Sprawozdanie 10

Grupa A3:

inż. Michał Liss

inż. Marceli Sokólski

inż. Piotr Krzystanek

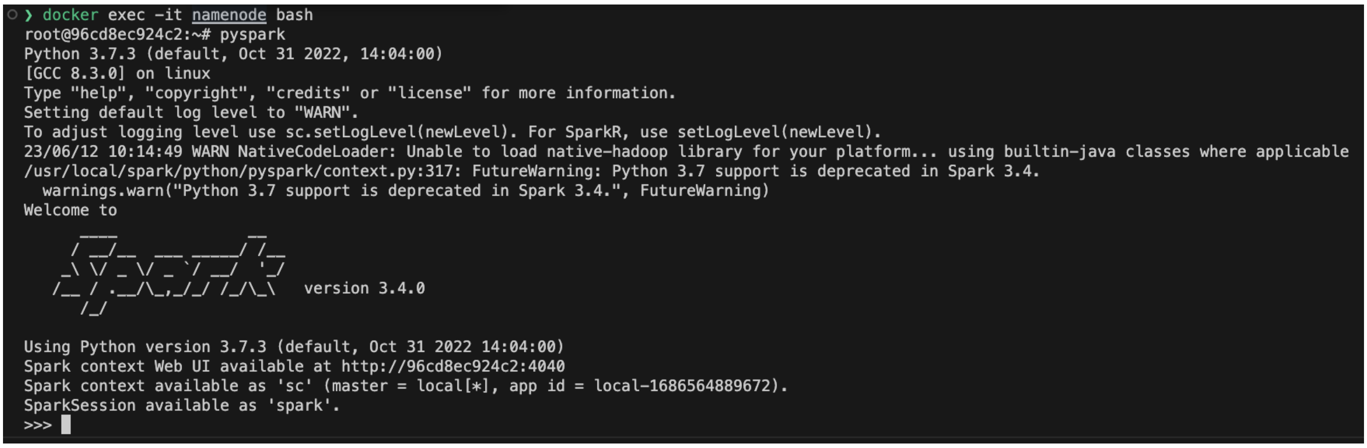
Instalacja spark w kontenerze namenode

```
RUN tar -xzf spark-3.4.0-bin-hadoop3-scala2.13.tgz
RUN mv spark-3.4.0-bin-hadoop3-scala2.13 /usr/local/spark
RUN rm spark-3.4.0-bin-hadoop3-scala2.13.tgz
ENV SPARK_HOME=/usr/local/spark
ENV PATH="${PATH}:${SPARK_HOME}/bin"
```

Instalacja python w pozostałych węzłach klastra

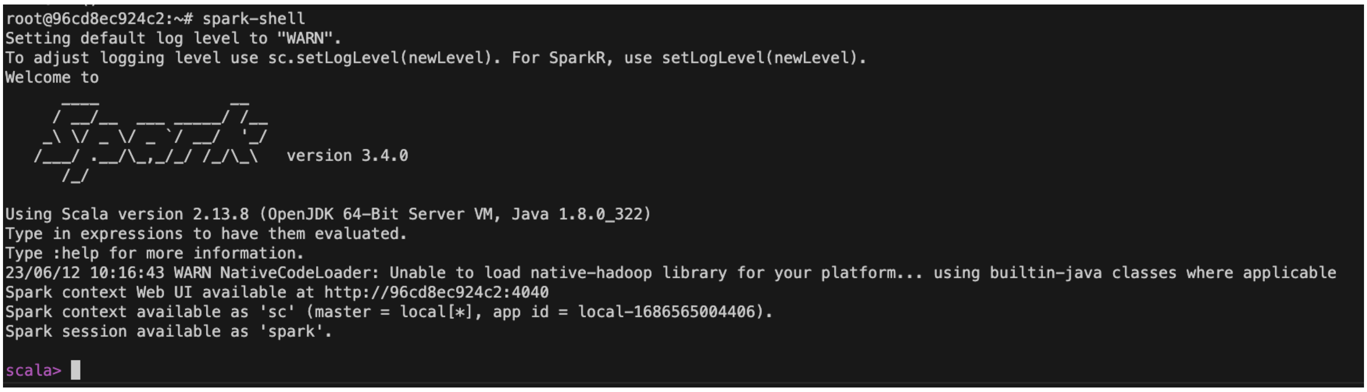
```
RUN apt-get update && DEBIAN_FRONTEND=noninteractive apt-get install -y --no-install-recommends python3.6
RUN update-alternatives --install /usr/bin/python python /usr/bin/python3 1
```

pyspark



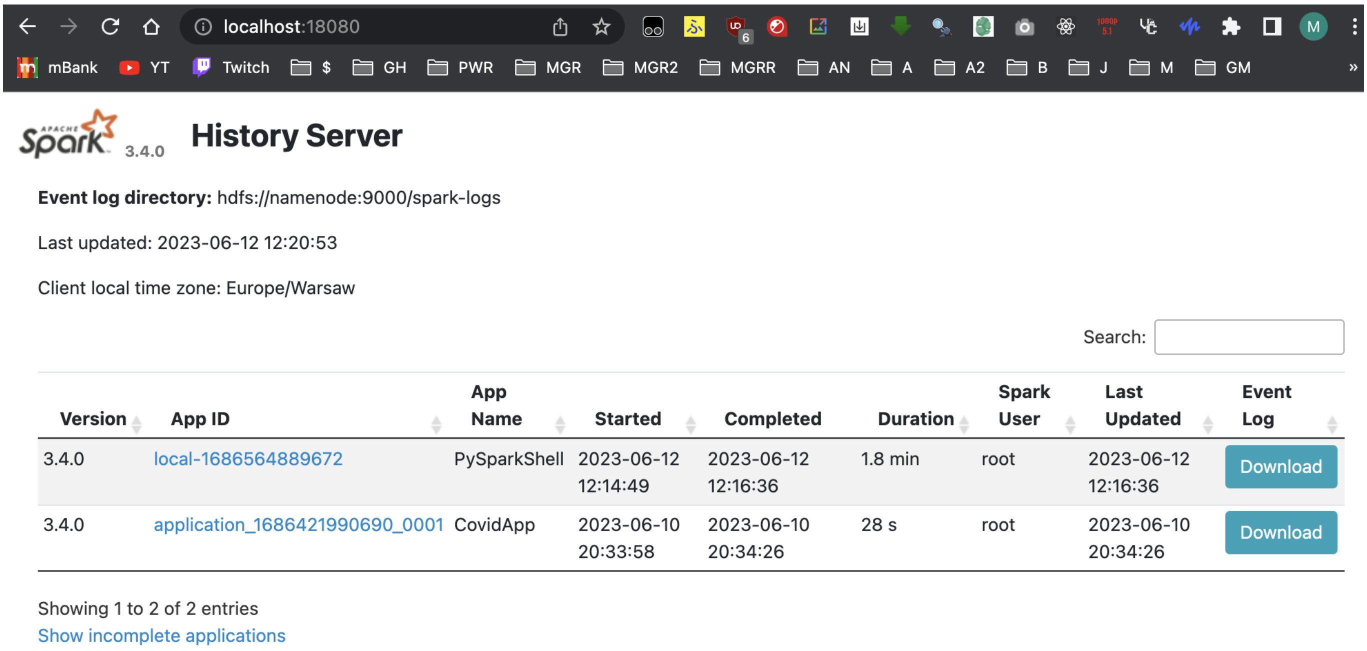
```
❯ docker exec -it namenode bash
root@96cd8ec924c2:~# pyspark
Python 3.7.3 (default, Oct 31 2022, 14:04:00)
[GCC 8.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/06/12 10:14:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/usr/local/spark/python/pyspark/context.py:317: FutureWarning: Python 3.7 support is deprecated in Spark 3.4.
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.7.3 (default, Oct 31 2022 14:04:00)
Spark context Web UI available at http://96cd8ec924c2:4040
Spark context available as 'sc' (master = local[*], app id = local-1686564889672).
SparkSession available as 'spark'.
>>> 
```

spark-shell



```
root@96cd8ec924c2:~# spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Scala version 2.13.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_322)
Type in expressions to have them evaluated.
Type :help for more information.
23/06/12 10:16:43 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://96cd8ec924c2:4040
Spark context available as 'sc' (master = local[*], app id = local-1686565004406).
Spark session available as 'spark'.

scala> 
```

historyserver



History Server

Event log directory: hdfs://namenode:9000/spark-logs

Last updated: 2023-06-12 12:20:53

Client local time zone: Europe/Warsaw

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.4.0	local-1686564889672	PySparkShell	2023-06-12 12:14:49	2023-06-12 12:16:36	1.8 min	root	2023-06-12 12:16:36	Download
3.4.0	application_1686421990690_0001	CovidApp	2023-06-10 20:33:58	2023-06-10 20:34:26	28 s	root	2023-06-10 20:34:26	Download

Showing 1 to 2 of 2 entries

Show incomplete applications

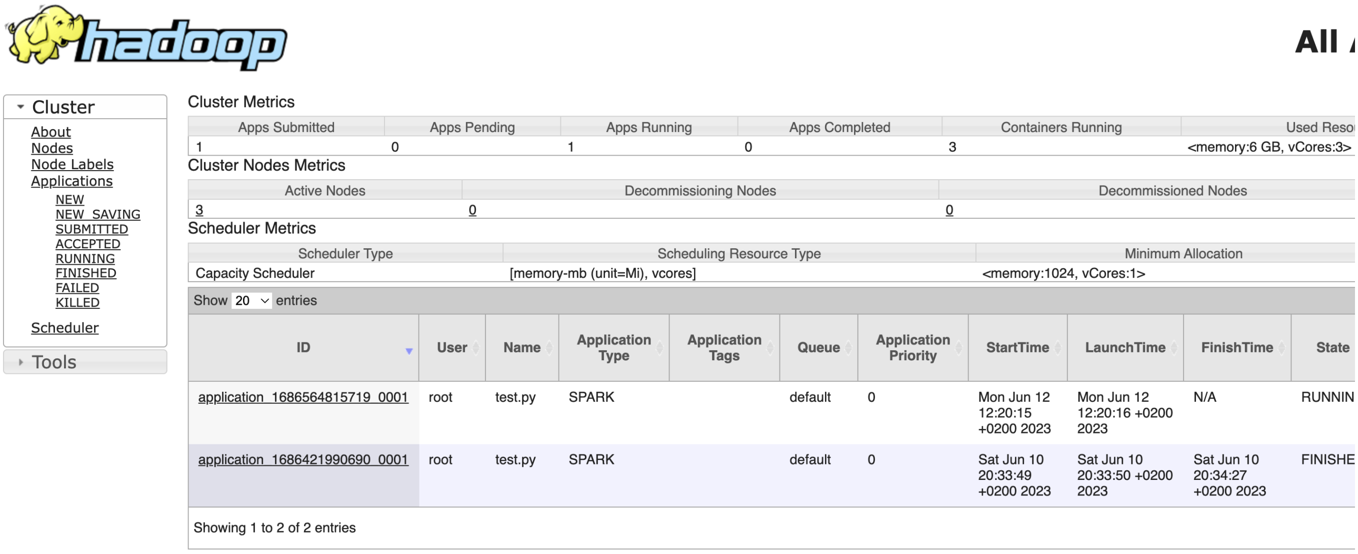
Przykładowy skrypt (pyspark + pyspark.sql)

```python
from pyspark.sql import SparkSession

covid = "/datasets/covid-dataset.jsonl"
spark = SparkSession.builder.appName("CovidApp").getOrCreate()

df = spark.read.json(covid)
df.select("location").write.csv('/spark-result/dataframe-select',
header=True)

spark.read.json(covid).createOrReplaceTempView("covid")
spark.sql("SELECT location FROM covid").write.csv('/spark-result/sql-
select', header=True)

spark.stop()
```

Uruchomienie z poziomu jupyter notebook

```
_ = run_in_master("spark-submit --master yarn --deploy-mode cluster
/data/master_volume/spark_scripts/test.py")
_ = merge_results("/spark-result/sql-select")
print_hdfs_output("/spark-result/sql-select")
```

Uruchomiona aplikacja na hadoopie



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Reso
1	0	1	0	3	<memory:6 GB, vCores:3>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
3	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State
application_1686564815719_0001	root	test.py	SPARK		default	0	Mon Jun 12 12:20:15 +0200 2023	Mon Jun 12 12:20:16 +0200 2023	N/A	RUNNIN
application_1686421990690_0001	root	test.py	SPARK		default	0	Sat Jun 10 20:33:49 +0200 2023	Sat Jun 10 20:33:50 +0200 2023	Sat Jun 10 20:34:27 +0200 2023	FINISHE

Showing 1 to 2 of 2 entries