Reg. No. :

Question Paper Code : 20397

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2023.

Fifth Semester

Computer Science and Engineering

CCS 334 — BIG DATA ANALYTICS

(Common to : Computer Science and Design/Computer Science and Engineering (Artificial Intelligence and Machine Learning)/Computer and Communication Engineering/Electrical and Electronics Engineering/Artificial Intelligence and Data Science/Computer Science and Business Systems and Information Technology)

(Also common to Minor Degree)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Distinguish Big Data processing and distributed processing.

2. Differentiate inter and trans firewall analytics.

3. What is the main advantage of using schemaless databases?

4. Summarize the key characteristics of the data model in Cassandra.

5. Define MapReduce workflows in the context of data processing.

6. What is the primary role of YARN in a Hadoop ecosystem?

7. In the context of Hadoop, what is the purpose of Hadoop Pipes?

8. Why is ensuring data integrity crucial in Hadoop distributed systems?

9. How does HBase differ from traditional relational databases in terms of data storage and access patterns?

10. Explain the primary purpose of HiveQL queries in the Hive ecosystem.

PART B — (5 × 13 = 65 marks)

11. (a) Elaborate the significance of the three V's (volume, velocity, and variety) in the context of big data.

Or

(b) List the role and implications of crowdsourcing analytics in today's data-driven landscape.

12. (a) Explore how graph databases handle huge data and its unique capabilities in data management and analytics.

Or

(b) Explain master-slave replication and consistency in big data distributed systems.

13. (a) Discuss the components involved in the anatomy of a MapReduce job run.

Or

(b) List the Relational-Algebra Operations. Illustrate the application of MapReduce by providing detailed explanations of two instances.

14. (a) Explain generic methods and classes in Java. Give a procedure to stop Java serialisation.

Or

(b) Elaborate the impact of seamless Hadoop integration on enhancing data processing and analytics.

15. (a) Examine HBase's real-world uses and benefits as a scalable and versatile NoSQL database.

Or

(b) Narrate the salient points on data manipulation in Hive using HiveQL.

PART C — (1 × 15 = 15 marks)

16. (a) Provide a conclusion by presenting insights into the distinct factors that organizations should carefully evaluate when choosing between MongoDB and Cassandra to meet the specific requirements of their applications. Discuss the same.

Or

(b) Explain the complex design principles and architecture of the Hadoop. Distributed File System (HDFS) to comprehend its functions and components.

———————

www.BrainKart.com

Anna University

for Affilated Engineering College - 2021 Regulation

AID (Artificial Intelligence & Data Science Engineering)

1st Semester | 2nd Semester | 3rd Semester | 4th Semester

5th Semester | 6th Semester | 7th Semester | 8th Semester

Click on Subject/Paper under Semester to enter.

1st Semester

- Professional English - I - HS3152
- Matrices and Calculus - MA3151
- Engineering Physics - PH3151
- Engineering Chemistry - CY3151
- Problem Solving and Python Programming - GE3151

2nd Semester

- Professional English - II - HS3252
- Statistics and Numerical Methods - MA3251
- Engineering Graphics - GE3251
- Physics for Information Science - PH3256
- Basic Electrical and Electronics Engineering - BE3251
- Data Structures Design - AD3251

3rd Semester

- Discrete Mathematics - MA3354
- Digital Principles and Computer Organization - CS3351
- Database Design and Management - AD3391
- Design and Analysis of Algorithms - AD3351
- Data Exploration and Visualization - AD3301
- Artificial Intelligence - AL3391

4th Semester

- Environmental Sciences and Sustainability - GE3451
- Probability and Statistics - MA3391
- Operating Systems - AL3452
- Machine Learning - AL3451
- Fundamentals of Data Science and Analytics - AD3491
- Computer Networks - CS3591

5th Semester

- Deep Learning - AD3501
- Data and Information Security - CW3551
- Distributed Computing - CS3551
- Big Data Analytics - CCS334
- Elective 1
- Elective 2

6th Semester

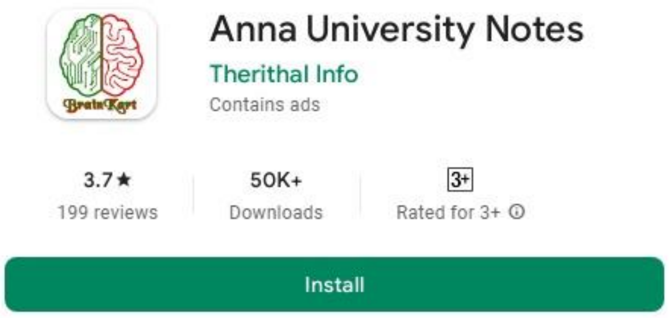
- Embedded Systems and IoT - CS3691
- Open Elective-1
- Elective-3
- Elective-4
- Elective-5
- Elective-6

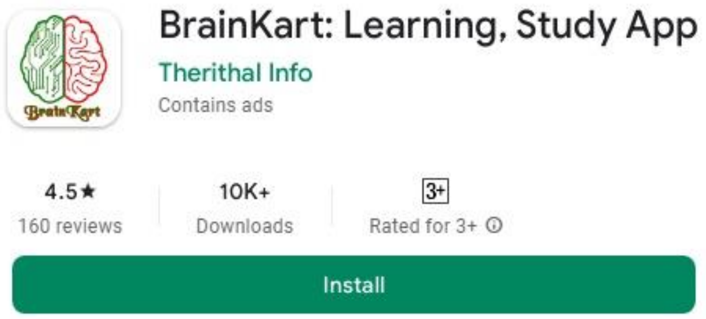
7th Semester

- Human Values and Ethics - GE3791
- Open Elective 2
- Open Elective 3
- Open Elective 4
- Management Elective

8th Semester

- Project Work / Intership





All Computer Engg Subjects - [B.E., M.E.,] (Click on Subjects to enter)

Programming in C	Computer Networks	Operating Systems
Programming and Data Structures I	Programming and Data Structure II	Problem Solving and Python Programming
Database Management Systems	Computer Architecture	Analog and Digital Communication
Design and Analysis of Algorithms	Microprocessors and Microcontrollers	Object Oriented Analysis and Design
Software Engineering	Discrete Mathematics	Internet Programming
Theory of Computation	Computer Graphics	Distributed Systems
Mobile Computing	Compiler Design	Digital Signal Processing
Artificial Intelligence	Software Testing	Grid and Cloud Computing
Data Ware Housing and Data Mining	Cryptography and Network Security	Resource Management Techniques
Service Oriented Architecture	Embedded and Real Time Systems	Multi - Core Architectures and Programming
Probability and Queueing Theory	Physics for Information Science	Transforms and Partial Differential Equations
Technical English	Engineering Physics	Engineering Chemistry
Engineering Graphics	Total Quality Management	Professional Ethics in Engineering
Basic Electrical and Electronics and Measurement Engineering	Problem Solving and Python Programming	Environmental Science and Engineering

Reg. No. : □□□□□□□□□□□□

Question Paper Code : 50417

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2024.

Fifth/Sixth Semester

Computer Science and Engineering

CCS 334 — BIG DATA ANALYTICS

(Common to Computer Science and Design/Computer Science and Engineering (Artificial Intelligence and Machine Learning) Computer and Communication Engineering/Electrical and Electronics Engineering/Artificial Intelligence and Data Science/Computer Science and Business Systems/Information Technology)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What characteristics will define a dataset as big data?

2. Analytic professionals need permissions to utilize the enterprise data warehouse. In such case, suggest an alternate mechanism that is ideal for data exploration.

3. What are NoSQL databases? Give example.

4. Why Cassandra data model is very popular among developers?

5. What is the role of mini reducer in Map reduce?

6. How YARN supports the notion of resource reservation?

7. List out the applications for which HDFS does not work well.

8. Mention the necessity for serialization in Hadoop and present the default serialization framework supported by Hadoop.

9. Write a short note on HiveQL queries.

10. Mention the data types in Hive.

PART C — (1 × 15 = 15 marks)

16. (a) The Indian government has decided to use big data analytics to optimize bus transport management. The State Bus Transport Authority division has planned to collect and disseminate real-time data to identify the causes of transport delays. For this purpose the appropriate data is collected from various sources such as bus timetables, inductive-loop traffic detectors, closed-circuit television cameras and GPS updates from the city buses. This allows traffic controllers to see the current status of the entire bus network. Elaborate on the different types of data that are being generated in this scenario, devise a big data Ecosystem for Bus transport and also explain the key roles for the new big data ecosystem with a diagram.

Or

(b) A health researcher wants to predict "VO2max", an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment, as well as requiring individuals to exercise to their maximum (i.e., until they can no longer continue exercising due to physical exhaustion). This can put off individuals who are not very active/fit and those who might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO2max based on attributes that can be measured more easily and cheaply. To this end, a researcher recruited participants to perform a maximum VO2max test, but also recorded their "age", "weight" and "heart rate". The researcher wants to store the recorded data of size 2.5GB in a Big data environment. Assume the default block size to be 128 MB with the replication factor as 3. Calculate the number of blocks needed for storing this dataset in HDFS. Illustrate and explain the sequence of events on how to use the methods provided by FileSystem API while reading a file.