

SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur - 603 203

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

QUESTION BANK



VI SEMESTER

1904011 – BIG DATA ANALYTICS

Regulation – 2019

Academic Year 2024 – 2025 (Even Semester)

Prepared by

Ms. R. LAKSHMI, Assistant Professor(Sel.G) / AI & DS

UNIT 1 INTRODUCTION TO BIG DATA

Big Data – Definition, Characteristic Features – Big Data Applications - Big Data vs Traditional Data - Risks of Big Data - Structure of Big Data - Challenges of Conventional Systems - Web Data – Evolution of Analytic Scalability - Evolution of Analytic Processes, Tools and methods - Analysis vs Reporting - Modern Data Analytic Tools

PART – A

Q.No	Question	Competence	Level
1	What is Big Data?	Remember	BTL 1
2	Differentiate Big Data and Conventional Data.	Understand	BTL 2
3	List the various dimensions of growth of Big Data.	Remember	BTL 1
4	List the main characteristics of Big Data.	Remember	BTL 1
5	Illustrate the risk of big data.	Understand	BTL 2
6	What is web data?	Remember	BTL 1
7	List the sources of big data.	Remember	BTL 1
8	Illustrate the challenges in big data.	Understand	BTL 2
9	Why domain expertise is required for any type of Data Analytics?	Remember	BTL 1
10	Give reason: “Web Data is the most popular Big Data”.	Understand	BTL 2
11	Interpret & Justify “Accuracy in big data is beneficial”	Understand	BTL 2
12	Give the structure of big data.	Understand	BTL 2
13	Give the list of bigdata applications.	Understand	BTL 2
14	List the challenges of convectional system.	Remember	BTL 1
15	Tell the role of analytical scalability in big data.	Remember	BTL 1
16	Outline the structure of bigdata .	Understand	BTL 2
17	Illustrate how big data can be represented.	Understand	BTL 2
18	Relate the importance of analysis vs reporting	Remember	BTL 1
19	Name the technologies used to handle big data.	Remember	BTL 1
20	What is a sand box.	Remember	BTL 1
21	Give the traditional analytical architecture.	Understand	BTL 2
22	Outline the issues in dealing with huge amount of data for analysis purpose.	Understand	BTL 2
23	Differentiate data warehouse architecture vs MPP Architecture.	Understand	BTL 2
24	What do you understand by the term data privacy.	Remember	BTL 1

PART-B

Q.No.	Question	Competence	Level
1	Examine the main features of a big data in detail. (13)	Analyze	BTL 4
2	(i) List the main characteristics of big data. (4) (ii) Analyze the big data architecture with a neat schematic diagram. (9)	Analyze	BTL 4

3	Identify the risk in handling big data and elaborate on it. (13)	Apply	BTL 3
4	Examine the structure of big data representation (13)	Apply	BTL 3
5	(i) Point out the features of Massive parallel processing system. (5) (ii) Explain the use of Massive Parallel Processing system in big data analytics (8)	Analyze	BTL 4
6	Identify the challenges faced by traditional system and explain with suitable example . (13)	Apply	BTL 3
7	Analyze in detail the analysis tools and reporting tools used in Big-data. (13)	Analyze	BTL 4
8	(i) What is a analytical data set ? (3) (ii) Explain the types of analytical data set (10)	Analyze	BTL 4
9	(i) Analyze about the need of web data. (6) (ii) Examine what does the web data reveal. (7)	Analyze	BTL 4
10	(i) Evaluate how big data are effectively filtered. (6) (ii) Explain how big data are mixed with traditional one. (7)	Evaluate	BTL 5
11	Analyse the Evolution Tools and Method in big data. (13)	Apply	BTL 3
12	(i) Assess the difficulties faced by conventional systems. (5) (ii) Explain the differences between big data from the traditional one.(8)	Evaluate	BTL 5
13	Sketch and summarize how the analytical scalability is handled in big data. (13)	Apply	BTL 3
14	(i) Point out some of the web data in current action today. (6) (ii) Analyse the modern tools for big data analysis. (7)	Analyze	BTL 4
15	Compare and contrast the analysis and reporting methods and tools.(13)	Analyze	BTL 4
16	Summarize the importance of analytical sandbox in detail. (13)	Evaluate	BTL 5
17	Illustrate the Evolution of Analytical Scalability. (13)	Apply	BTL 3
PART - C			
1.	Generalize in detail about the challenges of the Bigdata (15)	Create	BTL 6
2.	Justify the Statement “Web Data is the Most Popular Big Data” with reference to data analytic professional. (15)	Evaluate	BTL 5
3.	Comment on the statement “Is the “Big” Part or the “Data” Part More Important “ in the term bigdata. (15)	Evaluate	BTL 5
4.	Develop the role of Analytic Sandbox and its benefits in the Analytic Process. (15)	Create	BTL 6
5.	Explain the evolution of Analytical Scalability. (15)	Evaluate	BTL 5

UNIT II HADOOP FRAMEWORK

Distributed File Systems - Large-Scale File System Organization – HDFS concepts - MapReduce Execution, Algorithms using MapReduce, Matrix-Vector Multiplication – Hadoop YARN

PART – A

Q.No	Question	Competence	Level
1	What is hadoop?	Remember	BTL 1
2	Show how does Map-Reduce computation execute.	Remember	BTL 1
3	List the key advantages in hadoop.	Remember	BTL 1
4	What is hadoop YARN?	Understand	BTL 2
5	List the core concepts of HADOOP.	Remember	BTL 1
6	Define MAP REDUCE concepts.	Remember	BTL 1
7	Demonstrate how a key value pair is formed.	Understand	BTL 2
8	Summarize the importance of DFS.	Understand	BTL 2
9	What is HDFS?	Remember	BTL 1
10	Outline the features of HDFS.	Understand	BTL 2
11	Discuss about name node .	Understand	BTL 2
12	What are the goals of HDFS?	Remember	BTL 1
13	Give an outline about data node.	Understand	BTL 2
14	Illustrate about shuffle and sort algorithm	Understand	BTL 2
15	What are the advantages of HDFS?	Remember	BTL 1
16	List out the hadoop applications.	Remember	BTL 1
17	What is matrix multiplication?	Understand	BTL 2
18	Why the partitions are shuffled in map reduce.	Remember	BTL 1
19	Point out the steps in map reduce algorithm.	Remember	BTL 1
20	Illustrate about application manager in hadoop.	Understand	BTL 2
21	Give the application of distributed system.	Understand	BTL 2
22	Show the importance of resource manager in Hadoop.	Remember	BTL 1
23	Expand YARN.	Remember	BTL 1
24	Outline the advantages of Hadoop.	Understand	BTL 2

PART-B

Q.No.	Question	Competence	Level
1	List the features of Hadoop and assess the functionalities of Hadoop? (13)	Analyze	BTL 4
2	Inspect the various core components of the Hadoop and examine them in detail. (13)	Apply	BTL 3
3	Explain about Hadoop distributed file system architecture with neat diagram. (13)	Evaluate	BTL 5
4	Explain briefly on (i) Algorithms using MapReduce. (8) (ii) Advantages of MapReduce. (5)	Analyze	BTL 4

5	Compare and Contrast the Hadoop and Map R. (13)	Analyze	BTL 4
6	Analyse the steps of Map Reduce Algorithms. (13)	Analyze	BTL 4
7	Inspect the concepts of HDFS and examine them in detail. (13)	Apply	BTL 3
8	(i) Identify the map and reduce algorithm in detail (6) (ii) Illustrate the map reduce algorithm with an example (7)	Apply	BTL 3
9	Asses the importance and explain about the phases in map reduce with an example. (13)	Analyze	BTL 4
10	Analyze the Apache Hadoop YARN architecture. (7)	Analyze	BTL 4
11	(i) Examine Map Reduce framework in detail. Draw the architectural diagram (7) (ii) Define HDFS and analyze HDFS in detail. (6)	Apply	BTL 3
12	Discuss matrix vector multiplication in detail. (13)	Evaluate	BTL5
13	(i) Explain what is YARN. (6) (ii) Illustrate HADOOP YARN architecture with neat diagram. (7)	Apply	BTL 3
14	Use map reduce architecture to illustrate the concept of hadoop for the following example Welcome to Hadoop class Hadoop is interesting Hadoop is useful Hadoop is useful in big data (13)	Create	BTL 6
15	Analyse the distributed file system. (13)	Apply	BTL 3
16	(i) Discover the benefits of using HDFS. (8) (ii) Examine the benefits of HDFS in marketing application. (5)	Apply	BTL 3
17	Discuss HDFS data replication. (13)	Evaluate	BTL5

PART - C

1	Recommend a procedure to find the number of Occurrence of a word in a document. (15)	Evaluate	BTL 5
2	(i) Generalize with a neat sketch about processing of a job in Hadoop. (8) (ii) List the various operational modes of Hadoop cluster configuration and explain in detail about configuring / installing the hadoop in local/standalone mode (7)	Create	BTL 6
3	Summarize how google file system differs from the Hadoop file system and explain the google file system architecture with a neat sketch. (15)	Evaluate	BTL 5
4	Prepare a Map-Reduce Algorithm to get the Dot Product of two Large Vectors. Assuming Only non-zero elements of those vectors are given in input files and output file should show only non-zero entries(assuming two vectors are same size) (15) ex: v1=[5 4 0 1 2] v2=[4 2 1 0 6] file1: file2: output: (0,5) (0,4) (0,20) (1,4) (1,2) (1,8) (3,1) (2,1) (4,12) (4,2) (4,6)	Create	BTL 6
5	Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. Using Hadoop and Map Reduce, write a program to count the occurrence of pre dominant key words. (15)	Create	BTL 6

UNIT III - DATA ANALYSIS

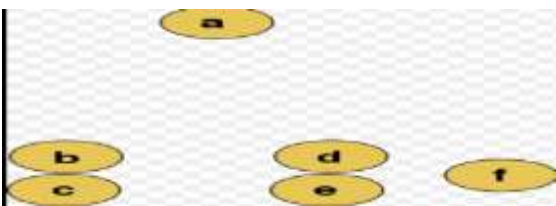
Statistical Methods: Regression modelling, Multivariate Analysis - Classification: SVM & Kernel Methods - Rule Mining - Cluster Analysis, Types of Data in Cluster Analysis, Partitioning Methods, Hierarchical Methods, Density Based Methods, Grid Based Methods, Model Based Clustering Methods, Clustering High Dimensional Data - Predictive Analytics – Data analysis using R.

PART – A

Q.No.	Question	Competence	Level
1	Define data analysis.	Remember	BTL 1
2	Show what classification is.	Remember	BTL 1
3	Tell about support-vector machines.	Remember	BTL 1
4	Define regression.	Understand	BTL 2
5	List out the different types of regression	Remember	BTL 1
6	Define multivariate analysis.	Remember	BTL 1
7	List the types of clustering.	Remember	BTL 1
8	Interpret the importance of classification in data analysis.	Understand	BTL 2
9	What is kernel?	Remember	BTL 1
10	Illustrate about rule mining.	Understand	BTL 2
11	Give the types of kernels.	Understand	BTL 2
12	What is Multiple Linear Regression?	Remember	BTL 1
13	Illustrate what is predictive analysis.	Understand	BTL 2
14	Compare and contrast regression and correlation.	Understand	BTL 2
15	What is clustering?	Understand	BTL 2
16	List the types of clustering.	Remember	BTL 1
17	What is SVM?	Understand	BTL 2
18	Outline about classification and clustering.	Understand	BTL 2
19	Point out the importance of clustering in data analysis.	Understand	BTL 2
20	Summarize about R	Understand	BTL 2
21	Give an outline on density based clustering..	Understand	BTL 2
22	What is the grid based clustering?	Remember	BTL 1
23	List the types of hierarchical clustering.	Understand	BTL 2
24	Show the partitioning methods in clustering.	Remember	BTL 1

PART-B

Q.No.	Question	Competence	Level
1	(i) What is regression? Inspect the various the types of regression. (5) (ii) Examine the purpose of using Regression Modeling in Data Analysis. (8)	Analyze	BTL 4
2	Describe in detail about Multivariate Analysis techniques with suitable example. (13)	Analyze	BTL 4

3	(i) Describe SVM in detail. (7) (ii) List out and explain some of the applications of SVM in detail. (6)	Apply	BTL 3
4	Explain about kernel methods in detail. (13)	Analyze	BTL 4
5	Develop a short note on types of data in clustering and its Importance. (13)	Apply	BTL 3
6	(i) Discuss in detail about the rule mining. (6) (ii) Explain in detail about association rule mining (7)	Analyze	BTL 4
7	(i) Examine clustering in data analysis. (3) (ii) Illustrate density based and Grid based clustering in detail. (10)	Analyze	BTL 4
8	Describe how clustering is used in high dimensional data. (13)	Apply	BTL 3
9	Illustrate the approaches of clustering with example (13)	Apply	BTL 3
10	Analyze about model based clustering and high dimensional clustering in detail. (13)	Analyze	BTL 4
11	Assess about the partitioning method of clustering in detail. (13)	Analyze	BTL 4
12	Explain k-means clustering algorithm with an example. (13)	Evaluate	BTL 5
13	(i) What is prediction? Identify how prediction helps in data analysis. (5) (ii) Explain in detail about predictive analysis (8)	Apply	BTL 3
14	Analyze the different hierarchical clustering techniques (13)	Apply	BTL 3
15	Examine about model based clustering. (13)	Analyze	BTL 4
16	Explain the density based clustering with a neat diagram. (13)	Analyse	BTL 4
17	Summarize grid based clustering in detail. (13)	Evaluate	BTL 5
PART – C			
1	Comment the statement in detail: “Data Analysis is not a decision-making system, but a decision-supporting system”. (15)	Analyze	BTL 6
2	Create a Regression Model for “happy people get many hours of sleep” using your own data and what kind of inferences it provides. (15)	Create	BTL 6
3	Summarize hierarchical clustering in detail. Analyse the given diagram and draw the dendrogram using hierarchical clustering algorithm. (15) 	Evaluate	BTL 5
4	Compose the K-means partitioning algorithm using the given data. Cluster the following eight points (with (x, y) representing locations) into three clusters: (15) A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)	Create	BTL 6
5	Summarize predictive analysis with some application. (15)	Evaluate	BTL 5

UNIT IV MINING DATA STREAMS

Streams: Concepts – Stream Data Model and Architecture - Sampling data in a stream - Mining Data Streams and Mining Time-series data - Real Time Analytics Platform (RTAP) Applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions

PART – A

Q.No.	Question	Competence	Level
1	List the main characteristics of stream sources.	Remember	BTL 1
2	What is a data stream?	Remember	BTL 1
3	Why data stream management is relevant and necessary in data Mining ?	Remember	BTL 1
4	What is meant by data stream ?	Remember	BTL 1
5	What is Sampling data in a stream?	Remember	BTL 1
6	List out the few challenges of data stream mining algorithms.	Remember	BTL 1
7	Differentiate between DBMS and DSMS.	Understand	BTL 2
8	Interpret the statement “Filtering a Data Stream”.	Understand	BTL 2
9	Give the applications of DSMS.	Understand	BTL2
10	Define Real-Time Analysis.	Understand	BTL 2
11	Show how to deal with infinite streams.	Understand	BTL 2
12	Define Time Series Data.	Remember	BTL 1
13	Show what examples you can find for stream sources.	Understand	BTL 2
14	What is called Data Stream Mining?	Remember	BTL 1
15	Difference between RTAP (real time analytics platform) and RTSA (real time sentiment analysis).	Understand	BTL 2
16	State the reason why we need RTAP.	Understand	BTL 2
17	What is Real Time Analytics Platform (RTAP)?	Remember	BTL 1
18	Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. Infer the third moment of this stream?	Understand	BTL 2
19	Tell about real time data.	Remember	BTL 1
20	What information is used to substitute the view of streams over databases?	Remember	BTL 1
21	List the importance of social media analytics.	Remember	BTL 1
22	Give the reasons why do we need RTAP.	Remember	BTL 1
23	Outline about prediction and forecasting.	Understand	BTL 2
24	Illustrate about time series data.	Understand	BTL 2

PART-B

Q.No.	Question	Competence	Level
1	(i) Define data stream. (3) (ii) Describe the Data Stream model with a neat architecture Diagram. (10)	Analyze	BTL 4
2	Illustrate briefly about the sources of data stream. (13)	Apply	BTL 3
3	Based on what you know, Analyse how would you partition the	Analyze	BTL 4

	following bit stream into buckets 1001011011101? Find all of them? (13)		
4	Explain about the stream data model and its architecture. (13)	Analyze	BTL 4
5	Analyse and write a short note on Aurora system model. (13)	Analyze	BTL 4
6	(i) Explain Sampling in Data Streams. (5) (ii) Explain the sampling types in detail (8)	Analyze	BTL 4
7	Describe about Aurora query model. (13)	Analyze	BTL 4
8	Examine how mining is done with data streams. (13)	Apply	BTL 3
9	(i) Describe briefly how to count the distinct elements in a stream. (9) (ii) What do you mean by count-distinct problem. (4)	Analyze	BTL 4
10	Construct short notes on (i) Sliding window concept (7) (ii) Land mark window concept (6)	Apply	BTL 3
11	Illustrate how would you describe the various windowing approaches to data stream mining. (13)	Apply	BTL 3
12	(i) List the methods for analyzing time series data. (7) (ii) Analyze about the several types of motivation and data analysis available for time series? (6)	Analyze	BTL 4
13	(i) Illustrate what approaches are used to estimate the moments. (8) (ii) Examine the function cost of exact counts. (5)	Analyze	BTL 4
14	(i) Evaluate what is real time sentiment analysis. (5) (ii) Assess how the mining concept is used in real time sentiment analysis (8)	Evaluate	BTL 5
15	Illustrate can you identify the following? (13) Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form $h(x) = ax + b \text{ mod } 32$ for some a and b . You should treat the result as a 5-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is: (a) $h(x) = 2x + 1 \text{ mod } 32$. (b) $h(x) = 3x + 7 \text{ mod } 32$. (c) $h(x) = 4x \text{ mod } 32$	Apply	BTL 3
16	(i) Express what bloom filters are. (3) (ii) Summarize the relevance of bloom filters in data mining. (10)	Evaluate	BTL 5
17	Describe how is data analysis used in (i) stock market predictions. (7) (ii) weather forecasting predictions. (6)	Analyze	BTL 4
PART – C			
1	Evaluate the process of Data Stream Mining with suitable examples. (15)	Evaluate	BTL 5
2	Summarize data streaming algorithms in detail. Evaluate key stream mining problems and discuss the challenges associated with each problem. (15)	Evaluate	BTL 5
3	Generalize data stream management systems in detail. (15)	Create	BTL 6
4	Prepare a generic design for Realtime Analytics Platform (RTAP). Discuss your answer related to real time sentiment analysis. (15)	Create	BTL 6
5	Evaluate the Bloom Filter in detail with an algorithm. Apply this bloom filter algorithm in Adhar card(Unique Identification number) (15)	Evaluate	BTL 5

UNIT V BIG DATA FRAMEWORKS

Introduction to NoSQL – Aggregate Data Models – Hbase: Data Model and Implementations – Hbase Clients – Examples – .Cassandra: Data Model – Examples – Cassandra Clients – Hadoop Integration. Pig – Grunt – Pig Data Model – Pig Latin – developing and testing Pig Latin scripts. Hive – Data Types and File Formats – HiveQL Data Definition – HiveQL Data Manipulation – HiveQL Queries

PART – A

Q.No.	Question	Competence	Level
1	Define NoSQL database.	Remember	BTL 1
2	Interpret few key features of NoSQL.	Understand	BTL 2
3	Tell the components of Hadoop framework.	Remember	BTL 1
4	Differentiate between SQL and NoSQL.	Understand	BTL 2
5	What is the advantage of NoSQL?	Remember	BTL 1
6	Give the disadvantages of NoSQL.	Remember	BTL 1
7	What is HBase?	Remember	BTL 1
8	Show the advantage of Cassandra.	Remember	BTL 1
9	Who is generating big data and what are the ecosystem projects used for processing?	Remember	BTL 1
10	Illustrate the difference between HBase and Hive.	Understand	BTL 2
11	List the aggregate data models.	Remember	BTL 1
12	Express what is Pig in Hadoop.	Understand	BTL 2
13	What is Apache pig?	Remember	BTL 1
14	Illustrate the difference between Pig and Hive.	Understand	BTL 2
15	Outline the usage of Pig, Hive and HBase.	Understand	BTL 2
16	Give the features of Hive.	Understand	BTL 2
17	Define Pig, Hive and HBase	Remember	BTL 1
18	What is hive in Big Data?	Remember	BTL 1
19	What is Cassandra Client ?	Remember	BTL 1
20	List out the types of built-in operator in HIVE.	Remember	BTL 1
21	Differentiate between HIVE internal tables and external tables.	Understand	BTL 2
22	Mention the methods used in class HTABLE.	Remember	BTL 1
23	Infer the importance of Cassandra.	Understand	BTL 2
24	Define Hadoop Streaming.	Remember	BTL 1

PART-B

Q.No.	Question	Competence	Level
1	(i) Describe the key features of NoSQL. (7) (ii) List the advantages and disadvantages of NoSQL. (6)	Analyze	BTL 4
2	(i) Illustrate in detail about Hive data manipulation, queries, and data types. (8) (ii) Illustrate data definition in Hive. (5)	Analyze	BTL 4
3	Describe the system architecture and components of Hive and Hadoop. (13)	Analyze	BTL 4
4	Analyze briefly on aggregate data models with cluster and order relationship. (13)	Analyze	BTL 4
5	Evaluate two types of data storage medium in Hbase. (13)	Evaluate	BTL 5
6	(i) Describe about HBase in detail. (7) (ii) Explain Hbase clients in detail. (6)	Analyze	BTL 4

7	(i) Analyse how Cassandra is integrated with Hadoop. (ii) Explain the tools related to Hadoop.	(6) (7)	Analyze	BTL 4
8	Illustrate briefly on Hbase architecture with neat diagram	(13)	Analyze	BTL 4
9	Assess and write short notes on (i) Features of Hive. (ii) Limitations of hive.	(7) (6)	Analyze	BTL 4
10	Discuss about Cassandra clients and apply it to suitable example.	(13)	Apply	BTL 3
11	Compare and Contrast the Hbase and Hive.	(13)	Analyze	BTL 4
12	(i) Explain about Pig in detail. (ii) What is invoking a Grunt shell?	(7) (6)	Analyze	BTL 4
13	Describe about Pig data model in detail with neat diagram.	(13)	Analyze	BTL 4
14	Explain how to develop and test pig scripts for data processing.	(13)	Apply	BTL 3
15	Identify the difference between Apache Hive and Apache Hbase	(13)	Apply	BTL 3
16	Evaluate hive data types and file formats.	(13)	Evaluate	BTL 5
17	Illustrate in detail Hive Query Language.	(13)	Apply	BTL 3
PART - C				
1	Explain in detail about Hive Architecture and its Features.	(15)	Evaluate	BTL 5
2	Recommend a procedure to find the number of occurrences of a word in a document using HIVE.	(15)	Analyze	BTL 5
3	Explain in detail about Pig Architecture components. List out the key features of Pig.	(15)	Create	BTL 6
4	Formulate the query for the following: a. Create a database named “ Students” (3) b. Display a list of all databases.(3) c. Describe the databases (3) d. Alter the databases(2) e. Drop database(2) f. To make the database as current working directory.(2)	(15)	Create	BTL 6
5	Explain the features of Apache Cassandra? Explain in detail about Cassandra data model.	(5) (10)	Evaluate	BTL5