《数据库原理》大作业项目

2022年5月25日

## 一、选题概述

选题一：基于关系数据库理论，选择某个感兴趣的模块，对HSQLDB深入分析与优化，形成一篇数据库实现技术的学术报告。

选题二：基于给出的ThssDB框架，开发相关的代码模块，实现一个简单的关系数据库管理系统，并提供设计文档和用户文档（含运行方式）。

## 二、HSQLDB系统优化（选题一）

需要完成HSQLDB相关模块分析，进行性能优化并给出实验验证。

## **2.1**相关模块分析

自行确定HSQLDB的相关模块（如存储，索引，查询等），通过小组讨论，深入阅读该模块的代码结构，形成一份该模块的逆向软件工程设计文档，**可以采用UML类图、顺序图等方式**呈现该模块的设计核心思想。

## **2.2**优化方案与代码实现

基于上述分析工作，参照课程讲授的内容，分析现有设计的优点和存在的问题。

详细描述针对现有问题的优化方案，使用理论分析或者代价模型说明其优化原理。给出上述优化方案的实现路径，**可以采用UML类图、顺序图等方式**表述其详细设计，并在HSQLDB系统中进行代码实现。

## **2.3**性能实验分析

针对优化工作给出相应的性能分析实验方案，包括数据集、事务负载以及性能指标。报告需要给出实验目的、实验步骤和结果分析。如果优化有多个步骤，分析每一个步骤带来的收益。

## 三、ThssDB系统实现（选题二）

基于给出的ThssDB框架，完成SQL查询与并发事务两个模块。

## 3.1查询模块

查询引擎包括SQL解析和查询执行两部分。其中，parser包利用antlr4解析SQL语句，query包提供了查询模块的基础类。

接收到一个SQL语句后，parser生成语法树。

parser/ImpVisitor类处理语法树，语法树的根结点对应一个数据操纵语句即DML语句（select\_stmt, delete\_stmt等），parser/ImpVisitor类根据树根结点选择不同函数分别执行（visitSelect\_stmt调用query/QueryTable类，delete\_stmt调用schema/Manager类等）。ImpVisitor类中已经给出实现实例（Drop Table等），请完成剩余部分。

query/QueryTable类处理select\_stmt语句，扫描From子句中对应的表，生成查询结果，并以QueryResult的形式返回，返回值为一个Table。已给出程序框架，请完成剩余部分。

query/QueryResult类：create Table等语句的返回结果为一个Message，以表示执行是否成功，而不是一个Table。query/QueryResult类用来保留任意语句的执行结果。已给出程序框架，请完成剩余部分。

**需要实现的SQL语句（不区分大小写）：**

* CREATE TABLE tableName(attrName1 Type1, attrName2 Type2,…, attrNameN TypeN NOT NULL, PRIMARY KEY(attrName1))

实现“NOT NULL”和“PRIMARY KEY”这两个关键字，Type为Int，Long，Float，Double，String（必须给出长度）之一。  
示例: CREATE TABLE person (name String(256), ID Int not null, PRIMARY KEY(ID))

* DROP TABLE tableName

删除表

* SHOW TABLE tableName

展示某张表的模式信息(每一列的名字及其数据类型，格式自定)

* INSERT INTO [tableName(attrName1, attrName2,…, attrNameN)] VALUES (attrValue1, attrValue2,…, attrValueN)

字符串需要用单引号包围。  
示例：INSERT INTO person VALUES (‘Bob’, 15)  
INSERT INTO person(name) VALUES (‘Bob’)会提示字段ID不能为空

* UPDATE tableName SET attrName = attrValue WHERE attrName = attrValue
* DELETE FROM tableName WHERE attrName = attrValue
* SELECT tableName1.AttrName1, tableName1.AttrName2…, tableName2.AttrName1, tableName2.AttrName2,… FROM tableName1 [JOIN tableName2 [ON tableName1.attrName1 = tableName2.attrName2]] [ WHERE attrName1 = attrValue ]

上述语句中，Where子句仅包含一个比较运算，具体为‘<’、‘>’、‘<=’、‘>=’、‘=’、‘<>’。From子句中涉及2张表，On子句的限制与Where子句相同。DML语句中不包括嵌套语句。

**进阶项：**

1. 应用课程中介绍的查询优化技术

2. 支持多列主键；

3. where条件支持逻辑运算符（and/or）；

4. 实现三张表以上的join；

5. 实现outer join等其他类型的join；

6. 其他标准SQL支持的查询语法。

## **3.2**事务并发与恢复模块

实现简单的事务及恢复功能，支持小规模的并发。

**基础要求：**

1. 实现begin transaction和commit；采用普通锁协议，实现read committed隔离级别；在报告中通过实例说明隔离的有效性。

schema/Table类、schema/Database类、schema/Manager类包含属于不同数据层次的insert/delete/update等基础操作，需要在这些操作中支持事务，对于实现中额外添加的数据操作函数，也需要添加必要的事务处理。已给出程序框架，请完成剩余部分（可以参照schema/Manager.quit函数的锁设置方法）。

2. 实现事务的WAL机制，要求实现写log和读log，在重启时能够恢复记录的数据即可。

在schema/Manager.writeLog函数中已给出写log的方法，请补充读Log方法。

**进阶项：**

1. 实现多事务的并发和恢复机制；

2. 实现更高级的隔离级别（repeatable read/serializable）；

3. 实现rollback、savepoint等功能。

4. 实现MVCC协议。

## **3.3**说明

**需要在查询（query/parser 模块）、事务（schema模块）补充完善代码，**具体需要更改的类包括：parser/ImpVisitor**，**query/QueryResult, query/QueryTable**,** schema/Table, schema/Database, schema/Manager.

ThssDB代码框架仅作为参考**，可根据需要增加模块和类，鼓励进行系统架构创新与代码重构。**

## 四、评分规则

* **选题一：**
  + **基础分数：**文档或报告结构完整，逻辑通顺，包含必要的优化代码；满分100分。
  + **进阶分数：**使用benchmark等系统性试验方法如TPC-H等; 最多加分10分。
* **选题二：**
  + **基础分数：**代码完整正确；文档或报告结构完整，逻辑通顺；满分100分。
  + **进阶分数：**最多加分10分。

## 五、进度安排

* 项目周期：项目5-25开始，至6-26日结束。
* 项目节点：5-27组队完成；5-28上午9:00线上说明与答疑会；6-12项目进展分享与答疑。