25 F TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2075 Bhadra

Exam.	**Regular**		
Level	BE	**Full Marks**	80
Programme	BCT	**Pass Marks**	32
Year / Part	IV / II	**Time**	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. Why distributed computing is necessary for big data? [5]
2. Define DFS. How client writes data in HDFS? Explain with the help of suitable block diagram. [10]
3. The data in big data warehouse is called hybrid data. Explain with suitable examples. [10]
4. How GFS differ from other File Systems? List out five distinct differences. [5]
5. What is the main role of GFS Master during read and write processes? How data and control messages flow in GFS architecture. Explain with suitable flow diagram. [10]
6. Map Reduce is the heart of Hadoop eco-system? Define work flow of Map reduce with suitable examples. [10]
7. Clock synchronization in DFS may be the big challenge. How this clock synchronization problem can be solved? [10]
8. Hbase, Cassandra and MongoDB are called column-oriented NoSQL database? How row-oriented database differ from column-oriented database? Explain with suitable examples. [10]
9. Write short notes on: [5×2]

   a) Scoop and fiume
   b) Zookeeper
   c) Oozie
   d) Pig and Hive
   e) Client-Server and Master-Slave architecture

***

35T TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2074 Magh

	[redacted]	Back	
Exam.		Full Marks	80
Level	BE	Pass Marks	32
Programme	BEX, BCT	Time	3 hrs.
Year / Part	IV / II		

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. How big data differ from traditional data? List out five distinct differences? [1

2. What are the sources of structured, semi-structured and un-structured data in real-world? [

3. Define DFS. How client writes data in HDFS? Explain with help of suitable block diagram. [

4. Clock synchronization in DFS may be the big challenge. How this clock synchronization problem can be solved? [1

5. How data and control messages flow in GFS architecture. Explain with suitable flow diagram. [

6. How GFS provides fault tolerance. How it allows tolerating chunk servers failures? [

7. How word count job is performed for the following file in HDFS using a Map-reduce flow chart? [

File.txt(file size: 200MB)

Hi how are you
How is your job
How is your family
How is your brother
How is your sister
What is the time now
What is the strength of Hadoop

8. What are the differences between row and column oriented database? Why Hbase, Cassandra and MongoDB are called column oriented NoSQL database? [

9. Write short notes on: [4

i) Zookeeper and Oozie
ii) Pig and Hive

***

35F TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2074 Bhadra

Exam.	**Regular**		
Level	BE	**Full Marks**	80
Programme	BCT	**Pass Marks**	32
Year / Part	IV / II	**Time**	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. a) Explain with example about the distributed system in Big Data. [8]

   b) What is the role of Data Scientist? [4]

2. a) Explain the architecture of Google File System (GFS). [8]

   b) What is availability and fault tolerance in Google File System? [5]

3. a) Explain in brief Data Flow technique of Map-Reduce Framework. [8]

   b) What is Optimization and Data Locality in Map Reduce? [4]

4. Differentiate between structured and unstructured data and discuss the Taxonomy of NoSQL. [8]

5. Explain the components of Indexing and searching. [8]

6. a) Explain in brief five daemons of Hadoop. [8]

   b) What is the role of Hadoop Distributed File System in Hadoop? [4]

7. Write short notes on: [5×3]

   i) Elastic Search
   ii) Hbase Architecture
   iii) Functional Programming

***

35 F TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2073 Magh

Exam.	**New Back (2066 & Later Batch)**		
Level	BE	**Full Marks**	80
Programme	BEX, BCT	**Pass Marks**	32
Year / Part	IV / II	**Time**	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. Why do we need data analytics process? Explain the role of Distributed computing in Big data. [5+5]

2. Why do we have large and fixed sized Chunks in GFS? What can be the demerits of that design? [10]

3. How is MapReduce library designed to tolerate different machines (map/reduce nodes) failure while executing MapReduce job? [10]

4. For following data, list the input to/output from both the map and reduce functions for getting maximum marks of each college. [10]

Student Name	College Name	Final Marks in %
Ram	ABC	70
Sita	ABC	80
Hari	ABC	60
Gita	XYZ	90
Rita	XYZ	80
Shyam	PQR	90
Laxmi	PQR	70
Gopal	PQR	60

OR

What is the combiner function in mapreduce? Explain its purpose with suitable example. [10]

5. Explain the term NO-SQL. Explain CAP theorem with suitable block diagram. [3+7]

6. Describe the typical components involved in search application. [10]

7. What are different daemons in HADOOP cluster? Explain each in details. [3+7]

8. Write short notes on any two of following. [2×5]

   a) Shadow Master and Cluak services
   b) Analyzers available in Lucene
   c) Vertical and Horizontal Scalability

***

35 F TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2073 Bhadra

Exam.	**Regular**		
Level	BE	**Full Marks**	80
Programme	BEX, BCT	**Pass Marks**	32
Year / Part	IV / II	**Time**	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What are the current trends in big data analytics? What are the technical challenges and characteristics of big data? [10]
2. Explain the GFS Architecture. Why single master is not a bottleneck in GFS cluster. [5+5]
3. How does MAP-REDUCE work? Explain each step with suitable example. [5+5]
4. Discuss the architecture of Hbase in short. Explain eventual consistency and tunable consistency in context of Cassandra. [10]
5. Explain LUCENE architecture and its data indexing approach. [10]
6. What are the components of Hadoop? Explain each in brief. [10]
7. How do you find max and min occurrence of the words in a given text document. Explain. [10]
8. Write short notes on: (any two) [2×5]

   a) CAP theorem

   b) Role of Data Scientist in Big data

   c) Amazon cloud

***

35F TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2072 Magh

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / II	Time	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is a "Big Data"? How distributed systems help to solve the Big Data problems? [12]
2. Explain how master implements garbage-collection and detects stale replica in a GFS. [10]
3. Why do we have large and fixed sized Chunks in a GFS? What are the demerits of that design? [10]
4. How a MapReduce library designed to tolerate different machines (map/reduce nodes) failure while executing MapReduce job? [8]
5. For following data, list the input to/output from both the map and reduce functions for getting maximum marks of each college. [10]

Students Name	College Name	Final Marks in %
Ram	ABC	70
Sita	ABC	80
Hari	ABC	60
Gita	XYZ	90
Rita	AYZ	80
Shyam	PQR	90
Laxmi	PQR	70
Gopal	PQR	60

OR

What is the combiner function in mapreduce? Explain its purpose with suitable example.

6. What is the difference between a structured and unstructured data. Explain the eventual consistency and tunable consistency in context of Cassandra. [10
7. What is an elastic search? Explain various types of analyzers. [2+8]
8. What are the components of the Hadoop? For a hadoop cluster with 128 MB block size, how many mappers will hadoop mapreduce form while performing mapper function on 1 GB of data. Justify with explanation. [10

***

35E

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2072 Ashwin

Exam.		Regular	
Level	BE	Full Marks	80
Programme	BEX. BCT	Pass Marks	32
Year / Part	IV / II	Time	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What are the technical challenges and characteristics of a big data? Who are the data scientists, list out their roles and skills. [6+6]

2. With diagram, explain general architecture of Google File System. [10]

OR

   a) Why do we have single master in a GFS and millions of chunk servers? [4]

   b) A cluster contains 1500 machines, each having 500GB disc capacity. Calculate approximate the number of the chunck servers, the blocks and the total available size if default chunck replica is 3 and 5 respectively. [6]

3. a) What is a map reduce? Explain the execution overview of the map reduce. [6]

   b) Draw the output of mapreduce of the following lines: [4]
   "big users big volume data cloud contributes bid data"
   "facebook has big users facebook operates big data"

4. a) Explain a CAP theorem. [5]

   b) Differentiate between a RDBMS and a NoSQL Databases. [3]

5. Explain taxonomy of a NoSQL databases. Explain Cassendra database in brief. [10]

OR

   Using a MongoDB database,

   a) Create a collections named "posts", insert following records: [3]
   title: MongoDB, description: MongoDB is a NoSQL database, by: Tom,
   Comments: We use MongoDB for unstructured data, likes: 100

   i) Now write a query to search title of the post written by Tom. [3]

   ii) Write mapReduce function to count number of posts created by various users. [4]

6. What is the Lucene? Describe the typical components involved in the search application. [10]

7. Explain various components of Hadoop in brief. [10]

8. Write short notes on: (any two) [5×2]

   i) Combiner Functions
   ii) Fault tolerant systems
   iii) JSON
   iv) Unstructured data

***

5/14 1)

35E TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2071 Magh

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / II	Time	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt* ***All*** *questions.*
- ✓ *The figures in the margin indicate* ***Full Marks****.*
- ✓ *Assume suitable data if necessary.*

1. a) Suppose you are working in a company that works on analyzing data. How would you determine if they have hit the big data phenomenon? Also, how would you explain your boss about the need to shift into new domain? [7]

   b) Explain Vertical and Horizontal Scaling. Explain their significance in Big data. [8]

2. Why GFS creates multiple replicas for a chunk? Explain the GFS architecture with suitable block diagram. [3+7]

3. What are the beauties of functional programming? Differentiate the terms mapping and folding with suitable example. [4+6]

4. How NoSQL differs from traditional relational database? Explain. [7]

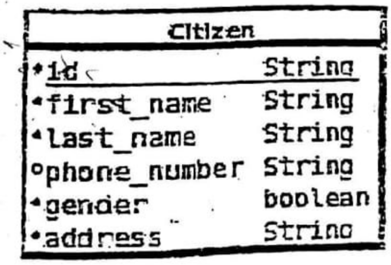
5. a) Explain Bigtable/Hbase as a NoSQL database. [7]

   b) Explain document search. Explain Elastic Search as a search engine technology. [8]

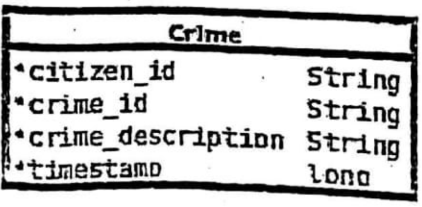
6. List components in a basic MapReduce job. Explain with diagram how does the data flow through Hadoop MapReduce. [8]

7. Below is a data schema of two files used to detect possible crime that can happen on citizens. Field id in citizen file is referenced by citizen_id in Crime file. Same citizen_id can occur multiple times in Crime file i.e. a citizen can do multiple crime at different times. [5]

   Write map/reduce program that finds count of crime for citizen. The output must be of the form (first_name + last_name) => {count of crime}



Citizen	
•id	String
•first_name	String
•last_name	String
○phone_number	String
•gender	boolean
•address	String



Crime	
•citizen_id	String
•crime_id	String
•crime_description	String
•timestamp	long

8. Write short notes on: (any two) [2×5]

   a) GFS Consistency Model

   b) CAP Theorem

   c) HADOOP and Amazon Cloud

***

35B TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2071 Bhadra

Exam.	**Regular / Back**		
Level	BE	**Full Marks**	80
Programme	BEX, BCT	**Pass Marks**	32
Year / Part	IV / II	**Time**	3 hrs.

Subject: - Big Data Technologies *(Elective II) (CT76507)*

✓ Candidates are required to give their answers in their own words as far as practicable.

✓ *Attempt* ***All*** *questions.*

✓ *The figures in the margin indicate* ***Full Marks****.*

✓ *Assume suitable data if necessary.*

1. What are the BIG-DATA challenges? Explain the data analytics process in terms of BIG-DATA. [3+7]

2. a) Explain the control flow of write mutation with diagram. [7]

   b) Explain the meta data stored by GFS master. [8]

3. a) Explain garbage collection implemented by GFS. Explain its purpose against implementing eager deletion for storage reallocation. [7]

   b) Explain CAP theorem and Eventual consistency. Also, explain the reason why some NoSQL databases like Cassandra sacrifice absolute consistency for absolute availability. [8]

4. How map-reduce works in distributed fashion? Describe the parallel efficiency of map-reduce with suitable block diagram. [3+7]

5. List out the HADOOP daemons. How HADOOP and GFS are similar interms of design architecture. [2+8]

6. Explain the term "NO-SQL". Justify "for distributed scenario normalization contradict the data availability". [3+7]

   OR

   Write down the map-reduce program to find the word frequency. [10]

7. What are the data indexing steps? Describe the components of search application. [3+7]

***

35B TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2070 Magh

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BCT	Pass Marks	32
Year / Part	IV / II	Time	3 hrs.

Subject: - Big Data Technology *(Elective II)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is a BIG DATA technology? Explain the data analytics process in big data technology. [3+7]
2. What are the common goals of GFS? Explain the general architecture of GFS. [4+6]
3. Explain the term "NO-SQL". Describe the **CAP** theorem and its implications on Distributed Databases like NO-SQL. [3+7]

   OR

   Write the equivalent **Elasticsearch** query to find the *unique* ***employee_id*** and corresponding ***salary.*** [10]
4. List out the basic components of map-reduce. Describe the parallel efficiency of map-reduce with appropriate diagram. [3+7]
5. Why Hadoop is important in BIG DATA? Explain the Master/Slave architecture of Hadoop. [4+6]
6. What is Lucene? Describe different type of analyzers available and its role in search engine development. [2+4+4]
7. Write down the map-reduce program for distributed **sort.** [10]

   OR

   How Hadoop and GFS are similar in terms of design architecture? [10]
8. **Write short notes on: (any two)** [2*5]
   a. Shadow Master
   b. Job Tracker and Task Tracker
   c. Data indexing process
   d. Hadoop and Amazon cloud

***

45K TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

Examination Control Division

2070 Bhadra

Exam.		Regular	
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / II	Time	3 hrs.

Subject: - Big Data Technologies *(Elective II)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt __All__ questions.*
- ✓ *__All__ questions carry equal marks.*
- ✓ *Assume suitable data if necessary.*

1. Explain the implications of "Big Data" in the current renaissance of computing. Describe the role of distributed system to solve the Big Data problems.
2. With diagram, explain general architecture of Google File System.
3. Why do we have single master in GFS managing millions of chunk servers? What are done to manage it without overloading single master?
4. List components in a basic MapReduce job. Explain with diagram how does the data flow through MapReduce.
5. Why does normalization fail in data analytics scenario? Explain CAP theorem.
6. Explain Eventual consistency and explain the reason why some NoSQL databases like Cassandra sacrifice absolute consistency for absolute availability.
7. Describe different components of enterprise search application.
8. You are asked to build a spam filter to know the words frequently used in the millions of spam emails received. Looping through all the documents using a single computer will be extremely time consuming. How do you speed it up by rewriting the program so that it distributes the work over several machines and processes at different phases/passes? Also describe how your algorithm will be applied to subset of data at different machines and then recombine it in next phase to produce the final output.

***