TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

DATABÁZE PRO BIG DATA
CVIČENÍ XI.
CASSANDRA

Lukáš Matějů

15.5.2024 | DPB

DNEŠNÍ CVIČENÍ

- zaměřené na základní práci s Cassandrou
  - tvorbu keyspace a tabulek
  - manipulaci s daty, import, filtrování
  - sekundární indexování, materialized view, uživatelské funkce
- veškeré potřebné soubory jsou dostupné na elearningu (cv11.rar)
  - python skript cv11.py s podrobným zadáním a init kódem
  - potřebná data pro import v messages_db.csv
    - zprávy z Discord serveru
    - room_id – jednoznačný identifikátor kanálu
    - speaker_id – jednoznačný identifikátor uživatele
    - time – čas odeslání zprávy
    - message – zpráva
- nezapomeňte si zapnout kontejner s Cassandrou

ÚLOHY

- detailněji popsané ve skriptu cv11.py

1. vytvořte keyspace dc a přepněte se do něj
   - SimpleStrategy, replication_factor 1
2. vytvořte tabulku message_db pro data z messages_db.csv
   - vhodně zvolte datové typy, jako primární klíč nastavte room_id a time
3. do tabulky message_db importujte data z messages_db.csv
4. vypište jednu zprávu
5. vypište 5 posledních zpráv v místnosti 1 odeslaných uživatelem 2
6. vypište počet zpráv odeslaných uživatelem 2 v místnosti 1
7. vypište počet zpráv v každé místnosti
8. vypište id všech místností (3 hodnoty)

BONUSOVÉ ÚLOHY

1. vytvořte materialized view pro tabulku messages, který bude obsahovat pouze čas, room_id a zprávu
   - vypište jeden výsledek z vytvořeného view
2. vytvořte vlastní funkci, která při výběru dat vrátí navíc příznak, zda vybraný text obsahuje nevhodný výraz
   - vyberte jeden výraz (nemusí být nevhodný) a otestujte
3. zjistěte čas odeslání nejnovější a nejstarší zprávy
4. zjistěte délku nejkratší a nejdelší zprávy na serveru
5. pro každého uživatele zjistěte průměrnou délku zprávy

- v celém cvičení by nemělo být použito ALLOW FILTERING
- Python musí sloužit jen k výpisu výsledků dotazů…