EDW Financial dataset - analýza a popis dat

Tomáš Chvosta, Petr Hanzl

Březen 2020

Obsah adresáře

Adresář, který byl poskytnut pro tuto úlohu obsahuje tyto soubory:

- **account.asc** - relace s údaji o účtech klientů
- **card.asc** - relace s údaji o kreditních kartách, které byly vydané ke konkrétním účtům
- **client.asc** - relace s údaji o jednotlivých klientech
- **disp.asc** - relace obsahující informace o klientech společně s jejich účtem
- **district.asc** - relace obsahující demografická data týkající se konkrétních demografických oblastí
- **loan.asc** - relace s údaji o půjčkách, které byly sjednány ke konkrétním účtům
- **order.asc** - relace s údaji, které se týkají jednotlivých platebních příkazů
- **trans.asc** - relace s údaji o jednotlivých transakcích

Relace account

Relace account obsahuje údaje o jednotlivých účtech. Můžeme si všimnout, že relace obsahuje id účtu, id pobočky, kde byl účet nejspíše vytvořen za období dlouhé 5 let od 1.1.1993 do 29.12.1997. Celkem se zde nachází 4500 záznamů. Každý záznam obsahuje následující atributy:

- **account_id** - Atribut sloužící jako unikátní identifikátor, tabulka tedy obsahuje 4500 unikátních hodnot tohoto atributu. Datový typ tohoto atributu je INTEGER. Nejnižší uložená hodnota, která je v tabulce uložena je 1, nejvyšší hodnota je 11382, průměrná hodnota je 2786.
- **district_id** - Jedná se o cizí klíč z relace demographic data.

- **frequency** - Tento atribut obsahuje informaci o frekvenci poplatků. Máme pouze 3 unikátní hodnoty tohoto atributu (POPLATEK MESICNE, POPLATEK TYDNE a POPLATEK PO OBRATU). Datový typ je VARCHAR. Nejdelší uložený řetezec má délku 16.
- **date** - Atribut představující datum vytvoření účtu. Datum je ve formátu YYMMDD. Celkem zde máme uloženo 1535 unikátních hodnot. Datový typ může být INTEGER nebo DATE. Nejnižší hodnota je 930101 (první vytvořený účet), nejvyšší je 971229 (poslední vytvořený účet).

Relace client

Relace client obsahuje údaje o jednotlivých klientech. Celkem se zde nachází 5369 záznamů. Každý záznam obsahuje následující atributy:

- **client_id** - Atribut sloužící jako unikátní identifikátor, tabulka tedy obsahuje 5369 unikátních hodnot tohoto atributu. Datový typ tohoto atributu je INTEGER. Nejnižší uložená hodnota, která je v tabulce uložena je 1, nejvyšší hodnota je 13998, průměrná hodnota je 3359.
- **birth_number** - Atribut představující rodné číslo klienta. Datum je ve formátu YYMMDD po muže a YYMM+50DD pro ženy. Z tohoto údaje lze tedy také vyčíst pohlaví klienta. Celkem zde máme uloženo 5019 unikátních hodnot. Datový typ může být INTEGER nebo DATE. Nejnižší hodnota je 110820 (nejstarší klient), nejvyšší je 875927 (nejmladší klient).
- **district_id** - Jedná se o cizí klíč z relace demographic data.

Relace disposition

Relace disposition obsahuje údaje spojující jednotlivé klienty s jednotlivými účty. Celkem se zde nachází 5369 záznamů. Každý záznam obsahuje následující atributy:

- **disp_id** - Atribut sloužící jako unikátní identifikátor, tabulka tedy obsahuje 5369 unikátních hodnot tohoto atributu. Datový typ tohoto atributu je INTEGER. Nejnižší uložená hodnota, která je v tabulce uložena je 1, nejvyšší hodnota je 13690, průměrná hodnota je 3337.
- **client_id** - Jedná se o cizí klíč z relace client.
- **account_id** - Jedná se o cizí klíč z relace account.
- **type** - Atribut popisující roli clienta vzhledem k danému účtu. Máme pouze 2 unikátní hodnoty tohoto atributu (OWNER a DISPONENT - pouze owner může žádat o půjčku). Datový typ je VARCHAR. Nejdelší uložený řetezec má délku 9.

Relace credit card

Relace credit card obsahuje údaje o kreditních kartách, které byly vydané ke konkrétním účtům. Celkem se zde nachází 892 záznamů. Každý záznam obsahuje následující atributy:

- **card_id** - Atribut, který slouží jako unikátní identifikátor, tabulka tedy obsahuje 892 unikátních hodnot tohoto atributu. Datový typ tohoto atributu je INTEGER. Nejnižší uložená hodnota, která je v tabulce uložena je 1, nejvyšší hodnota je 1247, průměrná hodnota je 481.
- **disp_id** - Jedná se o cizí klíč z relace disposition.
- **type** - Jedná se o atribut popisující roli typ kreditní karty. Máme pouze 3 unikátní hodnoty tohoto atributu (classic, junior a gold). Datový typ je VARCHAR. Nejdelší uložený řetezec má délku 7.
- **issued** - Atribut představující datum vydání karty. Datum je ve formátu YYMMDD HH:MM:SS. Datový typ tohoto atributu je DATE. Celkem zde máme uloženo 607 unikátních hodnot.

Relace demographic data

Relace demographic data obsahuje informace týkající se demografických charakteristik jednotlivých okresů v ČR. Celkem se zde nachází 77 záznamů. Každý záznam obsahuje následující atributy:

- **A1 - district id** - Jedná se o atribut představující identifikátor okresu. Celkem se zde nachází 77 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 1, nejvyšší je 77, průměrná hodnota je 39.
- **A2 - district name** - Jedná se o název okresu. Datový typ je VARCHAR. Nejdelší uložený řetězec má délku 19 (Rychnov nad Kneznou).
- **A3 - region** - Jedá se o označení části ČR. Datový typ je VARCHAR. Nejdelší uložený řetězec má délku 15 (central Bohemia).
- **A4 - inhabitants** - Atribut přdstavuje počet obyvatel v daném okrese. Celkem se zde nachází 77 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 42821, nejvyšší je 1204953, průměrná hodnota je 133885.
- **A5 - municipalities1** - Atribut představuje počet obcí s počtem obyvatel menším než 500. Celkem se zde nachází 53 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 0, nejvyšší je 151, průměrná hodnota je 49.

- **A6 - municipalities2** - Atribut představuje počet obcí s počtem obyvatel v rozmezí 500 až 1999. Celkem se zde nachází 36 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 0, nejvyšší je 70, průměrná hodnota je 24.

- **A7 - municipalities3** - Atribut představuje počet obcí s počtem obyvatel v rozmezí 2000 až 9999. Celkem se zde nachází 17 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 0, nejvyšší je 20, průměrná hodnota je 6.

- **A8 - municipalities4** - Atribut představuje počet obcí s počtem obyvatel větším než 9999. Celkem se zde nachází 6 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 0, nejvyšší je 5, průměrná hodnota je 2.

- **A9 - cities** - Atribut představuje počet měst v okrese. Celkem se zde nachází 11 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 1, nejvyšší je 11, průměrná hodnota je 6.

- **A10 - ratio of urban inhabitants** - Jedná se o atribut představující poměr městských obyvatel v procentech. Tabulka obsahuje celkem 70 unikátních hodnot. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 33.9, nejvyšší je 100.0, průměrná hodnota je 63.0.

- **A11 - average salary** - Jedná se o atribut představující průměrný plat v daném okrese. Tabulka obsahuje celkem 76 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 8110, nejvyšší je 12541, průměrná hodnota je 9032.

- **A12 - unemployment rate '95** - Atribut představuje míru nezaměstnanosti v roce 1995. Míra je uvedená v procentech. Tabulka obsahuje celkem 71 unikátních hodnot. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 0.29, nejvyšší je 7.34, průměrná hodnota je 3.1. Jedna hodnota není definovaná a místo ní se v tabulce vyskytuje symbol '?'.

- **A13 - unemployment rate '96** - Atribut představuje míru nezaměstnanosti v roce 1996. Míra je uvedená v procentech. Tabulka obsahuje celkem 73 unikátních hodnot. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 0.43, nejvyšší je 9.4, průměrná hodnota je 3.8.

- **A14 - enterpreneurs** - Atribut udává počet podnikatelé na 1000 obyvatel. Tabulka obsahuje celkem 44 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 81, nejvyšší je 167, průměrná hodnota je 116.

- **A15 - commited crimes '95** - Atribut představuje počet spáchaných trestných činů v roce 1995. Tabulka obsahuje celkem 76 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je

818, nejvyšší je 85677, průměrná hodnota je 4850. Jedna hodnota není definovaná a místo ní se v tabulce vyskytuje symbol '?'.

- **A16 - commited crimes '96** - Atribut představuje počet spáchaných trestných činů v roce 1995. Tabulka obsahuje celkem 76 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 888, nejvyšší je 99107, průměrná hodnota je 5031.

Relace loan

Relace loan obsahuje údaje o jednotlivých půjčkách. Celkem se zde nachází 682 záznamů. Každý záznam obsahuje následující atributy:

- **loan_id** - Jedná se o atribut představující identifikátor půjčky. Celkem se zde nachází 682 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 4959, nejvyšší je 7308, průměrná hodnota je 6172.
- **account_id** - Jedná se o cizí klíč z relace account.
- **date** - Atribut představuje datum uskutečnění půjčky. Datum je ve formátu YYMMDD. Celkem zde máme uloženo 559 unikátních hodnot. Datový typ může být INTEGER nebo DATE. Nejnižší hodnota je 930705 (první půjčka), nejvyšší je 981208 (poslední půjčka).
- **amount** - Atribut udává množství peněz, které byly poskytnuty v rámci půjčky. Předpokládáme, že tato hodnota je v Kč. Tabulka obsahuje celkem 645 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 4980, nejvyšší je 590820, průměrná hodnota je 151410.
- **duration** - Atribut představuje délku půjčky. Tabulka obsahuje pouze 5 unikátních hodnot (12, 24, 36, 48, 60) a tyto hodnoty udávají počet měsíců. Datový typ je INTEGER.
- **payments** - Atribut představuje výši měsíční splátky v Kč. Tabulka obsahuje celkem 577 unikátních hodnot. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 304.00, nejvyšší je 9910.00, průměrná hodnota je 4191.00.
- **status** - Jedná se o atribut, který značí stav splácení dané půjčky. Tabulka obdahuje pouze 4 unikátní hodnoty:
  1. A - smlouva ukončena, půjčka zaplacena
  2. B - smlouva ukončena, půjčka nezaplacena
  3. C - probíhající smlouva neukončena, zatím však bez problému
  4. D - probíhající smlouva, klient v dluhu

  Datový typ je CHARACTER.

Relace permanent order

Relace permanent order obsahuje údaje, které se týkají jednotlivých platebních příkazů. Celkem se zde nachází 6471 záznamů. Každý záznam obsahuje následující atributy:

- **order_id** - Jedná se o atribut představující identifikátor platebního příkazu. Celkem se zde nachází 6471 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 29401, nejvyšší je 46338, průměrná hodnota je 33778.
- **account_id** - Jedná se o cizí klíč z relace account.
- **bank_to** - Atribut představuje kód banky příjemce. Každý kód se skládá ze dvou písmen. Celkem zde máme uloženo 13 unikátních hodnot (UV, WX, QR, CD, IJ, GH, EF, OP, AB ,YZ, ST ,KL, MN). Datový typ je VARCHAR.
- **account_to** - Tento atribut reprezentuje číslo bankovního účtu příjemce. Celkem se v tabulce nachází 6446 unikátních hodnot. Není zaručeno, že se v tabulce nacházejí korektně zadané čísla bankovních účtů. Například jedna z hodnot (399) nesplňuje formát BBAN. Datový typ by mohl být INTEGER, ale je lepší zvolit VARCHAR.
- **amount** - Atribut udává množství peněz, které je v rámci platebního příkazu odesláno. Předpokládáme, že tato hodnota je v Kč. Tabulka obsahuje celkem 4412 unikátních hodnot. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 1.0, nejvyšší je 14882.0, průměrná hodnota je 3281.0.
- **k_symbol** - Atribut charakterizuje, o jaký typ platby se jedná. Celkem zde máme uloženy 4 unikátní hodnoty (LEASING, SIPO, POJISTNE, UVER). Datový typ je VARCHAR. Nejdelší řetězec má délku 8. Né vždy je tato hodnota korektně zadána a místo jedné z uvedených hodnot je v tabulce uložena mezera. S touto skutečný je nutné počítat.

Relace transaction

Relace transaction order obsahuje údaje, které se týkají jednotlivých transakcí. Celkem se zde nachází 1056320 záznamů. Každý záznam obsahuje následující atributy:

- **trans_id** - Jedná se o atribut představující identifikátor transakce. Celkem se zde nachází 1056320 unikátních hodnot. Datový typ je INTEGER. Nejnižší uložená hodnota toho atributu je 1, nejvyšší je 3682987, průměrná hodnota je 1335311.
- **account_id** - Jedná se o cizí klíč z relace account.

- **date** - Atribut představující datum uskutečnění transakce. Datum je ve formátu YYMMDD. Datový typ může být INTEGER nebo DATE. Nejnižší hodnota je 930101 (první transakce), nejvyšší je 981231 (poslední transakce).
- **type** - Atribut charakterizuje, zda se jedná o příjem nebo výdaj. Celkem zde tedy máme uloženy 2 unikátní hodnoty (PRIJEM a VYDAJ). Datový typ je VARCHAR. Nejdelší řetězec má délku 6.
- **operation** - Atribut charakterizuje, o kterou operaci se jedná. Celkem je v tabulce uloženo 5 unikátních hodnot (VYBER KARTOU, VKLAD, PREVOD Z UCTU, VYBER a PREVOD NA UCET). Ne vždy je tato hodnota korektně zadána a místo jedné z uvedených hodnot je v tabulce uložena prázdná hodnota. Datový typ je VARCHAR. Nejdelší řetězec má délku 12.
- **amount** - Atribut udává množství peněz v rámci jedné transakce. Předpokládáme, že tato hodnota je v Kč. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je 0.0, nejvyšší je 87400.0, průměrná hodnota je 5937.0.
- **balance** - Atribut udává zůstatek peněz po provedení transakce. Předpokládáme, že tato hodnota je v Kč. Datový typ je REAL nebo DOUBLE. Nejnižší uložená hodnota toho atributu je -41125.7, nejvyšší je 209637.0, průměrná hodnota je 38518.9. Můžeme si tedy všimnout, že tabulka obsahuje i záporné hodnoty tohoto atributu.
- **k_symbol** - Atribut udává nějakou bližší charakteristiku transakce. Celkem je v tabulce uloženo 7 unikátních hodnot (POJISTNE, SLUZBY, UROK, SANKC. UROK, SIPO, DUCHOD a UVER). Ne vždy je tato hodnota korektně zadána a místo jedné z uvedených hodnot je v tabulce uložena prázdná hodnota. Datový typ je VARCHAR. Nejdelší řetězec má délku 8.
- **bank** - Atribut představuje kód banky partnera. Každý kód se skládá ze dvou písmen. Datový typ je VARCHAR. Ne vždy je tato hodnota korektně zadána a místo jedné z uvedených hodnot je v tabulce uložena prázdná hodnota.
- **account** - Tento atribut reprezentuje číslo bankovního účtu partnera. Datový typ by mohl být INTEGER, ale je lepší zvolit VARCHAR. Ne vždy je tato hodnota korektně zadána a místo jedné z uvedených hodnot je v tabulce uložena prázdná hodnota.