1. Why are three replicas placed on only two racks in the HDFS default Block Placement Strategy?

1. To cut the inter-rack write traffic this generally improves write performance
2. The chance of rack failure is far less than the chance of node failure
3. It does not reduce the aggregate network bandwidth used when reading data since a block is placed in only two unique racks rather than three.

2. Which of the following are correct about Rebalancing?

1. Main purpose is uniform disk utilization in the cluster
2. Usually runs when the new data nodes are added
3. To place different blocks of a file in the same data node
4. Rebalancing helps in blocks replication

3 .What are different techniques used by HDFS to prevent failure of master node

1. Active and Stand-By nodes and Leader election
2. b.Splitting the files to multiple blocks.
3. By increasing the replication factor

4 .Rationale behind dividing a file into multiple chunks

1. To prevent Data Loss
2. To make the system fault tolerant.
3. c.To reduce latency

5. Hadoop is written in

1. C
2. C++
3. java
4. Python

6 .Which of the following are the features of HDFS?

1. High availability
2. Scalable
3. Fault Tolerant
4. Replicable
5. All of the above

7. Which of the following is true about metadata?

1. Metadata shows the structure of HDFS directories/files
2. Metadata contain information like number of blocks, their location, replicas
3. FsImage&EditLogs are metadata files
4. All of the above

8. Which among the following are the duties of the Data Nodes

1. Manage file system namespace
2. Stores meta-data
3. Regulates client’s access to files
4. Perform read-write operation as per request for the clients

9. The name node knows that the data node is active using a mechanism known as

1. Active Signal
2. Active Pulse
3. health Signal
4. Heartbeat

10. The HDFS architecture in Hadoop has been inspired by originated as

1. Google Distributed File system
2. Yahoo Distributed File System
3. Network File System
4. None of the above

11. Which of the following is a framework-specific entity that negotiates resources from the Resource Manager?

1. Node Manager
2. Resource Manager
3. Application Master
4. All of the above

12. Apache Hadoop YARN stands for

1. Yet Another Reserve Negotiator
2. Yet Another Resource Network
3. Yet Another Resource Negotiator
4. None of the above

13. What is Yarn?

1. Storage Component
2. Batch Processing Engine
3. Resource Management Component
4. Security Component

14. Which is the slave daemon of Yarn?

1. Node Manager
2. Resource Manager
3. Application Master

15. From the following sequence of operations, Select the correct order of details of Initiating Jobs/Apps in YARN.

1. Application Master on boot-up registers with the Resource Manager.
2. Application Master negotiates appropriate resource containers.
3. ResourceManager negotiates a container to start  
   the ApplicationMaster
4. Client Program submits the app.
5. Once the application is complete, all the necessary work has been finished, the application Master deregisters with the Resource manager and shuts down, allowing its own container to be repurposed.
6. During the app execution, client that submitted the app communicates directly with the ApplicationMaster to get status, progress  
   updates etc.

1. 4,3,1,6,2,5
2. 4,3,1,2,6,5
3. 4,3,1,5,2,6
4. 4,3,1,6,5,2

16. Which of these describes best the key problem that Big Data Engineering is trying to solve?

1. Finding patterns from large amounts of data
2. Efficiently storing large amounts of data
3. Creating large amounts of data
4. Efficiently storing and processing large amounts of data

17. Fault Tolerance in RDD is achieved using?

1. Immutable nature of RDD
2. DAG(Directed Acyclic Graph)
3. lazy-evaluation
4. none of these

18.  Which of the following is/are **FALSE** about Block Placement Strategy?

1. One replica on a local node.
2. Second and third replicas on different nodes but in a remote rack.
3. Second and third replicas on different nodes but in the same rack.
4. Additional replicas, if any, get placed on nearest free machines.
5. There cannot be less than 2 replicas.
6. Client always reads from the nearest replica, when locality is important.

19. Which of the following is/are master daemons of Hadoop?

1. Name Node
2. Secondary Name Node
3. Application master
4. Resource Manager
5. Data Node

20. Choose all the correct operations about Spark.

1. Actions are lazy and transformations lead to results
2. Transformations are lazy and actions lead to results.
3. Spark is a fast and expressive cluster computing system
4. The primary abstraction for spark is RDD.
5. RDDs can change once constructed.

21. Which of the following are true about Spark?

1. Spark Shares memory RAM instead of a hard disk of multiple nodes.
2. Spark enables distributed data structures
3. Write programs in form of transformation on distributed datasets
4. Spark does not build the RDD using parallel transformations

22 .Mark all applicable characteristics about YARN.

1. Multi-tenancy
2. Access Control
3. Improved Cluster Utilization
4. Application Awareness
5. Context Awareness
6. Improved Cluster De utilization
7. Interference in data storage

23 .Which of the following statements are true?

1. YARN’s distributed resource management is programmed on MR2.
2. MR2 uses YARN for distributed computation.
3. YARN uses HDFS for its own storage.
4. HDFS must request resources through YARN.

24 .Mark all options that are applicable to Hadoop.

1. More distributed than parallel
2. Programs and data have similar life cycles
3. Aims to be programmer-friendly
4. Aims to minimize talking between nodes

25. Which of the following are true about the following Statement??

Distributed computing while dealing with its ability for failure should have the following Characteristics.

1. Performance should not degrade gracefully with partial failure of the system.
2. Loss of data is sometimes okay whenever there is a failure.
3. Partial failures should not affect outcome.
4. Recovered components should be able to rejoin the system without needing a reboot.

26 .Which of the following are true about SparkSQL ?

1. Spark SQL is a Spark module for structured data processing.
2. A DataFrame is a Dataset organized into named columns.
3. Spark DataFrames are not as optimized as RDDs
4. Spark DataFrames can be used to run SQL Queries.
5. Spark DataFrames are not fit into memory and only the tabular data is fit into memory.

 27 .Arrange the following in the order of their execution.

1. Resource Manager Contacts related Node Manager
2. Resource Manager allocates a Container
3. Client submits an application to Resource Manager
4. Node Manager launches Container for running Application Master

1. 4, 1, 3, 2
2. 1, 3, 2, 4
3. 3, 2, 1, 4
4. 3, 4, 2, 1

28 .What are the functionalities of Spark core

1. Memory management
2. Fault management
3. Cluster management

|  |  |
| --- | --- |
|  |  |

29. Data in \_\_\_\_\_\_\_\_\_\_\_ bytes size is called Big Data.

a. Tera

b. Giga

c. Peta

d. Meta

30. Transaction data of the bank is?

1. Structured data
2. Unstructured data
3. Both A and B
4. None of these

31. The overall percentage of the world's total data that has been created just within the past two years is?

* 1. 80%
  2. 86%
  3. 90%
  4. 98%

32. By 2027, the volume of data produced digitally will reach:

* 1. TB
  2. YB
  3. ZB
  4. EB

33. For drawing insights for business what are needed?

* 1. Collecting the data
  2. Storing the data
  3. Analyszing the data
  4. All of the above

34. \_\_\_\_\_\_\_\_\_\_\_\_is an open-source framework for storing data and running application on clusters of commodity hardware.

* 1. HDFS
  2. Hadoop
  3. MapReduce
  4. Cloud

35. Data Set is the:

* 1. Tweets stored in a flat file
  2. A collection of image files in a directory
  3. An extract of rows from a database table stored in a CSV formatted file
  4. All of the above

36. Data variety refers to:

* 1. Multiple schemas
  2. Multiple formats and types of data
  3. Multiple Data Models
  4. None of these

37. Big data analysis does the following except:

* 1. Collects data
  2. Spreads data
  3. Organizes data
  4. Analyzes data

38. The new source of big data that will trigger a Big Data revolution in the years to come is:

* 1. Business transactions
  2. Social media
  3. Transactional data and sensor data
  4. RDBMS

39. The unit of data that flows through a Flume agent is:

* 1. Log
  2. Row
  3. Event
  4. Record

40. The word 'Big Data' was coined in the year:

* 1. 2000
  2. 1978
  3. 1996
  4. 1998

41. The feature of big data that refers to the quality of the stored data is \_\_\_\_\_\_.

* 1. Variety
  2. Volume
  3. Variability
  4. Veracity

42. Input to the \_\_\_\_\_\_\_\_\_\_ is the sorted output of the mappers.

* 1. Reducer
  2. Mapper
  3. Shuffle
  4. None of these

43. A \_\_\_\_\_\_\_\_ serves as the master and there is only one NameNode per cluster.

* 1. Data node
  2. Name node
  3. Data block
  4. Replication

44. Apache Kafka is an open-source platform that was created by?

* 1. LinkedIn
  2. Facebook
  3. Google
  4. IBM

45. What are the main components of Big Data?

* 1. MapReduce
  2. HDFS
  3. YARN
  4. All of the above

46. HDFS works in a \_\_\_\_\_\_\_\_\_\_ fashion.

* 1. master-worker
  2. master-slave
  3. worker/slave
  4. All of the above

47. HDFS works in a \_\_\_\_\_\_\_\_\_\_ fashion.

* 1. HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file
  2. HDFS is suitable for storing data related to applications requiring low latency data access
  3. HDFS is suitable for storing data related to applications requiring high latency data access
  4. None of these

48. HDFS provides a command line interface called \_\_\_\_\_\_\_\_\_\_ used to interact with HDFS.

* 1. HDFS Shell
  2. FB Shell
  3. DFSA Shell
  4. None

49. Random Forest has \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ as base learning models

1. Multiple Decision Trees
2. Bagging
3. Entropy
4. None of These

50. \_\_\_\_\_\_\_\_\_\_\_\_ helps to improve Machine Learning results by combining several models.

1. Machine Learning
2. Bagging
3. Entropy
4. Ensemble Learning
5. What are the main goals of AI?
6. To Create Expert Systems
7. To Implement Human Intelligence in Machines
8. Both A and B
9. None of the Above
10. Which of the following is not the type of AI?
11. Reactive machines
12. Unlimited memory
13. Theory of mind
14. Self-awareness
15. Which of the following is an application of AI?
16. Gaming
17. Expert Systems
18. Vision Systems
19. All of the above
20. What is Artificial intelligence?
21. Putting your intelligence into Computer
22. Programming with your own intelligence
23. Making a Machine intelligent
24. Playing a Game
25. The Science And Engineering Of Making Intelligent Machine Is Called?.
26. Ai
27. Machine Learning
28. Deep Learning
29. Data Mining
30. Google Recommendations Are Called?
31. Ai
32. Machine Learning
33. Data Mining
34. Data Where Housing
35. Clustering Is A Problem Of?
36. Regression
37. Classification
38. Un-Supervised Learning
39. Reinforcement

1. The Process Of Deriving Meaningful Information From A Lot Data Is Called?.
2. Extraction
3. Data Mining
4. Data Where Housing
5. Abstraction
6. Nlp Stands For?
7. Neural Language Processing
8. Natural Language Processing
9. Network Language Processing
10. None Linear Problem
11. The Process Of Splitting The Whole Data Into Smaller Chunks Is Called?
12. Splitting
13. Tokenization
14. Sorting
15. Division
16. Logistic regression is a ……….. regression technique that is used to model data having a ……….. outcome.
17. linear, numeric
18. linear, binary
19. nonlinear, numeric
20. nonlinear, binary
21. To Overcome The Limitation Of Stemming —— Is Used?
22. Lemmatization
23. Normalization
24. Erosion
25. Spanning
26. A Set Of Commonly Used Words In Any Language, Not Just English?.
27. Higher Language
28. Stop Words
29. Predefine Words
30. Common Words
31. Which Programming Language Is Best Choice For Development Of Ai?
32. Python
33. Java
34. Dot .Net
35. Mat-Lab
36. There Are How Many Steps In Machine Learning To Solve A Problem?.
37. 3 Steps
38. 5 Steps
39. 7 Steps
40. 9 Steps
41. In Supervised Learning There Is Given.
42. Both Inputs & Outputs
43. Only Inputs
44. Only Outputs
45. None
46. In Unsupervised Learning There Is Given.
47. Both Inputs & Outputs
48. Only Inputs
49. Only Outputs
50. None
51. A term used to describe the case when the independent variables in a multiple regression model are correlated is
52. regression
53. correlation
54. multicollinearity

1. In Unsupervised Learning Type Of Data Is.
2. Labelled Data
3. Unlabeled Data
4. Both
5. None
6. Which Algorithm Is Example Of Supervised Learning?.
7. K-Nearest Neighbours
8. K-Means
9. Both
10. None
11. Which Algorithm Is Example Of Unsupervised Learning?.
12. K-Means
13. K-Nearest Neighbours
14. Both
15. None
16. The First Al Chat-Bot Name Which Introduced In 1961?.
17. Alexa
18. Siri
19. Eliza
20. Cortana

73. Unsupervised Learning is used for solving \_\_\_\_\_\_\_\_\_ problems

1. Clustering
2. Regression
3. Classification
4. None Of Them

74. Supervised Learning is used for solving \_\_\_\_\_\_\_\_\_ problems

1. Clustering
2. Regression
3. Classification
4. None Of Them

#### 75. Bootstrapping allows us to

1. choose the same training instance several times
2. choose the same test set instance several times
3. build models with alternative subsets of the training data several times

d) test a model with alternative subsets of the test data several times

76. Training is under \_\_\_\_\_\_\_\_\_\_\_ in Unsupervised Learning

1. External Supervision
2. Internal Supervision
3. No Supervision
4. None Of Them

77. In Supervised Learning Output Is.

1. Known
2. Unknown
3. Both A & B
4. None Of Them

78. Output is \_\_\_\_\_\_\_\_\_\_\_in Unsupervised Learning

1. Known
2. Unknown
3. Both A & B
4. Understand Patterns And Discover Output

79. What Is Example Of Regression?

1. Predict Stock Market Price
2. Spam And Non-Spam Emails
3. Transactions That Are Fraud In Nature
4. None Of Them

80. What Is Example Of Classification?

1. Predict Stock Market Price
2. Spam And Non-Spam Emails
3. Transactions That Are Fraud In Nature
4. None Of Them

81. What Is Example Of Clustering?

1. Predict Stock Market Price
2. Spam And Non-Spam Emails
3. Transactions That Are Fraud In Nature
4. None Of Them

82. Which Is Regression Algorithm?

1. Linear Regression
2. Logistic Regression
3. Both A & B
4. None Of Them

83. Which Is Classification Algorithm?

1. Linear Regression
2. Logistic Regression
3. Both A & B
4. None Of Them

84. Output of Regression Model Is.

1. Categorical Quantity
2. Continuous Quantity
3. Both A & B
4. None Of Them

#### 85. Which of the following algorithms is an example of the ensemble learning algorithm?

1. Decision Tree
2. Neural Networks
3. Random Forest
4. Support Vector Machine

86. Decision Tree Is Which Type Of Machine Learning Algorithm.

1. Supervised Learning
2. Q-Learning
3. Unsupervised Learning
4. Semi Supervised Learning

87. in Decision Tree Root Node Represent

1. Input
2. Output
3. Predictor Variable
4. None Of Them

88. in Decision Tree Each Leaf Node Represent.

1. Input
2. Output
3. Predictor Variable
4. None Of Them

#### 89. Exploration of numerical data can be best done using

1. Boxplots
2. Histograms
3. Scatter plot
4. All the above

90. If You Want To Avoid Over-Fitting You Use.

1. K-Means
2. Random Forest
3. Cnn
4. None Of Them

91.A Hospital creates two groups of 10 volunteers each, to conduct a drug trial/experiment. These two groups were then given the drug in a controlled safe environment. All their health parameters were then monitored and observed. Upon completion of the tests, the volunteers are were grouped into 4 labels : FEASIBLE, NOT FEASIBLE, PREVENTABLE, NOT PREVENTABLE. You are now given the task of automating the task of labelling using the health parameter. Which method would you choose to automate ?

1. Supervised method
2. Unsupervised method
3. Semi-supervised method
4. Cannot be solved

93. Machine Learning (ML) is the study of computer algorithms that learn and improve automatically from experience (historical data) without being explicitly programmed.

1. True
2. False

94. Given the transactions and minimum support (s) = 60%, how many frequent 3-itemsets are there?

| **Tid** | **Items bought** |
| --- | --- |
| 10 | Book, Eraser, Scale |
| 20 | Book, Pencils, Scale, Eraser |
| 30 | Book, Scale, Glue |
| 40 | Book, Eraser, Glue, scribbling pad |
| 50 | Eraser, Pencils, Scale, Glue, scribbling pad |

a. 0 b. 1 c. 2 d. 3

**Solution**: There is no unique combination of 3-itemset in the transaction table

95.Consider a database with 4 transactions:

Transaction 1: {cheese, bread, milk}

Transaction 2: {soda, bread, milk}

Transaction 3: {cheese, bread}

Transaction 4: {cheese, soda, juice}

.Given, Minimum Support = 0.25. Which rule has a confidence equal to 50%?

1. {bread} => {milk}
2. {bread, milk} => {cheese}
3. {juice} => {soda}
4. {bread} => {cheese}

#### 96. Which of the following is a performance measure for regression?

1. Accuracy
2. Recall
3. Error rate
4. RMSE

97. Which two of the following data points

A = [3, 6, 7], B=[3, 4, 8] and C = [3, 3, 7], are closest to each other using Euclidean distance measure?

1. A and C
2. B and C
3. A and B
4. Equidistant

98.In K-means algorithm, K refers to \_\_\_\_\_.

1. The K nearest points used for clustering
2. K iterations before stopping the clustering process
3. K number of clusters into which the data points are clustered into
4. None of the above

99.Decision Trees can only be used for classification tasks.

a. True b. False

100. How does a Decision Tree make Decision boundaries?

1. It calculates the impurity in all possible feature spaces and chooses the split that gives least impurity or maximum information gain.
2. It uses a distance metric to see which feature spaces are far from each other and then creates a split based on maximum distance.
3. It uses a sorted index of features based on Euclidean distance and creates splits accordingly.
4. None of the above

101. What type of data can Decision Trees handle?

1. Nominal
2. Ordinal
3. Continuous
4. All of the above

102. Decision Trees follow recursive partitioning of the feature space.

a. True b. False

103. Random forest model can be built by aggregation of hundreds of trees. Which of the following statement is true about any individual tree in Random forest?

1. Individual tree is built on all the features
2. Individual tree is built on a subset of the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on all observations given in the data
5. 1 and 3 b. 2 and 3 c. 1 and 2 d. 3 and 4

104. A/An \_\_\_\_\_\_\_\_\_\_\_ is a single layer neural network used as a classifier to learn and process training data one example at a time.

a. synapse b. hidden layer c. perceptron d. ANN

105. In the given network, how many hidden layers are there?

Diagram

Description automatically generated

* 1. 0 b. 1 c. 2 d. 3

106. A multiple regression model has:

1. Only one independent variable
2. More than one independent variable
3. More than one dependent variable
4. None of the above

#### 107. Data used to build a machine learning model

1. validation data
2. training data
3. test data
4. hidden data

#### 108. Which of the following statements regarding outliers is correct?

1. Outliers should be identified and removed from a dataset
2. Outliers should be part of the training dataset but should not be present in the test data
3. Outliers should be part of the test dataset but should not be present in the training data
4. The nature of the problem determines how outliers are used

#### 109. Which of the following are not classification problems?

1. Predicting price of house
2. Predicting patient has tumour
3. Predicting who will win the title in football league
4. All the above

**FILL IN THE BLANKS:**

1. \_\_\_\_\_\_Skewness\_\_\_\_\_\_\_ is a measure of asymmetry in descriptive statistics
2. \_\_\_\_\_\_**Standard deviation** (SD)\_\_\_\_\_\_\_ is a measure of dispersion in descriptive statistics
3. The formula of Min Max scaler is \_\_\_**Xsc=X−XminXmax−Xmin**\_\_\_\_\_\_\_\_\_\_\_\_\_
4. The hypothesis test used to check whether Average price paid by male and female customers is equal is \_\_\_\_\_\_\_\_\_\_\_\_two-sample t-test \_\_\_\_\_\_\_\_
5. The hypothesis test used to check association between 2 non numeric categorical variables is \_\_\_\_\_\_\_\_Chi-squared Test \_\_\_\_\_\_\_\_\_\_\_\_
6. \_\_\_\_\_\_Box Plots\_\_\_\_\_\_\_\_\_\_\_ is a univariate plot created using quartiles Q1, Q2, Q3 and IQR.
7. Normal Distribution is also called as \_\_**Gaussian distribution**\_\_\_\_\_\_\_\_\_\_
8. Scatter plot is a \_\_\_\_\_XY\_\_\_\_\_\_\_\_\_ plot
9. \_\_\_\_outlier\_\_\_\_\_\_ is an observation that lies outside the normal data range.
10. \_\_\_\_\_\_**cube root, and log**\_\_\_\_\_\_\_ transformation is used for correcting positive skewness
11. \_\_*d*​=∣*a*1​−*b*1​∣+∣*a*2​−*b*2​∣­= ­-1\_\_\_\_\_\_will be the Manhattan Distance between the two data points A[1,3] and B[2,3]
12. \_\_\_\_\_\_\_\_\_\_\_\_\_\_kind of data is present on web pages, blogs, scanned documents.
13. Multiple Decision Tree Combine To Make A \_\_Random Forest\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
14. Output of Classification Model is \_\_\_\_\_\_\_discrete\_\_\_\_\_\_\_\_ quantity.
15. Training is under \_\_\_\_\_**labelled dataset**\_\_\_\_\_\_ in Supervised Learning.
16. In Supervised Learning Type of Data is \_\_\_\_\_\_\_\_\_**labeled data**\_\_\_\_\_\_\_\_\_\_\_\_.
17. Normalize Words Into Its Base Form Or Root Form is known as \_\_\_\_**NLP**\_\_\_\_\_\_\_.

SHORT QUESTIONS:

1. Discuss splitting methods Gini and Entropy & Information Gain
2. Discuss 5 Data pre-processing techniques
3. Big data Ecosystem
4. Decision Tree

LONG QUESTIONS:

1. Define Big data? Explain the characteristics and 3Vs of Big data?
2. Discuss in detail Confusion matrix and various metrics calculated from it.
3. Discuss HDFS architecture Namenode, datanode
4. Explain the concept of neural networks in detail.