# 数据库项目描述文档

本学期数据库项目的题目是基于大数据的数据库查询优化系统设计，采取的数据集是工业标准TPC-H。

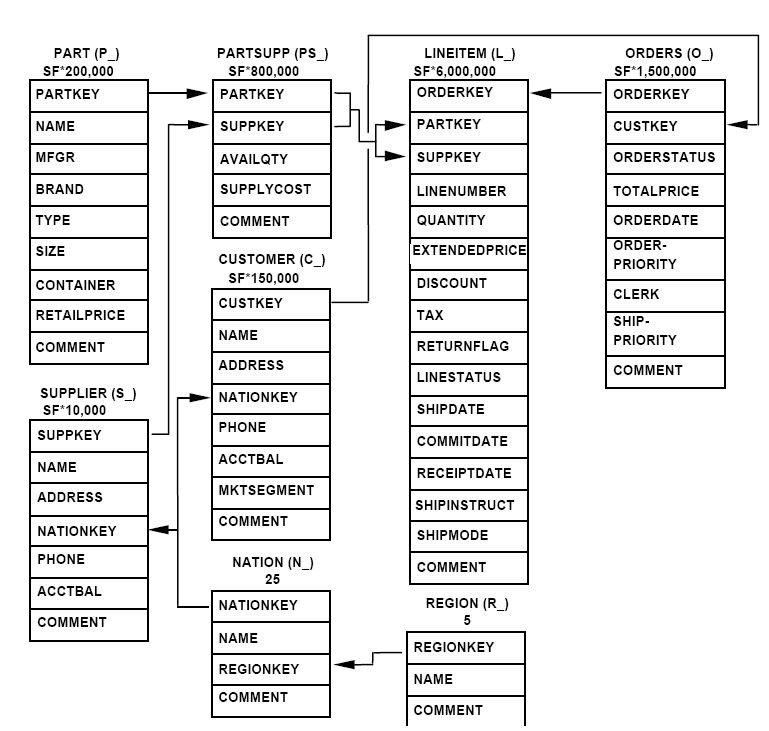
学生们需要根据原始数据生成自己的数据库存储系统以及查询优化系统，包括简单的sql解析等。

要求：所有小组的最终代码在6月11日24:00点之前提交，并提交一份英文版的设计文档，最终在TA处进行答辩。

1.1 据格式：

TPC-H数据库的组成被定义为由八个单独的表（基本表）组成。这些表的列与列之间的关系举例如图1：TPC-H模式。表中的数据类型以及类型的定义详见附录。

图1：TPC-H模式



图例：

·每个表名后面的括号内为这个表的列名的前缀;

·箭头指出表与表之间的一对多的关系;

·每个表名下方的数字或公式表示表的行数。一些是SF(Scale Factor)中的因子，用来获得数据库的大小。在LINEITEM表中的行数是近似值

## 1.2 SQL

本课程设计要求学生设计的数据库系统能执行以下的sql语句，并在尽可能短的时间内输出最终结果。

#### 1.2.1 范围查询以及聚集优化

查询功能定义：

select

l\_returnflag,

l\_linestatus,

sum(l\_quantity) as sum\_qty,

sum(l\_extendedprice) as sum\_base\_price,

avg(l\_quantity) as avg\_qty,

avg(l\_extendedprice) as avg\_price,

avg(l\_discount) as avg\_disc,

count(\*) as count\_order

from

lineitem

where

l\_shipdate <= date '1998-12-01' - '[DELTA]' day

group by

l\_returnflag,

l\_linestatus

order by

l\_returnflag,

l\_linestatus;

价格摘要报告查询提供了给定日期的运送的所有行的价格摘要报告，这个日期在数据库包含的最大的运送日期的60－120天以内。查询列出了扩展价格、打折的扩展价格、平均数量、平均扩展价格和平均折扣的总和。这些统计值根据RETURNFLAG 和LINESTATUS进行分组，并按照RETURNFLAG 和LINESTATUS的升序排列。每一组都给出所包含的行数。

替换参数：

替换参数（这里是[DELTA]）的值必须被产生以用来形成可执行查询文本：

DELTA在区间[60, 120]内随机选择。

#### 1.2.2 Join操作优化

查询功能定义：

select

l\_orderkey,

o\_orderdate,

o\_shippriority

from

customer,

orders,

lineitem

where

c\_mktsegment = '[SEGMENT]'

and c\_custkey = o\_custkey

and l\_orderkey = o\_orderkey

and o\_orderdate < date '[DATE]'

and l\_shipdate > date '[DATE]'

group by

l\_orderkey,

o\_orderdate,

o\_shippriority

order by

revenue desc,

o\_orderdate;

替换参数：

日期（[DATE]）在[1995-03-01, 1995-03-31]中随机选择。

[SEGMENT]在 [AUTOMOBILE，BUILDING，FURNITURE，MACHINERY，HOUSEHOLD]中随机选择

#### 1.2.3 TopN SQL优化

查询功能定义：

select top n \* from ORDERS where o\_orderdate< date '[DATE]'

order by o\_totalprice

替换参数：

日期（[DATE]）在[1995-03-01, 1995-03-31]中随机选择。

## 2 检验标准

* 课程设计的评分标准划分为5块
  1. simpleDB的存储系统（有压缩算法额外加分） 20%
  2. Join 的处理 20%
  3. 范围查询问题 20%
  4. top N问题 20%
  5. 代码规范+文档描述 20%

## Appendix

## A数据类型定义

下列数据类型将应用于每个表的列的清单中：

·Identifier

注释：一个默认的数据类型是整形。然而在SF大于300时，一些值将超出所支持的4个字节的整数。用户必须使用一些其他的数据类型比如8个字节的整数、小数或者字符串类型来实现。

·Integer意思是必须为整数（比如值的增长为1），取值范围为-2,147,483,646到2,147,483,647。

·Decimal意思是必须能够描述取值范围从-9,999,999,999.99到+9,999,999,999.99内、数值增长为0.01的所有有理数。

·Big Decimal是扩展的Decimal数据类型，它具有的附加特性是它必须足够大以至于能够描述存放在临时表中创建的查询变量的总数。

·Fixed Text，size N是用来存储一个固定长度为N的字符串类型。

注释：如果字符串本身小于长度N，那么剩余的空间必须被存储在数据库中，或者数据库自动加上一些空间使得CHAR\_LENGTH()函数的返回值为N。

·Variable text, size N是该列可以存储变量长度最长为N的字符串变量。被定义为“Variable text, size N”的列可以和定义为“fixed text, size N”一样执行。

·Date是一个可以被描述为YYYY-MM-DD的值，它所有的字符均为数字。一个日期必须能够描述连续的14年里的每一天，但是对日期的内部描述没有特殊的要求。

## B 表的规划

以下的列表定义了每个表所需要的结构（列的清单）。主键的注释及外键引用仅仅只是为了说明，而不是指定实现要求，例如完整性约束。

PART表的规划

列名 数据类型需求 注释

P\_PARTKEY identifier SF\*200,000

P\_NAME variable text, size 55

P\_MFGR fixed text, size 25

P\_BRAND fixed text, size 10

P\_TYPE variable text, size 25

P\_SIZE integer

P\_CONTAINER fixed text, size 10

P\_RETAILPRICE decimal

P\_COMMENT variable text, size 23

主键: P\_PARTKEY

SUPPLIER表的规划：

列名 数据类型需求 注释

S\_SUPPKEY identifier SF\*10,000

S\_NAME fixed text, size 25

S\_ADDRESS variable text, size 40

S\_NATIONKEY identifier 外键引用N\_NATIONKEY

S\_PHONE fixed text, size 15

S\_ACCTBAL decimal

S\_COMMENT variable text, size 101

主键: S\_SUPPKEY

PARTSUPP表的规划

列名 数据类型需求 注释

PS\_PARTKEY identifier 外键引用P\_PARTKEY

PS\_SUPPKEY identifier 外键引用S\_SUPPKEY

PS\_AVAILQTY integer

PS\_SUPPLYCOST decimal

PS\_COMMENT variable text, size 199

**主键**: PS\_PARTKEY, PS\_SUPPKEY

CUSTOMER表的规划

列名 数据类型需求 注释

C\_CUSTKEY identifier SF\*150,000

C\_NAME variable text, size 25

C\_ADDRESS variable text, size 40

C\_NATIONKEY identifier 外码参照N\_NATIONKEY

C\_PHONE fixed text, size 15

C\_ACCTBAL decimal

C\_MKTSEGMENT fixed text, size 10

C\_COMMENT variable text, size 117

**主键**: C\_CUSTKEY

ORDERS表的规划

列名 数据类型需求 注释

O\_ORDERKEY identifier 少量的计算SF\*1,500,000

O\_CUSTKEY identifier 外键引用C\_CUSTKEY

O\_ORDERSTATUS fixed text, size 1

O\_TOTALPRICE decimal

O\_ORDERDATE date

O\_ORDERPRIORITY fixed text, size 15

O\_CLERK fixed text, size 15

O\_SHIPPRIORITY integer

O\_COMMENT variable text, size 79

**主键**: O\_ORDERKEY

注释：并不是所有的顾客都会有订单。事实上，数据库中大约1/3的用户不会有任何订单。这些订单随机的分配给2/3的用户。

LINEITEM表的规划

列名 数据类型需求 注释

L\_ORDERKEY identifier 外键引用O\_ORDERKEY

L\_PARTKEY identifier 外键引用P\_PARTKEY，

和L\_SUPPKEY混合引用(PS\_PARTKEY, PS\_SUPPKEY)

L\_SUPPKEY identifier 外键引用S\_SUPPKEY,

和L\_PARTKEY混合引用外键(PS\_PARTKEY, PS\_SUPPKEY)

L\_LINENUMBER integer

L\_QUANTITY decimal

L\_EXTENDEDPRICE decimal

L\_DISCOUNT decimal

L\_TAX decimal

L\_RETURNFLAG fixed text, size 1

L\_LINESTATUS fixed text, size 1

L\_SHIPDATE date

L\_COMMITDATE date

L\_RECEIPTDATE date

L\_SHIPINSTRUCT fixed text, size 25

L\_SHIPMODE fixed text, size 10

L\_COMMENT variable text size 44

混合的主键: L\_ORDERKEY, L\_LINENUMBER

NATION表的规划：

列名 数据类型需求 注释

N\_NATIONKEY identifier 组合25 nations

N\_NAME fixed text, size 25

N\_REGIONKEY identifier 外键引用R\_REGIONKEY

N\_COMMENT variable text, size 152

主键: N\_NATIONKEY

REGION表的规划：

列名 数据类型需求 注释

R\_REGIONKEY identifier 组装5 regions

R\_NAME fixed text, size 25

R\_COMMENT variable text, size 152

主键: R\_REGIONKEY