



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

TECHNOLOGIE PRO BIG DATA

CVIČENÍ IX.

APACHE FLINK

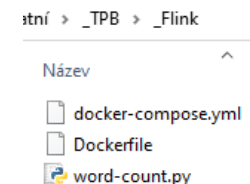
Lukáš Matějů

7.12.2023 | TPB



PŘÍPRAVA CVIČENÍ

- Apache Flink
 - cvičení je založené na oficiálním Docker [image](#)
 - pro simulaci clusteru je používán Docker Compose
 - jak na to?
 1. stáhněte si z elearningu soubor s daty k dnešnímu cvičení
 2. rozbalte jej do libovolného adresáře a v konzoli se tam přesuňte
 3. nejprve je potřeba postavit custom image, který rozšíří Flink image o PyFlink
 - na detaily se můžete podívat do souboru Dockerfile nebo do oficiální [dokumentace](#)



```
D:\Prezentace\_ostatní\_TPB\_Flink>docker build --tag pyflink:latest .
```

```
FROM flink:1.14.0
```

```
# install python3: it has updated Python to 3.9 in Debian 11 and so install Python 3.7 from source  
# it currently only supports Python 3.6, 3.7 and 3.8 in PyFlink officially.
```

```
RUN apt-get update -y && \  
apt-get install -y build-essential libssl-dev zlib1g-dev libbz2-dev libffi-dev && \  
wget https://www.python.org/ftp/python/3.7.9/Python-3.7.9.tgz && \  
tar -xvf Python-3.7.9.tgz && \  
cd Python-3.7.9 && \  
./configure --without-tests --enable-shared && \  
make -j6 && \  
make install && \  
ldconfig /usr/local/lib && \  
cd .. && rm -f Python-3.7.9.tgz && rm -rf Python-3.7.9 && \  
ln -s /usr/local/bin/python3 /usr/local/bin/python && \  
apt-get clean && \  
rm -rf /var/lib/apt/lists/*
```

```
# install PyFlink  
RUN pip3 install apache-flink==1.14.0
```

PŘÍPRAVA CVIČENÍ

- Apache Flink

- jak na to?

- 4. spusťte Docker Compose

`docker-compose up -d`

- jedná se o rozšířenou verzi oficiálního [docker-compose.yml](#)
- současný adresář je mapovaný na /files pomocí volumes
- cluster se skládá z 1 JobManagera a 2 TaskManagerů
- jako image je pyflink:latest postavený v předchozím kroku

```
D:\Prezentace\_ostatní\TPB\Flink>docker-compose up -d
[+] Running 3/3
- Container flink-jobmanager-1 Started
- Container flink-taskmanager-2 Started
- Container flink-taskmanager-1 Started
```

```
root@dfb4fceb0dafb:/files# ls
Dockerfile  docker-compose.yml  word-count.py
```

```
version: "2.2"
services:
  jobmanager:
    image: pyflink:latest
    ports:
      - "8081:8081"
    command: jobmanager
    environment:
      - |
        FLINK_PROPERTIES=
        jobmanager.rpc.address: jobmanager
    volumes:
      - "../files:rw"
```

```
taskmanager:
  image: pyflink:latest
  depends_on:
    - jobmanager
  command: taskmanager
  deploy:
    replicas: 2
  environment:
    - |
      FLINK_PROPERTIES=
      jobmanager.rpc.address: jobmanager
      taskmanager.numberOfTaskSlots: 2
  volumes:
    - "../files:rw"
```

PŘÍPRAVA CVIČENÍ

- Apache Flink
 - jak na to?
 5. přepněte se do běžícího kontejneru JobManagera
`docker exec -i -t flink-jobmanager-1 /bin/bash`

```
D:\Prezentace\_ostatní\_TPB\_Flink>docker exec -i -t flink-jobmanager-1 /bin/bash
root@dfb4fce0dafb:/opt/flink#
```

- jméno JobManagera můžete zjistit z `docker ps -a`

```
D:\Prezentace\_ostatní\_TPB\_Flink>docker ps -a
```

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS	NAMES
c7bfdab40895	pyflink:latest	"/docker-entrypoint..."	16 minutes ago	Up 15 minutes	6123/tcp, 8081/tcp	flink-taskmanager-1
0dcf3583f466	pyflink:latest	"/docker-entrypoint..."	16 minutes ago	Up 15 minutes	6123/tcp, 8081/tcp	flink-taskmanager-2
dfb4fce0dafb	pyflink:latest	"/docker-entrypoint..."	16 minutes ago	Up 15 minutes	6123/tcp, 0.0.0.0:8081->8081/tcp	flink-jobmanager-1

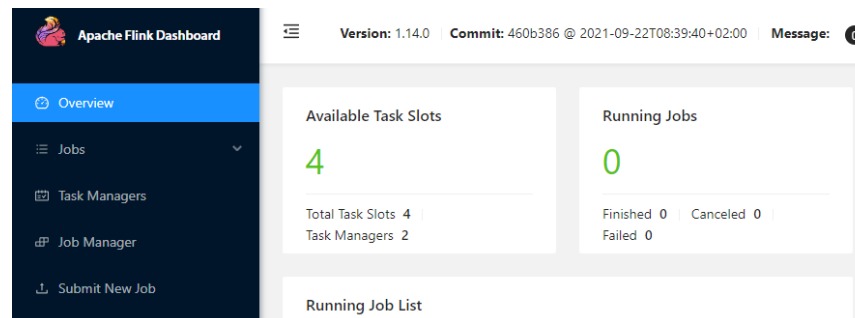
- nyní by již mělo být vše připravené k použití...

PŘÍPRAVA CVIČENÍ

- Apache Spark
 - funguje vše jak má?
 1. v kontejneru se přesuňte do /files a vypište si obsah adresáře
 - cd /files
 - ls
 - pokud nevidíte soubor docker-compose.yml, něco je špatně...

```
root@dfb4fce0dafb:/opt/flink# cd /files/  
root@dfb4fce0dafb:/files# ls  
Dockerfile  docker-compose.yml  word-count.py
```

2. v prohlížeči přejděte na adresu <http://localhost:8081/>
 - pokud nevidíte webové rozhraní, něco je špatně...



PŘÍPRAVA CVIČENÍ

- Apache Spark
 - funguje vše jak má?
 3. spustíte skript
 - python word-count.py
 - skript vypíše četnost slov v souboru, případně v uloženém vstupu

```
root@dfb4fce0dafb:/files# python word-count.py
Executing word_count example with default input data set.
Use --input to specify file input.
Printing result to stdout. Use --output to specify output path.
(a,5)
(Be,1)
(Is,1)
(No,2)
(Or,1)
(To,4)
(be,1)
(by,2)
(he,1)
(in,3)
(is,2)
(my,1)
(of,14)
```

- pokud se skript neprovede, něco je špatně...

DNEŠNÍ CVIČENÍ

1. počty slov začínající na jednotlivá písmena abecedy v textu
 - vstupem je adresář, jehož soubory jsou zpracovávány dávkově
 - program počítá četnost slov začínajících na jednotlivá písmena abecedy
 - slova jsou převedena na malá písmena
 - slova nezačínající na znaky A-Z jsou odfiltrována
 - výstup je předán na konzoli, kde je i vypisován
 - jako testovací soubor můžete použít knihu z předchozích cvičení
2. **BONUS:** stejnou úlohu vyřešte i pro proudové zpracování
 - vybraný adresář je v pravidelných intervalech monitorován
 - při vložení souboru dojde k automatickému vyhodnocení

```
6> (v,234)
1> (h,1213)
8> (d,1310)
3> (b,2091)
4> (p,1900)
6> (x,3)
```

```
5> (w,2479)
2> (i,3510)
7> (a,5219)
8> (f,1635)
3> (e,1243)
2> (j,257)
```