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:

A public transport network includes various types of transportation modes such as buses, trams, metros, and suburban trains. Each transportation mode has its unique identification code and operates in one or more cities. Each city belongs to a specific province and region. The province and region information are also stored.
Transportation modes are further divided into unique routes. Each route may have several stops or stations along its path, and each stop has a unique identifier. Each route has some available services (AC, WIFI, SpecialSeats)
Tickets purchased by passengers are recorded. There are 4 types of tickets available: "Single Ride", "Daily Pass", "Weekly Pass", and "Monthly Subscription". Each ticket type has a different price and validity period. Additionally, tickets can be purchased with different discounts based on the passenger type: "Adult", "Student" (ages 14-24), "Senior Citizen" (ages 65 and above), and "Child" (under 14).
The system also stores information on how tickets are purchased. Tickets can be purchased in several ways: online via the website or mobile app, at ticket vending machines located at stations, at authorized sales points (e.g., kiosks), or directly from the driver/conductor.
Passengers can transfer between different transportation modes using a single ticket, and each journey (or trip) is recorded in the system. The record includes the start time, the starting stop, the end stop, the duration of the journey, the transportation modes used, and the date. Transfers between different lines or modes are also logged.
The company is interested in statistics on the average revenue and duration of the journey.

The analysis 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 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 transportation mode and for each month of the year, analyze: the average daily number of tickets, the cumulative number of tickets from the beginning of the year, and the percentage of tickets using each transportation mode over the total number of tickets in that month.

    b. Considering journeys from 2022, separately for each mode and city, analyze: the average journey duration, the total revenue generated from that city, the percentage of total revenue contributed by each route for the corresponding mode in a city, and assign a rank to each route within its transportation mode based on the total revenue generated in decreasing 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.

NOTE

For any issue, please write an email with the following metadata:
To: daniele.regecambrin@polito.it, davide.napolitano@polito.it
Object: [DSTBD] Bug Homework 1
Body: <Description of the issue>

Email without the previous format could be missed.