Data Science and Database Technology

Politecnico di Torino

Homework 1

The Urban Transport Analytics Division of the National Public Transport Authority is interested in analyzing revenue from public transportation services.

Specifically, they would like the analyses to address the following features:

An urban public transport network comprises several lines. Each line is operated by a single transportation mode. There are various transportation modes, such as buses, trams, metros and city trains. Each line operates in a specific city. Each city belongs to a specific province and region. Each line is characterized by the services available (air conditioning, WIFI, special seats).
Passengers purchase a ticket for each route on which they travel. The route has a departure stop and a destination stop and uses one line.
Tickets purchased by passengers are registered. Four types of tickets are available: 'Type A', 'Type B', 'Type C' and 'Type D'. Each ticket type has a price. In addition, tickets can be purchased with different discounts depending on the type of passenger: 'Adult', 'Student' (age between 14 and 24), 'Elderly' (age 65 and over) and 'Child' (age under 14).
The system also stores information on how to purchase tickets. Tickets can be purchased in several ways: online via the website or mobile application, at ticket vending machines located in stations, at authorized points of sale (e.g. kiosks) or directly from the driver/driver.

The analysis on journeys must be carried out considering the following details:

- Transportation mode (e.g., bus, metro, tram, suburban train), route, start and end stops, and services
- City, province, and region where the journey occurs
- Ticket type (single ride, daily pass, weekly pass, monthly subscription),purchase method (online, vending machine, authorized sales point, driver) and ticket discount (student, regular, child, senior)
- Journey details: day, month, bimester, trimester, year, timeslot and if the timeslot is peak or non-peak hours.

Homework Tasks

1. Design the data warehouse to address the previous specifications and to efficiently answer all the provided frequent queries. Draw the conceptual schema of the data warehouse and the logical schema (fact and dimension tables).

2. Write the following frequent queries using the extended SQL language:

   a) Separately for each means of transport and for each month, analyse:
      - the average daily number of tickets,
      - the cumulative number of tickets since the beginning of the year, and
      - the percentage of tickets used for each mode of transport compared to the total number of tickets in that month.

   b) Considering trips from 2022 onwards, separately for each line and city, we analyse:
      - the average trip duration,
      - the total revenue generated by that city,
      - the percentage of total revenue contributed by each line compared to the city's total, and
      - we rank each line within its city according to the total revenues generated by the line in descending order.
      -

3. Create and update a materialized view with CREATE MATERIALIZED VIEW and CREATE MATERIALIZED VIEW LOG in ORACLE

   <u>Frequent Queries of Interest:</u>

   - Separately for each transportation mode and for each month, analyze the average daily number of tickets.
   - Separately for each transportation mode and for each month, analyze the cumulative number of tickets from the beginning of the year.
   - Separately for each transportation mode and for each month, analyze the total number of tickets sold, the total revenue, and the average revenue.
   - Separately for each transportation mode and for each month, analyze the total number of tickets sold, the total revenue, and the average revenue for the year 2024.
   - Analyze the percentage of tickets related to each transportation mode and month over the total number of tickets of the month for each transportation mode.

a. Define a materialized view with CREATE MATERIALIZED VIEW useful to reduce the response time of the reported frequent queries.
b. Define the materialized view logs with CREATE MATERIALIZED VIEW LOG for each table where you deem it necessary. For which tables is it useful to keep track of logs? Identify all and only the necessary tables. Furthermore, for each table identify all and only the attributes for which it is necessary to keep track of the variations.
c. Specify which operations (e.g. INSERT on a specific table) cause an update of the defined MATERIALIZED VIEW

4. Update and management of views via Trigger Assuming that the CREATE MATERIALIZED VIEW command is not available, create the materialized views defined in the previous exercise and define the update procedure starting from changes on the fact table created by means of a trigger.

    a. Create the structure of the materialized view with CREATE TABLE VM1 (...)
    b. Specify an example of statement to populate the VM1 table with the necessary records using the statement INSERT INTO VM1 (...) (SELECT ...)
    c. Write the triggers necessary to propagate the changes (insertion of a new record) made in the FACTS table to the materialized view VM1.
    d. Specify which operations (e.g. INSERT) trigger the trigger created in 4.c.