Database Project

May 19th,2015

DEADLINE June 19th,23:59pm

1. 目标：

论文 Sinew 介绍的 Catalog 和 Serializer 模块的简单实现

2.编程语言和实验系统：

c/c++，不允许使用 STL; Linux

3. 问题描述：

实现简单版本的 catalog，记录所需信息（下面会具体定义），同时实现 JSON 格式到论文中 serializer 中介绍的格式（下面会给出定义）的相互转换。

4. 具体要求：

1.实现 insert filename 命令，其中文件中的数据为 JSON 记录，每行 1 条记录；

2.实现 check catalog 命令，返回 catalog 中的数据；

Catalog 规范：以 table 形式记录 Catalog，Catalog 只需记录如下信息：

例：

_id	Key_name	Key_type`	count
1	url	text	1
2	age	int	4

3.实现 find A = B 命令，其中 A 对应 JSON 记录中的 key，B 对应 JSON 记录中的 Value，输出所有符合 key = value 的 JSON 记录，没有则输出 NONE；

5. 实验数据

实验数据实验 Nobench 程序生成,规模为 10w 条记录。(附件给的 nobench 程序需要把其中的 nltk 源更换为 github 上的源才能正查使用，具体解决办法请自行在网上搜索）

6. 实现要求

1.注意：对于每一条输入的 JSON 记录，需要进行格式转化，变为论文中介绍过的如下格式：

<table>
<tr><td colspan="2">#attributes_num</td><td>aid0</td><td>aid1</td><td>...</td><td>aid(n-1)</td></tr>
<tr><td>offs0</td><td>offs1</td><td>...</td><td>...</td><td>offs(n-1)</td><td>len</td></tr>
<tr><td colspan="6">data</td></tr>
</table>

转化为如上格式后，注意把它以二进制文件的形式保存起来，文件名自定义，同时应实现对这个格式的解析，把数据还原为 JSON 格式的数据记录，而不是简单的把实际数据做一个简单的文件读写（这种实现记为 0 分）。

2.使用 c 语言中的 fread，fwrite 进行二进制文件读写，规定每个数据页的大小为 8Kb，即对文件数据的读写单位为 8Kb 大小的页。注意：要求记录之间要连续存储。对于当前页面存储不下整条数据记录的情况,请实现跨页面存储,即一条记录可能一部分存储于一页的末尾，而另一部分存储于下一页的开始。另外，同一条记录的属性之间也要求连续存储，中间不要有 padding（特别在用 C/C++中的 struct 时尤其要注意，编译器有可能加入一些 padding）。

3.写字符串的时候，注意不要包括'\0'。

4.写文件的时候，遇到整数均用 int 类型，也就是 32 位进行存储。我们检查的时候会检查文件的大小以确认你们是否正确实现。

5.（可选）考虑如何加快查询速度，例如对文件记录建索引，提取一条记录中某条属性时如何加快速度（如基于属性 id 的二分查找）等。

7. 提交说明

所有文件压缩到一个 zip 文件中，命名格式为“组号_组长姓名_project”,组号信息见各自班级的分组名单，要求提供的内容如下：

1.README

写明所提交文件的内容和应用程序的执行方法。

2.源代码

放在 src 文件目录下。

3.Makefile：编译程序

5.实验报告：

可用中文或英文写，必须为 pdf 格式，请以“组号_report”命名，应包含以下内容：

(a)实验环境，工具

(b)实验思路

(c)实验过程设计，如何对 JSON 格式进行格式的转换化为变长记录，如何进行解析

(d)代码结构，可根据需要使用简单的伪代码和流程图，请勿大篇幅粘贴代码

(e)实验结果，测试计划（实例分析查询所需时间）

(f)实验心得体会

(g)小组信息和分工