BookKeeper

BookKeeper

BookKeeper is a system to reliably log streams of records. It is designed to store write ahead logs, such as those found in database or database like applications. In fact, the Hadoop NameNode inspired BookKeeper. The NameNode logs changes to the in-memory namespace data structures to the local disk before they are applied in memory. However logging the changes locally means that if the NameNode fails the log will be inaccessible. We found that by using BookKeeper, the NameNode can log to distributed storage devices in a way that yields higher availability and performance. Although it was designed for the NameNode, BookKeeper can be used for any application that needs strong durability guarantees with high performance and has a single writer.

In BookKeeper, servers are "bookies", log streams are "ledgers", and each unit of a log (aka record) is a "ledger entry". BookKeeper is designed to be reliable; bookies, the servers that store ledgers can be byzantine, which means that some subset of the bookies can fail, corrupt data, discard data, but as long as there are enough correctly behaving servers the service as a whole behaves correctly; the meta data for BookKeeper is stored in ZooKeeper.

BookKeeper achieves high availability and strong durability guarantees by replicating ledger entries across multiple bookies. The ledgers themselves are striped across the bookies for high performance.

The BookKeeper data model is a flat namespace of ledgers identified by a long. The ledgers are append only and writable by a single client. The basic operations of BookKeeper are: createLedger to create a new ledger available for writing, openLedger to read from an existing ledger, addEntry, removeEntry, and closeLedger. Once a ledger is closed it becomes read-only.

Documentation:

- 3.2 Documentation

What is going on:

- Bookie recovery (BookieRecoveryPage)
- Bookie registration and failure detection (BookieRegPage)
- Ledger deletion (LedgerDeletionPage)
- Performance numbers(BookKeeperPerfPage)