

## Zápočtový test

Na vypracování testu máte 120 minut. Očekávanými a hodnocenými výstupy jsou:

1. Textový soubor s názvem *váš\_login.txt* (např. *pascep.txt*) obsahující zdrojový kód příkazů, kterými jste provedli zadání. Pokud máte zároveň něco zjistit nebo vypsát, zaznamenejte to přímo do souboru, ideálně jako komentář. **Tento soubor na konci své práce zašlete mailem na adresu [jan.hucin@profinit.eu](mailto:jan.hucin@profinit.eu) a podepište se svým celým jménem.**
2. Existence a vlastnosti tabulky, kterou jste vytvořili při plnění zadání. To zhodnotíme přímo na clusteru.

V jednotlivých oblastech testu se hodnotí každý úkol nebo jeho část. Pokud si s nějakou částí zadání nebudete vědět rady, můžete ji přeskočit nebo zadání splnit bez této části, počet bodů se pak přiměřeně sníží.

Výchozí data jsou v databázi *fel\_bigdata*, tabulka *transakce*. Tabulka obsahuje karetní platby na čerpacích stanicích od 1. 2. 2017 do 31. 1. 2018. Sloupce mají postupně tento význam:

id klienta – id stanice – den platby – hodina platby – částka v Kč – příznak, zda jde o tzv. oblíbenou (častou) stanici klienta – příznak hlavní oblíbené stanice klienta – zem. šířka bydliště klienta – zem. délka bydliště klienta – kód státu – obec – název stanice – číslo clusteru – zem. šířka polohy stanice – zem. délka polohy stanice – vzdálenost stanice od bydliště klienta v km

### Vytvoření tabulky Hive (8 bodů)

Zkontrolujte, že máte založenou databázi Hive, a pokud ne, založte ji (název = váš login). **Dále pracujte se svou databází.** Vytvořte managed (interní) tabulku Hive *transakce2* s formátem Parquet a kompresí GZIP. Tabulka bude mít stejné názvy sloupců a jejich typy jako tabulka s výchozími daty až na tyto rozdíly:

- oba příznaky oblíbené stanice budou mít v nové tabulce typ boolean;
- v nové tabulce bude navíc sloupec *cze* typu boolean, podle kterého bude nová tabulka partitionovaná.

(Nevíte-li, jak zjistit typy u existující tabulky, zvolte je aspoň podle vlastního uvážení co nejvhodněji. Neshoda typů se považuje za malou chybu.)

Z výchozí tabulky zkopírujte do nové tabulky všechny záznamy se dnem platby mezi 1. 3. 2017 a 30. 7. 2017. Sloupec *cze* bude mít hodnotu True pro záznamy s hodnotou „cze“ ve sloupci *country*, jinak False. Dejte pozor na správné převedení příznaků oblíbené stanice.

### Analytika (22 bodů)

V této části budete pracovat s daty nově vytvořené tabulky Hive (viz předchozí oddíl), a to **pouze s hodnotou True ve sloupci *cze***. (Pokud se vám nepodaří tabulku správně vytvořit, můžete pracovat s daty výchozí tabulky s výše uvedeným omezením na den a stát platby. Považuje se to však za malou chybu.) Úlohy jsou na sobě nezávislé, předpokládá se použití Hive nebo Sparku (konkrétní volba záleží na vás).

1. Kolik unikátních klientů je v datech?
2. Jaký podíl klientů (v %) má aspoň jednu transakci v sobotu časně ráno (s číslem hodiny od 5 do 7)?
3. Jaký podíl klientů (v %) nemá žádnou ze svých transakcí na oblíbené stanici?
4. Jaký je nejvyšší počet různých oblíbených stanic pro jednoho klienta? Zjistěte pro klienta s nejvyšším počtem různých oblíbených stanic průměr GPS souřadnic pro tyto stanice (každou stanici počítejte jen jednou).
5. Rozdělíme transakce podle vzdálenosti stanice od bydliště do pásem po 5 km. Vyloučíme záznamy s vyšší vzdáleností než 150 km nebo s neuvedenou vzdáleností. Jaký počet transakcí byl proveden v jednotlivých pásmech a ve kterém pásmu to bylo nejvíce? Liší se to pro stanice v západní části ČR (*pos\_gps\_lon* <= 14) a pro stanice ve východní části ČR (*pos\_gps\_lon* >= 17)?
6. Kterých osm stanic má nejvyšší počty plateb a kterých osm nejvyšší počty unikátních klientů?
7. Zjistěte průměrnou zaplacenou částku na stanicích Shell a na stanicích Eurooil (názvy jsou součástí textu ve sloupci *name*), porovnejte průměry mezi sebou.
8. Jaká jsou nejčastější slova v názvech stanic (každou stanici počítejte jen jednou)? Liší se to pro stanice v západní a ve východní části ČR (viz úlohu 5)?
9. Spočítejte Jaccardův koeficient podobnosti (počet prvků průniku děleno počtem prvků sjednocení) mezi množinami klientů s aspoň jednou platbou na stanicích s id 8776 a 8777.