压缩态内存数据库实时算法设计与实现

系统概要设计说明书

V1.0

小组名称：Never give up
小组口号：Make the change
指导教师：赵振刚老师
文档撰写人：刘勇
文档撰写时间：2013 年 11 月 15 日

团队分工记录表

项目名称	学号	姓名	分工
压缩态内存数据库实时算法设计与实现	SG13225025	蓝鸿翔	
	SA13226282	刘勇	系统概要设计说明书编写
	SG13225022	范亚林	

目录

1. 前言

1.1 编写目的

本文档适用于压缩态内存数据库实时算法设计与实现项目，是基于压缩态内存数据库实时算法设计与实现的需求分析编写的。

压缩态内存数据库是数据放在内存中直接操作,同时根据数据库的列属性特性进行压缩的数据库。它能够满足目前电信行业和金融行业日益增长的对实时数据处理的需求，同时解决了数据量和存储空间矛盾的严峻问题。

1.2 预期读者和阅读建议

本说明书的主要目的是明确所要开发的系统应具有的功能、结构和接口，使开发人员能清楚地了解系统的整体框架，并在此基础上进一步设计和开发，为软件开发范围、业务处理规范提供依据，也是应用软件进行最终验收的依据。

该报告的可能的读者有：

- 用户；
- 开发人员；
- 项目经理；
- 测试人员；
- 文档编写人员；
- 等等。

其应用范围见表 1：

读者分类	目的
用户	了解本文档对系统的理解是否和他们要求的一致
开发人员	理解整体架构，进行编码
项目经理	理解用户，在设计时和开发人员交流
系统测试人员	了解接口，为测试提供参考
文档人员	编写用户使用和操作手册

表1：本文档应用范围

1.3 参考资料

[1] 荣垂田.一个内存数据库模型的设计与实现 [D].辽宁：中国科学院沈阳计算技术研究所，2008
[2] 王珊 肖艳芹等.内存数据库关键技术研究. 计算机应用[J] Vol.27 2007(10):2354-2355
[3] 数据库系统概念 Abraham Silberschatz ,Henry E Korth S.Sudarshan 著，杨冬青 唐世渭等译，机械工业出版社,2003:244
[4] James E Kurose , Keith W. Ross 著，陈鸣译.计算机网络-自顶向下方法与 Internet 特色[M].北京：机械工业出版社，2006
[5] 黄 鹏,李占山等.基于列存储数据库的压缩态数据访问算法[D].吉林：吉林大学，2009
[6] 马洪连 杨 波等.一个适用于内存数据库系统的多维索引结构[D]. 辽宁：大连理工大学，2003
[7] 俞甲子，石凡，潘爱民. 程序员的自我修养: 链接、装载与库[M].电子工业出版社,2009:32-90

2. 设计概述

2.1 限制和约束

■ 技术条件：

本小组的选题压缩态内存数据库实时算法设计与实现将与主流的数据库产品中使用的实时算法进行对比，以增强算法效率，节省时间为目标。该选题的技术难点有以下几点：

（1）在较高的压缩比的情况下实现较优的存取性能；
（2）自主压缩学习算法的设计；
（3）字典压缩字典选取范围和粒度选取的难度；
（4）如何进行列间关联；
（5）如何在列式存储压缩的情况下高效完成数据的修改，插入，删除；
（6）如何在在树形结构中兼顾压缩效率和存取效率。

■ 时间限制：

由于时间有限和本小组人员数目的限制，所以工作量较大。

■ 开发环境：

宿主机操作系统：Ubuntu 12.04
开发工具：VIM, GDB
对比数据库：Sqlite
编程语言：C/C++

2.2 设计原则和设计要求

该系统的设计过程中我们遵循以下原则和标准：

- 健壮性：
  该系统中每个函数都需要有参数检查，尤其对指针的处理需要检查。

- 容易理解：
  所有函数的编写都用英文名直接可以看出。

- 程序简便：
  在编写代码中做到逻辑清晰，易读性强。

- 标准化原则：
  在结构上实现开放，基于业界开放式标准，符合国家和信息产业部的规范。

- 可扩展性：
  该系统中各个功能预留一定参数接口，以便后面扩展功能时使用。

- 系统易操作性要求：
  该系统在设计上，可能没有用户直接操作界面，需要使用命令行的形式，应尽可能做到操作简便，如多条命令时可以在 linux 下编写 shell 脚本。

- 系统可维护性要求；
  系统中代码的注释要做到详细，并达国标规定，注释率为代码量的 25%～50%，可重用性要强。

- 等等。

3. 系统逻辑设计

3.1 系统结构设计

该系统的整体功能模块应分为数据导入模块、列属性统计模块、压缩算法调度模块、数据压缩模块、数据操作模块、数据导出模块，各个模块的交互情况从图 1 中看出。。

该系统的整体架构，如下图所示：

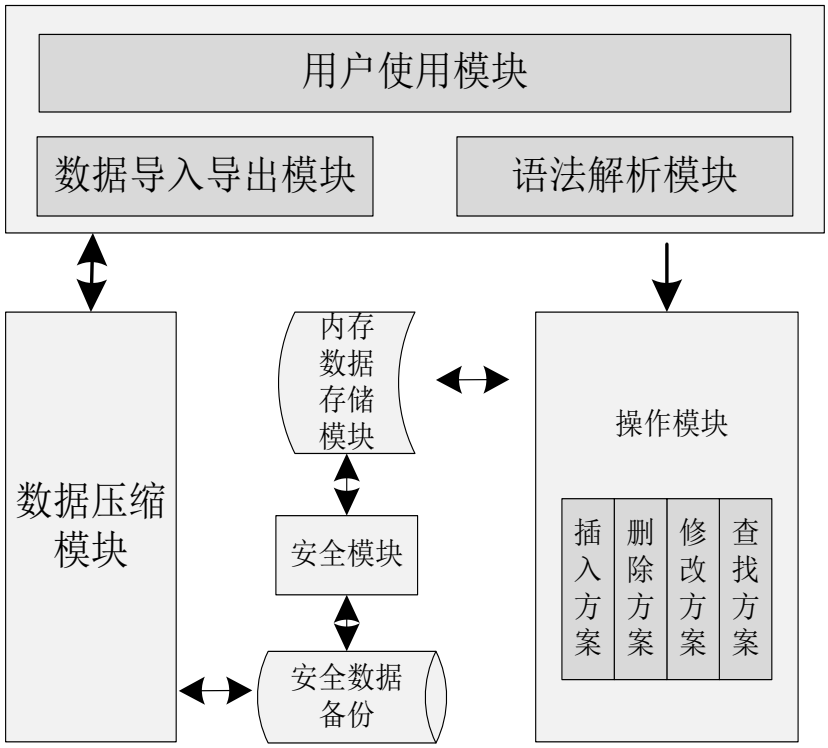

图 1：系统整体架构设计

该系统的业务逻辑处理过程如图 2 所示：

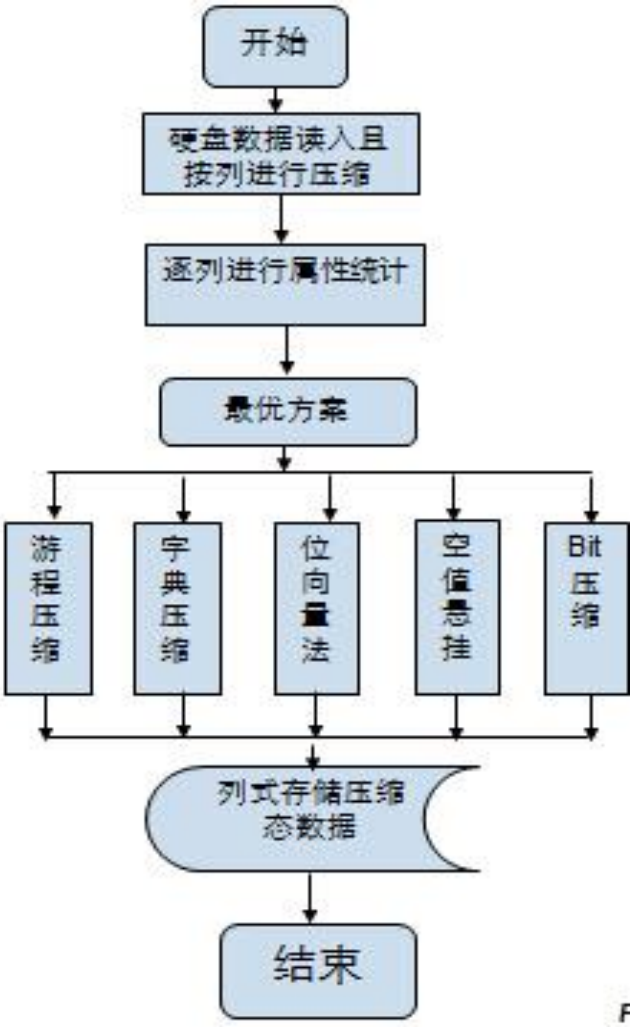

图 2：系统的业务逻辑流程图

执行流程如下：

（1） 用户执行命令将数据库文件导入内存；
（2） 该系统对每列数据属性进行统计；
（3） 根据每列数据属性的不同，计算不同压缩方案的空间消耗；
（4） 根据以上 5 种压缩方案的不同，由调度算法选择最佳的压缩方案；
（5） 系统存储列式的压缩态数据；
（6） 等待用户操作，此时，用户可以根据需要进行数据的增删改查操作；
（7） 用户操作结束，系统等待用户下一步操作
（8） 用户没有其他任务要做，可将原始态、压缩态数据导出。

3.2 系统模块间交互关系设计

- **数据导入模块：**

  输入：csv 格式的硬盘表格数据

  该 csv 文件满足一下条件：

  首行 ： 存放该 csv 文件的行数 rows 和列数 cols；

  次行 ： 存放每列数据的属性；

  剩下 ： 都是数据本身；

  输出：内存表格数据

- **属性统计模块：**

  输入：内存表格数据；

  功能：hash 统计每列数据属性,包括每列数据相同字段不同字段空值字段的个数；

  输出：各个字段的值，以便后面使用。

- **策略选择模块**

  输入：每列数据的属性特征，即属性统计模块 hash 统计的结果；

  功能：计算采用不同压缩方案的空间消耗情况；

  输出：无。

- **数据压缩模块：**

  输入：csv 格式文件的每列原始数据；

  功能：调用相应压缩算法进行压缩；

  输出：压缩态数据。

- **数据操作模块**

  输入：主键值；

  功能：根据主键值查询相应记录，并对该记录进行相应操作；

  输出：根据操作的不同，返回结果也不尽相同，如果是查询，返回查询的记录。如果是插入删除修改不返回结果，只返回操作是否成功。

3.3 系统接口设计

3.3.1 用户界面设计规则

尽可能设计成面向用户的通用 SQL 接口，符合用户需求的、美观大方的用户界面，如果时间不允许，编写成 shell 脚本从而简化多个命令行的操作，方便用户使用。

3.3.2 内部接口设计

各模块之间相互独立，低耦合高内聚。根据系统的整体架构图进行设计。
各模块根据文档内部控制域值提取其所需的文档。

3.3.3 外部接口设计

与硬件之间的接口：无
与软件之间的接口：linux 库，sqlite3 接口。

4. 出错处理设计

出错处理：在错误发生时，给出出错的原因。

5. 系统维护设计

采用模块化的设计，方便维护。