

Technologie pro velká data, 1. domácí úkol

Termín odevzdání: 31. 12. 2019

Forma odevzdání: zaslat e-mailem (jan.hucin@profinit.eu) zdrojový SQL (HiveQL) kód analytického dotazu a jeho výsledek. Hive tabulka se bude hodnotit tak, jak je, není tedy nutné zasílat zdrojové kódy pro vytváření tabulky.

Domácí úkol obsahuje tyto hodnocené kroky:

- zkopírování dat, rozbalení, nahrání do HDFS (1 bod)
- proces vytvoření interní Hive tabulky z dat (3 body)
- analytický dotaz nad Hive tabulkou (1 body)

Za neúplné nebo jen částečně správné řešení se uděluje přiměřeně nižší počet bodů s granularitou půl bodu.

Zdrojová data

Jsou na edge node (lokální filesystem): `/home/pascepel/fel_bigdata/data/chess_ratings.zip`. Ze souborů zabalených v zipu použijte data pouze za některé tři po sobě jdoucí měsíce (zvolte si sami).

Finální tabulka Hive

Finální tabulka bude umístěna ve vaší databázi a bude mít jméno `chess_ratings`. Požadované vlastnosti:

- interní tabulka
- obsahuje pouze pole `name`, `fed`, `sex`, `rat`, `gms`, `bday`, `year`, `mon`
- formát ORC
- komprese ZLIB
- partitioning podle pole `sex`
- vyřadit záznamy s prázdným `name` a záznamy s hodnotou pole `sex` jinou než „F“ nebo „M“

Pomocné tabulky a další mezikroky udělejte zcela podle svého uvážení, důležitá je finální tabulka.

Analytický dotaz nad Hive tabulkou

Z finální tabulky vypište pro každý z měsíců (`mon`) pět žen s nejvyšším ratingem (`rat`), a to jediným SQL dotazem. Můžete k tomu použít Hive analytické funkce (viz např. <http://www.hadoopinsight.com/blog/hive/hive-analytic-functions/>).