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 2022 – 2023 (Even Semester)

Prepared by

Dr.R.Deepa, Assistant Professor (O.G)

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 advantages of big data.	Remember	BTL 1
4	List the main characteristics of Big Data.	Remember	BTL 1
5	Illustrate the risk of big data.	Apply	BTL 3
6	What is web data?	Remember	BTL 1
7	List the sources of big data.	Remember	BTL 1
8	Analyse the challenges in big data.	Analyze	BTL 4
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	Justify "Accuracy in big data is beneficial"	Evaluate	BTL 5
12	Give the structure of big data.	Understand	BTL 2
13	Write down the characteristics of big data applications.	Understand	BTL 2
14	Analyse the challenges of conventional system.	Analyze	BTL 4
15	Can you generalize the role of analytical scalability in big data.	Create	BTL 6
16	Analyse the structure of bigdata .	Apply	BTL 3
17	Write down the four computing resources of big data storage.	Apply	BTL 3
18	Assess the importance of analysis vs reporting	Evaluate	BTL 5
19	Analyse the technologies used to handle big data.	Analyze	BTL 4
20	Generalize what is a sand box.	Create	BTL 6
21	Give the traditional analytical architecture.	Understand	BTL 2
22	Analyse what are the challenges in data.	Analyze	BTL 4
23	Differentiate big data vs traditional data.	Apply	BTL 3
24	Summarize the data privacy.	Evaluate	BTL 5

PART-B

Q.No.	Question	Competence	Level
1	Define Bigdata? Describe the main features of a big data in detail. (13)	Remember	BTL 1
2	(i) List the main features of big data. (4) (ii) Describe the benefits of big data processor. (9)	Remember	BTL 1

3	Formulate what is the risk in handling big data. (13)	Create	BTL 6
4	Describe the nature of big data representation (13)	Remember	BTL 1
5	Describe the nature of big data and its application. (13)	Analyze	BTL 4
6	Analyse the challenges faced by traditional system. (13)	Apply	BTL 3
7	Describe in detail the analysis tools and reporting tools used in Big-data. (13)	Understand	BTL 2
8	(i) What is a analytical data set ? (3) (ii) Explain the types of analytical data set (10)	Analyze	BTL 4
9	(i) Summarize what is web data. (6) (ii) Discuss how bigdata are mixed with traditional one. (7)	Understand	BTL 2
10	Describe how big data are effectively filtered. (13)	Remember	BTL 1
11	Analyse the Evolution Tools and Method in big data. (13)	Apply	BTL 3
12	Illustrate the evolution of analytic process. (13)	Evaluate	BTL 5
13	Summarize how the analytical scalability is handled in big data. (13)	Understand	BTL 2
14	Differentiate the Analysis and reporting methods and tools (13)	Understand	BTL 2
15	Summarize the importance of analytical sandbox in detail. (13)	Evaluate	BTL 5
16	Illustrate the Evolution of Analytical Scalability. (13)	Apply	BTL 3
17	List out few modern data analytic tools and explain. (13)	Remember	BTL 1
	PART – C		
1.	Distinguish between big data and traditional data. (15)	Create	BTL 6
2.	Explain about the challenges of convolutional system. (15)	Evaluate	BTL 5
3.	Explain the following A)Basic Steps of web analytics process . (5) B) Web analytics technologies. (5) C) Web analytics data sources. (5)	Evaluate	BTL 5
4.	List out the Big data applications. (15)	Create	BTL 6
5.	Explain in detail about evolution of analytic process and the role of analytical sandbox. (15)	Evaluate	BTL 5
	UNIT II HADOOP FRAMEWORK		
	Distributed File Systems - Large-Scale FileSystem Organization – HDFS concepts - MapReduceExecution, Algorithms using MapReduce, Matrix-Vector Multiplication – Hadoop YARN		
	PART – A		
Q.No	**Question**	**Competence**	**Level**
1	Define distributed file system.	Remember	BTL 1
2	List out data charateristics.	Apply	BTL 3
3	What is hadoop.	Create	BTL 6
4	List out some of the Hadoop distributions.	Understand	BTL 2
5	List out the terms related to HDFS.	Remember	BTL 1
6	Define data block.	Remember	BTL 1
7	When to use HDFS.	Analyze	BTL 4

8	List out the objectives and assumptions of HDFS.	Evaluate	BTL 5
9	Define name node and data node.	Remember	BTL 1
10	What are the benefits of using HDFS.	Create	BTL 6
11	List out the limitations of HDFS.	Understand	BTL 2
12	Why is there a need to divide the file into blocks?	Remember	BTL 1
13	What are the five core elements of bigdata organized by HDFS services?	Apply	BTL 3
14	What are HDFS daemons?	Analyze	BTL 4
15	List out the features of HDFS	Understand	BTL 2
16	Define metadata disk failures.	Remember	BTL 1
17	Define rack awareness.	Understand	BTL 2
18	What are the operations performed by map and reducer?	Apply	BTL 3
19	Write down the different stages in map reducer technique.	Analyze	BTL 4
20	List out the major components of Hadoop YARN.	Evaluate	BTL 5
21	How is an application submitted in Hadoop YARN?	Understand	BTL 2
22	Define rack awareness.	Apply	BTL 3
23	What are the salient features of Hadoop YARN?	Analyze	BTL 4
24	Difference between Hadoop YARN and map reduce.	Evaluate	BTL 5

PART-B

Q.No.	Question	Competence	Level
1	Explain about distributed file system. (13)	Remember	BTL 1
2	Discuss in detail about data organization. (13)	Understand	BTL 2
3	Difference between local file system and distributed file system. (13)	Evaluate	BTL 5
4	List out and explain HDFS key concepts. (13)	Understand	BTL 2
5	What is HDFS? List out the components of HDFS and its functions.(13)	Analyze	BTL 4
6	Explain in detail the advantages of HDFS. (13)	Analyze	BTL 4
7	Write down the steps involved in file read in HDFS architecture. (13)	Remember	BTL 1
8	Explain in detail about the workflow of map reduce and how to handle machine failure? (13)	Apply	BTL 3
9	List out the various terminologies used in Hadoop map reduce. (13)	Remember	BTL 1
10	Write in detail about map reduce functions. (13)	Analyze	BTL 4
11	Describe in detail about different stages of Map Reduce algorithms. (13)	Remember	BTL 1
12	Write an algorithm for matrix multiplication. A) Map function B) Reducer function (13)	Understand	BTL 2
13	List out the two important steps in map reduce and visualize matrix multiplication with an example. (13)	Apply	BTL 3
14	Write down the steps of workflow of Hadoop YARN. (13)	Create	BTL 6
15	List out some features of YARN in detail. (13)	Apply	BTL 3
16	Explain about Hadoop YARN architecture. (13)	Understand	BTL 2
17	Describe in detail about YARN and its components. (13)	Evaluate	BTL5

PART – C			
1	Describe in detail about HDFS architecture. (15)	Evaluate	BTL 5
2	Describe the following: A) Anatomy of file read. (5) B) Anatomy of file write. (5) C) Replica Placement Strategy. (5)	Create	BTL 6
3	How Hadoop map reduce works? (15)	Evaluate	BTL 5
4	Illustrate matrix vector multiplication. (15)	Create	BTL 6
5	Write down the steps involved in running a job using Hadoop YARN. (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.	Apply	BTL 3
3	Generalize support-vector machines.	Create	BTL 6
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.	Analyze	BTL 4
8	Assess the importance of classification in data analysis.	Evaluate	BTL 5
9	What is Kernel?	Remember	BTL 1
10	Generalize rule mining.	Create	BTL 6
11	Give the types of kernels.	Understand	BTL 2
12	What is Multiple Linear Regression?	Remember	BTL 1
13	Illustrate what is predictive analysis.	Apply	BTL 3
14	Differentiate regression and correlation.	Analyze	BTL 4
15	What is clustering?	Understand	BTL 2
16	List the types of regression.	Remember	BTL 1
17	What is SVM?	Understand	BTL 2
18	Relate classification and clustering.	Apply	BTL 3
19	Point out the importance of clustering in data analysis.	Analyze	BTL 4
20	Summarize what is R.	Evaluate	BTL 5
21	Evaluate density based clustering..	Evaluate	BTL 5
22	What is the grid based clustering?	Analyze	BTL 4
23	List the types of hierarchical clustering.	Understand	BTL 2

24	Show partitioning methods in clustering.	Apply	BTL 3
PART-B			
Q.No.	**Question**	**Competence**	**Level**
1	Define regression? List out the types of regression and the purpose of using Regression Modeling in Data Analysis. (13)	Remember	BTL 1
2	Describe in detail about Multivariate Analysis. (13)	Evaluate	BTL 5
3	Describe about SVM and some of its applications. (13)	Remember	BTL 1
4	Explain about kernel methods in detail. (13)	Analyze	BTL 4
5	Write a short note on the following. (i) Density based methods (3) (ii) Grid based methods (5) (iii) Model based clustering methods (5)	Understand	BTL 2
6	Explain in detail about the Rule Mining. (13)	Understand	BTL 2
7	(i) Examine clustering in data analysis. (3) (ii) Illustrate density based and Grid based clustering in detail (10)	Apply	BTL 3
8	Describe how clustering is used in high dimensional data. (13)	Remember	BTL1
9	Illustrate the approaches of clustering. (13)	Apply	BTL 3
10	Discuss model based clustering and high dimensional clustering in detail. (13)	Understand	BTL 2
11	Explain in detail about association rule mining. (13)	Remember	BTL 1
12	Explain about Predictive analysis with some applications (13)	Analyze	BTL 4
13	What is prediction? Generalize how prediction helps in data analysis. (13)	Create	BTL 6
14	Analyze the different hierarchical clustering techniques (13)	Apply	BTL 3
15	Discuss the model based clustering. (13)	Understand	BTL2
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	Createa 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	Explain hierarchical clustering and it types in detail.(15)	Evaluate	BTL 5
4	Explain K-means clustering with an example. (15)	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 streaming?	Remember	BTL 1
3	How does the streaming data pipeline works.	Analyze	BTL 4
4	Define Stream processor.	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	List out the three basic components in strreaming data architecture.	Understand	BTL 2
8	Analyse the statement “Filtering a Data Stream”.	Apply	BTL 3
9	Give the applications of data streams.	Understand	BTL2
10	Define Real-Time Analysis.	Understand	BTL 2
11	Show how to deal with mining time series.	Apply	BTL 3
12	Define Time Series Data.	Remember	BTL 1
13	Show what examples you can find for stream sources.	Apply	BTL 3
14	What is called Data Stream Mining?	Understand	BTL 2
15	Difference between RTAP (real time analytics platform) and RTSA (real time sentiment analysis)?	Analyze	BTL 4
16	Analyse why do we need Real Time Analytics Platform (RTAP).	Analyze	BTL 4
17	Evaluate the key component of real time analytics.	Evaluate	BTL 5
18	Evaluate the challenges of sentiment analysis work.	Evaluate	BTL 5
19	Generalize real time data.	Create	BTL 6
20	List out few examples of sentiment analysis.	Create	BTL 6
21	Evaluate the importance of social media analytics.	Evaluate	BTL 5
22	Give the reasons why do we need RTAP.	Understand	BTL 2
23	Analyse prediction and forecasting.	Analyse	BTL 4
24	Illustrate what is time series data.	Apply	BTL 3

PART-B

Q.No.	Question	Competence	Level
1	(i) Define data stream. Describe the Data Stream model with a neat architecture diagram. (13)	Remember	BTL 1
2	Illustrate briefly about the sources of data stream. (13)	Apply	BTL 3
3	Write a short note on the following: (i) Mining Data Streams (7) (ii) Mining Time-series data (8)	Analyze	BTL 4

4	Summarize the importance of Stream Data Model and its architecture. (13)	Understand	BTL 2
5	Analyze and write a short note on Aurora system model. (13)	Analyze	BTL 4
6	Explain Sampling data in a stream. (13)	Remember	BTL 1
7	Explain in detail about Real Time Analytics Platform. (13)	Understand	BTL 2
8	Generalize how mining is done with data streams. (13)	Create	BTL 6
9	Describe how data analysis used in Stock Market Prediction. (13)	Understand	BTL 2
10	Describe in detail about mining and time series data. (13)	Remember	BTL 1
11	Illustrate how would you describe the various windowing approach to data stream mining. (13)	Apply	BTL 3
12	(i) List the methods for analyzing time series data. (7) (ii) What are the several types of motivation and data analysis available for time series? (6)	Remember	BTL 1
13	Demonstrate case study on real time sentiment analysis for Live Social Feeds. (13)	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	Briefly explain time series analysis on forecasting pandemic spread, diagnosis and medication planning in healthcare applications. (13)	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 Weather Forecasting Predictions. (13)	Understand	BTL 2
	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	Describe the following: (i) Need of Real-time Sentiment Analysis. (7) (ii) Steps involved in Real-time Sentiment Analysis. (8)	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. – 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	Describe few key features of NoSQL.	Understand	BTL 2

3	Deduce the components of Hadoop framework.	Evaluate	BTL 5
4	Differentiate between SQL and NoSQL.	Analyze	BTL 4
5	What is the advantage of NoSQL?	Remember	BTL 1
6	Give the disadvantages of NoSQL.	Understand	BTL 2
7	What is HBase?	Remember	BTL 1
8	Show the advantage of Cassandra.	Apply	BTL 3
9	Who is generating big data and what are the ecosystem projects used for processing?	Create	BTL 6
10	Illustrate the difference between HBase and Hive.	Apply	BTL 3
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.	Apply	BTL 3
15	Classify the usage of Pig, Hive and HBase.	Analyze	BTL 4
16	Give the features of Hive.	Understand	BTL 2
17	Define Pig, Hive and HBase	Analyze	BTL 4
18	What is hive in Big Data?	Remember	BTL 1
19	What is Cassandra Client.	Evaluate	BTL 5
20	List out the types of built-in operator in HIVE.	Create	BTL 6
21	Differentiate between HIVE internal tables and external tables.	Apply	BTL 3
22	Defie the methods used in class HTABLE.	Analyse	BTL 4
23	Evaluate the importance of Cassandra.	Evaluate	BTL 5
24	Define Hadoop Streaming.	Understand	BTL 2

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)	Remember	BTL 1
2	(i) Illustrate in detail about Hive data manipulation, queries, and data types. (8) (ii) Illustrate data definition in Hive. (5)	Apply	BTL 3
3	Describe the system architecture and components of Hive and Hadoop. (13)	Remember	BTL 1
4	Explain briefly on aggregate data models with cluster and order relationship. (13)	Analyze	BTL 4
5	Generalize two types of data storage medium in Hbase. (13)	Create	BTL 6
6	(i) Describe about HBase in detail. (7) (ii) Explain Hbase clients in detail. (6)	Remember	BTL 1
7	(i) Analyse how Cassandra is integrated with Hadoop. (6) (ii) Explain the tools related to Hadoop. (7)	Apply	BTL 3
8	Summarize briefly on Hbase architecture with neat diagram (13)	Understand	BTL 2
9	Quote short notes on (i) Features of Hive. (7) (ii) Limitations of hive. (6)	Remember	BTL 1
10	Discuss about Cassandra clients. (13)	Understand	BTL 2
11	Compare and Contrast the Hbase and Hive. (13)	Analyze	BTL 4

12	(i) Explain about Pig in detail. (7) (ii) What is invoking a Grunt shell? (6)	Analyze	BTL 4
13	Describe about Pig data model in detail with neat diagram. (13)	Understand	BTL 2
14	Explain how to develop and test pig scripts for data processing. (13)	Evaluate	BTL 5
15	Difference between Apache Hive and Apache Hbase (13)	Understand	BTL 2
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: (15) 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)	Create	BTL 6
5	Explain the features of Apache Cassandra? (5) Explain in detail about Cassandra data model. (10)	Evaluate	BTL5