

TECHNOLOGIE PRO BIG DATA CVIČENÍ VIII. APACHE SPARK IV

Lukáš Matějů 12.11.2024 | TPB



PŘÍPRAVA CVIČENÍ

- pro dnešní cvičení je potřeba netcat
 - připojení k lokálnímu streamu dat přes TCP socket
 - umožní předávání zpráv na vybraném portu
 - v Bitnami Spark Dockeru ale není obsažený
 - z minulého cvičení ale již máte root oprávnění
 - instalace pomocí příkazu apt-get install netcat



 spuštění netcat na vybraném portu pomocí příkazu nc-l-k-p 9999



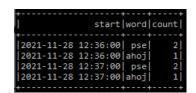
- netcat musíte spustit první
- port musí odpovídat portu nastavenému v Driver Programu ve Sparku

```
lines = spark.readStream.format("socket").option("host", "localhost").option("port", 9999).load()
```

v případě chybové hlášky o obsazenosti portu jen změňte číslo portu

DNEŠNÍ CVIČENÍ

- 1. počítejte četnost slov v textu vstupního streamu
 - jedná se o rozšířenou verzi úlohy z přednášky na Structured Streaming
 - stream bude číst textová data přes TCP socket z vámi vybraného portu
 - data si budete předávat pomocí netcat
 - program počítá četnost slov na vstupu
 - slova jsou před zpracováním převedena na malá písmena a je odstraněna interpunkce
 - slova jsou seřazena od nejčetnějšího po nejméně četná
 - výstup je předáván na konzoli, kde je i vypisován
 - doplňte funkcionalitu pro zpracování posuvným oknem
 - okno délky 30 sekund s posunem 15 sekund
 - bude potřeba nový sloupec s aktuálním časem
 - pro přehlednost zobrazte začátek i konec okna
 - výsledek setřiďte prvně podle začátku okna a následně podle četnosti







DNEŠNÍ CVIČENÍ

- 2. BONUS: četnost slov v souborech přidávaných do adresáře
 - úlohu z prvního cvičení upravte
 - vstupem tentokrát bude každý soubor vložený do vybraného adresáře
 - vyhodnocení by mělo být spuštěno po každém vloženém souboru
 - okna pro tento příklad ignorujte

