7. High-Water Mark

Let's learn about high-water mark and its usage.

We'll cover the following

Background#

Distributed systems keep multiple copies of data for fault tolerance and higher availability. To achieve strong consistency, one of the options is to use a leader-follower setup, where the leader is responsible for entertaining all the writes, and the followers replicate data from the leader.

Each transaction on the leader is committed to a write-ahead log (WAL), so that the leader can recover from crashes or failures. A write request is considered successful as soon as it is committed to the WAL on the leader. The replication can happen asynchronously; either the leader can push the mutation to the followers, or the follower can pull it from the leader. In case the leader crashes and cannot recover, one of the followers will be elected as the new leader. Now, this new leader can be a bit behind the old leader, as there might be some transactions that have not been completely propagated before the old leader crashed. We do have these transactions in the WAL on the old leader, but those log entries cannot be

recovered until the old leader becomes alive again. So those transactions are considered lost. Under this scenario, the client can see some data inconsistencies, e.g., the last data that the client fetched from the old leader may not be available anymore. In such error scenarios, some followers can be missing entries in their logs, and some can have more entries than others. So, it becomes important for the leader and followers to know what part of the log is safe to be exposed to the clients.

Definition#

Keep track of the last log entry on the leader, which has been successfully replicated to a quorum of followers. The index of this entry in the log is known as the High-Water Mark index. The leader exposes data only up to the high-water mark index.

Solution#

For each data mutation, the leader first appends it to WAL and then sends it to all the followers. Upon receiving the request, the followers append it to their respective WAL and then send an acknowledgment to the leader. The leader keeps track of the indexes of the entries that have been successfully replicated on each follower. The high-water mark index is the highest index, which has been replicated on the quorum of the followers. The leader can propagate the high-water mark index to all followers as part of the regular Heartbeat message. The leader and followers ensure that the client can read data only up to the high-water mark index. This guarantees that even if the current leader fails and another leader is elected, the client will not see any data inconsistencies.

Examples#

Kafka: To deal with non-repeatable reads and ensure data consistency, Kafka brokers keep track of the high-water mark, which is the largest

offset that all In-Sync-Replicas (ISRs) of a particular partition share. Consumers can see messages only until the high-water mark.

← Back

Next →

6. Segmented Log

8. Lease

Mark as Completed

Report an Issue