Module-3: NoSQL

3.1 Introduction to NoSQL, NoSQL Business Drivers

3.2 NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable)stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study

3.3 NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; NoSQL systems to handle big data problems.

Practice Questions

1. What is the primary motivation behind adoption of NoSQL databases?
2. What are the business drivers for NoSQL?
3. In what scenarios is NoSQL often preferred over SQL databases?
4. Describe characteristics of NoSQL database.
5. Elaborate the fundamental differences between NoSQL databases and traditional SQL databases, focusing on key characteristics that make NoSQL unique?
6. How would you articulate the core business drivers that lead organizations to consider adopting NoSQL databases over traditional relational databases?
7. In what scenarios would a company find it most advantageous to switch from a traditional SQL database to a NoSQL solution?
8. How do NoSQL databases align with the needs of businesses that deal with unstructured and rapidly changing data?
9. How NOSQL can be used to handle big data problems.
10. Discuss architectural patterns of NoSQL.
11. Describe the key-value store architecture pattern in NoSQL.
12. In what situations would you recommend using a key-value store over other NoSQL architecture patterns?
13. Can you describe a use case where a graph store significantly outperforms other NoSQL architectures due to its ability to represent and traverse relationships?
14. Provide an example scenario where a graph store NoSQL database would be more beneficial than other NoSQL architectures?
15. Provide examples of industries or applications where a column family store is particularly well-suited and explain why.
16. Discuss the querying capabilities of document stores. How do they handle complex queries and aggregations?
17. What does the CAP theorem state?
18. Define each of the three terms in the CAP theorem: Consistency, Availability, and Partition Tolerance. How are these terms specifically defined in the context of distributed systems?
19. A startup is considering different database options for their new project. They are curious about NoSQL databases but unsure about how they differ from traditional SQL databases. Provide them with a comprehensive overview of NoSQL, highlighting the fundamental differences and advantages of using NoSQL in certain scenarios.
20. A company has been using traditional relational databases for several years. They are facing challenges in handling the increasing volume of data and need to scale their system. Explain to them the key concepts of NoSQL databases and the business drivers that make NoSQL a suitable choice for their evolving needs.
21. Consider an e-commerce platform where users can make purchases and view product availability. In the event of a network partition, how would you ensure that users can still browse available products while maintaining data consistency?

Module-4: Mining Data Streams

4.1 The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing.

4.2 Sampling Data techniques in a Stream

4.3 Filtering Streams: Bloom Filter with Analysis.

4.4 Counting Distinct Elements in a Stream, Count Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements

4.5 Counting Ones in a Window: The Cost of Exact Counts,The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.

Practice Questions

1. What is a Data-Stream-Management System (DSMS)?
2. Explain with Block Diagram of Data-Stream-Management System.
3. Discuss the components of DSMS (Data Stream Management System) architecture.
4. Provide an example of a stream source.
5. Identify two major issues in stream processing.
6. What are the challenges of querying on large data stream?
7. How stream processing is different from batch processing?
8. Name a common technique for sampling in stream processing.
9. Why might Bloom Filters produce false positives?
10. How bloom filter is useful for big data analytics? Explain with one example.
11. Justify "Bloom Filter has false positives but no false negatives".
12. What is the Count Distinct Problem in stream processing?
13. How does the DGIM Algorithm handle query answering?
14. Explain the concept of decaying windows in stream processing.
15. What is Sliding window sampling technique for big data.
16. Consider size of Bloom filter m = 10 and two hash functions: h1(x) = x mod 10 and h2(x) = (3x+1) mod 10. Insert the element 5, 7, 12 and 13. Check, are elements 11, 15 and 8 presents? Which elements return false positive result? How we can reduce it.
17. What is Flajolet-Martin (FM) algorithm. Using FM algorithm, compute number of distinct elements in the set S = {1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1} with hash function h(x)=(6x+1) mod 5.
18. In the context of network security, you need to identify malicious IP addresses from a stream of incoming data. Explain how a Bloom Filter can assist in quickly identifying whether an IP address is part of a known blacklist. Discuss the implications of false positives and false negatives in this security application.
19. You are working on a social media analytics platform that needs to estimate the number of unique hashtags in a real-time stream. How could the Flajolet-Martin Algorithm be applied in this scenario? Discuss the considerations for combining estimates from multiple streams to improve accuracy.
20. Imagine you are developing a system for monitoring power consumption in a smart city. Explain how the Datar-Gionis-Indyk-Motwani (DGIM) Algorithm could be applied to estimate the number of occurrences of power spikes within a sliding window.

Module-5: Real-Time Big Data Models

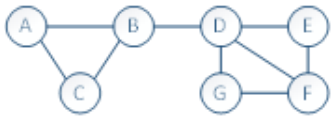
5.1 A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering

5.2 Case Study: Product Recommendation

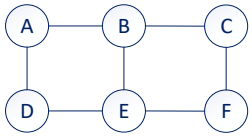
5.3 Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph

Practice Questions

1. What is social Network?
2. In a social network graph, what do nodes and edges represent?
3. Discuss the concept of clustering in the context of social-network graphs.
4. How do content-based filtering and collaborative filtering differ when employed in recommendation systems?
5. What role does community detection play in enhancing user experience on social media platforms?
6. In the context of a product recommendation, what factors might be considered in a content-based recommendation system?
7. What challenges might arise in implementing a product recommendation system for a large e-commerce platform?
8. Can collaborative filtering work well when there are new items in the system?
9. Can you provide examples of real-world applications where social network graphs are used?
10. In the context of a social network graph, what is a community?
11. What is edge betweenness? Where is it used?
12. How can the concept of communities be applied to enhance the analysis of social networks?
13. Outline the main steps of the Girvan-Newman algorithm for detecting communities in a social network graph.
14. A new e-commerce website is looking to implement a recommendation system to enhance user experience. Compare and contrast content-based recommendations and collaborative filtering, highlighting the advantages and potential challenges of each approach for this business.
15. Imagine you are building a recommendation system for an online streaming platform. Explain how content-based recommendations work and provide an example of how this approach could be applied to suggest movies or TV shows to users with specific preferences.
16. What you mean communities in social network. Detect communities in the below given social graph using Girvan-Newman (GN) algorithm.



17. Calculate edge betweenness for every edge in the below social graph. Also detect communities using Girvan-Newman (GN) algorithm.



Module-6: Data Analytics with R

6.1 Exploring Basic features of R, Exploring RGUI, Exploring Rstudio, Handling Basic Expressions in R, Variables in R, Working with Vectors, Storing and Calculating Values in R, Creating and using Objects, interacting with users, Handling data in R workspace, Executing Scripts, Creating Plots, Accessing help and documentation in R

6.2 Reading datasets and Exporting data from R, Manipulating and Processing Data in R, Using functions instead of script, built-in functions in R

6.3 Data Visualization: Types, Applications

Important Questions

1. List the basic features of R.
2. What are the different data structures in R. Explain with example
3. Explain the concept of vectors in R. How are they defined, and what operations can be performed on them?
4. Discuss the importance of objects in R. Provide examples of creating objects and illustrate how they are used in R programming.
5. How can R interact with users during script execution? Describe methods for taking user input and displaying output within R scripts.
6. Discuss the capabilities of R for data visualization. Provide examples of creating basic plots using R's plotting functions.
7. Explore some of the essential built-in functions in R. Provide examples of functions for basic statistical calculations and data summarization.
8. Define a vector in R. How can you create a numeric vector with values 1, 2, 3, 4, and 5?
9. Perform element-wise multiplication using R on two vectors: vec1 = c(2, 4, 6) and vec2 = c(1, 2, 3).
10. Define a matrix in R. Create a 3x3 matrix with values 1 to 9.
11. Define a factor in R. Create a factor vector with levels "Low," "Medium," and "High."
12. Your team is preparing a presentation for a client, and visual representation of data is crucial. Discuss the different types of data visualizations available in R and provide examples of when to use scatter plots, bar charts, and line graphs based on the nature of the data.
13. The marketing department wants to convey sales trends over the past year to stakeholders. Illustrate how you would choose and create an appropriate data visualization in R to effectively communicate this information. Emphasize the importance of selecting the right visualization type for the intended message.
14. Your manager has requested a summary report of specific metrics from a large dataset. Explain how you would manipulate and process data in R to extract the necessary information efficiently. Highlight the use of functions to streamline the analysis process.