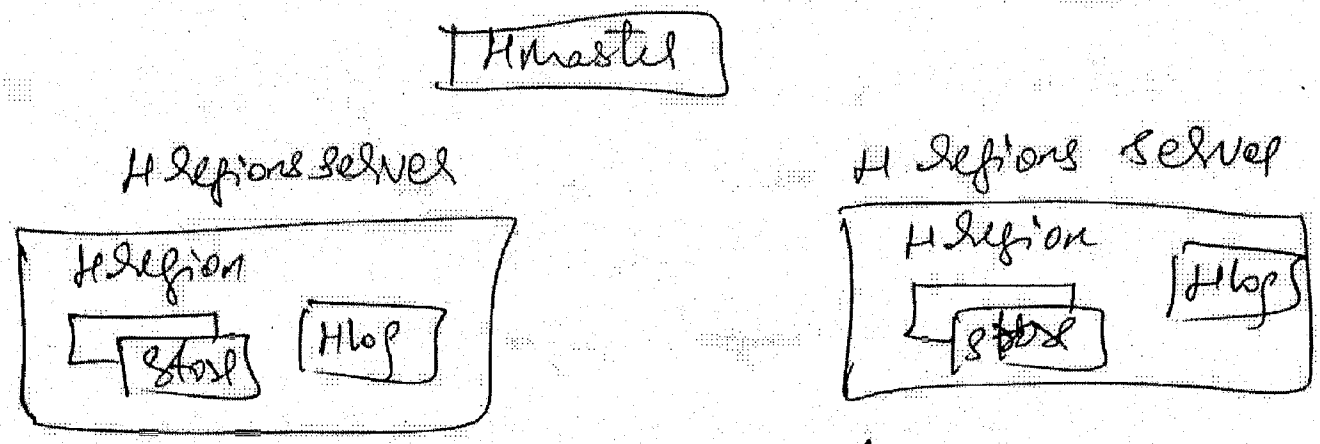
Characteristic lists

- It stores data into 2 forms row & column oriented.
- Contains master-slave relationship architecture.
- tables are partitioned into multiple regions. These regions are assigned to region servers.
- H master is responsible in doing functions such as performing administration, monitoring & managing clusters.
- H region server performs - Hosting & managing

22/6/20 Madeha.

Regions, splitting regions automatically, direct client communication, handling reads & write operation

- Hbase has components such as tables, rows, columns cells & versions



HBase Data Model

- It is suitable for converting traditional RDBMS to NoSQL base because
  - → data model is multidimensional way of arrangement so it is possible to process large data if RDBMS is converted to NO SQL.
  - → Data migration can be done.

Q5.
```
import org.apache.hadoop.hbase.client.Results;
import org.apache.hadoop.hbase.client.ResultScanner;
import org.apache.hadoop.hbase.client.Scan;
import org.apache.hadoop.hbase.util.Bytes;
Public class HBaseLoading
{
```

22/6/20 Madeha

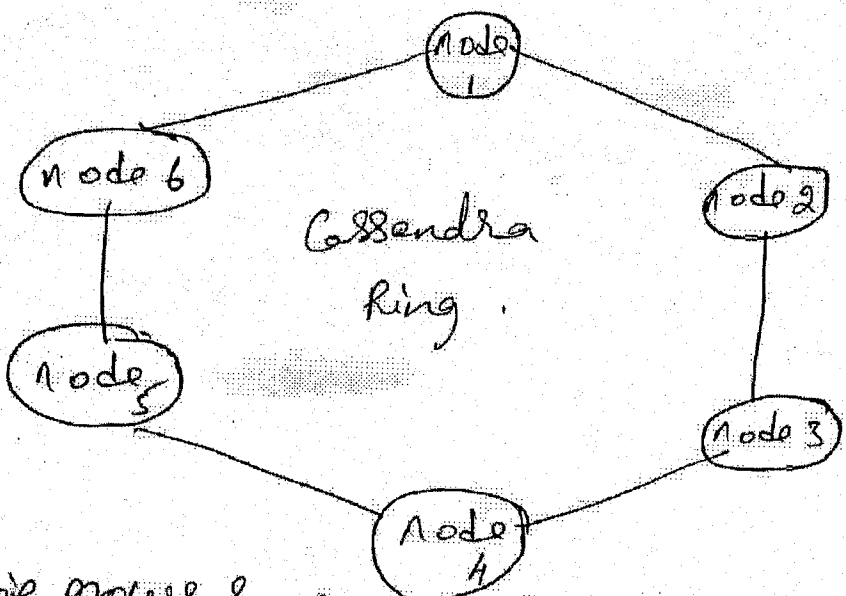
```
public static void main (String[] args[]) throws IOEx
                                                  -ception.
{
org.apache.hadoop.conf.Configuration Config =
    HBaseConfiguration.create();
Htable table = new HTable (Config, "guru99");
Put p= new put (Bytes.toBytes ("row1"));
P.add (Bytes.toBytes ("education"), Bytes.toBytes (
  "Col1"), Bytes.toBytes ("Big Data"));
P.add (Bytes.toBytes ("projects"), Bytes.toBytes ("Col2"),
  Bytes.toBytes ("HBase Tutorials"));
table.put(p);
```

Q10. In Cassandra, all nodes are the same & have peer to peer architecture & there is no concept of a master slave node.

- In Cassandra all nodes communicating with each other is gossip protocol.
- Gossip is the message system that Cassandra node uses to make their data consistent with each other.
- Gossip message that Cassandra nodes, virtual nodes used to make their data consistent

22/6/20 Madeha

with each other, & is used to enforce the replication factor in a cluster.



- The gossip process runs every second for every node & exchange state message with up to 3 other nodes in cluster
- Since, the whole process is decentralized, there is nothing or no one that Coordinate each node to gossip.
- Each node independently will always select one to ~~the~~ three peer to gossip with.
- In Cassandra the Communication b/w nodes is ~~of~~ like peer to peer Communication, where every node talks to the other.
- The gossip protocol is a method to resolve this Communication to a several chaos.

22/6/20 Madeha

1BY19SC503, Madeha Kausar, 18SCS21, Managing Big Data,
Computer Science, 2nd Sem, Dr. Anjan Krishnamurthy

(10)

• In Cassandra when one node talks to another, the node which is expected to respond, not only provides information about its status, but also provides information about the nodes that it has communicated with before.

22/6/20. Madeha.

B M S Institute Of Technology & Management

Dept of Computer Science & Engineering

BATCH:	2019	SEM:	II
SUBJECT:	Managing Big Data		
Faculty In-Charge:	Dr. Anjan K		

SESSION: Mar 2020-Jun 2020

PG COURSE

	TARGET		LEVEL
60%	STUDENTS MUST SCORE	60% & ABOVE	3-High
55%	STUDENTS MUST SCORE	60% & ABOVE	2-Moderate
50%	STUDENTS MUST SCORE	60% & ABOVE	1-Low

COURSE OUTCOMES	ATTAINMENT - I A	ATTAINMENT - VTU	OVERALL ATTAINMENT
CO1	3.00	0.00	1.20
CO2	3.00	0.00	1.20
CO3	3.00	0.00	1.20
CO4	3.00	0.00	1.20
			0.00

CLASS STRENGTH	7
SET IA TARGET	60%

CO No.		TEST-1										TEST-2										TEST-3											CO1			CO2			CO3			CO4			CO5			ALL COs
		2	3	1	2	2	2	3	4			2	2	3	3	2	3	3				4	3	3	4	3	3	3				5															THEORY EXT
	Question No.	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	#	1	2	3	4	5	6	7	8	9	10	Seminar & Assignment Assignment	TOTAL	PERCENT	Target >=55%	TOTAL	PERCENT	Target >=55%	TOTAL	PERCENT	Target >=55%	TOTAL	PERCENT	Target >=55%	TOTAL	PERCENT	Target >=55%	Grade Point	60
	MAXIMUM MARKS FOR QUESTION	6	6	6	6	6	6	6	6			6	6	6	6	6	6	6				10	10	10	10	10	10	10				20		%			%			%			%			%			
Sl No.	USN	Name																																													
1	1BY19SCS01	DEEPTHI M	4			4	6		4	6			6		5	6		6		5			7	7	7			8	6				20	0	0.00%		26	86.67%	Y	48	75.00%	Y	13	81.25%	Y	20	100.00%	Y		32
2	1BY19SCS02	GAUTHAM S K											5	4		5		3					8			7	8	7	8				20	0	0.00%		12	66.67%	Y	36	78.26%	Y	7	70.00%	Y	20	100.00%	Y		31
3	1BY19SCS03	MADEHA KAUSER		4		5		6	6	3			4		5	5			4	5			8	8	7		7	8	8				20	0	0.00%		15	83.33%	Y	67	77.91%	Y	11	68.75%	Y	20	100.00%	Y		30
4	1BY19SCS04	MEGHANA KUMAR K J	3					4	4				5	6		6		6	4														20	0	0.00%		24	80.00%	Y	14	77.78%	Y	0	0.00%		20	100.00%	Y		30
5	1BY19SCS05	PUJITHA J		3	6		6		4				6						3	6			7	6	7			5	7				20	6	100.00%	Y	12	100.00%	Y	41	64.06%	Y	7	70.00%	Y	20	100.00%	Y		34
6	1BY19SCS06	PEJASWINI A KANTANAVAR											5	5		5		3		3			7	8	6		8	7	7				20	0	0.00%		13	72.22%	Y	44	70.97%	Y	7	70.00%	Y	20	100.00%	Y		34
7	1BY19SCS07	VARSHINI N		5		1	6		4	5			3	6		5		5						6	7			7	7	7			20	0	0.00%		21	70.00%	Y	48	70.59%	Y	5	83.33%	Y	20	100.00%	Y		35
8																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
9																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
10																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
11																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
12																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
13																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
14																																	0	0.00%		0	0.00%		0	0.00%		0	0.00%		0	0.00%			
15																																	0	0.00%		0												
																																	Target >=55%	1			Target >=55%	7			Target >=55%	7			Target >=55%	6			Target >=55%	7			226
																																	CO1	100.00%			CO2	100.00%			CO3	100.00%			CO4	100.00%			CO5	100.00%			Avg 32.2857
																																																42.86%

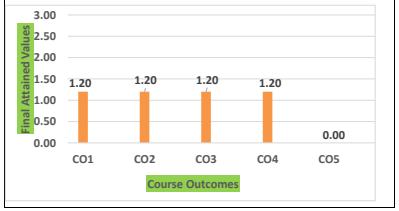
BATCH	2019
SEM	II
SUB	Managing Big Data
SESSION	Mar 2020-Jun 2020
Class Strength	7
Set Target I A	60%
Faculty	Dr. Anjan K
Target >=55%	
3	60%
2	55%
1	50%
No. of COs	4
Set Target Uni Theory	55.00%
No of students who were present in EXAM	7

	CO RESULT	PO1	PO2	PO3	PO4	PO5	PO6	PSO1	PSO2	Delivery Mechanism
CO1:Summarize the fundamentals and concepts of Big Data.	3									Lecture, PPT
CO2:Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	3	1					1			Lecture, PPT
CO3:Analyze methods and algorithms, to compare them to solve problems.	3				3	1	2			Lecture, Hands on, PPT
CO4:Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce	3						2			Lecture, Hands on, PPT
										Lecture, Hands on, PPT
SUM		1	0	0	3	1	5	0	0	
CIA*		1	0	0	3	1	1.7	0	0	

1 0 0 1 1 3 0 0

FINAL CO ATTAINMENT ANALYSIS

CO1	CO2	CO3	CO4	CO5
1.20	1.20	1.20	1.20	0.00



	CO RESULT	PO1	PO2	PO3	PO4	PO5	PO6	PSO1	PSO2
CO1:Summarize the fundamentals and concepts of Big Data.	1.20	0	0	0	0	0	0	0	0
CO2:Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured	1.20	0.4	0	0	0	0	0.4	0	0
CO3:Analyze methods and algorithms, to compare them to solve problems.	1.20	0	0	0	1.2	0.4	0.8	0	0
CO4:Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce	1.20	0	0	0	0	0	0.8	0	0
	0.00	0	0	0	0	0	0	0	0
SUM		0.4	0	0	1.2	0.4	2	0	0
DIRECT ATTAINMENT		0.4	0	0	1.2	0.4	0.7	0	0
INDIRECT ATTAINMENT		1			1	1	1	0	0

UNIVERSITY EXAM CONTRIBUTION

Attained all COs (=3) (meaning, more than 55% of the students were able to score more than 60% of the marks in university examination)

ACTION TAKEN

There was a gap for PO1. Research activity was conducted inclluded paper publication to overcome this gap. The outcome is: Students got more exposure towards research and development work to solve real life problem

CONCLUSION

High Attainment: PO3,PO4,PO5,PO6

Moderate Attainment: PO2

Low Attainment: PO1

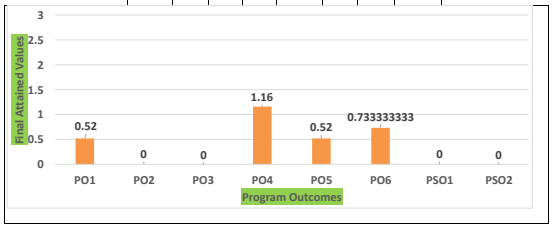
OBSERVATION

Not Attained: PO2, PO3 (to try to attain in the next academic year)

All the Moderate Attained POs to be improved and meets course threshold.

FINAL PO ATTAINMENT ANALYSIS

PO1	PO2	PO3	PO4	PO5	PO6	PSO1	PSO2
0.52	0	0	1.16	0.52	0.733	0	0



	PO1	PO2	PO3	PO4	PO5	PO6	PSO1	PSO2
FINAL ATTAINMENT	0.52	0	0	1.16	0.52	0.733333333	0	0