Data Warehouse

INSTRUCTIONS

1. Submit one ZIP file, `<student number>.zip` (for example: "`A012345L.zip`"), containing the following files to the folder "TPCH Submissions" 12 April at 18:30.
   - One PDF file "`schema.pdf`" containing the logical diagrams of the TPC-H schema and the TPC-H star schema. Make sure to write your **student number** in the PDF file.
   - One Python file "`tpch.py`". Make sure to write your **student number** the Python file.
2. After the deadline and until Friday 19 April at 18:30, you can submit to the folder "TPCH Late Submissions" in Luminus Files (penalties apply).

In this project, we use the TPC-H benchmark[1] to create a data warehouse system. TPC-H uses the example of a wholesale supplier business analytics application. You can read the selected excerpts from the TPC-H definition document in the `tpch.pdf` file on Luminus Files.

The TPCH star schema is created by combining the LINEITEM and ORDERS tables into one FACT_LINEORDER fact table, by dropping the PARTSUPP table and dropping the COMMENT attributes in all tables, by adding a DIM_DATE dimension table, by merging the NATION and REGION tables into DIM_CUSTOMER and DIM_SUPPLIER tables, by dropping P_RETAILPRICE (this is likely to change too frequently to be in a dimension; the part price is better determined for an order many days old as LO_EXTENDEDPRICE/LO_QUANTITY). This database is populated using the original TPC-H data.

Download the '`table.zip`", '`tpch.sql`" and "`tpch.py`" files from Luminus files.

[1] http://www.tpc.org/tpch

Question 1 [4 marks]

Create and populate the database following a star schema version of the TPC-H schema using the provided data in the folder `/home/cs4221/tpch-table` of your Virtual Machines or the downloaded `table.zip` from the Luminus Files and the SQL script '`tpch.sql`".

Submit a PDF file "`schema.pdf`" with the two logical diagrams representing the original TPC-H schema and the TPC-H star schema, respectively.

Question 2 [16 marks]

Write Python code that queries the data warehouse with the four following queries by modifying the template file "`tpch.py`".

1. Display the order key, part name, supplier name, order date and extended price for each line order for customer "`Customer#000000001`".
2. Display the total sum of the extended prices of all line orders for all possible grouping sets of customer region, customer nation and customer market segment for brand "`Brand#13`".
3. **Pricing Summary Report Query** (A modification from TPC-H Query 1)
   Display in ascending order of year and month of the order date the total number line orders, the corresponding total sums of their extended prices, discounted extended prices, discounted extended prices plus tax, average quantities, average extended prices, and average discounts for all possible grouping sets of year and month of the order date in this hierarchical order .
4. **Order Priority Checking Query** (A modification from TPC-H Query 4)
   This query determines how well the order priority system is working and gives an assessment of customer satisfaction. The Order Priority Checking Query counts the number of orders ordered in which at least one line item was received by the customer later than its committed date (receiptdate > commitdate). The query lists the count of such orders for each order priority sorted in ascending priority order.

Modify and submit the Python file "`tpch.py`" by writing the four SQL queries in the space indicated.

– END OF PAPER –