**Techno India NJR Institute of Technology**

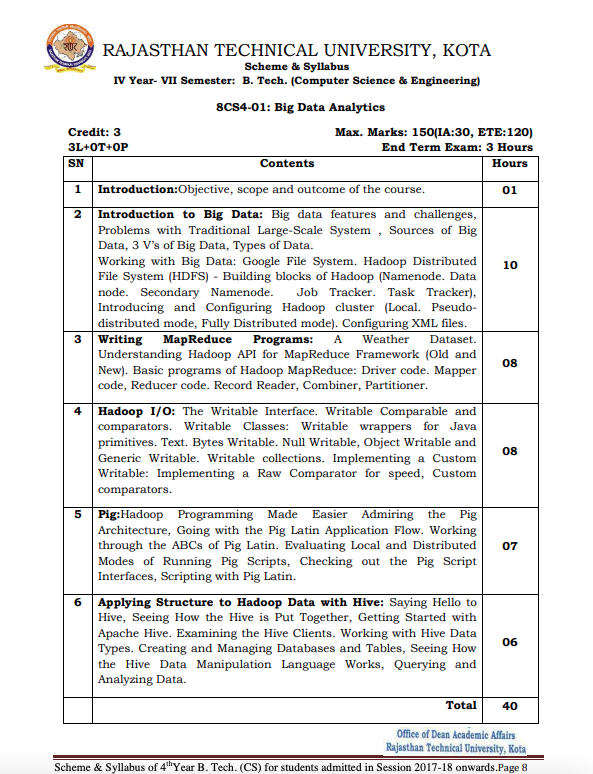


**Course File**

**Big Data Analytics (8CS4-01)**

Aditya Maheshwari

**Department of CSE**

****

**Course Overview:** The course has certain outcomes by virtue of which the students will get an idea of the subject Big Data Analytic.

**Course Outcomes:**

|  |  |  |
| --- | --- | --- |
| **CO No** | **Cognitive Level** | **Course Outcome (Theory)** |
| 1 | Comprehension | Student will be able to identified the business decisions which can be optimized and competitive advantage created with Big Data. |
| 2 | Application | Student will be able to design the database for the data analytics. |
| 3 | Application | Student will be able to write or design the script according to Hadoop architecture along with MapReduce paradigm. |
| 4 | Application | Student will be able to work with Hadoop script to manage the Big Data Analytics. |
| 5 | Application | Students will be able to write scripts with programming tools like PIG & HIVE in Hadoop eco system. |

**Prerequisites:**

1. Fundamentals of Database Management System.
2. Students should be efficient to write the script code.
3. Students should be able to implement the Data Analytics algorithm with Excel.
4. Students should be able to work with XML files.

**Course Outcome Mapping with Program Outcome (Theory):**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course Outcome** | **Program Outcome** | | | | | | | | | | | | | | |
| CO No. | **Domain Specific** | | | | | **Domain Independent** | | | | | | | **PSO** | | |
|  | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** | **PSO1** | **PSO2** | **PSO3** |
| **CO 1** | 3 | 3 | 1 | 0 | 3 | 1 | 0 | 2 | 0 | 0 | 0 | 2 | 3 | 0 | 3 |
| **CO 2** | 3 | 2 | 2 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 3 | 3 | 3 |
| **CO 3** | 3 | 3 | 1 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 3 | 3 | 3 |
| **CO 4** | 3 | 3 | 2 | 1 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 3 | 3 |
| **CO 5** | 2 | 2 | 3 | 2 | 3 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 3 | 3 | 3 |
| **1: Slight (Low), 2: Moderate (Medium), 3: Substantial (high)** | | | | | | | | | | | | | | | |

**Lecture plan based on Unit 1(Introduction)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit Mapping** |
| 0 | Scope and outcome of the course | 1 |
| 1 | Objective and Concept of Big Data Analytics, and how are they different from traditional Database Management systems. | 1 |
| 2 | Real-time applications use (case studies) of Big Data Analytics. | 1 |

**Lecture plan based on Unit 2 (Introduction to Big Data)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit Mapping** |
| 3 | Features & Challenges of Big Data Analytics. | 2 |
| 4 | Problems of Traditional Large - Scale System | 2 |
| 5 | Source of Big Data, 3 V’s of Big Data and Case Study | 2 |
| 6 | Types of Data, Structured, Semi Structured & Unstructured | 2 |
| 7 | Google File System and Case Study | 2 |
| 8 | Hadoop Distributed File System and Architecture | 2 |
| 9 | Building Blocks of Hadoop (Name Node, Data Node & Secondary Node) | 2 |
| 10 | Building Blocks of Hadoop (Job Tracker and Task Tracker) | 2 |
| 11 | Introducing & Configuring Hadoop Cluster | 2 |
| 12 | Configuring XML Files | 2 |

**Lecture plan based on Unit 3 (Writing MapReduce Programs)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit Mapping** |
| 13 | Understanding Hadoop API for MapReduce Framework | 3 |
| 14 | Basic Program of Hadoop MapReduce | 3 |
| 15 | MapReduce Program for Driver Code & Mapper Code | 3 |
| 16 | MapReduce Program for Reducer Code | 3 |
| 17 | MapReduce Program for Record Reader, Combiner & Partitioner | 3 |
| 18 | MapReduce Program for Weather Data. | 3 |
| 19 | MapReduce Program for Matrix Multiplication. | 3 |
| 20 | MapReduce Program for basic Word Count. | 3 |

**Lecture plan based on Unit 4 (Hadoop I/O)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit Mapping** |
| 21 | Understanding Writable Interface | 4 |
| 22 | Writable Comparable & Comparators | 4 |
| 23 | Writable Wrappers for Java Primitives | 4 |
| 24 | Writable : Text, Bytes, Null, Object and Generic | 4 |
| 25 | Writable Collections | 4 |
| 26 | Custom Comparators : Implementing a Raw Comparator for speed | 4 |
| 27 | Hadoop I/O Architecture | 4 |
| 28 | HDFS : File System | 4 |

**Lecture plan based on Unit 5 (Pig)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit based mapping** |
| 29 | Pig Architecture | 5 |
| 30 | Latin Application Flow : Pig | 5 |
| 31 | Working through the ABCs of Pig Latin | 5 |
| 32 | Running Pig Scripts : Local Mode | 5 |
| 33 | Running Pig Scripts : Distributed Mode | 5 |
| 34 | Pig Scripts Interface | 5 |
| 35 | Scripting with Pig Latin | 5 |

**Lecture plan based on Unit 6 (Applying Structure to Hadoop Data with Hive)**

|  |  |  |
| --- | --- | --- |
| **Lecture No.** | **Topic** | **Unit Mapping** |
| 36 | Hive Architecture | 6 |
| 37 | Getting Started with Apache Hive | 6 |
| 38 | Understand the Hive Clients | 6 |
| 39 | Hive Data Types and It’s working | 6 |
| 40 | Hive Data Manipulation Language : Databases | 6 |
| 41 | Hive Data Manipulation Language : Querying and Analyzing Data | 6 |

**Textbook –** Raj Kamal, Preeti Saxena, “Big Data Analytics – Hadoop, Spark and Machine Learning” 1st Edition, Mc Graw Hill

**Reference Sessions –** <https://www.ee.columbia.edu/~cylin/course/bigdata/>

**MOOC (Coursera)** - <https://www.coursera.org/learn/big-data-essentials>

**Lab Practical’s** - <https://github.com/technoindianjr/Big-Data-Analytics-Lab_8cs4-21>

**Assessment Methodology:**

1. Quiz/Viva
2. Practical exam in lab where they have to implement their skills to manage the big data for the given problem statement.
3. Midterm subjective paper where they have to write algorithm to perform different operations.
4. Final paper (subjective paper) at the end of the semester.

**Quiz Questions**

Q 1. Just collecting and storing information isn’t enough to product real business value. Big data analytics technologies are necessary to:

A Formulate eye-catching charts and graphs

**B Extract valuable insights from the data**

C Integrate data from internal and external sources

D Determine business goals and objectives

Q 2. The Method by which companies analyse customer data or other types of information in an effort to identify patterns and discover relationships between different data element is often referred to as:

**A Data Mining**

B Data Digging

C Customer Data Management

D Consumer Engagement

Q 3. Donal Farmer, principal at analytics consultancy TreeHive strategy, outlined six potential benefits big data has for organizations, except for:

A More agile supply chain operations

B Smarter recommendations and targeting

C Increased Market Intelligence

**D Consumer-Driven Product Innovation**

Q 4. What is the recommended best practice for managing big data analytics programs?

A Adopting data analysis tools based on a laundry list of their capabilities

B Letting go entirely of ‘old ideas’ related to data management.

**C Focusing on business goals and how to use big data analytics technologies to meet them**

Q 5. Big data developed the three V’s – Volume, Velocity & Variety in 2001. In the years since, the V’s have expanded to include Veracity and Value. Sometimes a sixth V is applied to big data which is:

**A Variability**

B Vector

C Vulnerability

D Volatile

Q 6. Companies that have large amounts of information stored in different systems should begin a big data analytics project by considering:

A The creation of a plan for choosing and implementing big data infrastructure technologies

**B The interrelatedness of data and the amount of development work that will be needed to link various data sources**

C The ability of business intelligence and analytics vendors to help them answer business questions in big data environments

D The database with the most information storage first and working through the storage systems sequentially

Q 7. True or False? For organizers that aren’t currently looking to do big data analytics, there is little or no benefit to examining the data they’re retaining and evaluating how it’s being used.

A True

**B False**

Q 8. What is the name of the programming framework originally developed by Google that supports the development of applications for processing large data sets in a distributed computing environment?

**A MapReduce**

B Hive

C ZookKeeper

D Google Cloud Dataproc

Q 9. True or False? Organizations are struggling with maintaining highly skilled data scientists and engineer due to market demands.

**A True**

B False

Q 10. Big data analytics doesn’t help an organizations:

A Better understand customers

B Increase shareholder dividends

C Refine marketing and advertising

**D Increase costs due to additional analytics investment**

**Viva Questions**

Q 1. What is a Big Data and where does it come from?

Q 2. What are the 7 V’s in the Big Data?

Q 3. Why business are using Big data for competitive advantage.

Q 4. How is Hadoop and Big data related?

Q 5. Explain the importance of Hadoop technology in Big data analytics.

Q 6. Explain the core components of Hadoop.

Q 7. What is the distributed processing in Hadoop?

Q 8. How is HDFS different from traditional NFS?

Q 9.What is data modelling and what is the need for it.

Q 10. What is the Data ingestion and Data Processing?

**Assignment - MapReduce (MapTask)**

|  |
| --- |
| Here, you will complete a back-end for a MapReduce system and test it on a couple MapReduce jobs: word count (provided), and meanCharsMR (you must implement). Template code is provided.  Specifically, you must complete:   1. **PartitionFunction (10 points)** Complete the partition function, making sure to use a hash that can handle: integers and strings. 2. **RunSystem (20 points)** Complete the “runSystem(self)” method which divides the data into chunks and schedules the running of mapTasks and reduceTasks. The are two places to complete: (1) Divide up the data into chunks according to num\_map\_tasks, and launch a map task per chunk. (2) Send each key-value pair to its assigned reducer. 3. **Combiner (15 points)** Edit the “MapTask” method to add support for running a Combiner. Look for “#<<COMPLETE>>”  within the method. Remember, a combiner runs the reduce task at the end of the map task in order to save communication cost of sending to multiple reducers. Note: main will run the WordCountBasicMR to test with and without the combiner. It is recommended that you look over the WordCountBasicMR to understand what it is doing.  You can assume your combiner code will only run on reducers that are both commutative and associative (see hint at bottom). 4. **Mean CharsMR (20 points)** Edit the “map” and “reduce” methods of “MeanCharsMR” to implement a map-reduce computation of the mean and standard deviation of the number of each character (a-z, case insensitive) per document (i.e. the mean is across all documents; the count of each character is per document). Consider each record (i.e. single key value pairs arriving at mapper) to be a single document. Reduce can return more than the mean, and standard deviation (hint: including other items will be helpful for the combiner to run).   Example: if one had three documents: [‘a bacd a’, ‘cda’, ‘bcd’], then the mean of each char would be: (‘a’: {‘mean’: 1.333 = (3+1+0) / 3, ‘std-dev’: 1.52 = sqrt(((3-1.333)^2 + (1-1.333)^2 + (0 - 1.333)^2)/2)}),  (‘b’: {‘mean’: 0.666 = (1+0+1)/3; … }), …  \*\***Do not use self.data from the mappers or reducers: they need to work with the key values that they are provided.\*\*** |
| **Template Code:**A template to be filled in with your code is provided here: [MRSystemSimulator2020\_lastname\_id.py](https://docs.google.com/document/d/103lb3trDwLrI0OCBxSFxotBrIMd9eiKVARlhw3ZInes/edit?usp=sharing) |

A picture containing logo

Description automatically generated

**TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY UDAIPUR**

**Computer Science and Engineering**

**B. TECH IV– YEAR (VIII Sem)**

**MID-TERM PAPER**

**SUBJECT – Big Data Analytics – 8CS4-21**

**Time: 1Hr 30 minutes + 15 Minutes for Submission Max. Marks: 40**

**Attempt any five questions.**  [5 \* 8 =40]

1. Why is Big Data hot now? Discuss one case study in brief. [CO1]

2. List out the key computing resources for Big Data in details. [CO2]

3. Explain Grid Computing VS Cluster Computing.[CO3]

4. Explain five V’s of Big Data. [CO4]

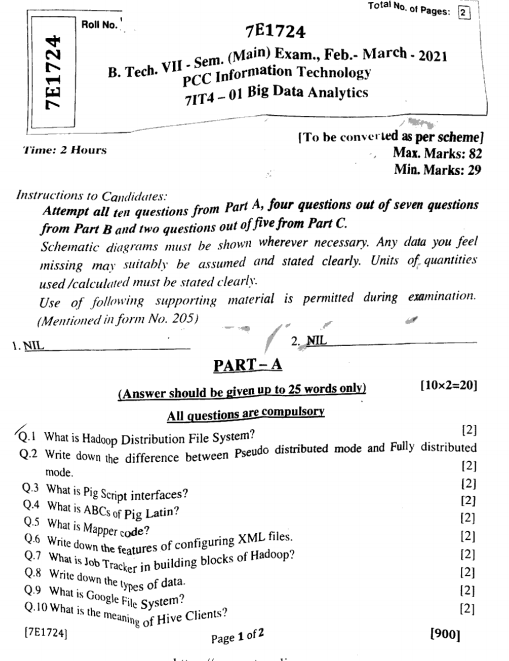
5. Explain Hadoop Architecture design in detail. [CO4]

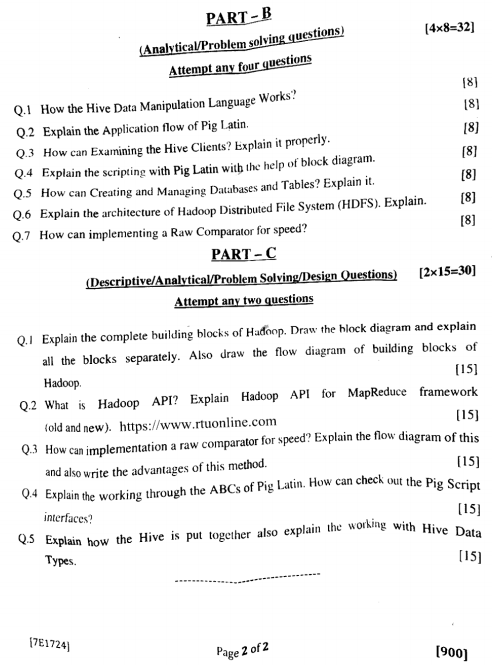
6. Explain MapReduce Processing Example i.e., Word Count Program. [CO3]

7. Explain HDFS Design. [CO2]

8. Explain Google File System Design. [CO4]

**Previous Year Question Paper**





**Result Analysis (Exam 2020 – 2017-21 Batch)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. No** | **Roll Number** | **Name of Student** | **BDA** | | |
| **Mr. Aditya M.** | | |
|  |  |  | **INT.** | **EXT.** | **TOT.** |
|  |  |  | **120** | **30** | **150** |
| 1 | 17ETCCS001 | AKMAL HUSSAIN | 77 | 25 | 102 |
| 2 | 17ETCCS002 | ANUJ SHARMA | 65 | 27 | 92 |
| 3 | 17ETCCS003 | APOORVA JINDAL | 108 | 28 | 136 |
| 4 | 17ETCCS004 | ARPITA KOTHARI | 113 | 29 | 142 |
| 5 | 17ETCCS006 | BHAVIT KANTHALIA | 108 | 30 | 138 |
| 6 | 17ETCCS007 | CHARUL SINGHVI | 113 | 25 | 138 |
| 7 | 17ETCCS008 | DEVYANI GUPTA | 94 | 27 | 121 |
| 8 | 17ETCCS009 | DHEERAJ DASHORA | 87 | 28 | 115 |
| 9 | 17ETCCS010 | DIVYA JAIN | 96 | 29 | 125 |
| 10 | 17ETCCS011 | EKANSH JAIN | 84 | 26 | 110 |
| 11 | 17ETCCS012 | GARVIT SOLANKI | 106 | 30 | 136 |
| 12 | 17ETCCS013 | HARSHIT KAWDIA | 84 | 25 | 109 |
| 13 | 17ETCCS015 | HIMANSHU JAIN | 96 | 27 | 123 |
| 14 | 17ETCCS016 | HIMANSHU TAK | 99 | 28 | 127 |
| 15 | 17ETCCS017 | JAYA GUPTA | 75 | 29 | 104 |
| 16 | 17ETCCS018 | JINISHA JAIN | 113 | 26 | 139 |
| 17 | 17ETCCS019 | KARTIK BOKADIA | 111 | 26 | 137 |
| 18 | 17ETCCS020 | KARTIK KUMAWAT | 118 | 25 | 143 |
| 19 | 17ETCCS021 | KARTIK PANCHAL | 77 | 27 | 104 |
| 20 | 17ETCCS022 | KHUSHBOO PADDIYAR | 101 | 28 | 129 |
| 21 | 17ETCCS024 | M SAJID MANSOORI | 94 | 29 | 123 |
| 22 | 17ETCCS025 | MAHIMA KOTHARI | 94 | 26 | 120 |
| 23 | 17ETCCS026 | MAHIMA SHARMA | 104 | 27 | 131 |
| 24 | 17ETCCS029 | MILAN PURBIA | 101 | 27 | 128 |
| 25 | 17ETCCS030 | MOHAMMED AFZAL RAZA | 94 | 28 | 122 |
| 26 | 17ETCCS031 | MOHIT AMETA | 99 | 29 | 128 |
| 27 | 17ETCCS033 | NAYAN SHARMA | 77 | 27 | 104 |
| 28 | 17ETCCS034 | NIDHI SHUKLA | 82 | 26 | 108 |
| 29 | 17ETCCS036 | NIKITA JAIN | 92 | 27 | 119 |
| 30 | 17ETCCS037 | NIKITA LILADHAR PANDE | 113 | 27 | 140 |
| 31 | 17ETCCS038 | NILESH SAHITYA | 99 | 28 | 127 |
| 32 | 17ETCCS039 | NIMISHA SHARMA | 106 | 29 | 135 |
| 33 | 17ETCCS040 | NISHANT JAIN | 87 | 26 | 113 |
| 34 | 17ETCCS041 | PARSHAVI BOLYA | 104 | 26 | 130 |
| 35 | 17ETCCS042 | PRACHI PANWAR | 89 | 27 | 116 |
| 36 | 17ETCCS044 | RAHUL CHOUDHARY | 89 | 28 | 117 |
| 37 | 17ETCCS046 | RAKSHIT JOSHI | 111 | 29 | 140 |
| 38 | 17ETCCS047 | RISHIKA JAIN | 104 | 26 | 130 |
| 39 | 17ETCCS048 | RONAK ARORA | 106 | 28 | 134 |
| 40 | 17ETCCS049 | RUCHIKA PUROHIT | 101 | 29 | 130 |
| 41 | 17ETCCS050 | SACHIN GARG | 89 | 26 | 115 |
| 42 | 17ETCCS051 | SIMRAN GERA | 108 | 26 | 134 |
| 43 | 17ETCCS052 | SUBHASH MEGHWAL | 89 | 27 | 116 |
| 44 | 17ETCCS053 | SWATI DEVPURA | 116 | 27 | 143 |
| 45 | 17ETCCS054 | TANISHKA JAIN | 99 | 28 | 127 |
| 46 | 17ETCCS055 | VARSHA CHOUDHARY | 101 | 29 | 130 |
| 47 | 17ETCCS056 | VEDPRAKASH GUPTA | 96 | 26 | 122 |
| 48 | 17ETCCS057 | VIDIT JAIN | 106 | 25 | 131 |
| 49 | 17ETCCS058 | VIKAS SONI | 113 | 27 | 140 |
| 50 | 17ETCCS059 | VIRENDRA SINGH | 84 | 28 | 112 |
| 51 | 17ETCCS060 | YASH MALASIYA | 94 | 29 | 123 |
| 52 | 17ETCCS300 | PRERNA PALIWAL | 104 | 30 | 134 |
| 53 | 17ETCCS301 | ASHI KOTHARI | 89 | 26 | 115 |
| 54 | 17ETCCS302 | SAKSHI MADRECHA | 118 | 27 | 145 |
| 55 | 17ETCCS303 | RATIKSHA KHATIK | 87 | 25 | 112 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  | **BDA** | | |
|  |  |  | **Mr. Aditya M.** | | |
|  |  |  | TOTAL | | 55 |
|  |  |  | PASS | | 55 |
|  |  |  | FAIL | | 0 |
|  |  |  | PASS% | | 100 |
|  |  |  | FAIL% | | 0 |

**MOOC Certifications**

**Graphical user interface, application, Teams

Description automatically generated**

**Text, letter

Description automatically generated with medium confidenceGraphical user interface, application

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface, text

Description automatically generated with medium confidenceGraphical user interface, application, Teams

Description automatically generated**