Techno India NJR Institute of Technology

Course File
Big Data Analytics (8CS4-01)

Aditya Maheshwari
Department of CSE

8CS4-01: Big Data Analytics

Credit: 3 **Max. Marks: 150(IA:30, ETE:120)**
3L+0T+0P **End Term Exam: 3 Hours**

SN	Contents	Hours
1	**Introduction:**Objective, scope and outcome of the course.	01
2	**Introduction to Big Data:** Big data features and challenges, Problems with Traditional Large-Scale System , Sources of Big Data, 3 V's of Big Data, Types of Data. Working with Big Data: Google File System. Hadoop Distributed File System (HDFS) - Building blocks of Hadoop (Namenode. Data node. Secondary Namenode. Job Tracker. Task Tracker), Introducing and Configuring Hadoop cluster (Local. Pseudo-distributed mode, Fully Distributed mode). Configuring XML files.	10
3	**Writing MapReduce Programs:** A Weather Dataset. Understanding Hadoop API for MapReduce Framework (Old and New). Basic programs of Hadoop MapReduce: Driver code. Mapper code, Reducer code. Record Reader, Combiner, Partitioner.	08
4	**Hadoop I/O:** The Writable Interface. Writable Comparable and comparators. Writable Classes: Writable wrappers for Java primitives. Text. Bytes Writable. Null Writable, Object Writable and Generic Writable. Writable collections. Implementing a Custom Writable: Implementing a Raw Comparator for speed, Custom comparators.	08
5	**Pig:**Hadoop Programming Made Easier Admiring the Pig Architecture, Going with the Pig Latin Application Flow. Working through the ABCs of Pig Latin. Evaluating Local and Distributed Modes of Running Pig Scripts, Checking out the Pig Script Interfaces, Scripting with Pig Latin.	07
6	**Applying Structure to Hadoop Data with Hive:** Saying Hello to Hive, Seeing How the Hive is Put Together, Getting Started with Apache Hive. Examining the Hive Clients. Working with Hive Data Types. Creating and Managing Databases and Tables, Seeing How the Hive Data Manipulation Language Works, Querying and Analyzing Data.	06
	Total	40

Office of Dean Academic Affairs
Rajasthan Technical University, Kota

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.

<u>Quiz Questions</u>

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

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]

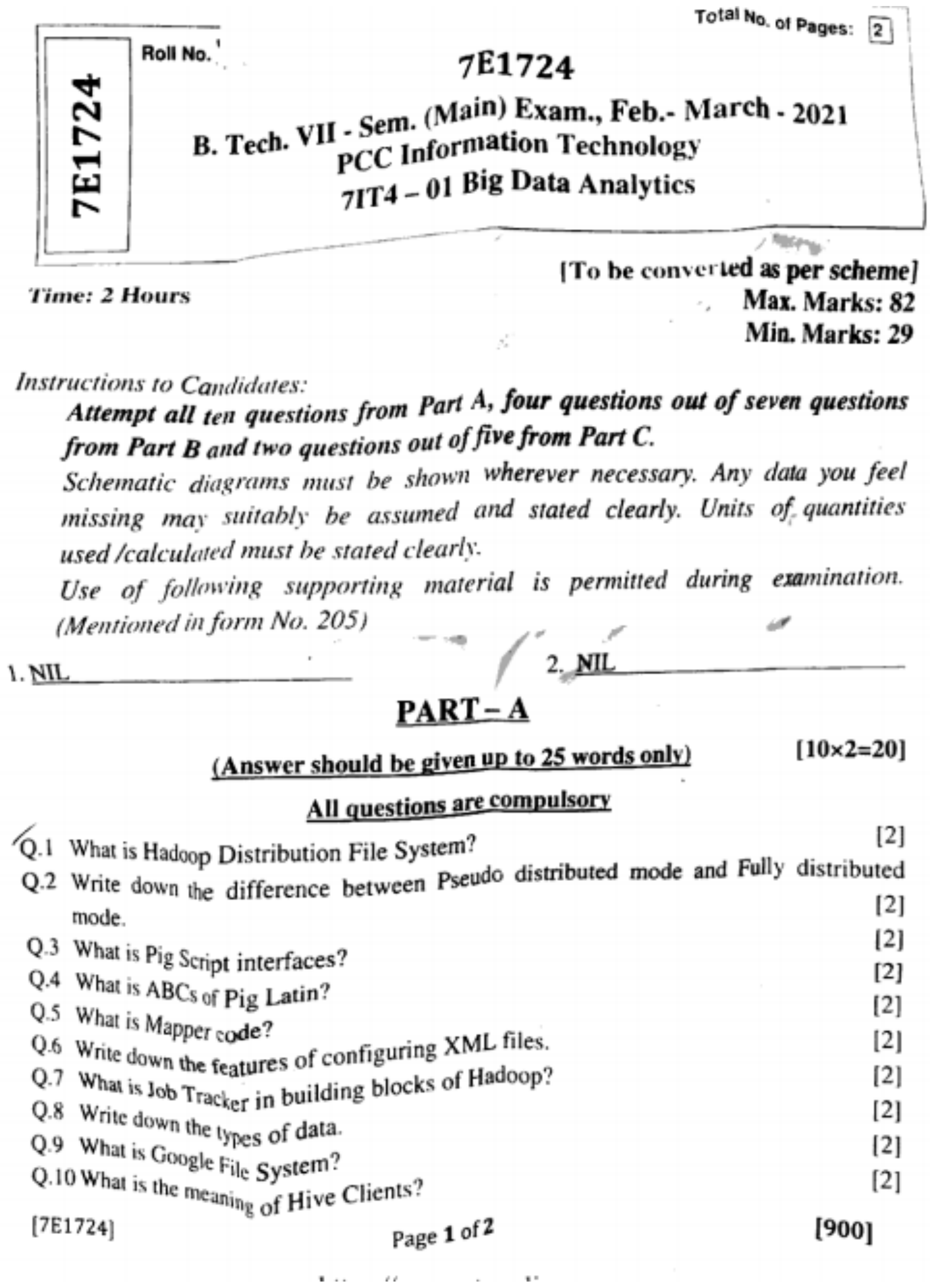
5. Explain Hadoop Architecture design in detail. [CO4]

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



7E1724

Roll No.

Total No. of Pages: 2

7E1724

B. Tech. VII - Sem. (Main) Exam., Feb.- March - 2021

PCC Information Technology

7IT4 – 01 Big Data Analytics

[To be converted as per scheme]

Time: 2 Hours

Max. Marks: 82

Min. Marks: 29

Instructions to Candidates:

Attempt all ten questions from Part A, four questions out of seven questions from Part B and two questions out of five from Part C.

Schematic diagrams must be shown wherever necessary. Any data you feel missing may suitably be assumed and stated clearly. Units of quantities used /calculated must be stated clearly.

Use of following supporting material is permitted during examination. (Mentioned in form No. 205)

1. NIL 2. NIL

PART – A

(Answer should be given up to 25 words only) **[10×2=20]**

All questions are compulsory

Q.1 What is Hadoop Distribution File System? [2]

Q.2 Write down the difference between Pseudo distributed mode and Fully distributed mode. [2]

Q.3 What is Pig Script interfaces? [2]

Q.4 What is ABCs of Pig Latin? [2]

Q.5 What is Mapper code? [2]

Q.6 Write down the features of configuring XML files. [2]

Q.7 What is Job Tracker in building blocks of Hadoop? [2]

Q.8 Write down the types of data. [2]

Q.9 What is Google File System? [2]

Q.10 What is the meaning of Hive Clients? [2]

PART – B

(Analytical/Problem solving questions) **[4×8=32]**

Attempt any four questions

Q.1 How the Hive Data Manipulation Language Works? [8]

Q.2 Explain the Application flow of Pig Latin. [8]

Q.3 How can Examining the Hive Clients? Explain it properly. [8]

Q.4 Explain the scripting with Pig Latin with the help of block diagram. [8]

Q.5 How can Creating and Managing Databases and Tables? Explain it. [8]

Q.6 Explain the architecture of Hadoop Distributed File System (HDFS). Explain. [8]

Q.7 How can implementing a Raw Comparator for speed? [8]

PART – C

(Descriptive/Analytical/Problem Solving/Design Questions) **[2×15=30]**

Attempt any two questions

Q.1 Explain the complete building blocks of Hadoop. Draw the block diagram and explain all the blocks separately. Also draw the flow diagram of building blocks of Hadoop. [15]

Q.2 What is Hadoop API? Explain Hadoop API for MapReduce framework (old and new). [15]

Q.3 How can implementation a raw comparator for speed? Explain the flow diagram of this and also write the advantages of this method. [15]

Q.4 Explain the working through the ABCs of Pig Latin. How can check out the Pig Script interfaces? [15]

Q.5 Explain how the Hive is put together also explain the working with Hive Data Types. [15]
