Raft 一致性算法

实现分布式一致性

Jiang Wu

EMC CTD

2016-06-25

大纲

Consensus(一致)

定义

一致，共识

应用

实现分布式系统的一致性 (consistency)

传统的一致性实现方法

- 锁 (行/表)
- 基于 version 的乐观锁
- 事务

在大并发情况下慢

其他的一致性问题

- 依赖特定数据库的实现

大纲

Concepts

- Node
- State: follower, candidate, leader
- Replicate: log, FSM
- Leader Election

Live demos

- thesecretlivesofdata.com/raft
- raft.github.io

大纲

Go raft

- github.com/hashicorp/raft
- github.com/hashicorp/raft-mdb
- github.com/hashicorp/raft-boltdb
- github.com/otoolep/hraftd

Demo of hraftd

- Clusterd key-value store
- Replicated through raft
- Leader election

vs. DB level consistency

Pros

- Lower learning curve, stick with Raft protocol
- Support cross DB replication
- Lightweight

Cons

- Have to write replication by hand, a lot of code to maintain

Migration

- Add Raft endpoints
- Reconstruct Raft log
- Start new nodes with new db
- Join new nodes with existing nodes
- Shutdown nodes with old db

Future

- Replicate through nodes with various backends
- Replicate across datacenters
- Tunable consistency