CS8091	BIG DATA ANALYTICS	L T P C 3 0 0 3

UNIT I **INTRODUCTION TO BIG DATA**

Evolution of Big data - Best Practices for Big data Analytics - Big data characteristics - Validating - The Promotion of the Value of Big data - Big data Use Cases- Characteristics of Big data Applications - Perception and Quantification of Value -Understanding Big data Storage - A General Overview of High Performance Architecture - HDFS - Map Reduce and YARN - Map Reduce Programming Model

PART-A CS8091.1

1. What is Big data?

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making

2. Define Big data analytics.

Big data analytics is the often complex process of examining large and varied data sets, or Big data, to uncover information such as hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.

3. List the categories of Big data applications.

a. Counting
b. Scanning
c. Modeling
d. Storing

4. What are the characteristics of Big data applications?

a. Data throttling
b. Computation-restricted throttling
c. Large data volumes
d. Significant data variety
e. Benefits from data parallelization

5. What are the key factors that an organization must consider before investing on Big data applications?

a. Feasibility
b. Reasonability
c. Value
d. Integrability
e. Sustainability

6. What are the benefits offered by Big data to an organization in increasing its value?

a. Increasing revenues
b. Lowering costs
c. Increasing productivity
d. Reducing risk

7. List the categories of Big data applications.

a. Business intelligence, querying, reporting, searching
b. Improved performance for common data management operations
c. Non-database applications
d. Data mining and analytical applications

8. Name the key computing resources commonly used on different appliances and frameworks.

a. Processing capability
b. Memory
c. Storage
d. Network

9. What is Hadoop?

Hadoop is an open source distributed processing framework that manages data processing and storage for Big data applications in scalable clusters of computer servers. It's at the center of an ecosystem of Big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning. Hadoop systems can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

10. **Write short notes on HDFS.**

HDFS attempts to enable the storage of large files, and does this by distributing the data among a pool of data nodes. A single name node (sometimes referred to as NameNode) runs in a cluster, associated with one or more data nodes, and provide the management of a typical hierarchical file organization and namespace. The name node effectively coordinates the interaction with the distributed data nodes. The creation of a file in HDFS appears to be a single file, even though it blocks “chunks” of the file into pieces that are stored on individual data nodes.

11. **Mention the contents of metadata maintained by the name node.**

The metadata maintained by the name node includes an enumeration of the managed files, properties of the files, and the file system, as well as the mapping of blocks to files at the data nodes.

12. **List some key tasks which can be implemented to overcome failure.**

a. Monitoring
b. Rebalancing
c. Managing integrity
d. Metadata replication
e. Snapshots

13. **What are the value proposition offered by HDFS from the perspective of Information Technology?**

a. decreasing the cost of specialty large-scale storage systems;

b. providing the ability to rely on commodity components;

c. enabling the ability to deploy using cloud-based services;

d. reducing system management costs.

14. **What is the role of JobTracker?**

The role of the JobTracker is to manage the resources with some specific responsibilities, including managing the TaskTrackers, continually monitoring their accessibility and availability, and the different aspects of job management that include scheduling tasks, tracking the progress of assigned tasks, reacting to identified failures, and ensuring fault tolerance of the execution.

15. **What is the role of TaskTracker?**

The role of the TaskTracker is much simpler: wait for a task assignment, initiate and execute the requested task, and provide status back to the JobTracker on a periodic basis.

16. **State the limitations of MapReduce model.**

a. The programming paradigm is nicely suited to applications where there is locality between the processing and the data, but applications that demand data movement will rapidly become bogged down by network latency issues.
b. Not all applications are easily mapped to the MapReduce model, yet applications developed using alternative programming methods would still need the MapReduce system for job management.

c. The allocation of processing nodes within the cluster is fixed through allocation of certain nodes as "map slots" versus "reduce slots." When the computation is weighted toward one of the phases, the nodes assigned to the other phase are largely unused, resulting in processor underutilization.

17. **How the limitations of MapReduce is overcome in future versions of Hadoop?**

The limitations of MapReduce is addressed in future versions of Hadoop through the segregation of duties within a revision called YARN. In this approach, overall resource management has been centralized while management of resources at each node is now performed by a local NodeManager. In addition, there is the concept of an ApplicationMaster that is associated with each application that directly negotiates with the central ResourceManager for resources while taking over the responsibility for monitoring progress and tracking status.

18. **Explain briefly about YARN.**

YARN stands for Yet Another Resource Negotiator. It is the resource management and job scheduling technology in the open source Hadoop distributed processing framework. YARN is one of Apache Hadoop's core components, and is responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.

19. **Explain briefly about Map Reduce.**

Map Reduce is a programming model which can be used to develop applications to read, analyze, transform, and share massive amounts of data for parallel, distributed computation involving massive datasets which can range from hundreds of terabytes to petabytes.

20. **List the two basic operations that are applied to data value pairs.**

a. Map
b. Reduce

21. **Define Map function.**

Map function performs an operation that describes the computation or analysis applied to a set of input key/value pairs to produce a set of intermediate key/value pairs

22. **Define Reduce function.**

Reduce function performs an operation in which the set of values associated with the intermediate key/value pairs output by the Map operation are combined to provide the results.

23. **List the five basic operations of Map Reduce programming model.**

a. Input data
b. Map
c. Sort/shuffle
d. Reduce
e. Output result

24. **What is rebalancing?**

It is a process of automatically migrating blocks of data from one data node to another when there is free space, when there is an increased demand for the data and moving it may improve performance (such as moving from a traditional disk drive to a solid-state drive that is much faster or can accommodate increased numbers of simultaneous accesses), or an increased need to replication in reaction to more frequent node failures.

25. **Write short notes about how integrity is managed in HDFS.**

HDFS uses checksums, which are effectively digital signatures, associated with the actual data stored in a file that can be used to verify that the data stored corresponds to the data shared or received. When the checksum calculated for a retrieved block does not equal the stored checksum of that block, it is considered an integrity error. In that case, the requested block will need to be retrieved from a replica instead.

	What are the components of Big data application development framework? a. a programming model and development tools; b. facility for program loading, execution, and for process and thread scheduling; c. system configuration and management tools.
27.	**Mention the benefits promoted by business intelligence and data warehouse tools.** a. Better targeted customer marketing b. Improved product analytics c. Improved business planning d. Improved supply chain management e. Improved analysis for fraud, waste, and abuse
28.	**Name some non database applications.** a. image processing b. text processing in preparation for publishing c. genome sequencing d. protein sequencing and structure prediction e. web crawling, and monitoring work-flow processes.
29.	**Mention some data mining and analytical applications.** a. social network analysis b. facial recognition c. profile matching d. other types of text analytics e. web mining
30.	**What are the benefits of data parallelization?** a. Reduced data dependencies, b. Improvement in the application’s runtime
	PART-B CS8091.1
1.	Explain in detail about the evolution of Big data.
2.	With the help of a table provide a sample framework for determining a score ranging from 0 to 4 for the factors that determine an organization readiness to Big data.
3.	Discuss in detail about the characteristics of Big data applications.
4.	Categorize Big data use cases with applications.
5.	Elaborate in detail about the applications suited to Big data analytics.
6.	Compare the performance characteristics of Big data with respect to hardware and software approaches.
7.	With the help of a neat diagram explain the organization of resources in a Big data platform.
8.	Elaborate in detail about HDFS.
9.	Explain MapReduce function and YARN in detail.
10.	Explain in detail about MapReduce programming model.
UNIT II	**CLUSTERING AND CLASSIFICATION** Advanced Analytical Theory and Methods: Overview of Clustering - K-means - Use Cases - Overview of the Method - Determining the Number of Clusters - Diagnostics - Reasons to Choose and Cautions .- Classification: Decision Trees - Overview of a Decision Tree - The General Algorithm - Decision Tree Algorithms - Evaluating a Decision Tree - Decision Trees in R - Naïve Bayes - Bayes‘ Theorem - Naïve Bayes Classifier.
	PART-A CS8091.2

Write short notes on clustering.

Clustering is a method often used for exploratory analysis of the data. In clustering, there are no predictions made. Rather, clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters. Clustering techniques are utilized in marketing, economics, and various branches of science.

2. **What is K- means?**

Given a collection of objects each with n measurable attributes, k-means is an analytical technique that, for a chosen value of k, it identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.

3. **List the steps of K- means algorithm.**

a. Choose the value of k and the k initial guesses for the centroids.
b. Compute the distance from each data point (x,y) to each centroid. Assign each point to the closest centroid. This association defines the first k clusters.
c. Compute the centroid, the center of mass, of each newly defined cluster from Step 2.
d. Repeat Steps 2 and 3 until the algorithm converges to an answer.

4. **Briefly explain how k-means analysis is used in image processing.**

Video is one example of the growing volumes of unstructured data being collected. Within each frame of a video, k-means analysis can be used to identify objects in the video. For each frame, the task is to determine which pixels are most similar to each other. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. With security video images, for example, successive frames are examined to identify any changes to the clusters. These newly identified clusters may indicate unauthorized access to a facility.

5. **Write short notes on Within Sum of Squares (WSS).**

WSS is denoted by the following equation

$$WSS = \sum_{i=1}^{M} d(p_i, q^{(i)})^2 = \sum_{i=1}^{M}\sum_{j=1}^{n}\left(p_{ij} - q_j^{(i)}\right)^2$$

It is the sum of the squares of the distances between each data point and the closest centroid. The term $q^{(i)}$ indicates the closest centroid that is associated with the ith point. If the points are relatively close to their respective centroids, the WSS is relatively small. Thus, if k + 1 clusters do not greatly reduce the value of WSS from the case with only k clusters, there may be little benefit to adding another cluster.

6. **Illustrate briefly about PAM and mention how is it implemented in R?**

Partitioning around Medoids (PAM) is a partitioning method. In general, a medoid is a representative object in a set of objects. In clustering, the medoids are the objects in each cluster that minimize the sum of the distances from the medoid to the other objects in the cluster. The advantage of using PAM is that the center of each cluster is an actual object in the dataset. PAM is implemented in R by the pam ( ) function included in the cluster R package.

7. **List the three types of classifiers.**

a. Logistic regression
b. Decision trees
c. Naïve Bayes

8. **Differentiate between true positives and false positives.**

True positives (TP) are the number of positive instances the classifier correctly identified as positive. False positives (FP) are the number of instances in which the classifier identified as positive but in reality are negative.

9. **Differentiate between true negatives and false negatives.**

True negatives (TN) are the number of negative instances the classifier correctly identified as negative. False negatives (FN) are the number of instances classified as negative but in reality are positive.

10. **Elaborate briefly about decision tree.**

A decision tree, also called a prediction tree uses a tree structure to specify sequences of decisions and consequences. Given input X= {x_1, x_2,…, xn}, the goal is to predict a response or output variable Y. Each member of the set {x_1, x_2,…, x_n } is called an input variable. The prediction can be achieved by constructing a decision tree with test points and branches. At each test point, a decision is made to pick a specific branch and traverse down the tree. Eventually, a final point is reached, and a prediction can be made. Each test point in a decision tree involves testing a particular input variable (or attribute), and each branch represents the decision being made.

11. **Discuss briefly about decision tree.**

A decision tree employs a structure of test points (called nodes) and branches, which represent the decision being made. A node without further branches is called a leaf node. The leaf nodes return class labels and, in some implementations, they return the probability scores. A decision tree can be converted into a set of decision rules.

12. **Write short notes on the varieties of decision trees.**

Decision trees have two varieties: classification trees and regression trees. Classification trees usually apply to output variables that are categorical often binary in nature, such as yes or no, purchase or not purchase, and so on. Regression trees, on the other hand, can apply to output variables that are numeric or continuous, such as the predicted price of a consumer good or the likelihood a subscription will be purchased.

13. **Demonstrate the structure of decision tree.**

A decision tree uses a tree structure to specify sequences of decisions and consequences. Internal nodes are the decision or test points. Each internal node refers to an input variable or an attribute. The top internal node is called the root. The decision tree is a binary tree in that each internal node has no more than two branches. The branching of a node is referred to as a split. The depth of a node is the minimum number of steps required to reach the node from the root. Leaf nodes are at the end of the last branches on the tree. They represent class labels-the outcome of all the prior decisions. The path from the root to a leaf node contains a series of decisions made at various internal nodes.

14. **What is decision stump?**

The simplest short tree is called a decision stump, which is a decision tree with the root immediately connected to the leaf nodes. A decision stump makes a prediction based on the value of just a single input variable

15. **List some decision tree algorithms.**

a. ID3,
b. C4.5
c. CART

16. **How will you avoid overfitting in decision tree?**

Overfitting in decision tree can be avoided by following any one of the two approaches listed below.

a. Stop growing the tree early before it reaches the point where all the training data is perfectly classified.
b. Grow the full tree, and then post-prune the tree with methods such as reduced-error pruning and rule-based post pruning.

17. **What do you know about naïve Bayes'?**

Naive Bayes is a probabilistic classification method based on Bayes' theorem (or Bayes' law) with a few tweaks. Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities. Bayes' law is named after the English mathematician Thomas Bayes.

18. **What do you mean by discretization of continuous variable?**

The input variables are generally categorical, but variations of the algorithm can accept continuous variables. There are also ways to convert continuous variables into categorical ones. This process is often referred to as the discretization of continuous variable.

19. **State Bayes' theorem.**

The conditional probability of event C occurring, given that event A has already occurred, is denoted as P(C|A) , which can be found using the formula,

$P(C|A) = P(A \cap C) / P(A)$

By applying some minor algebra and substitution of the conditional probability the most common form of the Bayes' theorem is given by

$P(C|A)= P(A|C)\cdot P(C) / P(A)$

where C is the class label $C \in \{c_1, c_2, ... c_n\}$ and A is the observed attributes $A= \{a_1, a_2, ... a_m\}$.Mathematically, Bayes' theorem gives the relationship between the probabilities of C and A, P(C) and P(A), and the conditional probabilities of C given A and A given C, namely P(C|A) and P(A|C).

20. **State the general form of Bayes' theorem.**

A more general form of Bayes' theorem assigns a classified label to an object with multiple attributes $A= \{a_1, a_2 ,.... ,a_m\}$ such that the label corresponds to the largest value of $P(c_i|A)$. The probability that a set of attribute values A (composed of m variables $a_1, a_2, \dots ,a_m$) should be labeled with a classification label c_i equals the probability that the set of variables $a_l, a_2, \dots , a_m$ given c_i is true, times the probability of c_i divided by the probability of $a_l, a_2, \dots , a_m$. Mathematically, this is shown as

$P(c_i |A)= P(a_l, a_2, \dots , a_m|c_i).P(c_i) / P(a_l, a_2, \dots , a_m), i = 1,2,\dots n$

21. **What is called numerical underflow problem? How is it solved?**

When looking at problems with a large number of attributes, or attributes with a high number of levels, using naïve Bayes' classifier the value of the product of $P(a_j|c_i)$ times $P(c_i)$ become very small in magnitude (close to zero), resulting in even smaller differences of the scores. This is the problem of numerical underflow, caused by multiplying several probability values that are close to zero. The solution to this problem is to compute the logarithm of the products, which is equivalent to the summation of the logarithm of the probabilities.

22. **How will you define the accuracy or overall success rate of a classifier?**

The accuracy {or the overall success rate) is a metric defining the rate at which a model has classified the records correctly. It is defined as the sum of TP and TN divided by the total number of instances.

Accuracy = ((TP+TN) / (TP+TN+FP+FN)) * 100%

23. **Mention the two simplifications made to the Bayes' theorem to extend it as the naïve Bayes' classifier.**

a. The first simplification is to use the conditional independence assumption.
b. The second simplification is to ignore the denominator $P(a_l, a_2, \dots , a_m)$

24. **Mention some classifiers other than decision trees and naïve Bayes'.**

a. Bagging
b. Boosting
c. Random forest
d. Support vector machines

25. **Write short notes on support vector machines.**

Support vector machines is another common classification method that combines linear models with instance based learning techniques. Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear decision function that separates them as widely as possible. SVM by default can efficiently perform linear classifications and can be configured to perform nonlinear classifications as well.

PART-B CS8091.2

1. Demonstrate the k-means algorithm to find k clusters.
2. Discuss in detail about k-means use cases.
3. Illustrate with an example using R to perform a k-means analysis.
4. Outline reasons to choose and cautions in k-means analysis.
5. Explain in detail about decision trees.
6. Explain the following Decision Tree Algorithm with examples:
   i) ID3
   ii) C4.5
   iii) CART
7. With the help of suitable example explain how to model decision trees in R?
8. How will you evaluate a decision tree? Explain in detail.
9. Explain Bayes theorem with an example.
10. Apply naïve Bayes classifier on a bank marketing dataset to predict if the clients would subscribe to a term deposit or not.

UNIT III ASSOCIATION AND RECOMMENDATION SYSTEM

Advanced Analytical Theory and Methods: Association Rules - Overview - Apriori Algorithm - Evaluation of Candidate Rules - Applications of Association Rules - Finding Association& finding similarity - Recommendation System: Collaborative Recommendation- Content Based Recommendation - Knowledge Based Recommendation- Hybrid Recommendation Approaches.

PART-A CS8091.3

1. Write short notes on association rules.

Association rules are an unsupervised learning method. They are a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset. The disclosed relationships can be represented as rules or frequent item sets. Association rules are commonly used for mining transactions in databases. Because of their popularity in mining customer transactions, association rules are sometimes referred to as market basket analysis.

2. Define itemset and k-itemset.

The term itemset refers to a collection of items or individual entities that contain some kind of relationship. This could be a set of retail items purchased together in one transaction, a set of hyperlinks clicked on by one user in a single session, or a set of tasks done in one day. An item set containing k items is called a k-itemset.

3. Write short notes on Apriori algorithm.

Apriori is one of the earliest and the most fundamental algorithms for generating association rules. It pioneered the use of support for pruning the itemsets and controlling the exponential growth of candidate itemsets. Shorter candidate item sets, which are known to be frequent item sets, are combined and pruned to generate longer frequent itemsets. This approach eliminates the need for all possible item sets to be enumerated within the algorithm, since the number of all possible itemsets can become exponentially large.

4. What do you know from the term support of an itemset? Give suitable examples.

Given an item set L, the support of L is the percentage of transactions that contain L. For example, if 80% of all transactions contain item set {bread}, then the support of {bread} is 0.8. Similarly, if 60% of all transactions contain itemset {bread, butter}, then the support of {bread, butter} is 0.6.

5. What is frequent itemset?

A frequent itemset has items that appear together often enough. The term "often enough" is formally defined with a minimum support criterion. If the minimum support is set at 0.5, any itemset can be considered a frequent item set if at least 50% of the transactions contain this itemset.

6. What is downward closure property? Give an example.

If an item set is considered frequent, then any subset of the frequent item set must also be frequent. This is referred to as the Apriori property or downward closure property. For example, if 60% of the transactions contain {bread, jam}, then at least 60% of all the transactions will contain {bread} or {jam}.

7. Define confidence.

Confidence is defined as the measure of certainty or trustworthiness associated with each discovered rule. Mathematically, confidence is the percent of transactions that contain both X and Y out of all the transactions that contain X.

$$\text{Confidence}(X \rightarrow Y) = \text{Support}(X \wedge Y) / \text{Support}(X)$$

8. What is called minimum confidence?

A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold. This predefined threshold is called the minimum confidence.

9. Define lift.

Consider X and Y be statistically independent items, lift is a measure of how X and Y are really related rather than coincidentally happening together.

$$\text{Lift}(X \rightarrow Y) = \text{Support}(X \wedge Y) / \text{Support}(X) * \text{Support}(Y)$$

10. Define leverage.

Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.

$$\text{Leverage}(X \rightarrow Y) = \text{Support}(X \wedge Y) - \text{Support}(X) * \text{Support}(Y)$$

11. State the applications of association rules.

a. Broad-scale approaches to better merchandising
b. Cross-merchandising between products and high-margin or high-ticket items
c. Physical or logical placement of product within related categories of products promotional programs

12. Write short notes on click stream analysis.

Clickstream analysis refers to the analytics on data related to web browsing and user clicks, which is stored on the client or the server side. Web usage log files generated on web servers contain huge amounts of information, and association rules can potentially give useful knowledge to web usage data analysts. For example, association rules may suggest that website visitors who land on page X click on links A, B, and C much more often than links 0, E, and F. This observation provides valuable insight on how to better personalize and recommend the content to site visitors.

13. Mention some approaches to improve Apriori's efficiency.

a. Partitioning
b. Sampling
c. Transaction reduction
d. Hash-based itemset counting
e. Dynamic itemset counting

14. What is user-based nearest neighbor recommendation?

Given a ratings database and the ID of the current (active) user as an input, identify other users (sometimes referred to as peer users or nearest neighbors) that had similar preferences to those of the active user in the past. Then, for every product p that the active user has not yet seen, a prediction is computed based on the ratings for p made by the peer users.

15. Mention the assumptions made in user-based nearest neighbor recommendation.

a. if users had similar tastes in the past they will have similar tastes in the future
b. user preferences remain stable and consistent over time.

16. What is cosine similarity measure?

It is a metric that measures the similarity between two n-dimensional vectors based on the angle between them. This measure is also commonly used in the fields of information retrieval and text mining to compare two text documents, in which documents are represented as vectors of terms. The similarity between two items a and b – viewed as the corresponding rating vectors $\vec{a}$ and $\vec{b}$ is formally defined as follows

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

17. **Write short notes on term frequency.**

Term frequency describes how often a certain term appears in a document (assuming that important words appear more often). To take the document length into account and to prevent longer documents from getting a higher relevance weight, some normalization of the document length should be done. We search for the normalized term frequency value TF(i, j) of keyword i in document j. Let freq(i, j) be the absolute number of occurrences of I in j. Given a keyword i, let OtherKeywords(i, j) denote the set of the other keywords appearing in j . Compute the maximum frequency maxOthers(i, j) as max(freq(z, j)), z ∈ OtherKeywords(i, j). Finally, calculate TF(i, j) as in

TF(i,j) = freq(i,j) / maxOthers(i, j)

18. **Write short notes on Inverse document frequency.**

Inverse document frequency is the second measure that is combined with term frequency. It aims at reducing the weight of keywords that appear very often in all documents. The idea is that those generally frequent words are not very helpful to discriminate among documents, and more weight should therefore be given to words that appear in only a few documents. Let N be the number of all recommendable documents and n(i) be the number of documents from N in which keyword i appears. The inverse document frequency for i is typically calculated as

IDF(i) = log (N/n(i))

19. **What is stemming or conflation?**

Stemming or conflation is a technique which aims to replace variants of the same word by their common stem.

20. **State the advantages of kNN- based methods.**

a. relatively simple to implement,
b. adapt quickly to recent changes,
c. requires a relatively small number of ratings to make a prediction of reasonable quality.

21. **Mention the limitations of pure content-based recommender systems.**

a. Shallow content analysis
b. Overspecialization
c. Acquiring ratings

22. **What are the two types of knowledge-based recommender systems?**

a. Constraint based systems
b. Case-based systems

23. **What is the difference between case based and constraint based recommender systems?**

Case-based recommender system focus on the retrieval of similar items on the basis of different types of similarity measures, whereas constraint-based recommender system rely on an explicitly defined set of recommendation rules.

24. **What is a conjunctive query?**

A conjunctive query is a database query with a set of selection criteria that are connected conjunctively.

25. **What are the three base hybridization designs?**

Monolithic, Parallelized, and Pipelined hybrids

PART-B IT6010.3

1. Demonstrate Apriori algorithm and the evaluation of candidate rules in detail.
2. With the help of an example explain in detail about user-based nearest neighbor recommendation system.
3. With the help of grocery store transactions demonstrate how to use R to perform association rule mining.
4. Discuss in detail about item-based nearest neighbor recommendation system.
5. Distinguish in detail about implicit and explicit ratings.
6. Outline data sparsity and cold start problem in detail.
7. What do you know about probabilistic recommendation approaches? Explain in detail.
8. Explain in detail about similarity based retrieval systems.
9. Discuss in detail about knowledge based recommendation system.
10. Elaborate the hybrid recommendation systems in detail.

UNIT IV **STREAM MEMORY**

Introduction to Streams Concepts – Stream Data Model and Architecture - Stream Computing, Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating moments – Counting oneness in a Window – Decaying Window – Real time Analytics Platform(RTAP) applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions. Using Graph Analytics for Big data: Graph Analytics

PART-A CS8091.4

1. Mention some stream sources.

a. Sensor data
b. Image data
c. Internet
d. Web traffic

2. Mention the types of queries.

a. Standing queries
b. Adhoc queries

3. What are standing queries?

Standing queries are those queries which are within the processor. These queries are permanently executing, and produce outputs at appropriate times.

4. What are adhoc queries?

Adhoc queries are arbitrary queries which are asked once. If we do not store all streams in their entirety, as normally we cannot, then we cannot expect to answer these arbitrary queries about streams.

5. What are the issues in stream processing?

Streams deliver elements very rapidly. We must process elements in real time, or we lose the opportunity to process them at all, without accessing the archival storage. Stream-processing algorithm should be executed in main memory, without access to secondary storage or with only rare accesses to secondary storage.

6. What are the contents of a Bloom filter?

a. An array of n bits, initially all 0's.
b. A collection of hash functions h_1, h_2. . . h_k. Each hash function maps "key" values to n buckets, corresponding to the n bits of the bit-array.
c. A set S of m key values

7. Mention the six rules that must be followed when representing a stream by buckets.

a. The right end of a bucket is always a position with a 1.
b. Every position with a 1 is in some bucket.
c. No position is in more than one bucket.
d. There are one or two buckets of any given size, up to some maximum size.

e. All sizes must be a power of 2.

f. Buckets cannot decrease in size as we move to the left (back in time).

8. **Mention the ways to deal stream data model.**

There are two ways to deal with the stream data model. The first strategy deals streams by maintaining summaries of the streams, sufficient to answer the expected queries about the data. The second approach maintains a sliding window of the most recently arrived data to deal streams.

9. **How will you create a sample of the stream?**

To create a sample of a stream that is usable for a class of queries, we identify a set of key attributes for the stream. By hashing the key of any arriving stream element, we can use the hash value to decide consistently whether all or none of the elements with that key will become part of the sample.

10. **What is the purpose of bloom filter?**

Bloom filter is a technique that allows us to filter streams so elements that belong to a particular set are allowed through, while most nonmembers are deleted. In this technique a large bit array, and several hash functions are used. Members of the selected set are hashed to buckets, which are bits in the array, and those bits are set to 1. To test a stream element for membership, we hash the element to a set of bits using each of the hash functions, and only accept the element if all these bits are 1.

11. **How will you get a reliable estimate the number of distinct elements in a stream?**

To estimate the number of different elements appearing in a stream, we can hash elements to integers, interpreted as binary numbers. 2 raised to the power that is the longest sequence of 0's seen in the hash value of any stream element is an estimate of the number of different elements. By using many hash functions and combining these estimates, first by taking averages within groups, and then taking the median of the averages, we get a reliable estimate.

12. **What do you mean by the term moments of stream?**

The k^{th} moment of a stream is the sum of the k^{th} powers of the counts of each element that appears at least once in the stream. The 0^{th} moment is the number of distinct elements, and the 1^{st} moment is the length of the stream.

13. **How will you estimate the second moment?**

A good estimate for the second moment, or surprise number, is obtained by choosing a random position in the stream, taking twice the number of times this element appears in the stream from that position onward, subtracting 1, and multiplying by the length of the stream. Many random variables of this type can be combined like the estimates for counting the number of distinct elements, to produce a reliable estimate of the second moment.

14. **Mention the way to estimate the higher moments?**

The technique for second moments works for k^{th} moments as well, as long as we replace the formula 2x-1 (where x is the number of times the element appears at or after the selected position) by $x^k - (x - 1)^k$.

15. **How will you estimate the number of 1s in a window?**

We can estimate the number of 1's in a window of 0's and 1's by grouping the 1's into buckets. Each bucket has a number of 1's that is a power of 2; there are one or two buckets of each size, and sizes never decrease as we go back in time. If we record only the position and size of the buckets, we can represent the contents of a window of size N with $O(\log^2 N)$ space.

16. **How will you determine the approximate numbers of 1's in the most recent k elements of a binary stream?**

If we want to know the approximate numbers of 1's in the most recent k elements of a binary stream, we find the earliest bucket B that is at least partially within the last k positions of the window and estimate the number of 1's to be the sum of the sizes of each of the more recent buckets plus half the size of B. This estimate can never be off by more than 50% of the true count of 1's.

17. How will you assure that the approximation to the true number of 1's is never off by more than 1/r?

By changing the rule for how many buckets of a given size can exist in the representation of a binary window, so that either r or r-1 of a given size may exist, we can assure that the approximation to the true number of 1's is never off by more than 1/r.

18. How will you recompute the weighted sum of elements when a new element arrives?

Rather than fixing a window size, we can imagine that the window consists of all the elements that ever arrived in the stream, but with the element that arrived t time units ago weighted by e^{-ct} for some time constant c doing so allows us to maintain certain summaries of an exponentially decaying window easily. For instance, the weighted sum of elements can be recomputed, when a new element arrives, by multiplying the old sum by 1-c and then adding the new element.

19. What is graph analytics?

Graph analytics is an analytics alternative that uses an abstraction called a graph model. It is an alternative to the traditional data warehouse model as a framework for absorbing both structured and unstructured data from various sources to enable analysts to probe the data in an undirected manner.

20. How will you enrich the meaning of the nodes and edges represented in the graph model?

a. Vertices can be labeled to indicate the types of entities that are related.
b. Edges can be labeled with the nature of the relationship.
c. Edges can be directed to indicate the "flow" of the relationship.
d. Weights can be added to the relationships represented by the edges.
e. Additional properties can be attributed to both edges and vertices.
f. Multiple edges can reflect multiple relationships between pairs of vertices.

21. What are the contents of a triple format?

A subject (the source point of the relationship), an object (the target), and a predicate (that models the type of the relationship).

22. What is a semantic database what is its advantage?

A collection of triples is called a semantic database, and this kind of database can capture additional properties of each triple relationship as attributes of the triple.

23. Mention the characteristics and factors of business problems that favors graph analytics solution.

a. Connectivity
b. Undirected discovery
c. Absence of structure
d. Flexible semantics
e. Extensibility
f. Knowledge is embedded in the network
g. Ad hoc nature of the analysis
h. Predictable interactive performance

24. Name the various types of graph analytics algorithmic approaches.

a. Community and network analysis
b. Path analysis
c. Clustering
d. Pattern detection and pattern analysis
e. Probabilistic graphical models
f. Graph metrics

25. What are the various factors that affect the performance of the graph analytics?

a. Unpredictability of graph memory accesses

b. Graph growth models

c. Dynamic interactions with graph

d. Complexity of graph partitioning

26. **Mention the three sets of features of a graph analytics platform.**

a. Ease of development and implementation,

b. Interoperability with complementary reporting and analysis technologies, and

c. System execution performance.

27. **What are the various general requirements of Big data applications?**

a. Seamless data intake

b. Data integration

c. Inferencing

d. Standards-based representation

28. **Mention some considerations for the operational aspects of the graph analytics platform.**

a. Workflow integration

b. Visualization

c. Complementariness

29. **What are the various expectations of the Big data platform?**

a. High-speed I/O

b. High-bandwidth network

c. Multithreading

d. Large memory

30. **Mention some dedicated appliances for graph analytics.**

RDF and SPARQL

PART-B IT6010.4

1. With the help of a neat sketch explain in detail about data stream management system.
2. Discuss the various examples of stream sources in detail.
3. Categorize and elaborate the various types of stream queries.
4. How will you perform sampling of data in a stream? Elaborate in detail.
5. Explain in detail about filtering stream.
6. Demonstrate the Flajolet-Martin algorithm of counting distinct elements in a stream.
7. Explain in detail about Alon-Matias-Szegedy algorithm for second moments.
8. Demonstrate Datar-Gionis-Indyk-Motwani algorithm.
9. Discuss in detail about some of the graph analytics use cases.
10. Explain in detail about technical complexity of analyzing graphs.
11. Elaborate the features of a graph analytics platform in detail.

UNIT V NOSQL DATA MANAGEMENT FOR BIG DATA AND VISUALIZATION

NoSQL Databases : Schema-less Models‖: Increasing Flexibility for Data Manipulation-Key Value Stores-Document Stores - Tabular Stores - Object Data Stores - Graph Databases Hive - Sharding –- Hbase – Analyzing Big data with twitter - Big data for E-Commerce Big data for blogs - Review of Basic Data Analytic Methods using R

PART-A CS8091.5

1. **What is NoSQL?**

NoSQL is an approach to database design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.

2. **What are the advantages of NoSQL data systems?**

NoSQL data systems hold out the promise of greater flexibility in database management while reducing the dependence on more formal database administration. NoSQL databases have more relaxed modeling constraints, which may benefit both the application developer and the end-user analysts when their interactive analyses are not throttled by the need to cast each query in terms of a relational table-based environment. NoSQL databases also provide for integrated data caching that helps reduce data access latency and speed performance.

3. **What is schema-less model and what are its advantages?**

Schema less modeling is a modeling scheme in which the semantics of the data are embedded within a flexible connectivity and storage model which provides for automatic distribution of data and elasticity with respect to the use of computing, storage, and network bandwidth in ways that don't force specific binding of data to be persistently stored in particular physical locations.

4. **What is a key value store?**

Key value store is a relatively simple type of NoSQL data store. It is a schema-less model in which values (or sets of values, or even more complex entity objects) are associated with distinct character strings called keys.

5. **What are the various core operations performed on a key value store?**

a. Get(key), which returns the value associated with the provided key.

b. Put(key, value), which associates the value with the key.

c. Multi-get(key1, key2,.., keyN), which returns the list of values associated with the list of keys.

d. Delete(key), which removes the entry for the key from the data store.

6. **What are the drawbacks of key value pair?**

a. The model will not inherently provide any kind of traditional database capabilities (such as atomicity of transactions, or consistency when multiple transactions are executed simultaneously)—those capabilities must be provided by the application itself.

b. As the model grows, maintaining unique values as keys may become more difficult, requiring the introduction of some complexity in generating character strings that will remain unique among a myriad of keys.

7. **What is a document store?**

A document store is similar to a key value store in that stored objects are associated (and therefore accessed via) character string keys. The difference is that the values being stored, which are referred to as "documents," provide some structure and encoding of the managed data.

8. **Name the various encodings utilized in document store for managing data.**

XML (Extensible Markup Language), JSON (Java Script Object Notation), BSON (which is a binary encoding of JSON objects), or other means of serializing data (i.e., packaging up the potentially linearizing data values associated with a data record or object).

9. **Distinguish between document store and key value store.**

While the key value store requires the use of a key to retrieve data, the document store often provides a means (either through a programming API or using a query language) for querying the data based on the contents. Because the approaches used for encoding the documents embed the object metadata, one can use methods for querying by example.

10. **What are tabular stores? Give an example.**

Tabular, or table-based stores are largely descended from Google's original Bigtable design to manage structured data. The HBase model is an example of a Hadoop-related NoSQL data management system that evolved from Bigtable.

11. **Write short notes on** Bigtable **NoSQL model.**

The Bigtable NoSQL model allows sparse data to be stored in a three-dimensional table that is indexed by a row key (that is used in a fashion that is similar to the key value and document stores), a column key that indicates the specific attribute for which a data value is stored, and a timestamp that may refer to the time at which the row's column value was stored.

	What are ACID properties? Atomicity, Consistency, Isolation, and Durability
13.	**Write short notes on object data stores.** Object data stores and object databases seem to bridge the worlds of schema-less data management and the traditional relational models. Approaches to object databases can be similar to document stores except that the document stores explicitly serializes the object so the data values are stored as strings, while object databases maintain the object structures as they are bound to object-oriented programming languages such as C11, Objective-C, Java, and Smalltalk.
14.	**Write short notes on graph databases.** Graph databases provide a model of representing individual entities and numerous kinds of relationships that connect those entities. More precisely, it employs the graph abstraction for representing connectivity, consisting of a collection of vertices (which are also referred to as nodes or points) that represent the modeled entities, connected by edges (which are also referred to as links, connections, or relationships) that capture the way that two entities are related. Graph analytics performed on graph data stores are somewhat different than more frequently used querying and reporting.
15.	**Mention the two key criteria for which NoSQL data management environment is engineered for.** a. fast accessibility b. scalability for volume
16.	**What is Hive?** Hive, is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.
17.	**What is the purpose of Hive?** Hive is specifically engineered for data warehouse querying and reporting and is not intended for use as within transaction processing systems that require real-time query execution or transaction semantics for consistency at the row level.
18.	**What is HBase?** HBase is a nonrelational data management environment that distributes massive datasets over the underlying Hadoop framework. HBase is derived from Google's BigTable and is a column-oriented data layout that, when layered on top of Hadoop, provides a fault-tolerant method for storing and manipulating large data tables.
19.	**Mention the basic operations of HBase.** a. Get (which access a specific row in the table), b. Put (which stores or updates a row in the table), c. Scan (which iterates over a collection of rows in the table), and d. Delete (which removes a row from the table).
20.	**What is R?** R is a programming language and software framework for statistical analysis and graphics. Available for use under the GNU General Public License, R software and installation instructions can be obtained via the Comprehensive R Archive and Network.
21.	**Mention the purpose of read.csv, head, summary and lm functions?** The read.csv () function is used to import the CSV file. The head ( ) function, is used for displaying the first six records. The summary () function provides some descriptive statistics, such as the mean and median, for each data column. The lm () function is used for linear regression
22.	**What is a generic function? Give an example.** A generic function is a group of functions sharing the same name but behaving

	differently depending on the number and the type of arguments they receive. Ex. summary () function
23.	**Mention some R graphical user interfaces.** a. RGui.exe, b. R commander c. Rattle d. RStudio
24.	**Name the four highlighted window panes of RStudio graphical user interface.** a. Scripts: Serves as an area to write and save R code b. Workspace: Lists the datasets and variables in the R environment c. Plots: Displays the plots generated by the R code and provides a straightforward mechanism to export the plots d. Console: Provides a history of the executed R code and the output
25.	**What is the purpose of save.image () and load.image () function?** R allows one to save the workspace environment, including variables and loaded libraries, into an .Rdata file using the save.image () function. An existing .Rdata file can be loaded using the load.image () function.
26.	**What are the various functions used for importing the dataset in R?** a. read.csv() function b. read.table()function c. read.delim() function d. read.csv2() function e. read.delim2() function
27.	**What are the various functions used for exporting the R dataset to an external file?** a. write.table(), b. write.csv() ,and c. write.csv2()
28.	**State the purpose of the R packages DBI and RODBC.** Sometimes it is necessary to read data from a database management system (DBMS). R packages such as DBI and RODBC are used for this purpose. These packages provide database interfaces for communication between R and DBMSs such as MySQL, Oracle, SQL Server, PostgreSQL, and Pivotal Greenplum.
29.	**What are the four major categories of attributes in R?** a. Nominal b. Ordinal c. Interval d. Ratio
30.	**What are the various data types supported by R?** a. Numeric b. Character c. Logical data types
31.	**What is the purpose of class and typeof functions?** The class () function represents the abstract class of an object. The typeof () function determines the way an object is stored in memory.
32.	**What is a vector in R?** Vectors are a basic building block for data in R. Simple R variables are actually vectors. A vector can only consist of values in the same class. The tests for vectors can be conducted using the is.vector() function.
33.	**What is the purpose of data frame?** Data frames provide a structure for storing and accessing several variables of possibly different data types.
34.	**What are contingency tables in R?**

In R, table refers to a class of objects used to store the observed counts across the factors for a given dataset. Such a table is commonly referred to as a contingency table and is the basis for performing a statistical test on the independence of the factors used to build the table.

35. Mention some functions for visualizing a single variable in R.

a. plot(data)
b. barplot(data)
c. dotchart(data)
d. hist(data)
e. plot(density(data))
f. stem(data)
g. rug(data)

36. What is hypothesis testing?

It is a common technique to assess the difference or the significance of the difference of the means from two samples of data. The basic concept of hypothesis testing is to form an assertion and test it with data. When performing hypothesis tests, the common assumption is that there is no difference between two samples. This assumption is used as the default position for building the test or conducting a scientific experiment. Statisticians refer to this as the null hypothesis (H_0). The alternative hypothesis (H_A) is that there is a difference between two samples.

37. What is confidence interval?

A confidence interval is an interval estimate of a population parameter or characteristic based on sample data. A confidence interval is used to indicate the uncertainty of a point estimate. If x is the estimate of some unknown population mean p, the confidence interval provides an idea of how close x is to the unknown p.

38. What is Wilcoxon rank sum test?

The Wilcoxon rank-sum test is a nonparametric hypothesis test that checks whether two populations are identically distributed.

39. What is ANOVA?

ANOVA is a generalization of the hypothesis testing of the difference of two population means. ANOVA tests if any of the population means differ from the other population means. The null hypothesis of ANOVA is that all the population means are equal. The alternative hypothesis is that at least one pair of the population means is not equal.

40. Mention the types of ANOVA.

a. One-way ANOVA
b. Two-way ANOVA
c. Multivariate ANOVA

PART-B IT 6010.5

1. Discuss in detail about key-value pair.
2. Elaborate document stores in detail.
3. What do you know about tabular stores? Explain in detail.
4. What is Bigtable? Discuss it in detail.
5. Elaborate in detail about HBase.
6. With the help of a neat sketch explain in detail about the architecture of HBase.
7. What is Hive? Explain in detail.
8. Illustrate in detail about the graphical user interface, data import and export operations in R.
9. Examine the attributes and data types in R.
10. Compare and contrast about visualizing a single variable and multi variable in R.
11. Explain in detail about Box-and-Whisker plot.

	Elaborate in detail about statistical methods for evaluation.