Kafka Most Asked Interview Questions

Q1. What is Apache Kafka?

Apache Kafka is an open-source stream-processing software platform developed by the Apache Software Foundation, written in Scala and Java. It is designed to handle data streams from multiple sources and deliver them to various consumers efficiently.

Q2. What are the main components of Kafka?

The main components of Kafka include topics, producers, consumers, brokers, partitions, and ZooKeeper. Each plays a crucial role in the messaging and storage system of Kafka.

Q3. What is a Kafka Broker?

A Kafka broker is a single Kafka server in a Kafka cluster. It receives messages from producers, assigns offsets to them, and commits the messages to storage on disk. A Kafka cluster consists of multiple brokers to ensure load balancing and fault tolerance.

Q4. What is a Kafka Producer?

A Kafka producer is an entity or application that publishes data to topics in the Kafka cluster. Producers send data to Kafka brokers, and each record consists of a key, a value, and a timestamp.

Q5. What is a Kafka Consumer?

A Kafka consumer is an entity or application that subscribes to one or more topics and processes the stream of records produced to them. Consumers pull data from Kafka brokers.

Q6. Explain Kafka Topics and Partitions.

A Kafka topic is a category or feed name to which records are published. Topics in Kafka are divided into partitions, which allow Kafka to parallelize processing by splitting the data across multiple brokers.

Q7. What is ZooKeeper in the context of Kafka?

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. In Kafka, ZooKeeper is used to manage the cluster metadata and to maintain the list of brokers, topics, and partitions.

Q8. How does Kafka ensure fault tolerance?

Kafka ensures fault tolerance through the use of replication. Each topic can be configured to replicate its partitions across multiple brokers. If one broker fails, other brokers can serve the data to ensure high availability.

Q9. What is offset in Kafka?

An offset in Kafka is a sequential id number assigned to each message within a partition. It is used to uniquely identify each message and track which messages have been consumed by maintaining the offset checkpoint.

Q10. How does Kafka handle load balancing?

Kafka handles load balancing using partitions across brokers in the cluster. Each partition can be consumed by only one consumer in a consumer group at any given time, which balances the load across consumers.

Q11. What is a consumer group in Kafka?

A consumer group in Kafka is a group of consumers that cooperate to consume data from one or more topics. The consumers in a group divide the topic partitions among themselves to share the workload.

Q12. Explain the concept of message retention in Kafka.

Kafka allows administrators to set a retention policy for messages, which determines how long messages should be kept before they are deleted. Retention can be based on time or size limits, or both.

Q13. What are the advantages of using Kafka over traditional message brokers?

Kafka offers advantages like high throughput, built-in partitioning, replication, and fault tolerance which make it suitable for large-scale message processing applications. It's designed to function as a distributed system, scalable both horizontally and vertically.

Q14. What is log compaction in Kafka?

Log compaction in Kafka is a feature that helps to prevent the data in a topic from growing indefinitely. It ensures that a compacted topic only retains the last message for each key within the topic's log, thus reducing space and overhead.

Q15. How can you secure a Kafka cluster?

Securing a Kafka cluster can be achieved through methods such as enabling TLS/SSL for data-in-transit encryption, using SASL for authentication, and implementing ACLs (Access Control Lists) to control permissions for topics, consumer groups, and Kafka brokers.

Q16. What is the role of the Kafka Connect API?

The Kafka Connect API is a framework for connecting Kafka with external systems such as databases, key-value stores, search indexes, and file systems. It is used for streaming data between Kafka and other systems in a scalable and reliable manner.

Q17. What is Kafka Streams?

Kafka Streams is a client library for building applications and microservices where the input and output data are stored in Kafka clusters. It enables real-time data processing and analysis.

Q18. What are Kafka MirrorMaker and its use?

Kafka MirrorMaker is a tool used for mirroring data between two Kafka clusters. It is commonly used for cross-datacenter replication of Kafka messages to ensure higher availability and disaster recovery.

Q19. What types of serializers are supported in Kafka?

Kafka supports several serializers out of the box, including String, Byte array, and Integer serializers. Custom serializers can be created by implementing the Serializer interface in Kafka.

Q20. How does Kafka handle rebalancing?

Kafka handles rebalancing in consumer groups to distribute topic partitions evenly among the consumers. This occurs when new consumers join a group, existing consumers leave the group, or when the topics being consumed change. Rebalancing ensures optimal distribution of partitions among consumers for load balancing.