Shivaji University, Kolhapur
Question Bank for Mar 2022 (Summer) Examination

Subject Code: 84719, Subject Name: Big Data Analytics

Sr. No.	Question	A	B	C	D
1	What are the main components of Hadoop?	MapReduce	HDFS	YARN	All of the above
2	The Big data analytics work on the unstructured data, where no specific pattern of the data is defined.	True	False	Can't Say	None of the above
3	Identify the incorrect big data Technologies.	Apache Pytorch	Apache Kafka	Apache Hadoop	Apache Spark
4	Identify among the options below which is general-purpose computing model and runtime system for Distributed Data Analytics.	HDFS	MapReduce	Oozie	All of the above
5	Big data analysis does the following except?	Spreads data	Analyze data	Organizes data	Collect data
6	What is NOT a characteristic of big data?	Volume	Variety	Vision	Velocity
7	Pig is a Hadoop-based open-source platform for analyzing the large-scale datasets via its own SQL-like language _____	Pig Latin	Pig German	Pig Roman	Pig Italian
8	The key aspect of the MapReduce algorithm is that if every Map and Reduce is independent of all other ongoing Maps and Reduces in the network, the operation will run in _____ keys and lists of data.	series on same	series on different	parallel on different	parallel on same
9	In Hadoop MapReduce, _____ is a Java class that comes with several methods to retrieve key and values by iterating them among the data splits.	Mapper	RecordReader	Reporter	RecordCollector
10	Which of the following scenarios	TaskTracker failure	JobTracker failure	NameNode failure	DataNode failure

	makes HDFS unavailable?				
11	Hadoop MapReduce is a popular_____ for easily written applications. It processes vast amounts of data (multi-terabyte datasets) in parallel on large clusters (thousands of nodes).	Spring framework	Java framework	Django framework	Web framework
12	Which is not a way to link R and Hadoop?	RHIPE	RHadoop	Hadoop streaming	RHDFS
13	The RHIPE package uses the _____ technique to perform data analytics over Big Data.	Divide and recombine	Divide and conquer	Integrate and recombine	None of the above
14	_____ phase of the data analytics lifecycle usually takes the longest time.	Phase 2: Data Preparation	Phase 3: Model Planning	Phase 4: Model Building	Phase 5: Communicate Results
15	The data analytics project life cycle stages in correct sequence are _____	Identifying the problem>designing the requirements> pre-processing data>performing analytics over data> visualizing data	Identifying the problem>designing the requirements> >performing analytics over data> pre-processing data > visualizing data	Identifying the problem > performing analytics over data >designing the requirements> pre-processing data> visualizing data	Identifying the problem> visualizing data >designing the requirements> pre-processing data>performing analytics over data
16	Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods? 1. Both methods can be used for classification task 2. Random Forest is used for classification whereas Gradient Boosting is used for regression task 3. Random Forest is used for regression whereas Gradient	1	2	2 and 3	1 and 4

	Boosting is used for Classification task 4. Both methods can be used for regression task				
17	In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest? 1. Individual tree is built on a subset of the features 2. Individual tree is built on all the features 3. Individual tree is built on a subset of observation 4. Individual tree is built on full set of observations	1 and 3	1 and 4	2 and 3	2 and 4
18	The primary Machine Learning API for Spark is now the _____ based API	Dataframe	Dataset	RDD	All of the above
19	Which of the following is a module for Structured data processing?	GraphX	MLib	SparkSQL	Spark R
20	SparkSQL translates commands into codes. These codes are processed by _____	Driver Nodes	Executor nodes	Cluster Manager	None of the above
21	SparkSQL plays the main role in the optimization of queries.	True	False	Can't Say	None is correct
22	Which of the following is not a SparkSQL query execution phases?	Analysis	Logical Optimization	Execution	Physical Planning
23	What is action in Spark RDD?	The ways to send result from executors to the driver	Takes RDD as input and produces one or more RDD as output.	Creates one or many new RDDs	All of the above

24	Which of the following is true about narrow transformation?	The data required to compute resides on multiple partitions.	The data required to compute resides on the single partition.	Both	None of the above
25	_____ is a distributed machine learning framework on top of Spark.	MLib	GraphX	Spark Streaming	RDDs
26	Which of following component of Spark runtime architecture provides resources to execute a task?	Cluster manager	Worker nodes	Driver program	Spark context
27	Among the following option identify the one which is not a type of learning.	Unsupervised learning	Reinforcement learning	Supervised learning	Semi unsupervised learning
28	Identify the type of learning in which labeled training data is used.	Unsupervised learning	Reinforcement learning	Semi unsupervised learning	Supervised learning
29	Machine learning is a subset of which of the following?	Artificial Intelligence	Deep Learning	Data Learning	None of the above
30	Which of the following machine learning techniques helps in detecting the outliers in data?	Classification	Clustering	Anomaly Detection	All of the above
31	Which of the following are common classes of problems in machine learning?	Regression	Classification	Clustering	All of the above
32	What is content based recommendation system?	Tries to recommend items based on profile built from their preferences	Similarity among items	Similarity among users buying, watching, or enjoying something	All of the above
33	Machine Learning is a field of AI consisting of learning algorithms that _____	At executing some task	Over time with experience	Improve their performance	All of the above
34	Which of the following machine learning algorithm is based upon the idea of bagging?	Decision tree	Random forest	Classification	Regression

35	Among the following options identify the one which is false regarding regression.	It is used for the prediction	It is used for interpretation	It relates inputs to outputs	It discovers casual relationships

UNIT – I

1. Define Big Data? Explain the Characteristics / V's of Bigdata?
2. Write a note on: Drivers for Big Data.
3. Explain different applications of Big Data.
4. Write a note on: Data Privacy Protection.
5. With neat diagram depict the Product Knowledge Hub in Big Data?
6. Write a short note on Location Based Services in Big Data?
7. Explain the architectural components of Big Data?
8. Explain Real-time Adaptive Analytics and Decision engine?
9. Explain Massively Parallel Processing (MPP) platforms.
10. Explain Unstructured Data Analytics and Reporting.

UNIT – II

1. Explain the features of R Language?
2. Explain different phases of MapReduce with an example?
3. What is HDFS? Explain the features of HDFS?
4. Explain the HDFS and MapReduce architecture.
5. List and explain different components of Hadoop.
6. Explain in detail the stages of Hadoop MapReduce data processing.
7. Explain in detail the dataflow of MapReduce with diagram.
8. Explain the limitations of MapReduce.
9. Explain the data mining techniques which are used to perform data modeling in R.
10. Mention different Hadoop installation modes? Explain each of them.

UNIT – III

1. Explain the architecture of RHIPE.
2. Explain RHadoop in detail.
3. Explain the architecture of RHadoop
4. Explain the working of RHadoop with example?
5. Explain the hstable reader function for Hadoop streaming.
6. Explain the hskeyval reader function for Hadoop streaming.
7. Explain the Hadoop streaming components?
8. Explain the format of Hadoop Streaming commands with each line?

UNIT – IV

1. Explain Data Analytics project life cycle stages.
2. Explain how data analytics problem for calculating the frequency of stock market changes can be solved using MapReduce.
3. Write a case study for predicting the auction sales price of heavy equipment to create a blue book for bulldozers.
4. Explain Poisson-approximation resampling technique on the Map of the MapReduce task.
5. How will data analytics help to identify the category of a web page of a website, which may categorize popularity wise as high, medium, or low (regular), based on the visit count of the pages.
6. Write steps to build and run the MapReduce algorithm with R and Hadoop integration for web page categorization problem.
7. Explain pre-processing and performing analytics over any data?
8. Explain how MapReduce problem is designed for computing the frequency of stock market changes.

UNIT – V

1. What is Resilient Distributed Dataset (RDD)? Explain transformations and actions in RDD. Explain RDD operations in brief?
2. Why Spark is preferred over Hadoop? Explain the limitations of Hadoop?
3. Explain how Spark overcomes the limitations of Hadoop.
4. Briefly explain the core components in Spark.
5. Explain the architecture of Spark.
6. What is SparkContext in Apache Spark?
7. What is a Directed acyclic graphs (DAG) in Spark, and how does it work?
8. What are Spark DataFrames? Why do we use them in Spark?
9. Explain Apache Spark RDD Operations in detail.
10. What are different types of RDD transformation? Explain functions in RDD transformation.
11. What are RDD actions? When they are used? Explain Spark actions.
12. What are the deployment modes in Spark? What is difference between client and cluster mode deployment?
13. What are the components of Spark architecture?
14. What is Spark core? What are the various functions of Spark core? Which is a component on the top of Spark core?
15. What are the components of Spark Streaming? What is Spark Streaming used for?

UNIT – VI

1. What is machine learning? Explain types of machine-learning algorithms.
2. Explain Supervised Machine Learning Algorithm.
3. Explain how Linear regression is performed using with R and Hadoop?
4. Explain how logistic regression is performed using with R and Hadoop?
5. Explain Unsupervised Machine Learning Algorithm.
6. Explain steps to performing clustering with R and Hadoop.
7. Explain Steps to generate recommendations in R.
8. What is recommendation algorithm? Explain two different types of Recommendations Algorithms.
9. How do you create a recommendation algorithm with R and Hadoop?
10. How one can use R and Hadoop together to generate recommendations from big datasets?