

TPC-H 数据生成与导入小结

黄一夫

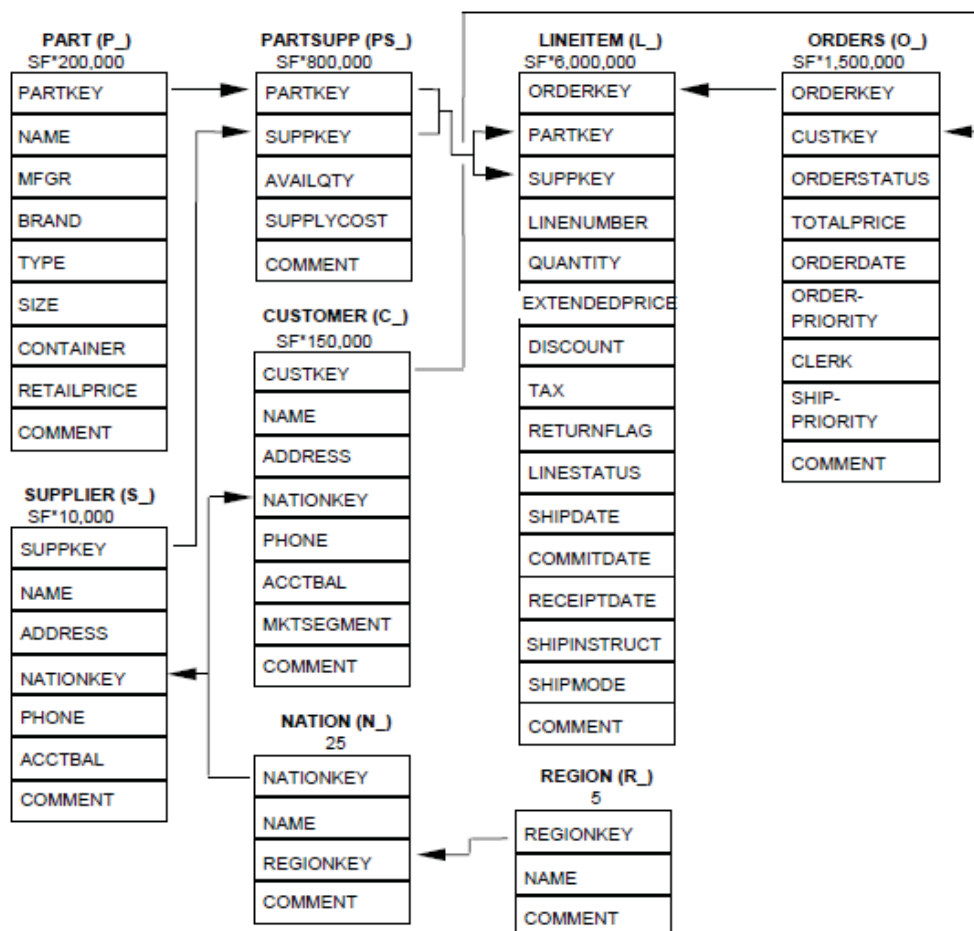
背景知识

TPC-H:

TPC-H(商业智能计算测试)是 TPC 的重要测试标准之一，主要用来模拟真实商业的应用环境。

TPC-H 用 3NF 实现了一个数据仓库,共包含 8 个基本关系/表,其中表 REGION 和表 NATION 的记录数是固定的(分别为 5 和 25),其它 6 个表的记录数,则随所设定的参数 SF 而有所不同,其数据量可以设定从 1GB ~ 3TB 不等。有 8 个级别供用户选择。

Figure 2: The TPC-H Schema



测试时,将 22 个复杂查询(SELECT)随机组成查询流,2 个更新(带有 INSERT 和 DELETE 的

程序段)操作组成一个更新流, 查询流和更新流并发执行数据库访问, 查询流数目随数据量增加而增加。

TPC-H 基准测试包括 22 个查询(Q1~Q22),其主要评价指标是各个查询的响应时间,即从提交查询到结果返回所需时间.TPC-H 基准测试的度量单位是每小时执行的查询数 (QphH@size), 其中 H 表示每小时系统执行复杂查询的平均次数, size 表示数据库规模的大小,它能够反映出系统在处理查询时的能力.

要了解更多背景知识, 请访问: <http://www.tpc.org/tpch/>

开发环境

操作系统:

Windows 7

开发工具:

Microsoft Visual Studio 2010

Oracle 11g R2

SQLPlus

SQLLoader

PL-SQL Developer

UltraEdit 17.00.0.1028

主要工作:

TPC-H 数据生成, 数据导入

实现步骤

1. 访问 <http://www.tpc.org/tpch/>

Current Version

The current version of TPC-H is version 2.14.2
Click on the preferred format below to download:

- PDF (1,720 KB)
- DOC (1,153 KB)
- with change bars from 2.14.0
- Tools
- DBGEN & Reference Data Set
 - zip (46,956 KB)
 - tgz file (37,892 KB)
- About Download Formats

下载介绍文档，阅读，了解背景知识，了解数据库的逻辑结构

下载 tpch 压缩包，解压，阅读 dbgen\README，了解 dbgen 的用法

2. 用 VS2010 打开 dbgen\tpch.sln，对 dbgen 工程进行组建，生成 dbgen.exe

```
1>----- Build started: Project: dbgen, Configuration: Debug Win32 -----
1> text.c
1> speed_seed.c
1> rng64.c
1> rnd.c
1> print.c
1> permute.c
1> load_stub.c
1> driver.c
1> build.c
1> bm_utils.c
1>d:\c_oracle\project\dbgen\bm_utils.c(544): warning C4996: '_strdup': The POSIX
1> c:\program files\microsoft visual studio 10.0\vc\include\string.h
1>d:\c_oracle\project\dbgen\bm_utils.c(561): warning C4101: 'remainder' : unre
1> Generating Code...
1> tpch.vcxproj -> D:\C_Oracle\Project\dbgen\Debug\dbgen.exe
===== Build: 1 succeeded, 0 failed, 0 up-to-date, 0 skipped =====|
```

3. 将生成的 dbgen.exe 从 dbgen\Debug 目录复制到 dbgen 目录下，目的是让

dbgen.exe 和 dists.dss 在同一目录，因为 dists.dss 将提供数据模板。打开 cmd，cd 到 dbgen 目录，执行命令：dbgen -s 1，表明在 dbgen.exe 存在的目录下生成 1G 的数据，至此请耐心等待一段时间。这 1G 的数据分为 8 个 tbl 文件存储，分别对应数据库中的 8 张表，字段和字段之间，元组与元组之间均采用 | 隔开

```

C:\Users\If>d:

D:\>cd D:\C_Oracle\Project\dbgen

D:\C_Oracle\Project\dbgen>dbgen -s 1
TPC-H Population Generator (Version 2.14.0)
Copyright Transaction Processing Performance Council 1994 - 2010

D:\C_Oracle\Project\dbgen>_

```

4. 用 SQLPlus@dss.ddl , DDL 是数据库模式定义语言 , 位于 dbgen 目录下 , 其中存放 8 张表的创建语句

```

C:\Users\If>sqlplus /nolog

SQL*Plus: Release 11.2.0.1.0 Production on 星期日 11月 13 10:50:35 2011

Copyright (c) 1982, 2010, Oracle. All rights reserved.

SQL> conn scott/tiger
已连接。
SQL> @D:\C_Oracle\Project\dbgen\dss.ddl

表已创建。

表已创建。

表已创建。

表已创建。

表已创建。

表已创建。

表已创建。

表已创建。

SQL> commit;

提交完成。

SQL>

```

5. 在 dbgen 父目录上创建新文件夹 , 取名 to_load , 将 dbgen 下当初生成的 8 个 tbl 文件移动到该 to_load 下。同时创建 8 个 ctl 控制文件 , 分别对应 8 个 tbl 文件。

例：lineitem.ctl 内容：

```
load data

INFILE 'lineitem.tbl'

INTO TABLE  LINEITEM

FIELDS TERMINATED BY '|'

(L_ORDERKEY,L_PARTKEY,L_SUPPKEY,L_LINENUMBER,L_QUANTITY,

L_EXTENDEDPRICE,L_DISCOUNT,L_TAX,L_RETURNFLAG,L_LINESTATUS,

L_SHIPDATE DATE "YYYY-MM-DD HH24:MI:SS",

L_COMMITDATE DATE "YYYY-MM-DD HH24:MI:SS",

L_RECEIPTDATE DATE "YYYY-MM-DD HH24:MI:SS",

L_SHIPINSTRUCT,L_SHIPMODE,L_COMMENT)
```

NOTE:

对于 DATE 字段，要指定其格式，在这里是"YYYY-MM-DD HH24:MI:SS"，否则会导
致使用默认 DATE 格式而出错

tbl 文件需要与 ctl 文件在同一目录下，否则导入出错

创建完毕后 cd 到 to_load 目录下，用 sqlldr 将数据导入都数据库中

例：

```
D:\C_Oracle\Project\dbgen>cd D:\C_Oracle\Project\to_load
D:\C_Oracle\Project\to_load>sqlldr scott/tiger control=customer.ctl,direct=y
SQL*Loader: Release 11.2.0.1.0 - Production on 星期日 11月 13 11:12:47 2011
Copyright (c) 1982, 2009, Oracle and/or its affiliates. All rights reserved.

加载完成 - 逻辑记录计数 150000。
D:\C_Oracle\Project\to_load>
```

NOTE:

添加参数 `direct=y` , 这样可以不用写 redo 日志 , 进而减少耗时

该过程需要将 8 个 tbl 数据文件全部导入数据库 , 需要耗费一定的时间

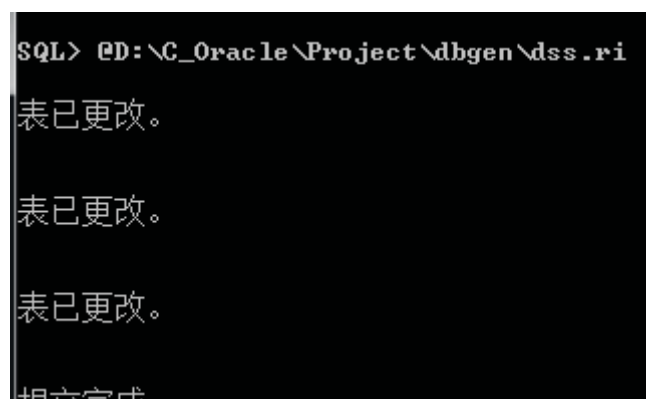
6. 用 `SQLPlus@dss.ri` 该文件内包含主外键约束的创建语句 , 原文件位于 `dbgen` 目录下 , 但添加外键的格式和数据库名称不正确 , 需要对其做一定的修改

例 :

```
ALTER TABLE NATION
```

```
ADD CONSTRAINT NATION_FK1 FOREIGN KEY (N_REGIONKEY) references
```

```
REGION(R_REGIONKEY);
```



```
SQL> ED:\C_Oracle\Project\dbgen\dss.ri
表已更改。
表已更改。
表已更改。
脚本完成
```

至此 , TPC-H 数据生成和数据导入完成 , 接下来可以使用提供的 22 个查询流更新流并行地访问数据库 , 通过记录时间来分析性能。

参考 :

1. <http://www.pilhokim.com/index.php?title=Project/EFIM/TPC-H>
 2. http://dsl.serc.iisc.ernet.in/projects/PICASSO/picasso_download/doc/Installation/tpch.htm
 3. http://tech.it168.com/a2011/0504/1186/000001186195_all.shtml
-