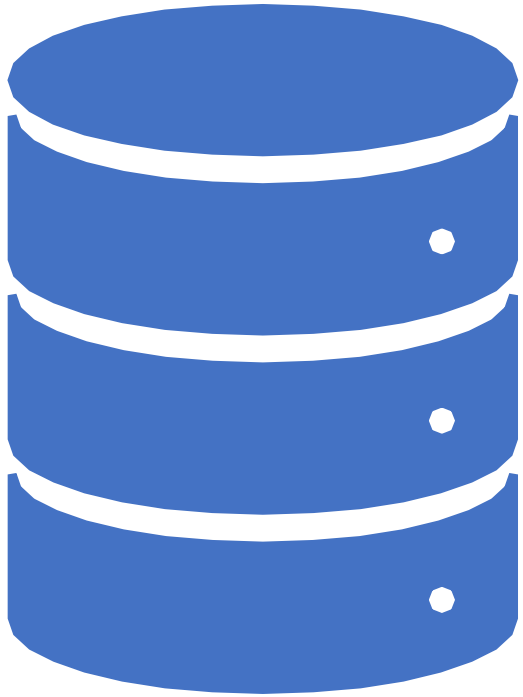

PA220: Database systems for data analytics

Home Assignment 1 (and general overview)

Vlastislav Dohnal

Home Assignment Overview

- Overall Objectives
  - design a DW, form analytical queries and optimize them
  - update DW with new data
- Methodology (procedure)
  - Split into 5 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 drivers to jot down events – refueling, loading/unloading cargo, rest time (sleeping);
    - allows the drivers and operators to communicate via messages;
    - periodically upload these data to server; and
    - periodically reports its status to server.

Domain of Data Warehouse

- Create a DW for information about status reports of the tracking app
  - Status of app (aka health) 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 a data-transmission log that contains app version, phone id, simcard id, transfer method and mobile network ID.
    - e.g., A38, 867897023525224, 230024100616400, “U”, “23106” (MCC of O2 Slovakia)
  - There is 212,062,680 report recs in total for the period of 216 months (2.1m last month), and 82,612,448 data-transmission recs in last 23 months (3.3m last month)
    - You will have a sample only ((-:

Domain of Data Warehouse

- DW should support analysis like:
  - Per program version, report
    - the number of different device (physical phones),
    - the number of different phone models,
    - the number of phone/app restarts (app_run_time / phone_run_time is zero (or close to)).
  - Per physical device, report the same (as per prog. ver.) plus:
    - the number of program versions.
  - By analogy, report the info per phone model.
  - Distribution (pie-chart) of program versions among physical devices
    - for varying time period
  - Distribution of phone models among physical devices.
    - How many phone of a particular model are used.

Assignment 1

- Analyze the data and report
  - number of unique values in attributes:
    - imsi, imei, device, gsmnet, method, program version, car key.
  - 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
  - Granularity of facts should be the reporting event
  - Describe measurements in the fact table
  - Describe designed dimensions and qualify their types in “SCD” (Slowly Changing Dimension)
- Instantiate the dimensional model in PostgreSQL
  - Create the dimension and fact tables.
  - Transform the input data to these tables.

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

You may use a student DBMS @ FI
https://www.fi.muni.cz/tech/unix/databases.html

Assignment 1 (cont.)

- Hand in to the IS vault:
  - report of unique values,
  - ERD of dimensional model (as PNG) plus the description of it
  - a script of create table command and other SQL commands to fill the dimensional model with input data (aka transformation script)

- Grading
  - values 2 pts, model 5 pts, script 3 pts
  - total 10 pts

Input Data Details

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

- Reports of app “health”
  - table service_log
- Attributes:
  - service_key – record ID
  - car_key – FK to cars (number)
  - time – timestamp (with time zone) when the record was created
  - app_run_time – hours elapsed since app has been started
    - starts from 0, so app restart can be detected by “a drop close to zero”
  - pda_run_time – hours elapsed since the phone has been booted
    - starts from 0, so the phone reboot can be detected by a “a drop to zero”
  - device – manufacture’s code name of the model
  - tracking_mode
    - 0 = AC/DC,
    - 2 = Bluetooth,
    - 4 or 1 = all-time
  - battery_level – charge status in %

Input Data Details

log_key (PK)	sim_imsi	time	car_key	pda_imei	gsmnet_id	method	program_ver
270819244	230024101003486	2017-02-01 00:59:23.182+01	3635	867721025715353	23106	U	A38
270819286	230021100851365	2017-02-01 00:59:57.199+01	2519	867897023542906	21670	U	A39
270819285	230024100623293	2017-02-01 00:59:56.864+01	2974	null	26203	U	A38
267258392	230024100885563	2017-01-03 10:29:02.9+01	1710	\N	00000	T	1.2

- Log of data-transfer connections
  - table conn_log
- Attributes:
  - log_key – record ID
  - sim_imsi – number of SIM card, can also be a random string
  - time – timestamp (with time zone) when the record was created
  - car_key – FK to cars (number)
  - pda_imei – unique ID of physical device, may not be available
  - gsmnet_id – MCC of GSM operator
  - method – either U (UDP) or T (TCP)
  - program_ver – version of SW, Axy or v.w