



東北大學 秦皇島分校
Northeastern University at Qinhuangdao

毕业论文

基于 LSM-Tree 结构和 Raft 算法的分布式存储
系统

院 别	计算机与通信工程学院
专业名称	计算机科学与技术
班级学号	1901-20197897
学生姓名	华令楠
指导教师	吕艳霞

2023 年 5 月 20 日

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：

日期：

基于 LSM-Tree 结构和 Raft 算法的分布式存储系统

摘 要

随着互联网的发展，网络用户的激增导致互联网服务提供公司要存储极大规模的数据，而对于一些复杂场景下的数据复制以及分布式系统下数据库的可靠性仍然是一个巨大的挑战。传统的互联网架构时代，单机数据库如 MySQL，Oracle 占领很大的市场份额，而单机数据库有很多缺点，其中一些缺点如下：成本高：它们通常比分布式数据库系统更加复杂和成本更高，因为它们需要专门的硬件、存储设备和管理软件。维护困难：由于它们是单独的系统，因此一旦出现问题，修复它们可能需要很长时间和高昂的成本。不易扩展：当需要增加新功能或存储容量时，单机数据库系统可能无法轻松地扩展。数据共享和备份困难：由于它们是单独的系统，数据不能轻松地在它们之间共享或备份。安全性差：由于它们是单独的系统，数据很容易受到黑客攻击和数据泄露。而分布式的数据存储恰恰解决了单机数据库的性能瓶颈问题和单机数据库在实现面向用户系统时的各种痛点。

由于一个可靠的分布式存储系统设计复杂，挑战较大，本文致力于系统的存储策略，数据压缩方法，日志复制同步过程，Raft 算法的可达性分析。我们利用 Golang 编程语言的天然并发的特性，来开发 Raft 共识算法和 LSM-tree 数据结构以实现分布式数据存储系统。我们的系统通过跨节点集群复制数据并使用 Raft 共识算法确保副本之间的一致性来提供容错性、可扩展性和高可用性。LSM-tree 数据结构通过优化磁盘访问和减少随机查找次数来实现高效的读写。通过不断精读原文和阅读参考并 demo 出 leveldb 的 LSM-Tree 实现以便尽最大限度复现 LSM-Tree 和 Raft 论文中提及的所有关键点，我们的评估表明，我们的系统在保持强大的一致性保证的同时实现了高性能和可扩展性。

关键词： 日志结构归并树，Raft 共识性算法，数据存储，分布式系统

Distributed storage system based on LSM-Tree structure and Raft algorithm

Abstract

With the development of the Internet, the surge of network users has led Internet service providers to store extremely large-scale data, but data replication in some complex scenarios and the reliability of databases in distributed systems are still a huge challenge. In the era of traditional Internet architecture, stand-alone databases such as MySQL and Oracle occupy a large market share, but stand-alone databases have many shortcomings, some of which are as follows: High cost: They are usually more complex and costly than distributed database systems because they require specialized hardware, storage devices, and management software. Difficult to maintain: Since they are separate systems, it can take a long time and be costly to fix if something goes wrong. Not easy to expand: When new functions or storage capacity need to be added, a stand-alone database system may not be easily expanded. Difficulty in data sharing and backup: Since they are separate systems, data cannot be easily shared or backed up between them. Poor security: Since they are separate systems, the data is vulnerable to hacking and data breaches. The distributed data storage just solves the performance bottleneck problem of the stand-alone database and the various pain points of the stand-alone database when implementing the user-oriented system.

Since the design of a reliable distributed storage system is complex and challenging, this paper focuses on the system's storage strategy, data compression method, log replication synchronization process, and the reachability analysis of the Raft algorithm. We use the natural concurrency of the Golang programming language to develop the Raft consensus algorithm and the LSM-tree data structure to implement a distributed data storage system. Our system provides fault tolerance, scalability, and high availability by replicating data across a cluster of nodes and using the Raft consensus algorithm to ensure consistency between replicas. The LSM-tree data structure enables efficient reads and writes by optimizing disk access and reducing the number of random lookups. By continuously intensively reading the original text and reading references and demoing the LSM-Tree implementation of leveladb in order to reproduce all the key points

mentioned in the LSM-Tree and Raft papers as much as possible, our evaluation shows that our system is maintaining strong consistency High performance and scalability are guaranteed at the same time.

Keywords: Log-Structured Merge-Tree, Raft consensus algorithm, Data Storage, Distributed System

目录

1	绪论	1
1.1	课题的背景和意义	1
1.2	分布式存储系统的发展状况	1
1.3	课题研究的主要方法及内容	2
1.4	论文组织结构	2
2	相关背景知识介绍	3
2.1	开发工具和环境	3
2.2	环信即时通讯云	3
2.3	涉及的开源库	3
3	系统需求分析	4
3.1	系统需求概述	4
3.2	功能需求分析	5
3.3	性能需求分析	6
4	系统总体设计	8
4.1	服务端总体设计	8
4.2	客户端总体设计	9
4.3	数据库总体设计	10
5	系统详细设计	12
5.1	基础层详细设计	12
5.1.1	网络通信的实现	12
5.1.2	日志记录的实现	13
5.1.3	加密解密的实现	13
5.1.4	缓存的实现	13
5.1.5	文件操作工具的实现	14
5.1.6	JSON 解析的实现	14

5.2	数据层详细设计	14
5.3	组件层详细设计	14
5.3.1	二维码扫描的实现	14
5.3.2	消息列表的实现	15
5.3.3	输入面板的实现	15
5.3.4	表情面板的实现	16
5.3.5	更多面板的实现	16
5.3.6	通讯录列表的实现	16
5.3.7	埋点统计的实现	16
5.4	表现层详细设计	17
5.4.1	Activity 基类的实现	17
5.4.2	Fragment 基类的实现	17
5.5	应用层详细设计	17
5.5.1	用户系统的实现	17
5.5.2	聊天功能的实现	17
5.5.3	好友管理的实现	18
5.5.4	群组管理的实现	18
5.5.5	通讯录的实现	18
5.5.6	应用列表的实现	18
5.5.7	个人中心的实现	18
6	系统部署与测试	20
6.1	服务端部署	20
6.2	客户端测试	20
6.2.1	单元测试	20
6.2.2	功能测试	20
6.2.3	深度兼容测试	20
	结论	21
	致 谢	22

附 录.....	23
附录 A	23
附录 B	23
参考文献.....	23

1 绪论

1.1 课题的背景和意义

分布式存储系统是一个重要且不断发展的领域，它涉及到许多学科和技术，如计算机科学、网络安全、数据库管理等。分布式存储系统的研究主要集中在三个方面：分布式存储系统的架构设计、底层协议的研究以及应用场景的拓展。其中，架构设计包括存储系统的硬件架构、软件框架等；底层协议的研究则包括各种存储接口、协议栈等；应用场景的拓展则涵盖了企业级、消费级、个人级等不同类型的存储系统。分布式存储系统的研究具有广泛的应用前景，它可以应用于企业级存储系统、云存储服务和分布式应用等方面。在企业级存储系统中，分布式存储系统可以用于存储大量的数据，提高数据存储的效率和可靠性。在云存储服务中，分布式存储系统可以用于存储云端的数据，提供高效的数据存储和管理服务。在分布式应用中，分布式存储系统可以用于存储和管理大量的数据，提高应用的可靠性和性能。在分布式存储系统的研究中，需要解决的关键问题包括数据一致性、数据持久性、数据备份和恢复等。在解决这些问题的过程中，需要采用一些新的技术和方法，如分布式数据库、分布式文件系统、分布式对象存储等。同时，需要考虑分布式存储系统的安全性和可靠性问题，采用一些新的安全技术和机制，如安全协议、数据加密、访问控制等，以保证分布式存储系统的安全性和可靠性。

1.2 分布式存储系统的发展状况

分布式存储系统的发展历史可以追溯到上世纪 90 年代，当时出现了一些基于局部存储器的分布式存储系统，如 Lustre 和 Xanadu 等。这些系统主要用于文件服务器等领域。随着网络技术的发展，基于网络的分布式存储系统出现了，如 Hadoop 和 HDFS 等。这些系统将数据存储分布在分布式节点上，并通过网络进行数据的访问和管理。近年来，随着云计算的发展，分布式存储系统开始广泛应用于云存储服务中。云存储服务将数据存储在云端，并通过互联网提供数据的访问和管理。这些系统通常采用分布式文件系统，如 AFS、HDFS 和 Gluster 等。随着大数据时代的到来，分布式存储系统的研究和应用也越来越广泛。基于分布式文件系统的分布式存储系统可以存储海量的数据，并支持高效的数据存储和管理。同时，基于块存储的分布式存储系统也得到了广泛的研究和应用，它可以实现数据的高效复制和同步，并支持大规模的数据存储和管理。总之，分布式存储系统的发展历史可以分为三个阶段：基于局部存储器的分布式存储系统、基于网络的

分布式存储系统和基于块存储的分布式存储系统。随着大数据时代的到来，分布式存储系统的研究和应用也将越来越广泛和深入。

1.3 课题研究的主要方法及内容

本课题主要工作是...

本课题主要包含以下几个方面内容：

- 1、调研主流即时通讯软件的功能...
- 2、深入研究...
- 3、设计实现...
- 4、实现整个...

1.4 论文组织结构

本文主要围绕相关技术选型，需求分析，系统整体设计、详细设计，部署与测试等方面来进行论述，共分为 6 章，各章内容如下：

第 1 章...

第 2 章...

第 3 章...

第 4 章...

第 5 章...

第 6 章...

为了更好的理解...

2.1 开发工具和环境

Android Studio 省略一段文字...

SDK (Software Development Kit) 省略一段文字...

对于移动端来说，省略一段文字...

环信成立于 2013 年 4 月，省略一段文字...，引用的用法... 环信^[9]。

1、网络请求库 Retrofit 2.0

Retrofit 是一个 Square 开发的类型安全的省略一段文字...

ButterKnife 是一个基于注解的视图绑定库省略一段文字...

LeakCanary 是一个监测省略一段文字...

Logger 是一个简单省略一段文字...

ZXing 是一个省略一段文字...

TinyPinyin 是一个省略一段文字省略一段文字省略一段文字省略一段文字省略一段
文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省
略一段文字

3 系统需求分析

3.1 系统需求概述

[illegible]

图 3.1 系统需求概述

[illegible]

5、性能需求 5

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

4 系统总体设计

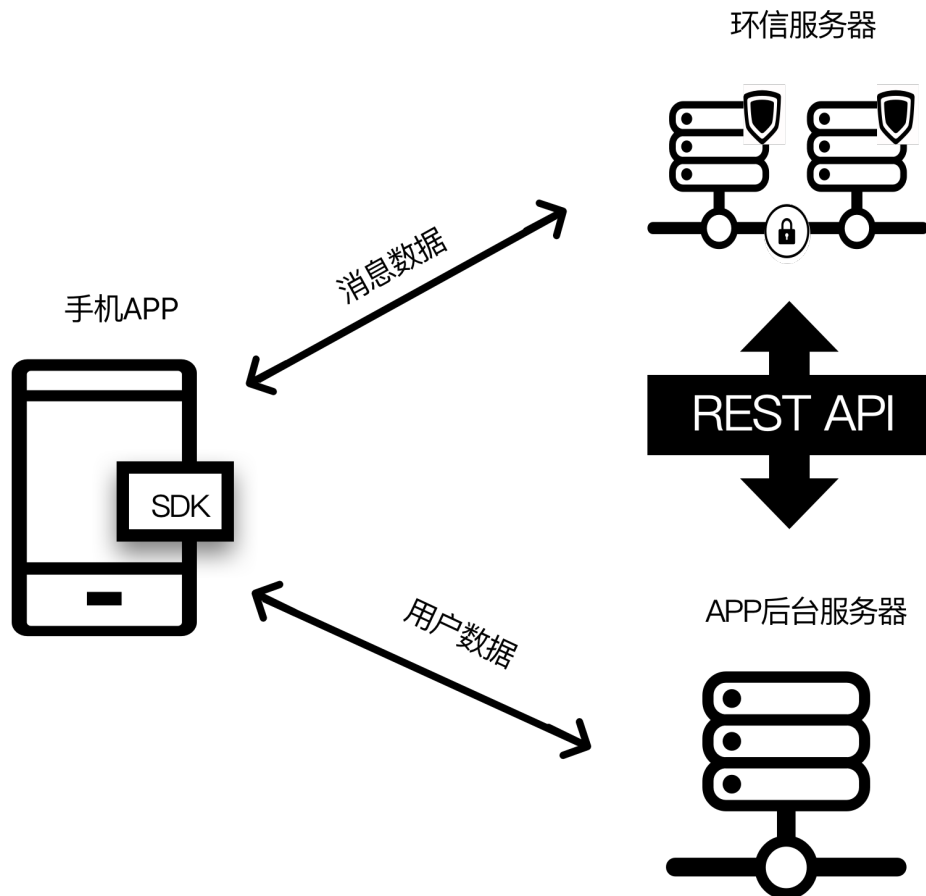
[illegible]

图 4.1 系统总体结构组成

4.1 服务端总体设计

服务端的实现省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省略一段文字省

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

4.2 客户端总体设计

客户端采用分层的架构进行设计，上层实现依赖于下层设计^[7]，自下而上分为基础层、数据层、组件层、表现层、应用层，整体设计结构如图 4.2 所示。



图 4.2 移动端总体设计结构

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

2、设计概述

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字
省略一段文字省

省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省，表格使用，
表格使用，表格使用，表格使用，表格使用，表格使用，如表 4.1 所示。

表 4.1 User 用户表

字段	数据类型	字段含义	约束条件
id	INT(11)	用户 ID	主键、非空、自增
im_id	VARCHAR(256)	环信 ID	唯一
account	VARCHAR(45)	用户名	唯一
nick_name	VARCHAR(100)	昵称	无
password	VARCHAR(256)	密码	非空
email	VARCHAR(45)	邮箱	无
mobile	VARCHAR(45)	手机号	唯一
sex	INT(11)	性别	无
signature	VARCHAR(512)	签名	无
avatar	VARCHAR(256)	头像	无
is_deleted	TINYINT(4)	删除标志	无

[illegible]

5.1.1 网络通信的实现

5.1.1 网络通信的实现

这里以代码清单 5.1 为例，代码块使用示例

代码清单 5.1 APIService

```
/**
 * 登录
 */
@POST("user/login")
Call<ApiResponse<LoginResponse>> login(@Body LoginRequest request);

/**
 * 获取单个用户详细信息
 */
@GET("user/{imid}/info")
Call<ApiResponse<UserInfo>> getUserInfo(@Path("imid") String imId);
```

5.1.2 日志记录的实现

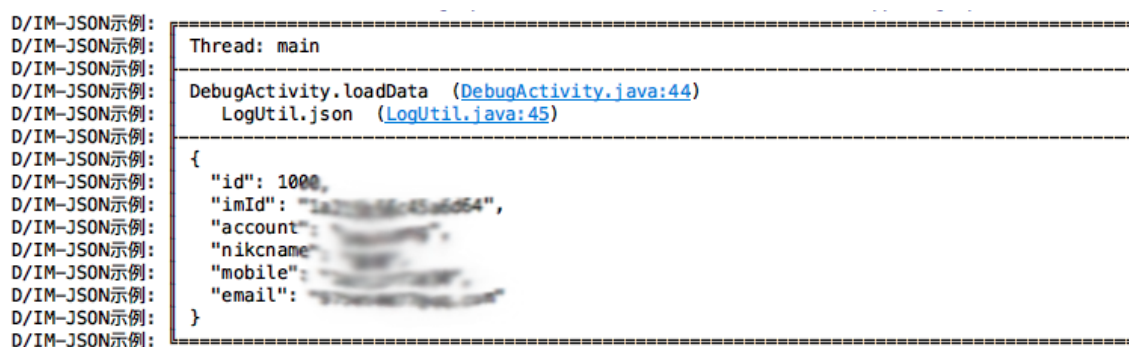
[illegible]

图 5.1 日志信息示例

5.1.3 加密解密的实现

[illegible]

5.1.4 缓存的实现

[illegible]

实现省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段
段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段

实现省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段
段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段

[illegible]

5.4 表现层详细设计

5.4.1 Activity 基类的实现

[illegible]

5.4.2 Fragment 基类的实现

[illegible]

5.5 应用层详细设计

5.5.1 用户系统的实现

[illegible]

5.5.2 聊天功能的实现

[illegible]

[illegible]

文字省略一段文字省省略一段文字省略一段文字省略一段文字省略一段文字省略一段
文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文字省略一段文
字省略一段文字省略一段文字省

[illegible]

结论

[illegible]

- 1、调研了...
- 2、调研...
- 3、设计...
- 4、依据...
- 5、对整个系统...

[illegible]

致 谢

[illegible]

Activities

Services

Broadcast receivers

Content providers

[illegible]

[illegible]

[illegible]