IN-SEMESTER EXAMINATION-I

BE (Semester-VII)

SUBJECT – BIG DATA ANALYTICS

Branch: IT
Div.: ALL
Duration: 60 Minutes
Instructions –

Date: 13/07/2026
Timing: 04.30 to 05.30PM
Maximum Marks: 20

1. All questions are compulsory.
2. Assume suitable data wherever necessary and state the assumptions made.
3. Diagrams / sketches should be given wherever necessary.
4. Use of logarithmic table, drawing instruments and non-programmable calculators is permitted.
5. Figures to the right indicate full marks.

Q.1		Answer any 5 of the following questions	Marks	Course Outcomes	Learning Levels
	a.	Analyze the significance of the 'Velocity' characteristic in a stock trading platform that processes thousands of transactions every second.	2	CO1	AN
	b.	Assess the storage challenges faced by a healthcare organization managing petabytes of patient records and medical images(List any two).	2	CO1	AN
	c.	Justify why customer reviews, product images, and videos stored by an online shopping platform are classified as unstructured data.	2	CO1	A
	d.	Analyze how Hadoop's fault tolerance mechanism ensures uninterrupted processing when a DataNode fails unexpectedly.	2	CO2	AN
	e.	Explain NoSQL drivers plays a key role in the emergence of NoSQL solutions.	2	CO2	AN
	f.	Examine the role of the reducer in generating the final output of a MapReduce job.	2	CO3	AN
	g.	Differentiate the implementation of Union and Intersection operations using MapReduce.	2	CO3	A
Q.2	a.	Interpret the Hadoop architecture and describe how it integrates with the Hadoop ecosystem for processing sales records.	5	CO2	AN
		OR			
	b.	Differentiate between ACID vs BASE properties with suitable example	5	CO2	AN
Q.3	a.	Evaluate the role of combiners in optimizing MapReduce jobs with a suitable example.	5	CO3	E

IN-SEMESTER EXAMINATION-II

BE (Semester-VII)

SUBJECT – Big Data Analytics

Branch: IT **Date: 01/09/2026**
Div.: ALL **Timing: 4:00 pm to 5:00 pm**
Duration: 60 Minutes **Maximum Marks: 20**

Instructions –

1. All questions are compulsory.
2. Assume suitable data wherever necessary and state the assumptions made.
3. Diagrams / sketches should be given wherever necessary.
4. Use of logarithmic table, drawing instruments and non-programmable calculators is permitted.
5. Figures to the right indicate full marks.

Q.1		Answer any 5 of the following questions	Marks	Course Outcomes	Learning Levels	PI
	a.	Interpret the characteristics of data streams and infer the stream operations performed in an online payment gateway processing thousands of transactions every second.	2	CO4	AN	4.4.1
	b.	Categorize the different data stream sources generated by traffic sensors, surveillance cameras, and weather stations in a smart city.	2	CO4	AN	1.6.1
	c.	Identify the major components of the DGIM algorithm and justify their significance in monitoring applications.	2	CO4	AN	4.6.1
	d.	Interpret the practical applications of the Canopy algorithm for grouping millions of telecom subscribers prior to detailed clustering.	2	CO5	AN	2.6.4
	e.	Justify the advantages of the PCY algorithm for performing market basket analysis on millions of supermarket transactions.	2	CO5	AN	1.7.1
	f.	A movie recommendation platform stores user ratings for thousands of movies. Illustrate the utility matrix for this scenario using a suitable example.	2	CO6	AN	1.6.1
	g.	A search engine encounters web pages that do not contain outgoing hyperlinks. Examine how such dead ends influence the PageRank computation process.	2	CO6	AN	4.4.1
Q.2	a.	Outline the execution of the Park–Chen–Yu (PCY) algorithm for frequent pattern mining on a large transactional dataset with limited main memory.	5	CO5	AN	1.2.1
		OR				

	b.	Demonstrate how the SON algorithm performs frequent itemset mining in a distributed e-commerce environment and justify how it eliminates false negatives.	5	CO5	A	2.8.1
Q.3	a.	A social networking application evaluates different community detection techniques. Compare direct discovery of communities with other community detection algorithms, highlighting their strengths and limitations.	5	CO6	AN	4.6.2
		OR				
	b.	An online community platform identifies groups of users with similar interests. Investigate how direct discovery of communities is performed in social network graph analysis.	5	CO6	AN	5.4.1

END SEMESTER EXAMINATION, OCT 2026

14 OCT 2026

B.E. SEMESTER VII (CBCGS-HME 2023)

Branch:	**IT**	**Q.P. Code:**	**E7C4O62-1**
Subject:	**Big Data Analytics**	**Duration:**	**2 hours**
Subject Code:	**PCC-IT702**	**Max. Marks:**	**60**

Instructions: **1. All sections are compulsory**
2. Figures to the right indicate full marks.
3. Assume suitable data if necessary and state the assumptions clearly.

Section-I **Short Answer Questions (Answer any 05 questions out of 06)** **(10 Marks)**
(Fundamental, Core Types)

Q. No.		Marks	CO	RBT Level	PI
1	Analyze the major storage challenges encountered when managing petabytes of data	2	CO1	AN	2.5.2
2	Identify the core Hadoop components required to build a Hadoop cluster for large-scale data processing and justify their roles.	2	CO2	AN	1.6.1
3	Evaluate a real-world application where MapReduce provides an efficient solution for large-scale data processing.	2	CO3	E	2.8.4
4	Analyze the significance of the major components of the DGIM algorithm.	2	CO4	AN	4.6.1
5	Apply frequent pattern mining to identify products commonly purchased together in an online shopping system.	2	CO5	A	1.7.1
6	Examine the role of clustering in social network graphs and identify a suitable clustering algorithm.	2	CO6	E	5.4.1

Section-II **Descriptive Answer Questions (Answer any 04 out of 06)** **(20 Marks)**
(Descriptive, Comprehension Types)

1	Apply the phases of the Big Data Analytics Lifecycle to a real-world data analytics problem and indicate the purpose of each phase.	5	CO1	A	2.5.2
2	Illustrate the HBase architecture and discuss how it enables efficient storage and retrieval of billions of data.	5	CO2	AN	1.2.1
3	Evaluate traditional parallel processing and distributed processing based on their methods, scalability, and limitations.	5	CO3	E	2.5.2
4	Apply the DGIM algorithm to estimate the number of 1s in a binary data stream over a sliding window. 2011 0101	5	CO4	A	1.2.1
5	Differentiate the Canopy clustering approach from Hierarchical Agglomerative Clustering with respect to computational complexity and memory requirements.	5	CO5	AN	1.2.1
6	Examine the structure of the web and assess the impact of spider traps on PageRank computation.	5	CO6	E	2.6.2

Section-III **Long Answer Question (Answer any 03 out of 05)** **(30 Marks)**
(Application, Analytical, Evaluation, Design Type)

1	Illustrate the YARN architecture and discuss how it efficiently allocates resources across a Hadoop cluster.	10	CO2	AN	5.4.1

2	Perform the Union operation using MapReduce for the given datasets and illustrate all intermediate processing steps. **Map Worker 1** Table 1: A 1, B 2; A 2, B 3; A 5, B 6 Table 2: A 2, B 3; A 4, B 4; A 6, B 7 **Map Worker 2** Table 1: A 6, B 7; A 8, B 3; A 7, B 6 Table 2: A 5, B 3; A 3, B 3; A 0, B 7	10	CO3	A	4.5.1
3	Compute the number of distinct elements in the given data stream using the Flajolet–Martin algorithm for the following hash functions: (a) (h(x)=(3x+1) mod 32) (b) (h(x)=(x+6) mod 32). 123 21 45	10	CO4	A	3.5.1
4	Design a dimensionality reduction strategy using PCA and justify the selection of principal components.	10	CO5	C	3.8.1
5	A movie streaming platform has historical ratings from several users. Analyze how collaborative filtering can be used to generate movie recommendations for a target user by examining rating patterns, determining similar users, and relating their preferences to the recommended movies.	10	CO6	AN	5.4.1