Modifications are in red.

PA220: Database systems for data analytics

Home Assignment 1 (and general overview)

Vlastislav Dohnal

Home Assignment Overview

- Overall Objectives
  - design a DW, load data, form analytical queries and create visualizations
  - update DW with new data
- Methodology (procedure)
  - Split into 4 individual assignments
  - Analyze the problem, propose a solution, instantiate it, execute it and measure metrics
    - Some assignments may not cover all these phases.
- Grading
  - Each assignment max. 10 points

Application Domain

- GPS tracking system of cars
  - Each car is equipped with a mobile device (Android) and an application.
  - The application
    - tracks movement of the car – records driving as well as stationarity;
    - allows the driver to enter events like refueling, loading/unloading cargo, rest time (sleeping);
    - allows the driver and operators to communicate via messages;
    - periodically uploads tracking data to server; and
    - periodically reports its status to server.
  - An application instance is uniquely identified by its ID (sim_imsi).
    - But it may change when the app is re-installed (on the same device).

Domain of Data Warehouse

- Create a DW for information about status of the tracking app and its network communication
  - Status of app is reported approximately every 10 minutes,
    - The report contains information about device model, app running time, phone running time.
    - E.g., HUAWEI Y600 U20, app running for 0.17 hrs (since app start), phone running for 112.67 hrs (since reboot)
  - There is also info about data-transmission log that contains *connection log rec id*, app version, sim card id (app instance unique ID) and mobile network ID.
    - E.g., 413085161, A38, 230024100616400, “23106” (MCC of O2 Slovakia)

  - There is about 2.1m report recs created per month, and 3.3m data-transmission recs created per month
    - **You have a sample only that (left-)joins both the source tables into one. ((-:**
    - **i.e. some records from data-transmission log do not have a pair from app status log since app statuses are not sent as often as regular (GPS) data, which results in a data-transmission log record.**

Domain of Data Warehouse

- DW must support the following analyses:
  - Per program version, report
    - the number of unique app instances,
    - the number of different phone models,
    - the number of phone/app restarts (app_run_time / phone_run_time is zero (or close to)).
  - Per app instance, report:
    - the number of program versions,
    - the number of different phone models,
    - the number of phone/app restarts.
  - By analogy, report the info per phone model.

  - Distribution (pie-chart) of program versions among app instances (sim_imsi)
    - for varying time period
  - Distribution of phone models among app instances.
    - Aka: How many devices of a particular model are used.

Assignment 1

- Analyze the data and report
  - number of unique values in attributes:
    - imsi, device, gsmnet, program version, car key, service_key, log_key,
    - check invalid values (NULL, suspicious zeroes, …)
  - look at it globally but also check for a shorter period, e.g., last month
- Design a dimensional model and create an ERD of it
  - Describe measurements in the fact table
  - Describe designed dimensions and qualify their types in "SCD" (Slowly Changing Dimension) and
    - *give reasons why you created a dimension and why it is of a specific SCD type.*
  - Discuss the granularity of your fact (w.r.t. the required queries in Assignment 2)
- Instantiate the dimensional model in PostgreSQL
  - Create dimension and fact tables.
  - Transform input data to these tables.

You may use a UML editor by Ondrej Novak
https://is.muni.cz/auth/th/np8o5/

You will use your own Pg server instance.
- install on your PC, or
- create a VM on stratus.fi.muni.cz (use template "PA220-homework")

Assignment 1 (cont.)

- Hand in to the IS vault one ZIP file containing:
  - report of unique values (values.txt),
  - ERD of dimensional model (as erd.png) plus the description (erd.txt), which will cover
    - individual measurements in the fact table and its granularity of facts,
    - individual dimension tables, where for each such a table
      - why you created it, and
      - qualify its types in “SCD” (Slowly Changing Dimension) and why this type.
  - a script of create table commands and other SQL commands to fill the dimensional model with input data (aka transformation script – load.sql)
- Grading
  - values 2 pts, model 5 pts, script 3 pts
  - total 10 pts

Input data

Illustrative sample data…

service_key (PK)	car_key	time	app_run_time	pda_run_time	device	tracking_mode	battery_level
129686177	2870	2017-01-01 01:00:00+01	41,97	41,98	HUAWEI Y530-U00	0	100
129686178	3749	2017-01-01 01:00:01+01	17,97	17,98	HUAWEI Y540-U01	0	97
129686179	3740	2017-01-01 01:00:01+01	227,02	227,03	VF695	0	100
129686181	3448	2017-01-01 01:00:01+01	5,12	39,65	Lenovo A6000	0	100
129686182	3838	2017-01-01 01:00:01+01	40,80	40,82	VF695	4	70

…

- Available in IS
  - https://is.muni.cz/auth/el/fi/podzim2022/PA220/um/data/pa220-data.zip
- Schema:

```
Column          |           Type           | Collation | Nullable | Default
----------------+--------------------------+-----------+----------+---------
service_key     | bigint                   |           |          |
car_key         | bigint                   |           |          |
time            | timestamp with time zone |           |          |
app_run_time    | numeric(6,2)             |           |          |
pda_run_time    | numeric(10,2)            |           |          |
device          | text                     |           |          |
tracking_mode   | text                     |           |          |
battery_level   | text                     |           |          |
log_key         | bigint                   |           |          |
sim_imsi        | character(15)            |           |          |
time_conn       | timestamp with time zone |           |          |
gsmnet_id       | character varying(6)     |           |          |
program_ver     | character varying(6)     |           |          |
```

Input Data Description

- service_key – record ID of app status log
- car_key – car ID
- time – timestamp (with time zone) when the app status record was received (in UTC)
- device – manufacture's code name of the phone model
- tracking_mode – GPS tracking mode
  - 0 = only in AC/DC, 2 = by Bluetooth, 4 = always (w/w.o. AC/DC), 1 = always (w/w.o. AC/DC) [obsolete]
- battery_level – battery charge level in %
- app_run_time – hours elapsed since app has been started on the device
  - starts from 0, so app restart can be detected by "a drop close to zero"
- pda_run_time – hours elapsed since the device has been booted
  - starts from 0, so the phone reboot can be detected by a "a drop to zero"

- log_key – record ID of data-transmission log
- sim_imsi – ID of app instance (may change upon app reinstallation), typically a random string
- time_conn – timestamp (with time zone) in UTC when the data transmission took place
- program_ver – version of application SW
  - typically, "Axy" (platform Android, SW ver. xy), or "m.n" (platform Windows, SW ver. m.n)
- gsmnet_id – MCC of GSM operator (Wikipedia)