NÁSKOK
DÍKY
ZNALOSTEM

PROFINIT

B0M33BDT
Technologie pro velká data

Supercvičení – Hive

Sergej Stamenov, Adéla Dragounová, Jan Hučín

31. 10. 2018

Osnova cvičení

- test
- opakování Hive & SQL
- import dat do tabulky Hive
- procvičování
- zadání domácího úkolu

PROFINIT

Opakování Hive & SQL

- tabulka = adresář (ne soubor)
- typy tabulek:
  - interní – vlastník adresáře je přímo Hive, umístěn v Hive warehouse
  - externí – vlastník adresáře je někdo jiný, umístěn kdekoliv
- partition u tabulky:
  - fyzické rozdělení dat do podadresářů
  - rychlejší vyhledání (podmínka WHERE v dotazu)
- které příkazy DML Hive umí a které ne?
  - insert ano (doporučen insert select)
  - update a delete standardně ne
- formáty a komprese
  - ORC, Avro, Parquet, text (CSV)
  - Gzip, Snappy, Zlib

Import dat do tabulky Hive

Viz příklad na přednášce.

Syntaxe CREATE TABLE:

```
CREATE [EXTERNAL] TABLE tablename (
    field type, ...
)
PARTITIONED BY (field type)
ROW FORMAT
    DELIMITED FIELDS TERMINATED BY string
    LINES TERMINATED BY string
STORED AS format
LOCATION hdfs_path
tblproperties("key" = "value", ...)
```

pole pro partitioning se uvede jen zde

PROFINIT

Import dat do tabulky Hive

Viz příklad na přednášce.

Syntaxe CREATE TABLE:

```
INSERT [OVERWRITE | INTO] TABLE tablename
PARTITION (field)
SELECT
  ...
FROM tablename
```


pole pro partitioning se uvede nahoře i zde

dynamický partitioning

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

PROFINIT

Procvičování

- https://github.com/stameser/BDT
- cviceni/03_HIVE
- vzorová řešení: branch solutions

PROFINIT

Domácí úkol

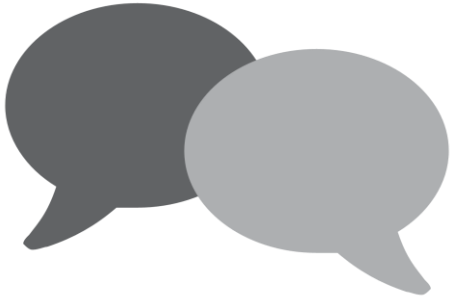
1. V rámci cvičení jste založili tabulku Hive s rozdělením na partitions podle čísla měsíce. Předpokládejme, že i v dalších letech budeme vždy na konci měsíce do tabulky doplňovat data, a to do nové partition, kterou ručně pojmenujeme podle měsíce a roku. Napište příkaz, který to bude provádět. (1 bod)
2. Předpokládejme, že interní (managed) tabulka ***T*** má v poli ***ym*** šestiznakový řetězec obsahující rok a měsíc ve formátu YYYYMM. Podle tohoto sloupce je u tabulky definován partitioning. Vždy jednou za měsíc chceme smazat partition za stejný měsíc předchozího roku. Napište příkaz, který to bude provádět. (1 bod)
3. Vytvořte pro Hive UDF na vzájemný převod teploty ve stupních Celsia na stupně Fahrenheita a opačně. Dodejte zdrojový kód funkce a příkaz Hive na vytvoření UDF. (3 body)

Termín: 31. 12. 2018

Řešení pošlete e-mailem na adresu jan.hucin@profinit.eu.

Případné dotazy pište na stejnou adresu.

PROFINIT



Diskuze

PROFINIT

Díky
za pozornost

PROFINIT
NÁSKOK DÍKY ZNALOSTEM

Profinit EU, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU