PKDD'99 Discovery Challenge

Курсова работа по “Складове от данни и бизнес анализ”

Зимен семестър, 2022 / 2023

Изготвено от:
Мирослав Дионисиев, 62390
Надежда Францева, 62391
Павел Сарлов, 62393

Преподавател: Тодор Кичуков

Януари 2023
София

Януари 2023
София

1. Описание на задачата

Имало едно време банка, предлагаща услуги. Сред тях: управление на акаунт, предлагане на заеми и т.н. Гореспоменатата банка се стреми да подобри своите услуги, като прави разлика между потенциално „добри“ и „лоши“ клиенти. Банковите мениджъри имат само бегли идеи за това на кого биха искали да предложат някои допълнителни услуги и кого да наблюдавате внимателно, за да опитат и минимизират риска и впоследствие - банковите загуби. За щастие банката съхранява данни за своите клиенти, техните сметки, извършените транзакции, отпуснатите заеми и издадените кредитни карти. Така че банковите мениджъри се надяват, че има начин, по който биха могли да се възползват от данните и да извлекат някои важни прозрения.

Описание на данните:

- релация ***account*** (4500 кортежа във файла **ACCOUNT.ASC**) - всеки запис описва статичните характеристики на акаунта
- релация ***client*** (5369 кортежа във файла **CLIENT.ASC**) - всеки запис описва характеристиките на клиента
- релация ***disposition*** (5369 кортежа във файла **DISP.ASC**) - всеки запис свързва клиент с акаунт
- релация ***permanent_order*** (6471 кортежа във файла **ORDER.ASC**) - всеки запис описва характеристиките на платежно нареждане
- релация ***transaction*** (1056320 кортежа във файла **TRANS.ASC**) - всеки запис описва една транзакция на един акаунт
- релация ***loan*** (682 кортежа във файла **LOAN.ASC**) - всеки запис описва заем, предоставен за дадената сметка
- релация ***credit_card*** (892 кортежа във файла **CARD.ASC**) - всеки запис описва кредитна карта, издадена към сметката
- релация ***demographic_data*** (77 кортежа във файла **DISTRICT.ASC**) - всеки запис описва демографските характеристики на областта

Всеки акаунт има както статични характеристики (напр. дата на създаване, адрес на клона) в релацията ***account***, така и динамични характеристики (напр. плащания, салдота) в релациите ***permanent_order*** и ***transaction***. Релацията ***client*** описва характеристиките на хората, които могат да боравят с акаунтите. Всеки клиент може да има няколко акаунта, няколко клиента могат да менажират един акаунт. Клиентите и акаунтите са свързани чрез релацията ***disposition***. Релациите ***loan*** и ***credit_card*** описват някои от услугите, които банката предлага на клиентите

си. Към един акаунт могат да бъдат издадени няколко кредитни карти, но максимум един заем може да се асоциира с акаунт. Релацията ***demographic_data*** дава публично достъпни данни за регионите (напр. процент на безработица), които могат да послужат като допълнителна информация за клиентите.

Януари 2023
София

2. Обосновка на избрания подход за изграждане на DWH

За реализация на текущата задача екипът ни се спря на хибридния подход на Дан Линстед. *Data Vault* моделът осигурява баланс между нуждата за зареждане на данните в реално време и пълната им интеграция в складът за данни. Също така е доста гъвкав и следователно подходящ за системи с променливи бизнес изисквания. Този модел е много стабилен във времето заради своята разделена структура и съдържание. Също така е мащабируемо решение, като позволява бързо интегриране на данните на различни платформи. Самият модел реагира добре на нови предметни области. Така може да се разширява и да се адаптира с минимално въздействие.

За текущата задача имаме само един единствен източник на данни, т.е. данните са предвидими и с ограничен обхват. Също така, предимствата на зареждането на данните в реално време и тяхната проследимост за текущата задача са малко излишни, но в случая на подобна система от реалността тези черти на склада от данни биха били от най-голямо значение, имайки предвид, че боравим с транзакции.

3. Описание на ETL процеса

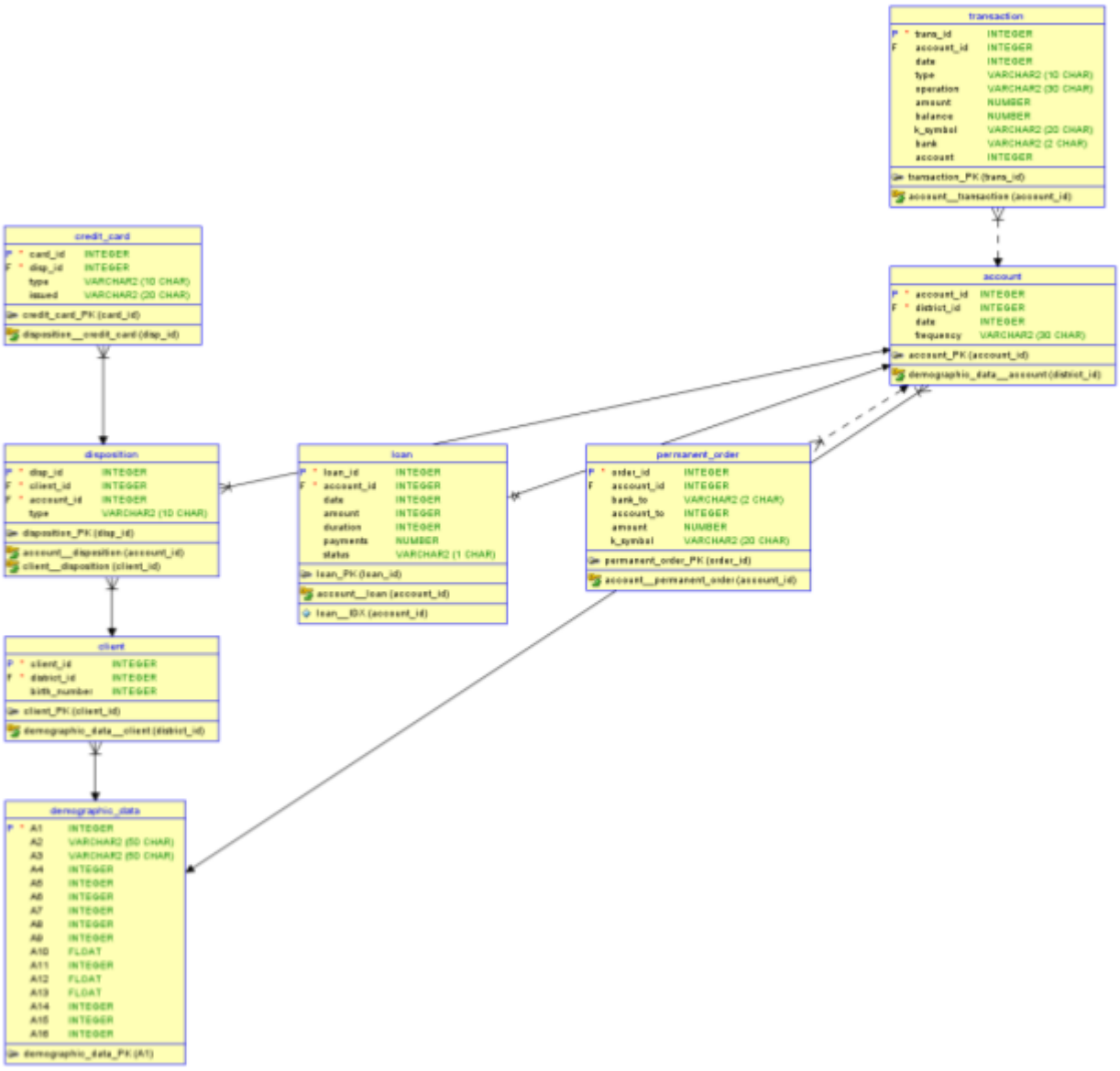
CBS (Core banking system) е единственият източник на данни в съответствие с целите на проекта. Данните в базата са заредени от предоставените *CSV* файлове. Всички *ETL* процеси се изпълняват веднъж (с други думи, данните се качват / актуализират само веднъж, а не ежедневно / седмично / месечно, както се случва в действителност).

- От *CBS* до *SA (Staging Area)* - всички записи от *CBS* таблиците се вмъкват в съответните *SA* таблици. Датата на изпълнение на заявката се добавя като ***staged_at*** във всяка *SA* таблица. Процеса на зареждане на данните в *SA* е осъществен чрез изпълнение на заявки, които извличат нужните данни от source базата и ги записват в съответните таблици в *SA*.
- От *SA* до *DV (Data Vault)* - за всяка от същностите в *SA* се създава хъб (*Hub)*, държащ сурогатен ключ (***id***), генериран спрямо брояч на постъпилите в таблицата кортежи, дата на зареждане на кортежа (***loaded_at***), източник на постъпилите данни (***source***) и бизнес ключа на самите данни (<***>_id***). Всеки хъб си има съответен сателит (*Satellite)*, който се свързва към хъба посредством външен ключ и включва характеристичните данни за съответната същност, както и споменатите дата и източник. Хъбовете са свързани посредством линкове (*Link*), които отново съдържат сурогатен ключ, дата и източник. Всеки линк представлява част от бизнес логиката на системата. Отделно има и референтни таблици, които съдържат фактологични данни. Те съдържат сурогатен ключ и съответстващите им данни.
- От *DV* до димeнсионен модел - преминаването от хъбове, линкове и сателити към димeнсии и факти е доста праволинеен - хъбовете и съответните им сателити стават димeнсии, а линковете и съответните им сателите стават факти. В случая нашите линкове нямат сателити. Референтните таблици по принцип също се превръщат в димeнсии, но за текущата задача решихме те да присъстват като колони в димeнсиите, които ги използват. Това се прави с цел да не се губят кортежи, при които липсва съответната референция.

Заявките за целия процес по създаването на базите, ETL процеса и извличането на докладите по-надолу, са в папката ***scripts/***. Файловете за самите модели са в папката ***models/***.

Януари 2023
София

4. Модел на входните данни



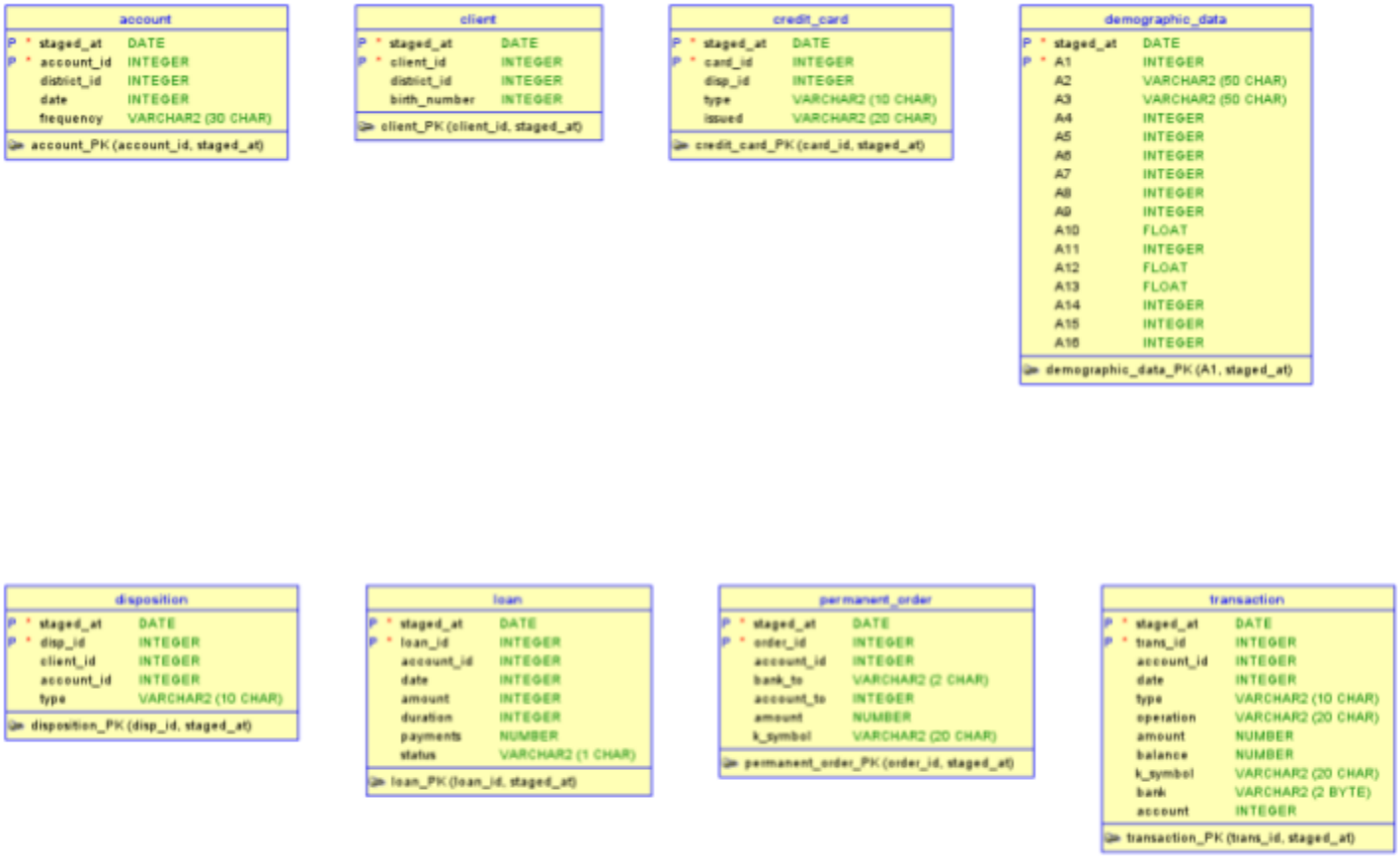
Фигура 1: Модел на източника

Януари 2023
София

5. Staging модел на данните

За модела на *SA* сме въвели следните промени в сравнение с модела *CBS*:

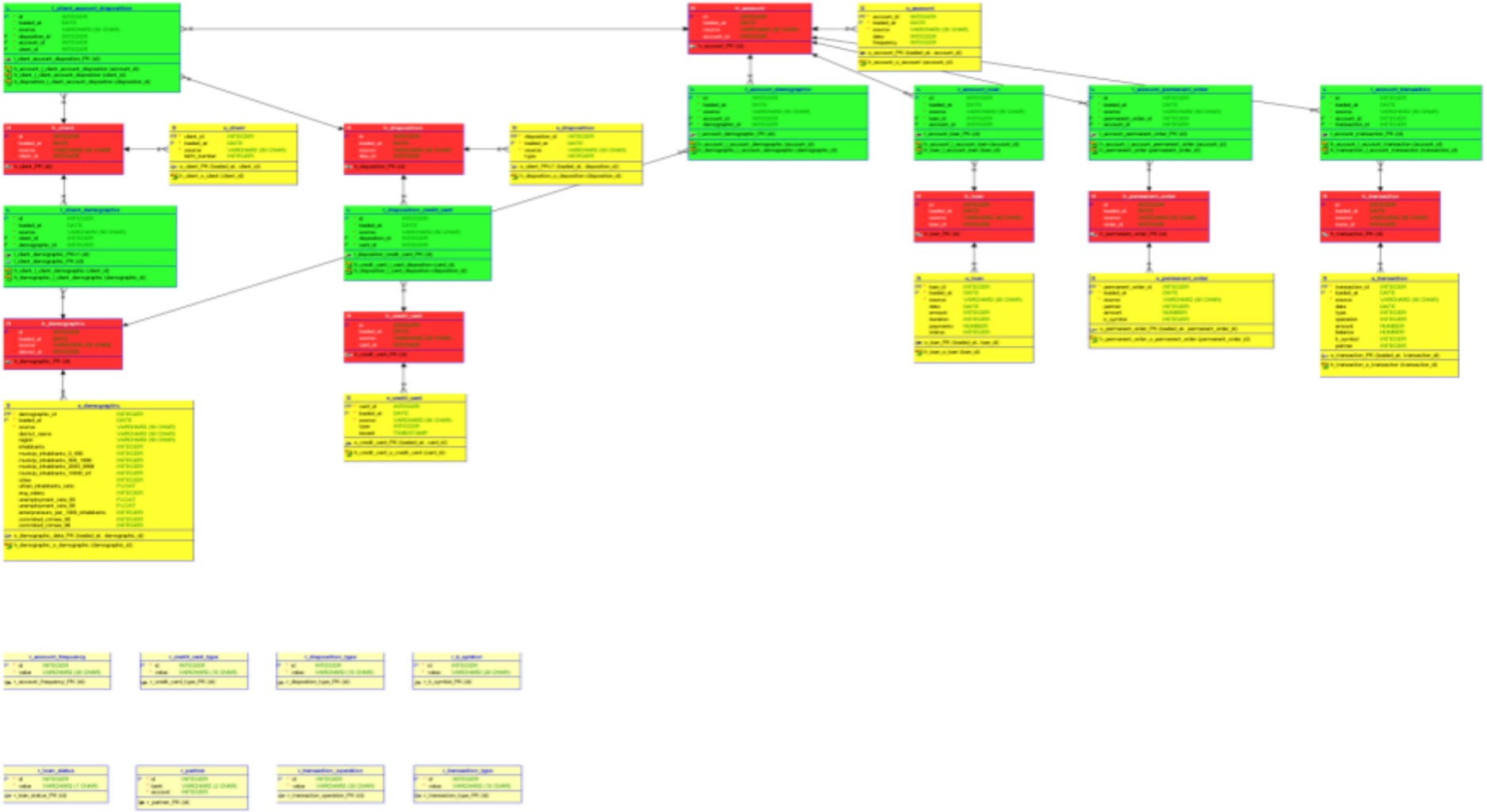
- Таблиците вече не са свързани една с друга.
- Няма ограничения в таблиците (с изключение на *PKs*).
- Ново поле - ***staged_at*** от тип DATE - се добавя към всяка таблица, така че се знае точно кога данните са били заредени/актуализирани за последен път.



Фигура 2: Staging Area модел

Януари 2023
София

6. Data Vault модел на данните



Фигура 3: Data Vault модел

Януари 2023
София

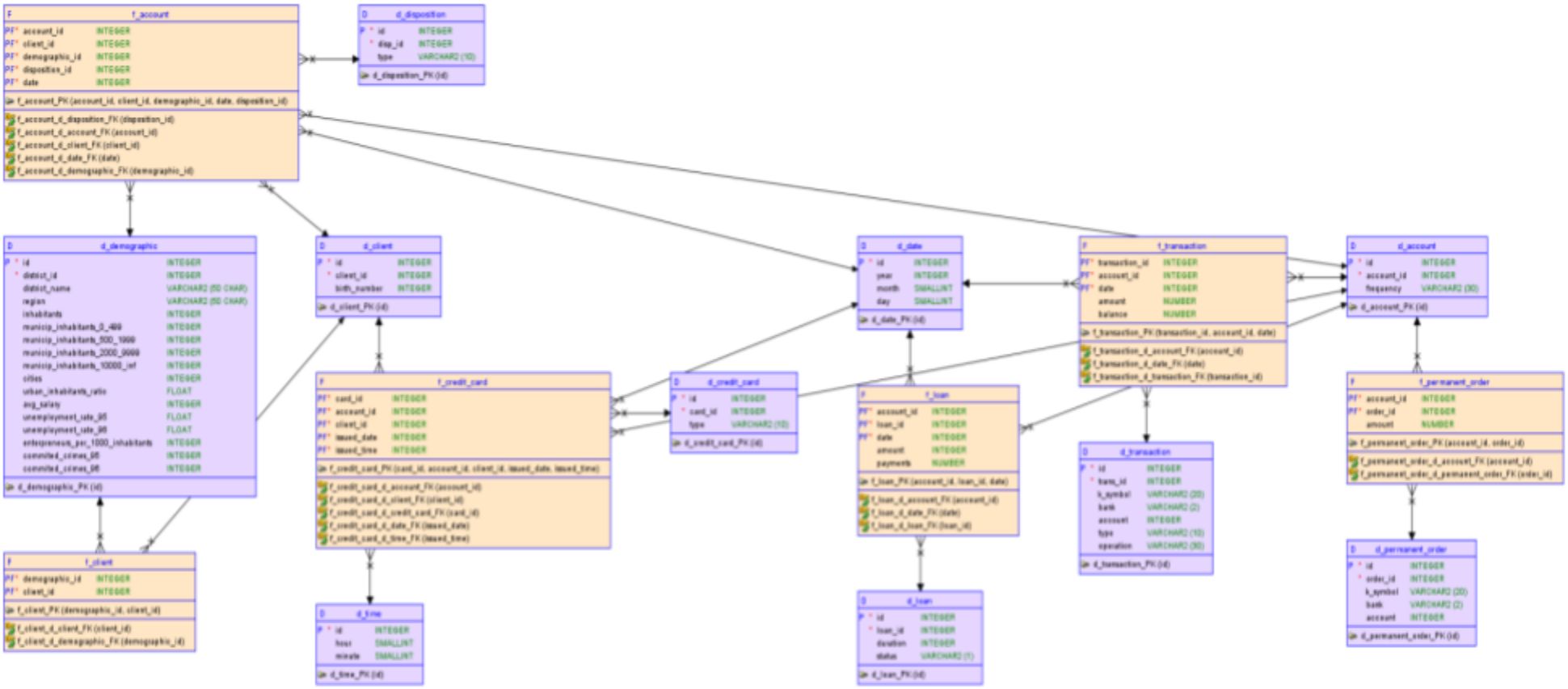
Хъбовете са таблиците в червено (или чиито имена започват с ***h_***). Сателитите са таблиците в жълто (или чиито имена започват с ***s_***). Линковете са таблиците в зелено (или чиито имена започват с ***l_***). Референтните таблици с тези без никакви релации (или чиито имена започват с ***r_***).

С цел по-добра яснота на данните са добавени някои изменения по имената на някои колони (най-вече за ***demographic_data***, където колоните бяха от типа ***a1***, ***a2***, ***a3…***). Новите наименования са съгласувани с обясненията в документа на предизвикателството.

Също така е извършено следното трансформиране на част от данните:

- Колоните ***date*** са извлечени като цели числа, които представляват дати във вида ***YYMMDD***. В *DV* вече са колони от тип ***DATE*.**
- За колоната ***issued*** при кредитните карти в пояснението на данните е упоменато, че ***issued*** е във вида ***YYMMDD***. Забелязахме обаче, че беше във вида ***'YYMMDD hh:mm:ss'***, което си е чиста времева марка. Затова и използвахме ***TIMESTAMP*** като тип.
- Фактологичните данни като ***k_symbol, operation, type*** и т.н. са пренесени в референтните таблици и ключовете за съответните стойности се реферират от сателитите, където е приложимо. Някои от кортежите съдържаха мръсни фактологични данни (например ***k_symbol*** при някои транзакции или липсваше, или беше някакъв низ с интервали). Подобни стойности са игнорирани и ключът в съответните сателити е *NULL*.

7.Димензионен модел



Фигура 4: Димензионен модел

Януари 2023
София

Димeнсиите са таблиците в лилаво (или чиито имена започват с ***d_***). Фактите са таблиците в бежово (или чиито имена започват с ***f_***).

Нашият димeнсионен модел се състои от 6 факт таблици, които се допълват от 10 димeнсии. Както споменахме по-горе, хъбовете със съответните им линкове стават на димeнсии, а линковете - на факти. Тъй като референтните данни са характеристични и поради липсата им в някои от кортежите като външни ключове сме ги преместили в димeнсиите, където намират приложение. Така няма да срещаме проблеми поради *NULL* външни ключове.

Предварително са направени следните димeнсии:

- ***d_date*** - раздробение на дата на година, месец, ден. Съдържа всички дати, които се срещат в данните.
- ***d_time*** - раздробение на време на час и минути. Съдържа 24 x 60 = 1440 кортежа за всеки час и всяка минута от денонощието.

За някои от фактите имаше числени данни, които могат да послужат за най-различни агрегации:

- ***f_transaction***
    - **amount** - стойността на изпълнената парична транзакция
    - **balance** - ново салдо на съответния акаунт
- ***f_permanent_order***
    - **amount** - стойността на поставената поръчка
- ***f_loan***
    - **amount** - размер на изтегления заем
    - **payments** - размер на месечните вноски по заема
    - **duration** - също е числена колона, но не намерихме причина да се използва в някаква агрегация, затова е изместена в димeнсията

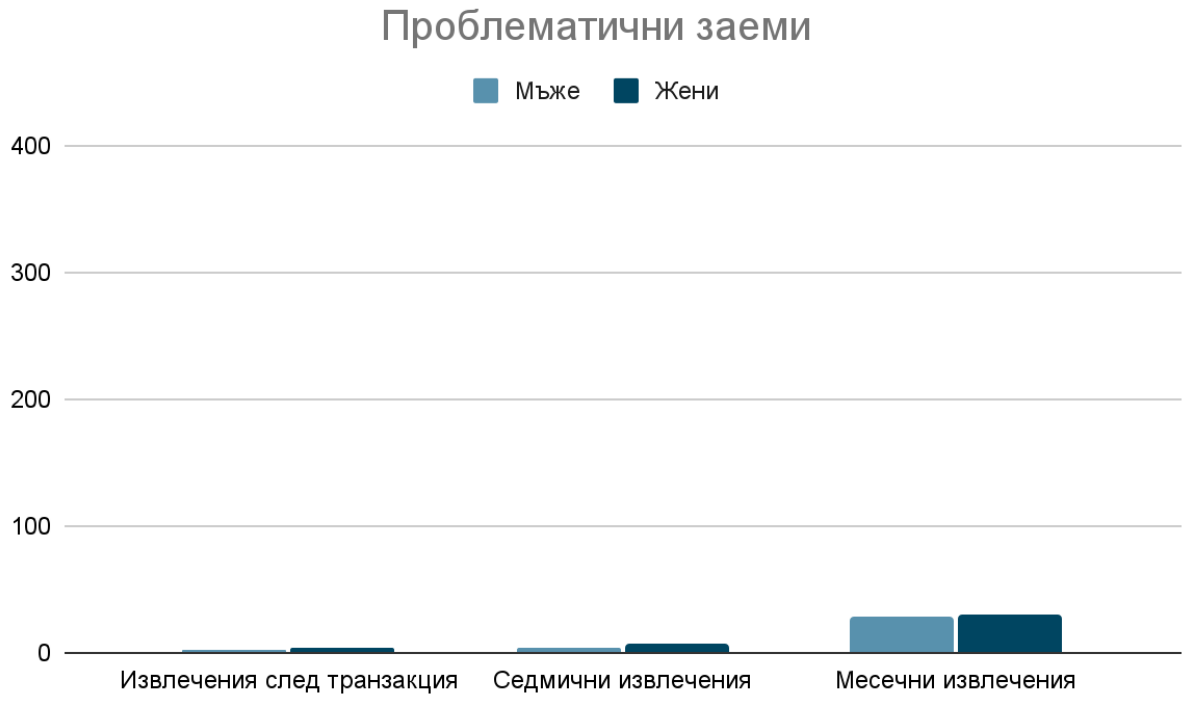
8. Полезни доклади

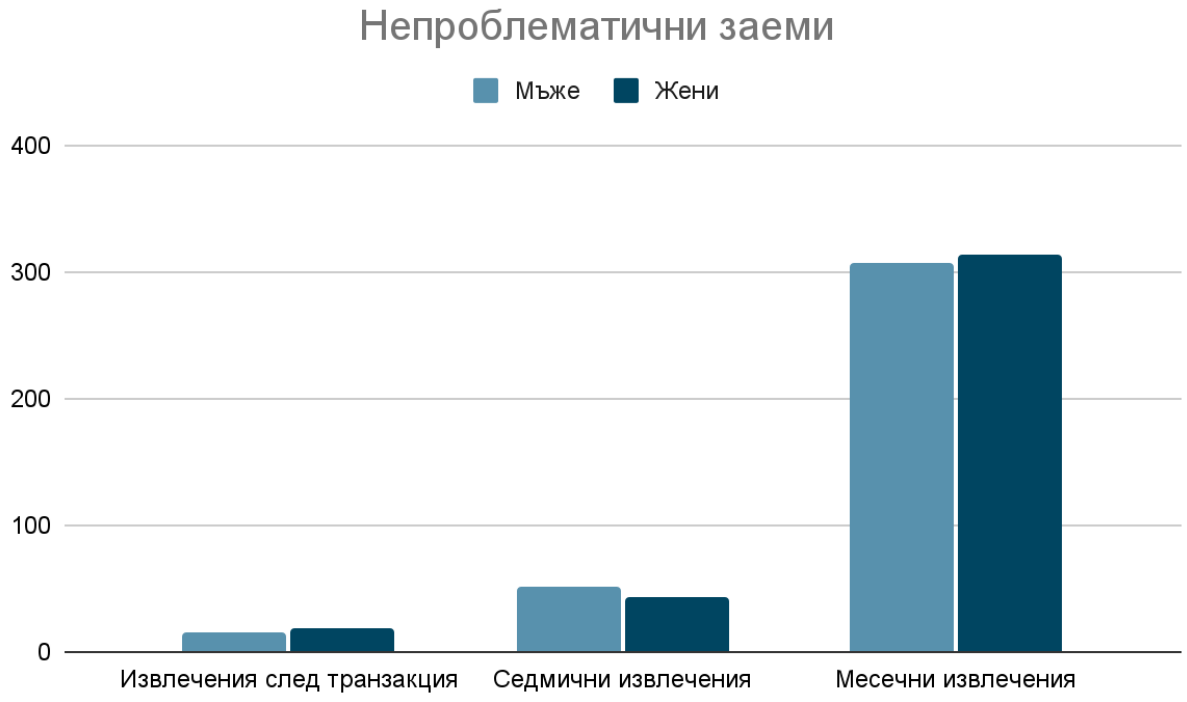
- Ако някога сте се чудели колко пари се изплащат за застраховки, ще разберем с една проста заявка към ***f_transaction*** таблицата, обединена с ***d_transaction***, като за застраховки търсим ***k_symbol = 'POJISTNE'***. В следната таблица можем да видим десетте акаунта с най-големи такива транзакции:

	account_id	max_amount
1	4386	12504.00
2	3592	10608.00
3	4312	9115.00
4	3448	8742.00
5	1733	8253.00
6	4079	7996.00
7	4349	7793.00
8	4121	7396.00
9	4400	7082.00
10	2229	6970.00

Таблица 1: Топ 10 акаунти с най-високи застраховки

- Нека да разгледаме заемите, групирани по пол на клиента, вид на заема (проблематичен - ***status*** е 'A' или 'C'; непроблематичен - ***status*** е 'B' или 'D'), както и честота на извлечения от съответния клиент:





Януари 2023
София

Както може да се види от горните графики повечето извлечения се правят месечно. Тоест можем да стигнем до извода, че клиенти, които по-рядко следят извлеченията си, са по-изкушени от това да поискат заем.

Не се забелязва корелация между изтеглените заеми и пола - и мъжете, и жените са еднакво вероятни да изтеглят заем. Също така, можем да кажем, че тенденцията е подобна и за двата вида заеми (проблематични и непроблематични). Тоест честотата на извлеченията не е определима за това дали клиентът ще изплати заема си.

Заявките за горните кратки доклади могат да бъдат намерени във файла ***reports.sql*** в папка ***scripts/***.

9. Използвани софтуерни инструменти

- **Oracle Data Modeler** - моделиране на логическите и релационните модели и генериране на SQL код
- **PostgreSQL** - прилагане на генерирания SQL код, като за целта се наложиха малки поправки поради разлики в използваната релационна база данни

Януари 2023
София