Week -4

1.Which of the following statements about Bloom filters is true?

a. Bloom filters guarantee no false negatives
b. Bloom filters use cryptographic hashing functions
c. Bloom filters may produce false positives but no false negatives
d. Bloom filters are primarily used for sorting large datasets

Option a: Bloom filters guarantee no false negatives - **Incorrect**. Bloom filters can produce false positives (indicating an element is present when it's not), but they guarantee no false negatives (indicating an element is absent when it's present).

Option b: Bloom filters use cryptographic hashing functions - **Incorrect**. While cryptographic hashing functions can be used, they are not a requirement. Bloom filters typically use multiple hash functions.

Option c: Bloom filters may produce false positives but no false negatives - **Correct**. This is the fundamental property of Bloom filters.

Option d: Bloom filters are primarily used for sorting large datasets - **Incorrect**. Bloom filters are primarily used for approximate membership testing.

2. How does CAP theorem impact the design of distributed systems?

A) It emphasizes data accuracy over system availability

B) It requires trade-offs between consistency, availability, and partition tolerance

C) It prioritizes system performance over data security

D) It eliminates the need for fault tolerance measures

Option A: It emphasizes data accuracy over system availability - Incorrect.
The CAP theorem does not prioritize data accuracy; rather, it highlights the

trade-offs between consistency, availability, and partition tolerance in distributed systems.

Option B: It requires trade-offs between consistency, availability, and partition tolerance - Correct. The CAP theorem states that in the presence of a network partition, a distributed system can only guarantee either consistency or availability, but not both.

Option C: It prioritizes system performance over data security - Incorrect. The CAP theorem does not address performance or security; it focuses specifically on the consistency, availability, and partition tolerance trade-offs in distributed systems.

Option D: It eliminates the need for fault tolerance measures - Incorrect. The CAP theorem does not eliminate the need for fault tolerance; in fact, it highlights the challenges that arise in maintaining consistency and availability when partitions occur.

3. Which guarantee does the CAP theorem consider as mandatory for a distributed system?

a. Consistency
b. Availability
c. Partition tolerance
d. Latency tolerance

Option a: Consistency - **Incorrect**. The CAP theorem states that it's impossible to achieve all three guarantees (Consistency, Availability, and Partition tolerance) simultaneously in a distributed system.

Option b: Availability - **Incorrect**. Availability is not guaranteed if partition tolerance is required.

Option c: Partition tolerance - **Correct**. The CAP theorem states that partition tolerance is essential for distributed systems, as network partitions are inevitable.

Option d: Latency tolerance - **Incorrect**. Latency tolerance is not explicitly mentioned in the CAP theorem.

4.What consistency level in Apache Cassandra ensures that a write operation is acknowledged only after the write has been successfully written to all replicas?

a. ONE
b. LOCAL_ONE
c. LOCAL_QUORUM
d. ALL

Option a: ONE - **Incorrect**. ONE requires only one replica to acknowledge the write.

Option b: LOCAL_ONE - **Incorrect**. LOCAL_ONE requires one replica within the same datacenter to acknowledge the write.

Option c: LOCAL_QUORUM - **Incorrect**. LOCAL_QUORUM requires a quorum of replicas within the same datacenter to acknowledge the write.

Option d: ALL - **Correct**. ALL requires all replicas to acknowledge the write before returning a response.

5. How does Zookeeper contribute to maintaining consistency in distributed systems?

A) By managing data replication

B) By providing a centralized configuration service

C) By ensuring data encryption

D) By optimizing data storage

Option A: By managing data replication – Incorrect. Zookeeper's role is more about coordination, not directly managing data replication.

Option B: By providing a centralized configuration service – Correct.

Zookeeper contributes to maintaining consistency in distributed systems by providing a centralized coordination and configuration service, ensuring consistent synchronization across distributed nodes.

Option C: By ensuring data encryption – Incorrect. Zookeeper doesn't handle encryption.

Option D: By optimizing data storage – Incorrect. Zookeeper is not involved in optimizing data storage.

Explanation:

Centralized Configuration: ZooKeeper acts as a centralized service where distributed applications can store and retrieve configuration information. This helps in ensuring that all nodes in the distributed system have consistent configuration settings, reducing the chances of configuration mismatches or inconsistencies.

6. A ___________ server is a machine that keeps a copy of the state of the entire system and persists this information in local log files.

a) Master
b) Region
c) Zookeeper
d) All of the mentioned

Option A: Master – Incorrect. The master server may manage parts of the system but does not persist the full system state in local logs.

Option B: Region – Incorrect. A region server manages a subset of data, but it doesn't maintain the full system state or persist it in logs.

Option C: Zookeeper – Correct. Zookeeper maintains a consistent view of the system's state and stores this information in local log files for fault tolerance and recovery.

Option D: All of the mentioned – Incorrect. Only Zookeeper is responsible for persisting the state of the entire system in local logs.

Explanation:

Master Server: A master server typically coordinates tasks within a cluster but doesn't necessarily store the entire system state.

Region Server: This term is often used in context of distributed databases like HBase, where region servers manage specific data partitions. They wouldn't hold the entire system state.

Zookeeper

Zookeeper is a centralized service that coordinates and manages distributed systems. It keeps a copy of the system's state and persists this information in local log files. This allows it to provide services such as naming, configuration management, and synchronization.

While a Master node might also have some state information, its primary role is often different, such as coordinating tasks or managing data. A Region node is typically a unit within a larger distributed system, and its role might involve managing specific data or tasks.

7.What is Apache Zookeeper primarily used for in Big Data ecosystems?

A) Data storage

B) Data processing

C) Configuration management

D) Data visualization

Option A: Data storage – **Incorrect.** Zookeeper is not designed for storing large amounts of data; its main purpose is coordination, not data storage.

Option B: Data processing – **Incorrect.** Zookeeper does not process data; it provides coordination services for distributed systems.

Option C: Configuration management – **Correct.** Zookeeper is primarily used for configuration management, leader election, and synchronization in distributed systems within Big Data ecosystems.

Option D: Data visualization – **Incorrect.** Zookeeper has no role in data visualization. Its function is more about system coordination and management.

8. Which statement correctly describes CQL (Cassandra Query Language)?

a. CQL is a SQL-like language used for querying relational databases
b. CQL is a procedural programming language used for writing stored procedures in Cassandra
c. CQL is a language used for creating and managing tables and querying data in Apache Cassandra
d. CQL is a scripting language used for data transformation tasks in Cassandra

Option A: CQL is a SQL-like language used for querying relational databases – Incorrect. While CQL is SQL-like, Cassandra is a NoSQL database, not a relational database.

Option B: CQL is a procedural programming language used for writing stored procedures in Cassandra – Incorrect. CQL is not a procedural language, nor is it used for writing stored procedures

.

Option C: CQL is a language used for creating and managing tables and querying data in Apache Cassandra – Correct. CQL is primarily used in Cassandra for creating, managing tables, and querying data.

Option D: CQL is a scripting language used for data transformation tasks in Cassandra – Incorrect. CQL is not a scripting language and is not designed for data transformation tasks.

9.Which aspect of CAP theorem refers to a system's ability to continue operating despite network failures?

A) Consistency

B) Accessibility

C) Partition tolerance

D) Atomicity

Option A: Consistency – Incorrect. Consistency refers to ensuring that all nodes see the same data at the same time, not handling network failures.

Option B: Accessibility – **Incorrect.** Availability refers to the system's ability to respond to requests, but does not specifically address network partitioning.

Option C: Partition tolerance – **Correct.** Partition tolerance refers to the system's ability to continue functioning even when network failures or partitions occur.

Option D: Atomicity – **Incorrect.** Atomicity is a concept related to transactions, ensuring that operations are fully completed or not at all, not related to network failures.

10. Why are tombstones used in distributed databases like Apache Cassandra?
    a. To mark nodes that are temporarily unavailable
    b. To mark data that is stored in multiple replicas
    c. To mark data that has been logically deleted
    d. To mark data that is actively being updated

Option a: To mark nodes that are temporarily unavailable - **Incorrect**. Tombstones are not used to mark unavailable nodes.

Option b: To mark data that is stored in multiple replicas - **Incorrect**. Tombstones are not used to mark data replication.

Option c: To mark data that has been logically deleted - **Correct**. Tombstones are used to mark data that has been deleted but still exists in the system for a certain period to prevent accidental overwrites.

Option d: To mark data that is actively being updated - **Incorrect**. Tombstones are not used to mark data that is being updated.