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 & MANAGEMENT

Department of Computer Science & Engineering

Calender of Events (CoE) 2019-20 (ODD Semester)

Month	Week	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Working Days	Events	Holidays
July	1st		29	30	31				2	29th July: Commencement of Higher Semester (3rd, 5th, & 7th Semester), 31st July: Jnanavardhan Commencement	
August						1	2	3	3	3rd Aug.: Commencement of First Semester Classes, First year welcome function	
	2nd	4	5	6	7	8	9	10	6	9th Aug.: Project Based Learning (PBL) and Final Year Project (FYP) Group Formation , 9th Aug PAC Meeting 1	
	3rd	11	12	13	14	15	16	17	4	15th Aug.: Independence Day Celebration	12/08/2019: Bakrid, 15/08/2019: Independence Day
	4th	18	19	20	21	22	23	24	6	23rd Aug.: PBL and FYP Synopsis Submission	
	5th	25	26	27	28	29	30	31	6	26th & 27th Aug.: Sports for Faculty Members , 30th Aug Industrial Visit -1, 30th Aug Dept. NBA Audit - 1, 31st Aug Invited Talk by Industry Expert/Alumni -1,	
September	6th	1	2	3	4	5	6	7	5	5th Sept.: Teachers' Day Celebration, 6th Sept PAC Meeting 2	02/09/2019: Ganesha Chathurthi
	7th	8	9	10	11	12	13	14	5	13th Sept Industrial Visit -1	10/09/2019: Last Day of Moharam
	8th	15	16	17	18	19	20	21	6	16th Sept.: PBL Review - 1, Engineer's Day (Sir. M. Visvesvaraya's Birthday) Celebration, 17th, 18th, & 19th Sept.: Internal Assessment (IA) Test - 1	
	9th	22	23	24	25	26	27	28	5	24th Sept.: Academic Monitoring / SMS Dispatch IA-1, 25th Sept.: Student's Feedback on Faculty - 1 , 27th Sept. Dept. NBA Audit - 2	28/09/2019: Mahalaya Amavasya
	10th	29	30						5		
October				1	2	3	4	5		5th Oct Invited Talk by Industry Expert/Alumni -2 , 4th Oct Industrial Visit -2, 4th Oct PAC Meeting 3	02/10/2019: Gandhi Jayanthi
	11th	6	7	8	9	10	11	12	4	11th Oct.: PBL Review - 2, 12th Oct.: Parents Teachers Association (PTA)	07/10/2019: Mahanavami/Ayudha Pooja, 08/10/2019: Vijayadasami
	12th	13	14	15	16	17	18	19	6	17th, 18th, & 19th Oct.: Internal Assessment (IA) Test - 2	
	13th	20	21	22	23	24	25	26	6	24th Oct.: Academic Monitoring / SMS Dispatch IA-2, Semaphore - 2019, 25th Oct.: Student's Feedback on Faculty - 2 , 25th Oct Dept. NBA Audit - 3	
	14th	27	28	29	30	31			4		29/10/2019: Balipadyami/Deepavali
November							1	2		1st Nov.: Kannada Rajyotsava Celebration , 2nd Nov Invited Talk by Industry Expert/Alumni -3	01/11/2019: Kannada Rajyotsava
	15th	3	4	5	6	7	8	9	6	4th Nov.: BMSIT&M Open Day, 5th & 6th Nov.: Final Year Project Review , 9th NovInvited Talk by Industry Expert/Alumni -4, 8th Nov PAC Meeting 4	
	16th	10	11	12	13	14	15	16	5	13th, 14th, & 16th Nov.: Internal Assessment (IA) Test - 3	15/11/2019: Kanakadasa Jayanthi
	17th	17	18	19	20	21	22	23	6	20th Nov.: Academic Monitoring / SMS Dispatch IA-3	
	18th	24	25	26	27	28	29	30	6	29th Nov.: Last Working Day (1st Semester), 30th Nov.: Last Working Day (3rd, 5th, & 7th Semester), 29th Nov Dept. NBA Audit - 4	
December		1	2	3	4	5	6	7		2nd Dec Subject Allotment Meeting, 6th Dec PAC Meeting 5	05/06/2019 Khutub-E-Ramzan
	19th	8	9	10	11	12	13	14			
	20th	15	16	17	18	19	20	21			
	21st	22	23	24	25	26	27	28		27th Dec Dept. NBA Audit - 5	25/12/2019: Christmas
	22nd	29	30	31							
Total Number of Working Days									96		

VTU Examination

Practical Examinations: 03-12-2019 to 13-12-2019 (I Sem B.E/B.Tech).
03-12-2019 to 13-12-2019 (III, V & VII Sem B.E/B.Tech)
03-12-2019 to 07-12.2019 (III & V Sem MCA)
Internship Viva Voce: 12-01.2020 to 19-01.2019 (III Sem M.Tech)

Theory Examinations: 16-12-2019 to 04-01-2020 (I Sem B.E/B.Tech)
16-12-2019 to 07-02-2020 (III, V, & VII Sem B.E/B.Tech)
09-12-2019 to 28-12-2019 (III & V Sem MCA)
27-12-2019 to 10-01-2020 (III Sem M.Tech)
Commencement of EVEN Semester: 27-01-2020 (II Sem B.E/B.Tech, II & IV MCA/M. Tech)
Commencement of EVEN Semester: 10-02-2020 (IV, VI, & VIII Sem B.E/B.Tech

**Student Centric Activity - Technical Talk/Seminar/Workshop/Quiz/HANDS-ON Session on Blended Learning (OBE Related activities) to be organized by the respective Departments.*

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA – BANGALORE - 64

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIVIDUAL TIME TABLE FOR THE ACADEMIC YEAR 2018 – 19(EVEN SEM)

Name: Dr.Anjan Krishnamurthy	Subject: CNC,MBD	Semester:VI-A, II M.TECH	Class room: BSN CR 101, BSN TR-401	W.E.F: 01-02-2019

	I 8. 30 – 9. 30	II 9. 30 – 10. 30	10. 30 – 10.50	III 10.50 – 11.50	IV 11.50 – 12.50	12.50 – 1.45	V 1.45 – 2.40	VI 2.40 – 3.35	VII 3.35 – 4.30
MONDAY				CNC					
TUESDAY		MBD			CNC				
WEDNESDAY	CNC				MBD				
THURSDAY	CPL-LAB-B1			MBD			DAA-LAB-D1		
FRIDAY	MBD			CNC					
SATURDAY									

Total Workload(2T+2L)	13 Hours

Time table officer

HoD

VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI
Scheme of Teaching and Examination – 2018-19
M.Tech COMPUTER SCIENCE AND ENGINEERING (SCS)
Outcome Based Education(OBE) and Choice Based Credit System (CBCS)

II SEMESTER

Sl. No	Course	Course Code	Course Title	Teaching Hours /Week		Examination				Credits
				Theory	Practical/ Field work/ Assignment	Duration in hours	CIE Marks	SEE Marks	Total Marks	
1	PCC	18SCS21	Managing Big Data	04	--	03	40	60	100	4
2	PCC	18SCS22	Advanced Algorithms	04	--	03	40	60	100	4
3	PCC	18SCS23	Cloud Computing	04	--	03	40	60	100	4
4	PEC	18SCS24X	Professional elective 2	04	--	03	40	60	100	4
5	PEC	18SCS25X	Professional elective 3	04	--	03	40	60	100	4
6	PCC	18SCSL26	Mini Project	--	04	03	40	60	100	2
7	PCC	18SCS27	Technical Seminar	--	02	--	100	--	100	2
			TOTAL	**20**	**06**	**18**	**340**	**360**	**700**	**24**

Note: PCC: Professional core, PEC: Professional Elective.

Professional Elective 2

Course Code under 18SCS24X	Course title
18SCS241	Advances in Storage Area Network
18SCS242	Agile Technologies
18SCS243	Business Intelligence and its Applications
18SCS244	Data Mining & Data Warehousing

Professional Elective 3

Course Code under 18SCS25X	Course title
18SCS251	Advances in Computer Graphics
18SCS252	Trends in Artificial Intelligence and Soft Computing
18SCS253	Object Oriented Software Engineering
18SCS254	Advances in Digital Image Processing

Note:
1. Technical Seminar: CIE marks shall be awarded by a committee comprising of HoD as Chairman, Guide/co-guide, if any, and a senior faculty of the department. Participation in the seminar by all postgraduate students of the same and other semesters of the programme shall be mandatory.
The CIE marks awarded for Technical Seminar, shall be based on the evaluation of Seminar Report, Presentation skill and Question and Answer session in the ratio 50:25:25.
2. Internship: All the students shall have to undergo mandatory internship of 6 weeks during the vacation of I and II semesters and /or II and III semesters. A University examination shall be conducted during III semester and the prescribed credit shall be counted in the same semester.Internship shall be considered as a head of passing and shall be considered for the award of degree. Those, who do not take-up/complete the internship shall be declared as failed and have to complete during the subsequent University examination after satisfying the internship requirements.

MANAGING BIG DATA
[As per Choice Based Credit System (CBCS) scheme]
(Effective from the academic year 2018 -2019)
SEMESTER – II

Subject Code	18LNI251 / 18SCE21 / 18SCN252 / **18SCS21** / 18SFC331 / 18SIT31 / 18SSE322	IA Marks	40
Number of Contact Hours/Week	04	Exam Marks	60
Total Number of Contact Hours	50	Exam Hours	03

CREDITS – 04

Course objectives: This course will enable students to
- 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

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	**10 Hours**

definition – HiveQL data manipulation – HiveQL queries. **RBT: L1, L2, L3**	

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

	LESSON PLAN	Date: 1/03/2019
	Sub Code & Name : 18SCS21- MANAGING BIG DATA	Page 1 of 6
	Semester : II | Program: M.Tech CSE	
	Academic Year: 2018-19	
	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 Analyze 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.

VTU 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**

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

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
Units	1			2				3			4				5	

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	3/3/2019	
Name and Affiliation	Dr. Anjan Krishnamurthy, Associate Professor, Dept. of CSE, BMSIT&M	
Date	03/03/2019	

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

Minutes of Meeting with PAC –Batch 2018-19

Date	05 -03-2019	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	Prof. Radhika K R	

Sl.No	Discussion	Action By/ Responsible	Action Taken
1	**Agenda:** Course Outcomes, CO – PO Mapping, Gap Identification for Managing Big Data (18SCS21)		

The **Course Outcomes (COs)** for Managing Big Data given in university curriculum are as follows:

Course Coordinator

CO No.	Course Outcome	BT Level
PCSE.121.1	Describe big data and use cases from selected business domains	K1
PCSE.121.2	Explain NoSQL big data management	K2
PCSE.121.3	Install, configure, and run Hadoop and HDFS Perform map-reduce analytics using Hadoop	K3
PCSE.121.4	Use Hadoop related tools such as HBase, Cassandra, Pig, and Hive for big data Analytics	K3

The Observations of the committee are as follows:

1. No COs is related to the research topic of Big Data concepts and practical aspects of the configuration and programming with Map-Reduce (Hadoop).
2. Emphasis of NoSQL usage was missing in CO and not highlighting its applications.
3. No COs is related to advance topics of conventional algorithms to Map reduce algorithm.

Gap identified:
Gaps are identified for PO1, PO2.

Redefined COs-

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

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

HOD

BMS INSTITUTE OF TECHNOLOGY, YELAHANKA, LORE-64

Batch: 2018-19 SEM: II

PG COURSE : COMPUTER SCIENCE AND ENGINEERING

AC. YEAR : 2018-2019

SL. NO	USN	NAME OF THE CANDIDATE
1.	1BY18SCS01	BHAGYASHREE A V
2.	1BY18SCS02	CHAITHRASHREE H S
3.	1BY18SCS03	DIVYASHREE S
4.	1BY18SCS04	FASIHA KAUSAR
5.	1BY18SCS05	KAVERI T HOMBAL
6.	1BY18SCS06	NAVEENKUMAR K V
7.	1BY18SCS07	P PRAJWALA
8.	1BY18SCS08	PURUSHOTHAM NAIDU V
9.	1BY18SCS09	RAJESHWARI N
10.	1BY18SCS10	RAMYA P L
11.	1BY18SCS11	RANJINI N
12.	1BY18SCS12	SNEHA S
13.	1BY18SCS13	SRIVATSA RAJU S
14.	1BY18SCS14	SUDHANSHU GUPTA
15.	1BY18SCS15	VIJAYALAKSHMI HOLIMATH

PRINCIPAL

BMS Inst. of Tech. & Mgmt
Doddaballapur Main Road
Avalahalli, Yelahanka, Bl'lore-64

12

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA – BENGALURU – 64
Department of Computer Science and Engineering

FIRST INTERNAL ASSESSMENT TEST, April-2019 (CBCS)

Subject: Managing Big Data (MBD) **Subject Code: 18SCS21** **Semester : 2nd Sem M.Tech CSE**

Max. Marks : 30 **Date: 11-04-2019** **Time: 9:30AM to 11:00AM**

Staff: Dr. Anjan Krishnamurthy

Answer ONE full question from Part A to Part C. Part D & E is compulsory.

			CO	BT
	Part A			
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.	*6 Marks*	CO2	K3
	OR			
2	Distinguish between Volunteer Computing and Grid Computing. Consider computational architecture viewpoints for the answer.	*6 Marks*	CO3	K2
	Part B			
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?	*6 Marks*	CO1	K2
	OR			
4	Explain CAP theorem. Distinguish between CAP concepts with ACID property.	*6 Marks*	CO2	K2
	Part C			
5	List and explain the broad classification of NoSQL data model. Provide appropriate examples with NoSQL DB to explain each data model.	*6 Marks*	CO2	K2
	OR			
6	Illustrate with examples the need for graph databases in storing Big Data.	*6 Marks*	CO2	K2
	Part D			
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.	*6 Marks*	CO3	K4
	Part E			
8	Write a columnar family DB for storing the customer and order details of the online e-commerce system using <key,value> pair mechanism.	*6 Marks*	CO4	K3

Course Outcomes (COs)

Students will be able to

CO 1	Summarize the fundamentals and concepts of Big Data.
CO 2	Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.
CO 3	Analyze methods and algorithms, to compare them to solve problems.
CO 4	Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map-Reduce.

Revised Bloom's Taxonomy

K1- Remembering, K2 - Understanding, K3 - Applying, K4 - Analyzing, K5 - Evaluating, K6 - Creating

*-- Blended Learning

HoD

Program Coordinator

Course Coordinator

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA – BENGALURU – 64
Department of Computer Science and Engineering

SCHEME FOR THE FIRST INTERNAL ASSESSMENT TEST, April-2019 (CBCS)

Subject: Managing Big Data (MBD) **Subject Code: 18SCS21** **Semester : 2nd Sem M.Tech CSE**

Max. Marks : 30 **Date: 11-04-2019** **Time: 9:30AM to 11:00AM**

Staff: Dr. Anjan Krishnamurthy

1*	Crowd source analytics (3m) NoSQL database for crowd source analytics (3m) Choice of NoSQL DB is Mongo DB	*6 Marks*
2	Differences between grid and volunteer computing (any three points with example) (3x2=6) Grid Computing • Nodes in the same cluster and processing is faster • Network bandwidth is shared between racks • Computation is performed with the organization. • Common resource facilities. Volunteer Computing • Computing resources are offered across globe. • Resources are geographically distributed. • Abundant computing resources. • Efficiency and security is less compared to grid	*6 Marks*
3	Formula and solving (4m) Final Answer (2m) Time taken = data size / speed Final answer is 37.03mins for 2 TB data and 19 mins for 1 TB data.	*6 Marks*
4	CAP theorem (2m) Any two differences between CAP and ACID (2x2=4)	*6 Marks*
5	NoSQL classification (2m) • Key/value store • Document store • Columnar • Graph • Relational Data model explanation (4x1=4m)	*6 Marks*
6	Graph database and data model of graph data (3m) Examples of graph database (3m)	*6 Marks*
7*	Details of the web analytics application (2m) Data semantics used in web (2m) MR approach for the web analytics (2m) Ex: Web analytics for an online store based customer purchase	*6 Marks*
8	Key,value pair system (2m) e-commerce system for columnar db (4m)	*6 Marks*

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Avalahalli, Doddaballapur Main Road, Bengaluru - 560064

SECOND INTERNAL ASSESSMENT TEST, MAY 2018 - 19

Name of the Course	**Managing Big Data**	**Course Code**	**18SCS21**
Branch & Semester	**2nd Sem M.Tech CSE**	**Date**	**08-05-2019 9:30AM to 11:00AM (MS)**
Name of the Course Coordinator	**Dr. Anjan Krishnamurthy**	**Max. Marks**	**30**

Note: *Answer **THREE** full questions from **Part A** and **Part B** questions are compulsory.*

Qn. No.	PART A	Marks	CO
1.	With a neat diagram, describe different types of distribution models. Give the complete categorization and examples for each type of distribution model.	06 M	**CO2 K2**
	OR		
2.	With neat diagram, show the anatomy of the file write. State the algorithm used for file write in the HDFS environment.	06 M	**CO2 K2**
3.	Discuss the scale-out in map/reduce environment? Give the roles of the Job-tracker and task-tracker in the overall process.	06 M	**CO3 K2**
	OR		
4.	Write code snippet to deal with the compressed data in Hadoop. Explain the need for data compression.	06 M	**CO3 K3**
5.	Give 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.	06 M	**CO3 K2**
	OR		
6.	Explain the process of clumping in data nodes. List the factors to be considered for arranging data in the nodes.	06 M	**CO2 K2**
	PART B		
7.	**Innovative question** Write the Hadoop pipe to determine max temperature using the unstructured data provided in weather mining data. Write C++ code for the same.	06 M	**CO3 K3**
8.	**Case Study Question** With sample weather mining data, illustrate the process of the map-reduce for obtaining minimum temperature? Write suitable map and reduce functions.	06 M	**CO3 K3**

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

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA – BENGALURU – 64
Department of Computer Science and Engineering

SCHEME FOR THE SECOND INTERNAL ASSESSMENT TEST, May-2019 (CBCS)

Subject: Managing Big Data (MBD) **Subject Code: 18SCS21** **Semester : 2nd Sem M.Tech CSE**

1*	Distribution models types (4m) Examples (2m)	*6 Marks*
2	Diagram for File write in HDFS (2m) Algorithm for HDFS (2m) HDFS calls for file write (2m)	*6 Marks*
3	Scale-out in Map Reduce (2m) Job tracker Role (2m) Task Tracker role (2m)	*6 Marks*
4		*6 Marks*
5	Network Topology of Hadoop (2m) Example (2m) Representations and metrics (2m)	*6 Marks*
6	Data clumping for data arrangement in racks (3m) Factors for data clumping (3m) • Locality • Bandwidth • Data type	*6 Marks*
7*	Strategy for determining max temperature (2m) C++ code (4m)	*6 Marks*
8	Design the MR module for temperature calculation. (2m) Map and Reduce functions for the same. (2+2=4m)	*6 Marks*

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Avalahalli, Doddaballapur Main Road, Bengaluru - 560064

THIRD INTERNAL ASSESSMENT TEST, Jun 2018 - 19

Name of the Course	**Managing Big Data**	**Course Code**	**18SCS21**
Branch & Semester	**2nd Sem M.Tech CSE**	**Date**	**03-06-2019 9:30AM to 11:00AM (MS)**
Name of the Course Coordinator	**Dr. Anjan Krishnamurthy**	**Max. Marks**	**30**

Note: *Answer **THREE** full questions from **Part A** and **Part B** questions are compulsory.*

Qn. No.	PART A	Marks	CO
1.	With neat diagram, explain how fair scheduling is performed on map reduce jobs?	06 M	CO3 K2
	OR		
2.	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.	06 M	CO3 K2
3.	Distinguish between HBase and RDBMS	06 M	CO2 K4
	OR		
4.	List and explain different failures that are possible in Map-Reduce 1 with the help of diagram.	06 M	CO3 K2
5.	With appropriate diagrams, give the anatomy of the job run on Map-Reduce 2 platform.	06 M	CO4 K2
	OR		
6.	Illustrate the process of the shuffle and sort in Map-Reduce platform. Use related diagram wherever required.	06 M	CO4 K2
	PART B		
7.	**Innovative question** 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.	06 M	CO4 K4
8.	**Case Study Question** Write a MapReduce function to insert column family into HBase Database.	06 M	CO3 K3

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

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA – BENGALURU – 64
Department of Computer Science and Engineering

SCHEME FOR THE THIRD INTERNAL ASSESSMENT TEST, Jun-2019 (CBCS)

Subject: Managing Big Data (MBD) **Subject Code: 18SCS21** **Semester : 2nd Sem M.Tech CSE**

1*	Diagram for Job Scheduling in Hadoop (2m) Explanation (4m)	*6 Marks*
2	Program of MR (3m) Advantages (2m) Limitation of MR (1m)	*6 Marks*
3	Any three differences between HBASE and RDBMS with focus on working principles and CAP theorem. (3x2=6)	*6 Marks*
4	Diagram (2m) Types of the failures (4m) • Task Failure • Task tracker Failure • Job tracker Failure • Name Node failure	*6 Marks*
5	MR Anatomy Diagram (3m) Explanation (3m) - YARN - Job Submission - Task Assignment - Execution of Task and Job	*6 Marks*
6	Diagram (2m) Shuffle operation (2m) Sort Operation (2m)	*6 Marks*
7*	RDBMS Tables (3m) Cassandra version of the tables (3m)	*6 Marks*
8	HBase Explanation (2m) Design of MR functions and tables in HBase (4m)	*6 Marks*

Second Semester M.Tech. Degree Examination, June/July 2019
Managing Big Data

Time: 3 hrs. Max. Marks: 100

Note: Answer any FIVE full questions, choosing ONE full question from each module.

Module-1

1 a. Explain about Big data analytics. (10 Marks)
b. Explain about unstructured data and implementing unstructured data management. (10 Marks)

OR

2 a. Explain briefly on Web analytics in Big data. (12 Marks)
b. Explain about crowd sourcing analytics and Mobile business intelligence. (08 Marks)

Module-2

3 a. Explain about Aggregate data models and key value document data models. (12 Marks)
b. Explain about Schema less databases in detail. (08 Marks)

OR

4 a. Explain about distribution models in detail. (12 Marks)
b. Explain about composing map-reduce calculations. (08 Marks)

Module-3

5 a. *Explain about Hadoop streaming and Hadoop pipes in detail.* (12 Marks)
b. *Explain about HDFS design and its concepts.* (08 Marks)

OR

6 a. Explain about file based data structure. (12 Marks)
b. Explain about Avro. (08 Marks)

Module-4

7 a. Explain detail about anatomy of Mapreduce Job Run. (12 Marks)
b. Explain about unit test with MR unit in detail. (08 Marks)

OR

8 a. Explain about YARN and job scheduling in MapReduce applications. (12 Marks)
b. Explain about types of MapReduce. (08 Marks)

Module-5

9 a. Explain about Hbase clients and its example. (10 Marks)
b. Explain about Cassandra data model and its example. (10 Marks)

OR

10 a. Explain about Pig data model in detail. (12 Marks)
b. Explain about Hive QL data in detail and its Queries. (08 Marks)

* * * * *

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

Assignment list for MBD 18SCS21 – 2019

Sl. No.	USN	Student Name	Topic Name	RBT	CO	PO1	PO2	PO3	PO4	PO5	PO6
1.	1BY18SCS01	Bhagyashree A V	Convergent Analytical Tools	K5	CO4		✓				✓
2.	1BY18SCS02	Chaitrashree H S	Storing massive small files in Hadoop	K5	CO4		✓				✓
3.	1BY18SCS03	Divyashree S	Hadoop Framework and Cluster	K5	CO4		✓				✓
4.	1BY18SCS04	Fasiha Kausar	Big Data Analysis Using Hadoop Framework	K5	CO4		✓				✓
5.	1BY18SCS05	Kaveri T Hombal	Load Balancing Tools For Hadoop	K5	CO4		✓				✓
6.	1BY18SCS06	Naveen Kumar K V	RDF data transfer and query on Hadoop Framework	K5	CO4		✓				✓
7.	1BY18SCS07	P Prajwala	Intrinsic Security Issues	K5	CO4		✓				✓
8.	1BY18SCS08	Purushotham Naidu V	Enhancing Performance And Efficiency	K5	CO4		✓				✓
9.	1BY18SCS09	Rajeshwari N	Hadoop Distributed File System	K5	CO4		✓				✓
10.	1BY18SCS10	Ramya PL	Data Flow Framework	K5	CO4		✓				✓
11.	1BY18SCS11	Ranjini N	Securing Cloud Using Fog Computing	K5	CO4		✓				✓
12.	1BY18SCS12	Sneha S	Emerging Real Time Streaming Analytics	K5	CO4		✓				✓
13.	1BY18SCS13	Srivatsa Raju S	Structural Analysis of HPC's	K5	CO4		✓				✓
14.	1BY18SCS14	Sudhanshu Gupta	A Structure For Hadoop-Compatible Private Data Evaluation	K5	CO4		✓				✓

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA, BENGALURU – 560064
Department of Computer Science and Engineering

Date: 3rd Apr 2019

Course Name: Managing Big Data
Faculty: Dr. Anjan Krishnamurthy
Type of Assignment: Literature Survey with research Publications

Assignment list for MBD 18SCS21 – 2019

Sl. No.	USN	Student Name	Topic Name	RBT	CO	PO1	PO2	PO3	PO4	PO5	PO6
1.	1BY18SCS01	Bhagyashree A V	Convergent Analytical Tools	K5	CO4	✓	✓				✓
2.	1BY18SCS02	Chaitrashree H S	Storing massive small files in Hadoop	K5	CO4	✓	✓				✓
3.	1BY18SCS03	Divyashree S	Hadoop Framework and Cluster	K5	CO4	✓	✓				✓
4.	1BY18SCS04	Fasiha Kausar	Big Data Analysis Using Hadoop Framework	K5	CO4	✓	✓				✓
5.	1BY18SCS05	Kaveri T Hombal	Load Balancing Tools For Hadoop	K5	CO4	✓	✓				✓
6.	1BY18SCS06	Naveen Kumar K V	RDF data transfer and query on Hadoop Framework	K5	CO4	✓	✓				✓
7.	1BY18SCS07	P Prajwala	Intrinsic Security Issues	K5	CO4	✓	✓				✓
8.	1BY18SCS08	Purushotham Naidu V	Enhancing Performance And Efficiency	K5	CO4	✓	✓				✓
9.	1BY18SCS09	Rajeshwari N	Hadoop Distributed File System	K5	CO4	✓	✓				✓
10.	1BY18SCS10	Ramya PL	Data Flow Framework	K5	CO4	✓	✓				✓
11.	1BY18SCS11	Ranjini N	Securing Cloud Using Fog Computing	K5	CO4	✓	✓				✓
12.	1BY18SCS12	Sneha S	Emerging Real Time Streaming Analytics	K5	CO4	✓	✓				✓
13.	1BY18SCS13	Srivatsa Raju S	Structural Analysis of HPC's	K5	CO4	✓	✓				✓

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
YELAHANKA, BENGALURU – 560064
Department of Computer Science and Engineering

14.	1BY18SCS14	Sudhanshu Gupta	A Structure For Hadoop-Compatible Private Data Evaluation	K5	CO4	✓	✓				✓

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

Rubrics

Dimension	Maximum Marks	High	Medium	Low
Introduction		Position and exceptions, if any, are clearly stated. Organization of the argument is completely and clearly outlined and implemented.	Position is clearly stated. Organization of argument is clear in parts or only partially described and mostly implemented.	Position is vague. Organization of argument is missing, vague, or not consistently maintained.
	5	4-5 pts	2-3 pts	0-1 pts
Research		Research selected is highly relevant to the argument, is presented accurately and completely – the method, results, and implications are all presented accurately; Theory is relevant, accurately described and all relevant components are included; relationship between research and theory is clearly articulated and accurate.	Research is relevant to the argument and is mostly accurate and complete – there are some unclear components or some minor errors in the method, results or implications. Theory is relevant and accurately described, some components may not be present or are unclear. Connection to theory is mostly clear and complete, or has some minor errors.	Research selected is not relevant to the argument or is vague and incomplete – components are missing or inaccurate or unclear. Theory is not relevant or only relevant for some aspects; theory is not clearly articulated and/or has incorrect or incomplete components. Relationship between theory and research is unclear or inaccurate, major errors in the logic are present.
	5	4-5 pts	2-3 pts	0-1 pts
Conclusions		Conclusion is clearly stated and connections to the research and position are clear and relevant. The underlying logic is explicit. 4-5 pts	Conclusion is clearly stated and connections to research and position are mostly clear, some aspects may not be connected or minor errors in logic are present.	Conclusion may not be clear and the connections to the research are incorrect or unclear or just a repetition of the findings without explanation. Underlying logic has major flaws; connection to position is not clear.
	5	4-5 pts	2-3 pts	0-1
Report Writing		Paper is coherently organized and the logic is easy to follow. There are no spelling or grammatical errors and terminology is clearly defined. Writing is clear and concise and persuasive.	Paper is generally well organized and most of the argument is easy to follow. There are only a few minor spelling or grammatical errors, or terms are not clearly defined. Writing is mostly clear but may lack conciseness.	Paper is poorly organized and difficult to read – does not flow logically from one part to another. There are several spelling and/or grammatical errors; technical terms may not be defined or are poorly defined. Writing lacks clarity and conciseness.
	5	4-5 pts	2-3 pts	0-1

Grading policies:

- The last date for the submission of the assignment is on or before 5th Jun 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.

Feed Back and Analysis:

- Students have implemented real time Hadoop system to demonstrate the data processing in BigData environment. The implementation helped them to learn new concept enabling research work and publication in this domain. Through this assignment students are enabled to attain CO4, PO1, PO2 and PO6.

Course Coordinator Signature:

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

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

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.

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

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

YELAHANKA, BENGALURU – 560064

Department of Computer Science and Engineering

Assignment Evaluation for MBD (18SCS21)

Sl. No.	USN	Student Name	Introduction (5 marks)	Literature survey (8 marks)	New directions of the work (2 marks)	Report Writing (5 marks)	Total (20 marks)
1.	1BY18SCS01	Bhagyashree A V	3	7	2	3	15
2.	1BY18SCS02	Chaitrashree H S	5	6	2	4	17
3.	1BY18SCS03	Divyashree S	5	8	2	4	19
4.	1BY18SCS04	Fasiha Kausar	5	6	2	4	17
5.	1BY18SCS05	Kaveri T Hombal	5	8	2	4	19
6.	1BY18SCS06	Naveen Kumar K V	4	7	2	3	16
7.	1BY18SCS07	P Prajwala	5	7	2	4	18
8.	1BY18SCS08	Purushotham Naidu V	4	7	2	3	16
9.	1BY18SCS09	Rajeshwari N	5	6	2	4	17
10.	1BY18SCS10	Ramya PL	3	7	2	4	16
11.	1BY18SCS11	Ranjini N	4	7	2	2	15
12.	1BY18SCS12	Sneha S	4	7	2	4	17
13.	1BY18SCS13	Srivatsa Raju S	3	7	2	3	15
14.	1BY18SCS14	Sudhanshu Gupta	3	7	2	4	16

Course Coordinator

VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI

HELLO, 1BYSCS

B.M.S. INSTITUTE OF TECHNOLOGY BENGALURU

LOGOUT! (logout.php)

Dashboard (dashboard.php) Subject - Faculty Allotment (subject_mapping.php) IA Marks Entry (internal_marks_home.php)

HOD Dashboard (hod_dashboard.php) Instruction (instruction.php)

INTERNAL MARKS ENTRY

Regular Sem

2nd Semester

SubjectCode

18SCS21 - Managing Big Data

Choose Faculty

1BYCS0012355 - ANJAN KRISHNAMURTHY

GO AHEAD

NOTE: Please Click on Save Button Before Going To Next Page of Marks Entry.

IA Entry for 18SCS21 - Managing Big Data

TOTAL
14

SAVED
14

REMAINING
0

Show 25 entries Search:

Sl.No.	USN	Student Name	Attendence	Marks Scored	Max Marks	Status
1	1BY18SCS01	BHAGYASHREE A V	Present	33	40	Frozer
2	1BY18SCS02	CHAITHRASHREE H S	Present	30	40	Frozer
3	1BY18SCS03	DIVYA SHREE S	Present	35	40	Frozer
4	1BY18SCS04	FASIHA KAUSAR	Present	34	40	Frozer
5	1BY18SCS05	KAVERI T HOMBAL	Present	34	40	Frozer
6	1BY18SCS06	NAVEENKUMAR K V	Present	30	40	Frozer

Sl.No.	USN	Student Name	Attendence	Marks Scored	Max Marks	Status
7	1BY18SCS07	P PRAJWALA	Present	32	40	Frozer
8	1BY18SCS08	PURUSHOTHAM NAIDU V	Present	34	40	Frozer
9	1BY18SCS09	RAJESHWARI N	Present	32	40	Frozer
10	1BY18SCS10	RAMYA P L	Present	30	40	Frozer
11	1BY18SCS11	RANJINI N	Present	30	40	Frozer
12	1BY18SCS12	SNEHA S	Present	37	40	Frozer
13	1BY18SCS13	SRIVATSA RAJU S	Present	31	40	Frozer
14	1BY18SCS14	SUDHANSHU GUPTA	Present	30	40	Frozer

Showing 1 to 14 of 14 entries

Previous 1 Next

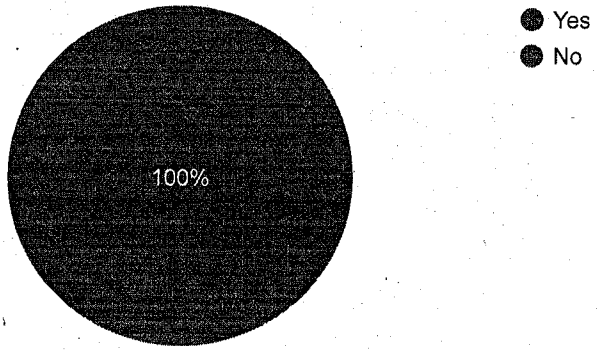
NOTE:
▶ The values are already Submitted and Frozen!

Course End Survey for MBD (18SCS21)

11 responses

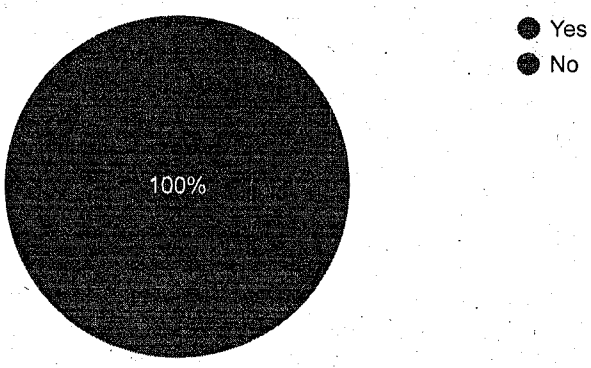
Did the course allow you to independently think to solve problems related to Big Data leading to research work?

11 responses



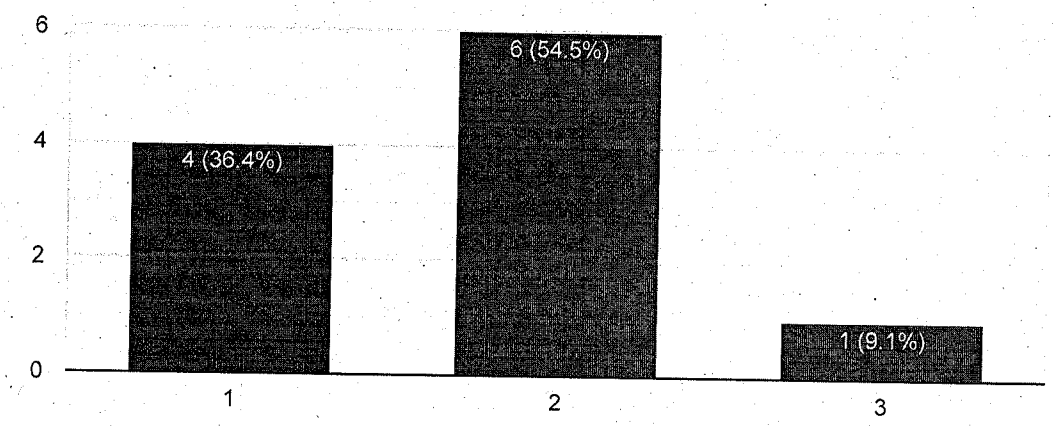
Did the course give you an ability to articulate, present, write reports or documents?

11 responses



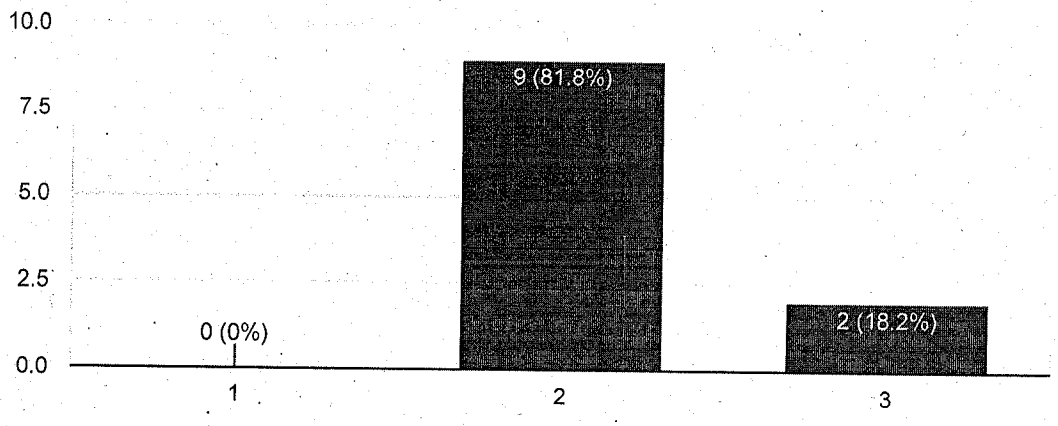
Rate the level of your mastery over the course before taking it.(1- Low, 2- Medium, 3 -High)

11 responses



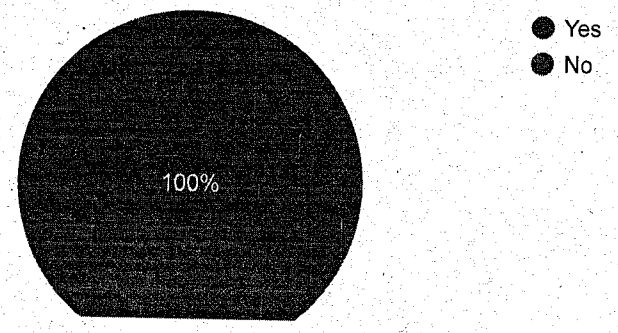
Rate the level of your mastery over the course after taking it.(1- Low, 2- Medium, 3 -High)

11 responses



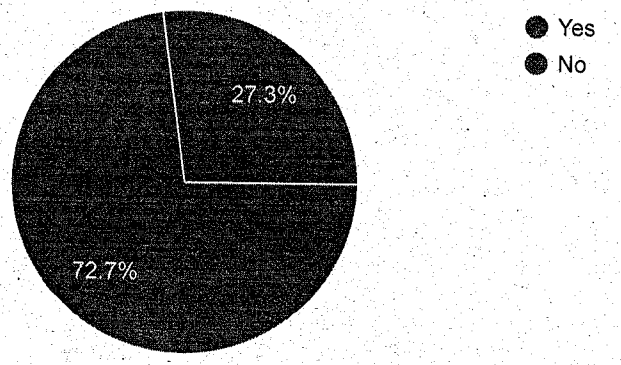
Are the topics in this course are appropriately assisted you in identifying solution to engineering problems?

11 responses



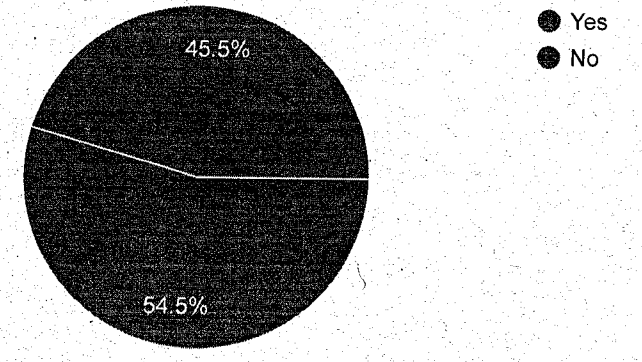
Are you able to do research work in the field of computer science aligned with your course where the work showcases your leadership, integrity and professional ethics?

11 responses



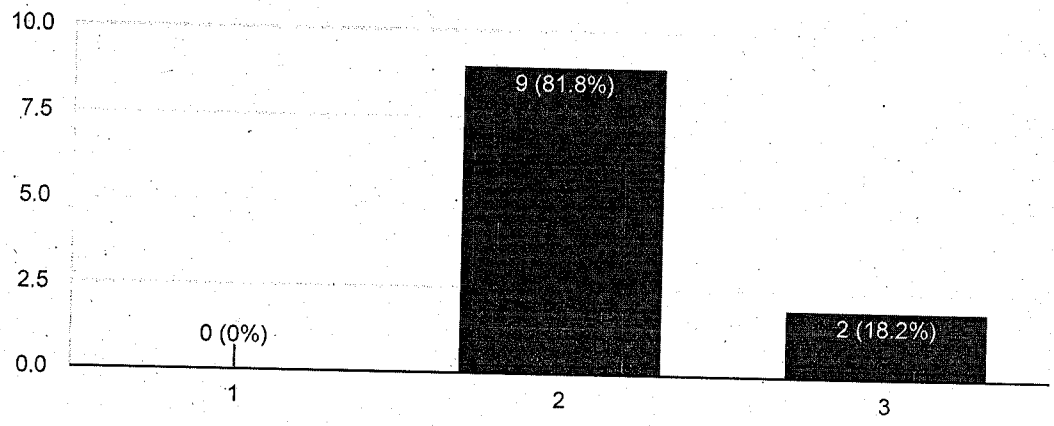
Did make use of the any research tools for the implementation of algorithms?

11 responses



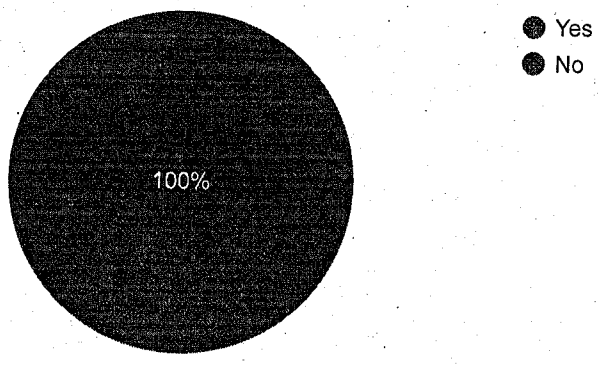
To what extent you grade the quality of contents in this subject?

11 responses



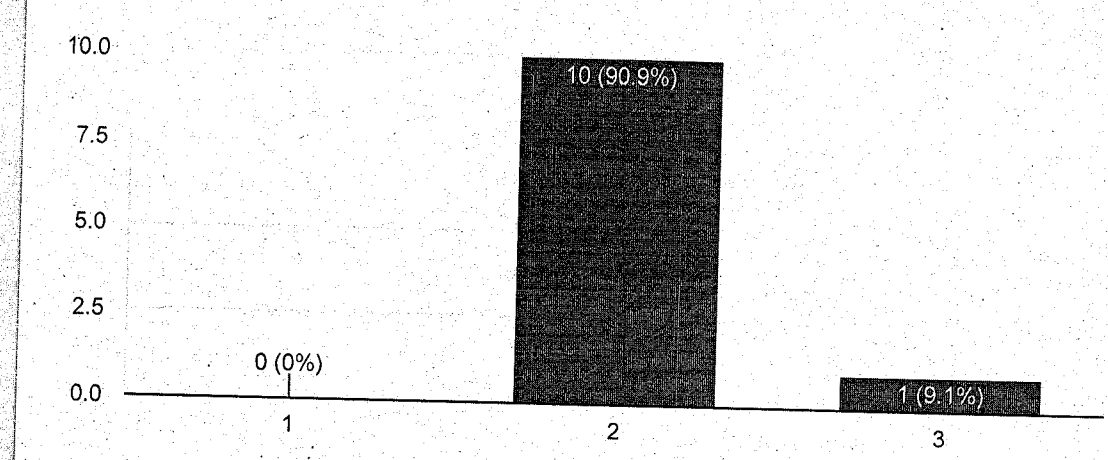
Do you feel topics included in this course will give good background for higher education?

11 responses



Rate the level of the knowledge improvement after the successful completion of this course.(1-Low, 2- Medium, 3 -High)

11 responses



International Journal of Engineering Applied Sciences and Technology,
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 221-223
Published Online December 2019 in IJEAST (http://www.ijeast.com)

ENHANCING PERFORMANCE AND EFFICIENCY FOR BIG DATA ANALYTICS APPLICATION IN HADOOP MAPREDUCE ENVIRONMENT

Purushotham Naidu V
M.Tech Scholar,
Department of Computer Science and Engineering
BMS Institute of Technology and Management
Doddaballapura Main Road, Avalahalli,
Yelahanka, Bengaluru-560064
Karnataka, India

Anjan K Koundinya
Associate Professor and PG Coordinator
Department of Computer Science and Engineering
BMS Institute of Technology and Management
Doddaballapura Main Road, Avalahalli, Yelahanka,
Bengaluru-560064
Karnataka, India

Abstract - **Hadoop finds its area in Big Data Analytics for analysing huge amounts of data. Hadoop implements MapReduce to distribute data to different clusters. Data compression is adopted in order to reduce the memory space occupied by the data. The concept of MapReduce performance with Data Compression focuses on a number of compression codecs of Hadoop cluster such as snappy, gzip, lz4, bzip2 and deflate. The Big data analytics in health care faces good benefits and also with all its associated components focuses with the proposal of a big data health care architecture. Big data analytics is an emerging field for extraction of closely connected information from very huge data-sets and focuses on the improvement of decision making with improved decision making. The educational system and academic trends of students needs to map up with the current trends in technological advancement which accumulates large amount of data which is unstructured and needs to be analysed. Data mining tools are required to obtain information with meaning by converting unstructured data to structured data. Data has become necessary part of every individual, industry, economy, business function and organization. As this data set increases, selecting the relevant information becomes a tedious task. The on-command and on-demand nature of digital universe gives creation of a data category called the Big Data because of its sheer volume, variety and velocity. It proposes computational and analytical challenges which includes measurement errors, scalability and storage bottleneck and noise accumulation.**

Index Terms – Big Data, Hadoop, MapReduce, HDFS, Data Mining.

I. INTRODUCTION

The field of Big Data has attained fast expansion by providing various tools to manage, accumulate and analyse data to make better decisions. Big data in health care focuses on huge benefits by improving the quality of detecting diseases at earlier stages and effective in the medical care. Cloud computing also plays an important role which provides an on demand close services for processing, analysing and storing huge amount of data. Big data includes many tools and platforms for analytics in health care such as Cassandra,

Hadoop Distributed File System (HDFS), MapReduce, Mahout, HBase, PIG, Hive and Zookeeper.

The study focuses on two scenarios, firstly data compression and map output are utilised. The execution time is better for input file with raw-text as snappy and deflate only. Next the results are compressed for map output which does not increase the performance compared to the uncompressed data. Secondly, compressed input file of bzip2 is utilised with the uncompressed MapReduce. The bzip2 files as input are compatible to word count job like raw text file. This can save storage space more than 70 percent of raw text file. Hadoop benchmarks which also has other applications for large output file of reduce phase are also focused.

The Big data in health care can be used to understand Structured, Semi Structured or Unstructured data and handle in an efficient way. Enhanced measures will be considered to cure the diseases. The field of Cloud computing is also being used to process huge (big) data in health care. Cryptographic techniques can be adopted and implemented to achieve a good frame work for sharing of sensitive information with patients on cloud with improved security.

II. METHODOLOGY

The compression codecs of Hadoop cluster as mentioned in the abstract are configured in an XML configuration file as codec properties. The research includes the utilisation of data compressed with word count MapReduce as follows:

A. *Scenario 1 (Map output compression)*: Raw-file, gzip and bzip2 of 4.8 GB was used as an input file. The MapReduce process utilises the compression codec's such

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 213-216
Published Online December 2019 in IJEAST (http://www.ijeast.com)

EMERGING REAL TIME STREAMING ANALYTICS PROCESSING USING HADOOP FRAMEWORK

Sneha. S
M.Tech Scholar,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Anjan K Koundinya
Associate Professor and PG Coordinator,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Abstract— ***Sensors, machines, vehicles, cell phones, web-based social networking systems and other constant sources are creating persistent stream of information. This information is utilized by the organizations to get the advantage from those information. A real time streaming system should address the issues of researchers, developers and records focus activities groups without requiring complex code for incorporation of numerous outsider tools. As there is increment in measure of information that is produced and gathered, statistical analysis wants adaptable, flexible, and high performance tool to analyze and obtain only the necessary data from the large growing data in a required timely manner. Hadoop Distributed File System (HDFS) is one of the file system to store huge measure of information. HDFS can oversee and keep up information in a dispersed manner. Real Time Streaming data can be put away into noSQl databases, for example, Mongo DB and Hive. Enormous information investigation can be performed on information put away on Hadoop distributed file system utilizing Apache Hive, Tez , Storm, Flume and Apache Presto. Hive is a environment which is over Hadoop (Map Reduce), and gives more significant level language to apply to the Hadoop's fundamental part Map Reduce to process the information. The key focal points of this methodology are it can equipped for processing and saving of the enormous measure of information. It additionally can adapt to the a large number of client demands at the same time. It can give the scalability to the machine is increasingly attractive with the guide of including new nodes. Incorporating the Visualization equipment with Big Data projects will give the gigantic picture to the clients to see the bits of knowledge of the Big data. It can offer the analytical reports for giving the big view about the file system..***

Keywords— **Real Time Streaming, HDFS, Hive, Tez, Storm, Flume, Storm, Apache Presto, Mongo DB, Big Data**

I. INTRODUCTION

As we are moving towards digitization, the amount of data being made and amassed is developing and expanding radically. Investigation of this enormously developing data will turn into an difficult task if we utilize the current tools. We expect restructuring to connect the gap between data being created and data that might be analysed accurately.
Enormous big data tools and innovations offer openings and requesting circumstances in having the option to investigate data successfully to understand client needs, increase an upper hand in the commercial center and build up association's business undertaking. Data management architectures have created from the data warehousing model to increasingly convoluted structures that address more noteworthy necessities, comprising of real-time and batch processing, structured and unstructured data, high-speed transactions etc. Investigating huge fact sets calls for colossal process limit that can change long essentially dependent on the amount of input data and the kind of analysis.

1.1 Real-Time Streaming

Internet of Things (IOT) has made a fresh out of the box new age as it is utilized all over the place. A huge measure of detecting devices(sensors) gather/produce different type of certainties after some time for a colossal scope of fields and applications. In view of the idea of the application, these gadgets will bring about gigantic or quick/continuous information streams. Applying examination over such information streams to find new actualities, anticipate future bits of knowledge, and settle on oversee decisions is a basic procedure that makes IoT a value stage for offices and a personal satisfaction improving technology[3]. For example, robots were being used for a considerable length of time in assembling, yet now they have extra sensors so we can perform quality confirmation, never again essentially meeting. For quite a long time, mechanical measures have been typical

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 168-170
Published Online December 2019 in IJEAST (http://www.ijeast.com)

STRUCTURAL ANALYSIS OF HPC'S FOR BIG DATA ANALYTICS

Srivatsa Raju S
Department of Computer Science and Engineering
B.M.S Institute of Technology & Management,
Bengaluru, Karnataka, India

Dr. Anjan K Koundinya
Department of Computer Science and Engineering
B.M.S Institute of Technology & Management
Bengaluru, Karnataka, India

Abstract— **Data analytics possess challenges in various scenarios, designing a system is one of them. This paper provides an insightful preview of the challenges and possible outcome one needs to consider. The High-Performance Computing system for analysis of complex computer-intensive data needs to be focused on architecture design, network hardware, clustering, I/O systems, Metadata, cost and how this analysis can be used in Machine Learning and Big Data analytics techniques.**

Keywords— **High-Performance computing, Architecture, Network Hardware, Clustering, I/O systems, Machine Learning.**

I. INTRODUCTION

This paper sheds light about how HPC is involved in the generation of huge data in every second and scaling data centers according to Moore's law, where the performance and functionality of digital informatics grows twice the present every 2 years with cost, energy requirements and area. This paper also shares insights about possible technologies which can be utilized for better performance characteristics for Big data analytics with architecture, cost-efficient network, clusters, I/O variability, Metadata Management, Platform selection, resource management for HPC and Machine Learning point of view.

II. HPC'S ARCHITECTURE SCALABILITY

As the size of information and data grows on the internet exponentially the existing infrastructures take a toll on managing this data specially the emerging accelerating technologies for the field of Machine learning the workload push the limits of the computing infrastructure for a better bandwidth and energy efficient technologies for the upscales.

The Moore's law where the growth doubles for every 2 years are transforming the existing architectures for the ready to use networking and storage devices which are not anymore a statistically fulfilling solution to respond the needs of being scalable and dynamically adaptable memory fabrics that uses the classics of photonics which can shape the future of vast datacentre and HPC architectures. [1]

On the basis of such considerations, the algorithm uses a different color image multiplied by the weighting coefficients of different ways to solve the visual distortion, and by embedding the watermark, wavelet coefficients of many ways, enhance the robustness of the watermark.

III. COST-EFFECTIVE ROUTING FOR HPC NETWORKS

With the scaled-out data centers and HPC which has large volume of nodes it is critical to have efficient network of distributed computing and memory resources, these kinds of network require high computation bandwidth, capacity, and internetwork parallelism, which makes a challenge to meet cost efficient, energy efficient, and reliability.

Building these networks with higher radix switches, where the signal rates are costlier than the packets this technology is cost efficient than the lower radix routers with low bisection bandwidth and path diversity. The Multiport Binding Tile-based Router (MBTR) proves to be effective and best alternative to off the shelf routers. [2]

IV. CONTAINER CLUSTERS FOR HPC'S

In the adaptive environment the cloud technologies support the microservices-based applications for scalability, dynamic and manageability by container concept for workloads with resources infrastructure High-Performance Data Analytics (HPDA) to process higher volumes of data generated from different applications. The utilization of SmartX Intelligence Cluster for running containerized HPDA workloads which can provide Hyper-converged style resources with integrated network support, storage and computing.

With the SmartX Intelligence the usage of parallel file system with high-performance within the work node has increase in the performance with the best integration of software for deep learning workloads. [3]

V. I/O VARIABILITY FOR HPC STORAGE SYSTEM

With the storage devices shared between numerous applications and managed in the best way, the I/O is often a major concern to be addressed which can be resolved by implementing messaging-based re-routing together with throttling at mid-level which can solve QoS-less HPC storage system and runtime scheduling that can be scalable.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 8, ISSN No. 2455-2143, Pages 119-122
Published Online December 2019 in IJEAST (http://www.ijeast.com)

A DEEP DIVE INTO LOAD BALANCING TOOLS FOR HADOOP APPLICATION MANAGEMENT

Kaveri T Hombal
M. Tech Scholar,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management
Yelahanka, Bengaluru

Anjan K Koundinya
Associate Professor and PG Coordinator,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Abstract-**Hadoop has become an important tool for the researchers and scientists in order to store and analyze huge amount of data. This huge data is placed in Hadoop th the help of Hadoop Distributed File System (HDFS). Block placement policy is employed in HDFS to split a really huge file into blocks and place these block across the cluster in an exceedingly distributed manner. Basically, Hadoop and HDFS are designed to works expeditiously on the consistent cluster. However during this era of networking, we cannot think about having solely a cluster of consistent nodes. So, there's the necessity of storage policy which will work expeditiously on each consistent still because of the heterogeneous cluster. Thus, the need of applications which will be executing in a time-efficiently manner and supporting consistent still because *the heterogeneous setting will be sufficient. In Hadoop data*1.**
MapReduce is programming framework for writing Map-Reduce applications which enables them to run on the distributed platform in parallel. MapReduce permits the applications to run on Hadoop environment.

Hadoop uses HDFS block placement policy to place the data blocks on nodes. Hadoop cluster gets unbalanced every now d then, because of overutilization of few nodes against the less used nodes or recently created other new nodes with no blocks hold on them. To resolve this case, Hadoop encompasses an inherent tool known as HDFS Balancer.

*KeyWords- **Hadoop, Load Balancing, Data Blocks, HDFS, Storage.***

I. INTRODUCTION

Hadoop [1] has been widely used because of its ability to make utilization general computers. Hadoop has essentially 3 most vital constituents: The Hadoop Distributed File system (HDFS), MapReduce and Yet Another Resource Negotiator (YARN) [3]. HDFS permits nodes to store information on the distributed cluster. HDFS is known to be adherent and it can be deployed on any machine.
Hadoop stores the data blocks on the various different nodes of the cluster. The policy of storing the blocks in Hadoop, helps to distribute the data block uniformly amongst the cluster nodes. This policy enables the data blocks to be placed in an consistent manner with the subsequent approach:

- Splitting the large number of files into blocks and replicates the blocks with respect to the replication issue that has been already outlined within the hdfs-site.xml file.
- That data node itself. Otherwise, place it any of the data node of the cluster.
- The next exact copy of the block are going to be placed on different racks of the nodes if offered, or even placed on a similar rack of the initial replica.
- Third copy needs to be placed on any node of the rack where the second copy is already placed.

II. HDFS BLOCK PLACEMENT POLICY

By making use of block placement policy offered in HDFS, block are mapped to a process within the same node by Data Locality, however typically once you're addressing huge data, there is a need of mapping the data block to processes over multiple nodes. To influence this Hadoop has a functionality to copy that data block wherever · mappers are running. This creates various performance degradation particularly on heterogeneous cluster because of I/O delay or congestion in network. In order to balance the data block on specific nodes i.e. custom block placement here we use an efficient algorithm, solely by dividing total number of nodes among two classes like: homogenous vs. heterogeneous or high performing vs. low performing nodes. This policy helps to attain load balancing among the nodes and that we will place data blocks truly wherever we would like our data to be placed for the processing.

III. HDFS BALANCER

Sometimes within the cluster, load imbalance might occur whenever the load is not uniformly distributed on the nodes. In case of Hadoop, another reason behind cluster imbalance is because of its elasticity. We can also add data nodes into existing cluster at any time. This

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 9, ISSN No. 2455-2143, Pages 283-285
Published Online January 2020 in IJEAST (http://www.ijeast.com)

CONVERGENT ANALYTICAL TOOLS FOR BIG DATA APPLICATIONS IN HADOOP ENVIRONMENT

Bhagyashree A V
M.Tech Scholar,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Anjan K Koundinya
Associate Professor and PG Coordinator,
Dept. Of. Computer Science & Engineering,
BMS Institute of Technology & Management,
Yelahanka, Bengaluru

Abstract- **Big Data Analytics offers an about interminable wellspring of business and instructive understanding, that can prompt operational improvement and new open doors for organizations to give undiscovered income crosswise over pretty much every industry. From use cases like client personalization, to hazard relief, to misrepresentation discovery, to inward tasks investigation, and the various new use cases emerging close every day, the value covered up in organization information has organizations hoping to make a front-line examination activity. Finding an incentive inside crude information presents numerous difficulties for IT groups. Each organization has various needs and various information resources. Business activities change rapidly in a regularly quickening commercial center, and staying aware of new orders can require readiness and versatility. In addition, a fruitful Big Data Analytics activity requires tremendous figuring assets, innovative framework, and exceptionally talented faculty. These troubles can make various exercises flop before they pass on regard. Beforehand, a nonappearance of enlisting power and access to computerization made a certified age scale examination movement past the compass of most associations: Big Data was too much expensive, with a ton of issue, and no sensible ROI. With the climb of conveyed registering and new headways in figure resource the load up, Big Data gadgets are more accessible than some other time in ongoing memory.**

Keywords- Big data, analytical tools, Apache Hadoop, Storm, Hive, Pig

I. INTRODUCTION

Big data implies the datasets which can't be perceived, acquired, oversaw, examined, and handled by present devices. Various meanings of huge information have been given by various clients of Big Data and various experts of Big Data like research researchers, information examiners, and specialized professionals. Big Data Analysis essentially includes logical techniques for huge information, precise engineering of huge information, and huge information digging and programming for examination. Information examination is the most significant advance in Big Data, for investigating important qualities, giving recommendations and choices. Potential qualities can be investigated by information examination. In any case, investigation of information is a wide region, which is dynamic and is mind boggling.

II. BIG DATA ANALYSIS

Big data examination alludes to the technique of dissecting enormous volumes of information, or huge information. This huge information is accumulated from a wide assortment of sources, including interpersonal organizations, recordings, computerized pictures, sensors, and deals exchange records. The point in examining this information is to reveal examples and associations that may some way or another be undetectable, and that may give significant bits of knowledge about the clients who made it. Through this knowledge, organizations might have the option to increase an edge over their adversaries and settle on prevalent business choices.

Modern programming projects are utilized for big data investigation, yet the unstructured information utilized in huge information examination may not be appropriate to ordinary information distribution centers. Big data's high preparing prerequisites may likewise make customary information warehousing a poor fit. Accordingly, more up to date, greater information investigation conditions and advancements have raised, including Hadoop, Map Reduce and Cassandra (No-sql) databases. These innovations make up an open-source programming structure that is utilized to process gigantic informational collections over grouped frameworks.

III. ANALYTICAL TOOLS OF BIG DATA

3.1. Apache Hadoop:

The Apache Hadoop programming library is a system that takes into consideration the appropriated preparing of enormous informational collections crosswise over bunches of

B M S Institute of Technology & Management

Dept of Computer Science & Engineering

BATCH:	2018	SEM:	II
SUBJECT:	Managing Big Data		
Faculty In-Charge:	Dr. Anjan K		

SESSION: Mar 2019-Jun 2019

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		3.00
CO2	3.00	3.00	3.00
CO3	3.00	3.00	3.00
CO4	3.00	3.00	3.00
CO5: CASE STUDY & ASSIGNMENT	#DIV/0!		#DIV/0!

CLASS STRENGTH	14
SET IA TARGET	60%

		TEST-1										TEST-2										TEST-3											CO1			CO2			CO3			CO4			CO5			ALL COs
	CO No.	2	1	1	2	2	2	3	4			2	2	3	3	3	2	3	1			3	3	2	3	4	4	4	3			4															
	Question No.	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	Seminar & Assignment Assignment	TOTAL	PERCENT %	Target >=60%	TOTAL	PERCENT %	Target >=60%	TOTAL	PERCENT %	Target >=60%	TOTAL	PERCENT %	Target >=60%	TOTAL	PERCENT %	Target >=60%	THEORY EXT Grade Point	60
	MAXIMUM MARKS FOR QUESTION	6	6	6	6	6	6	6	6			6	6	6	6	6	6	6	6			6	6	6	6	6	6	6	6			20															
Sl No.	USN	Name																																													
1	1BY18SCS01	BHAGYASHREE A.V	4			4	6		4	6			5	6		6		5	6			6			6	6	6					15	6	100.00%	Y	19	79.17%	Y	33	91.67%	Y	33	86.84%	Y	0	0.00%		39
2	1BY18SCS02	CHAITRASHREE H.S										4		5		3			5			5				6	5		3			17	5	83.33%	Y	4	66.67%	Y	16	66.67%	Y	28	87.50%	Y	0	0.00%		24
3	1BY18SCS03	DIVYA SHREE S		4		5		6	6	3			5	5			4	5	4			4			6	6	4		3			19	8	66.67%	Y	20	83.33%	Y	16	88.89%	Y	22	84.62%	Y	0	0.00%		35
4	1BY18SCS04	FASIHA KAUSAR	3					4	4			6		6		6		4	5			4		6	6	4	3					17	5	83.33%	Y	13	72.22%	Y	33	78.57%	Y	27	84.38%	Y	0	0.00%		40
5	1BY18SCS05	KAVERI T HOMBAL		3	6		6		4								3	6	6			5	5		6		5	4				19	15	83.33%	Y	14	77.78%	Y	19	79.17%	Y	30	93.75%	Y	0	0.00%		43
6	1BY18SCS06	NAVEEN KUMAR KV										5		5		3		3	5			6			4	6	3					16	5	83.33%	Y	5	83.33%	Y	21	70.00%	Y	25	78.13%	Y	0	0.00%		32
7	1BY18SCS07	PRAJWALA P		5		1	6		4	5			6			5		5				3			3		6	3				18	8	66.67%	Y	13	72.22%	Y	20	66.67%	Y	32	84.21%	Y	0	0.00%		26
8	1BY18SCS08	PURUSHOTHAM NAIDU V		5	6		3		5	3			3			5			4	6	6		6		6	6	6					16	17	94.44%	Y	16	66.67%	Y	22	91.67%	Y	37	84.09%	Y	0	0.00%		46
9	1BY18SCS09	RAJESWARI N											5			2			2	3	6		5	6			6		3			17	6	100.00%	Y	13	72.22%	Y	13	54.17%		29	90.63%	Y	0	0.00%		35
10	1BY18SCS10	RAMYA P.L					2		3	2			6			5		3		2	2		5			6		4	3			16	2	33.33%		8	66.67%	Y	27	64.29%	Y	28	73.68%	Y	0	0.00%		39
11	1BY18SCS11	RANJINI N											6					5	4	3			6		5			5	4			15	3	50.00%		11	91.67%	Y	19	79.17%	Y	26	81.25%	Y	0	0.00%		45
12	1BY18SCS12	SNEHA S		5	6	6	6		6	6				5	6		6		1	5			5			6	6		5	6			17	16	88.89%	Y	17	94.44%	Y	36	85.71%	Y	34	89.47%	Y	0	0.00%		41
13	1BY18SCS13	SRIVATSA RAJU S		5	6		6		5	2			3			4		1			4		5			6	3	6	4			15	15	83.33%	Y	9	75.00%	Y	21	70.00%	Y	30	68.18%	Y	0	0.00%		43
14	1BY18SCS14	SUDHANSHU GUPTA	3			3	3		3	1			3			3		4		4			5	5		6		5	4			16	0	0.00%		17	56.67%		23	63.89%	Y	28	73.68%	Y	0	0.00%		40
																																	Target >=60%	11		Target >=60%	13		Target >=60%	13		Target >=60%	14		Target >=60%	0		528
																																	CO1	78.57%		CO2	92.86%		CO3	92.86%		CO4	100.00%		CO5	#DIV/0!		Avg. 37.71429
																																														64.29%

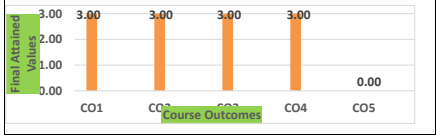
BATCH	2018
SEM	II
SUB	Managing Big Data
SESSION	Mar 2019-Jun 2019
Class Strength	14
Set Target I A	60%
Faculty	Dr. Anjan K
Target >=60%	
3	60%
2	55%
1	50%
No. of COs	4
Set Target Uni Theory	60.00%
No of students who were present in EXAM	14

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

Delivery Mechanism
Lecture, PPT
Lecture, PPT
Lecture, Hands on, PPT
Lecture, Hands on, PPT
Lecture, Hands

FINAL CO ATTAINMENT ANALYSIS - RM&IPR - 18RMI17 - 2018-19 - 2018 Batch

CO1	CO2	CO3	CO4	CO5
3.00	3.00	3.00	3.00	#DIV/0!

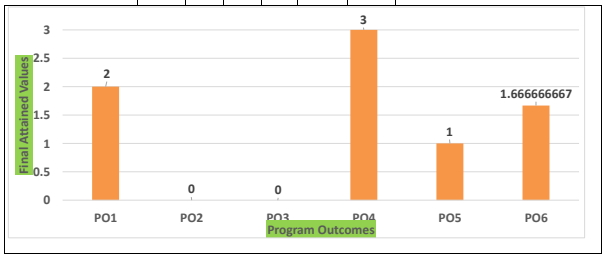


	CO RESULT	PO1	PO2	PO3	PO4	PO5	PO6
CO1:Summarize the fundamentals and concepts of Big Data.	3.00	0	0	0	0	0	0
CO2:Apply non-relational databases (NoSQL) techniques for storing and processing large volumes of structured and unstructured data.	3.00	1	0	0	0	0	1
CO3:Analyze methods and algorithms, to compare them to solve problems.	3.00	0	0	0	3	1	2
CO4:Evaluate efficient big data solutions for various application using novel platform architectures of Hadoop and Map Reduce	3.00	0	0	0	0	0	2
SUM		1	0	0	3	1	5
FINAL ATTAINMENT		1	0	0	3	1	1.7

UNIVERSITY EXAM CONTRIBUTION
Attained all COs highly (=3) (meaning, 64 % of the students have scored more than 55% of marks)
ACTION TAKEN
Research Assignments given to achieve PO1, PO2.
CONCLUSION
High Attainment: -PO4-
Moderate Attainment: -PO6-
OBSERVATION
PO2 PO3 NOT ATTAINED

FINAL PO ATTAINMENT ANALYSIS - RM&IPR- 18RMI17 - 2018-19 - 2018 Batch

PO1	PO2	PO3	PO4	PO5	PO6
2	0	0	3	1	1.7



	PO1	PO2	PO3	PO4	PO5	PO6
FINAL ATTAINMENT	1	0	0	3	1	1.6666667