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

TECHNOLOGIE PRO BIG DATA
CVIČENÍ X.
APACHE FLINK II

Lukáš Matějů

28.11.2024 | TPB

DNEŠNÍ CVIČENÍ

- úlohu řešte výhradně pomocí Table API & SQL

1. nalezněte deset filmů s nejvíce nejlepšími hodnoceními (5 *)
   - k dispozici máte soubory *u.data* (hodnocení) a *u-mod.items* (filmy)
     - data z MovieLens (datové sady), info viz přednášky
   - vytvořte skript vracející 10 filmů s nejvíce hodnoceními známkou 5
     - skript pojmenujte *most-hyped-movies.py*
   - výsledný seznam vraťte setříděný podle počtu nejlepších hodnocení
     - počet také vypisujte
   - ze souboru *u-mod.items* zjistěte k nalezeným filmům pravá jména
   - k finální tabulce přidejte sloupec počítající poměr nejlepších hodnocení daného filmu vůči všem hodnocením daného filmu
     - sestupně setřiďte podle tohoto sloupce

BONUSOVÁ ÚLOHA

- zpracování a analýza proudu dat (iDNES článků) z Kafky
  - úlohu řešte výhradně pomocí Table API & SQL
- nejprve je nutné provést přípravu na cvičení
  - z elearnigu si stáhněte balík dat a rozbalte je do nového adresáře
  - v adresáři je docker-compose.yml vytvářející nový cluster
    - Flink (1 JM + 2 TM), Kafka a Zookeaper
    - Kafka a Flink jsou postaveni podle odpovídajících Dockerfiles
    - Kafka spouští skript start-stream.sh streamující články z idnes.json v náhodných intervalech
    - Flink obsahuje skript pyflink-kafka-example.py čtoucí Kafka proud s výpisem do konzole
    - váš aktuální adresář je opět přístupný ve Flinku pod /files
    - soubor idnes.json obsahuje ukázková data
  - postavte images Kafky a Flinku

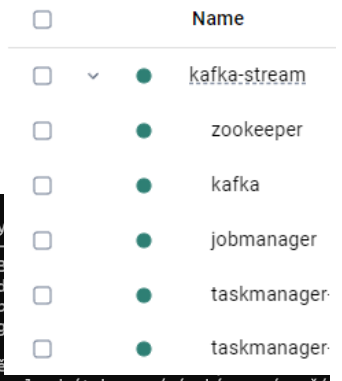
    docker-compose build

BONUSOVÁ ÚLOHA

- nejprve je nutné provést přípravu na cvičení
  - spusťte cluster
    - s přehledem streamovaných článků z Kafky v dané konzoli
      docker-compose up --scale taskmanager=2
    - klasicky na pozadí
      docker-compose up --scale taskmanager=2 -d
  - přepněte se do kontejneru Flinku
    docker exec -i -t jobmanager /bin/bash
  - spusťte skript pyflink-kafka-example.py
    - budou se vám postupně vypisovat streamované články

```
1> +I[Konec týdne přinese déšť a ochlazení. Objeví se přízemní mrazíky]
1> +I[Hnutí bez výrazných tváří. Jihomoravské ANO hledá „vlastního Grolicha“]
```

  - soubor idnes.json nahraďte svými data
    - ujistěte se, že je váš json soubor validní
  - nyní jste připraveni na samotné cvičení…



kafka
dnes.cz/zpravy
rael-iran-joe-
jo","title":"B
řekl Biden. Úd
ntent":["Hrozb
východního reg
cký prezident
ude stát ještě
ží odradit Izrael od útoku na íránská ropná zaří
zení, odpověděl, že nevyjednává na veřejnosti.",
"Biden ve čtvrtek prohlásil, že Washington disku
tuje s Tel Avivem o možných izraelských úderech
na íránskou ropnou infrastrukturu. Přímé odpověd
i na otázku, zda je proti nim, se ale vyhnul. Ji
ž o den dříve řekl, že nepodporuje možný útok na
íránská jaderná zařízení, o němž uvažují někteř
í izraelští politici.","Írán v úterý vyslal témě
ř 200 raket na Izrael, který krátce před tím zah
ájil pozemní operace na jihu Libanonu, kde se sn
aží zlikvidovat Teheránem podporované hnutí Hizb
alláh.","Tel Aviv Íránu opakovaně pohrozil za út
ok odvetou, která bude pro Írán „bolestivá“. Pod
le Bidena však není odvetná akce otázkou nejbliž
ších hodin. Dá se také zabránit úplné blízkových
odní válce.","„Nemyslím si, že bude válka mezi v
šemi. Myslím, že tomu můžeme zabránit,“ řekl Bid
en novinářům. „Ale ještě je třeba udělat hodně,“
dodal. Zopakoval při tom, že USA budou nadále p
odporovat Izrael."],"tags":["Izrael ve válce","Í
rán","Joe Biden","Tel Aviv","válka","Izrael"],"p
hoto_count":4,"time":"2024-10-04T06:55:00+02:00"
,"comment_count":12,"category":"zpravy"}

BONUSOVÁ ÚLOHA

- a nyní již samotné zadání…
  - vhodně navrhněte vstupní tabulku
  - do tabulky vkládejte již přímo jednotlivé informace o článcích
    - o parsování dat z formátu json se musí postarat přímo Flink
    - přidejte časovou informaci – vyberte si mezi odesláním z Kafky a přijetím do Flinku
  - do konzole vypisujte název zpracovávaného článku
  - detekujte aktivní články s více než 100 komentáři
    - zapisujte do souboru ve formátu název článku;počet komentářů
  - detekujte články s datem přidání na iDNES mimo pořadí
    - články jsou Kafkou streamovány sekvenčně
    - uvažujte čas přidání článků sestupný nebo vzestupný (podle toho, jak jste data stahovali)
    - detekujte články, které toto pravidlo porušují
    - zapisujte do souboru ve formátu název článku; datum článku; předchozí datum

BONUSOVÁ ÚLOHA

- a nyní již samotné zadání…
  - data také zpracovávejte posuvným oknem nad vámi přidaným časem
    - s délkou 1 minuta a posunem 10 vteřin
  - počítejte celkový počet článků přidaných v daném okně a kolik z nich obsahuje v textu článku výraz válka
    - zapisujte do souboru ve formátu začátek okna; konec okna; počet článků; počet výskytů
  - všechny výpisy jsou prováděny zároveň jediným skriptem
    - 1x konzole, 3x soubory
    - použijte pipeline

```
# Write results to sinks
pipeline = table_env.create_statement_set()
pipeline.add_insert("article_count_sink", article_counts)  # Write article counts to file
pipeline.add_insert("high_priority_sink", high_priority_articles)  # Write high-priority articles to file
pipeline.add_insert("article_title_console_sink", article_titles)  # Print titles to the console

# Execute the pipeline
pipeline.execute().wait()
```

BONUSOVÁ ÚLOHA

- problémy po restartu clusteru…
  - Kafka nestreamuje data

```
kafka          | [2024-12-01 18:38:43,953] WARN [AdminClient clientId=adminclient-1] Connection to node -1 (kafka/172.18.0.3:9092) could n
ot be established. Broker may not be available. (org.apache.kafka.clients.NetworkClient)
kafka          | [2024-12-01 18:38:44,058] WARN [AdminClient clientId=adminclient-1] Connection to node -1 (kafka/172.18.0.3:9092) could n
ot be established. Broker may not be available. (org.apache.kafka.clients.NetworkClient)
kafka          | [2024-12-01 18:38:44,160] WARN [AdminClient clientId=adminclient-1] Connection to node -1 (kafka/172.18.0.3:9092) could n
ot be established. Broker may not be available. (org.apache.kafka.clients.NetworkClient)
```

  - jedním řešením je smazání clusteru a jeho znovuvytvoření

	Name	Container ID	Image	Port(s)	CPU (%)	Last star	Actions
	kafka-stream	-	-	-	N/A	14 secon	
	zookeeper	d691014f5493	confluentin	2181:2181	N/A	15 secon	
	kafka	4ea21df0e2f8	kafka-strea	9092:9092	N/A	15 secon	
	jobmanager	0eca158835b3	kafka-strea	8081:8081 Show all ports (2)	N/A	15 secon	
	taskmanager-	8d50118e7690	kafka-strea		N/A	14 secon	
	taskmanager-	96bc811bcd3c	kafka-strea		N/A	14 secon	

  - kdo vymyslí lepší řešení, získá bonusový bod