# Analiza dużych zbiorów danych – Laboratorium

## Laboratorium 1 - Wstęp do Apache Spark

## **1. Przygotowanie środowiska**

Mamy (co najmniej) 3 sposoby na uruchomienie Apache Spark oraz Jupyter Notebook:

### W laboratorium 3.27d

Uruchamiamy w terminalu:

jupyter-notebook

### Lokalne środowisko oparte na Docker

Środowisko można skonfigurować na własnym komputerze, przy czym wymagane jest ok. 4GB miejsca na dysku.

docker run -p 8888:8888 -p 4040:4040 -v $HOME:/home/jovyan/work jupyter/all-spark-notebook

Jupyter: <http://localhost:8888/>

Spark: <http://localhost:4040/> (after starting SparkSession)

**Dla zainteresowanych**: można uruchomić obraz z repozytorium

<https://hub.docker.com/repository/docker/malawski/adzd>

docker run -p 8888:8888 -p 4040:4040 -v $HOME:/home/jovyan/work malawski/adzd

W tym kontenerze dodana jest obsługa Amazon S3.

Hasło: adzd

### Spark na Google Collab:

Można stworzyć własne środowisko na Google Collab <https://colab.research.google.com/>

oraz dodać co niego Spark 3.1.2:

<https://colab.research.google.com/drive/1EcotODzgSnLozSH3hDuBfZr06gJXY8I0?usp=sharing>

<https://medium.com/analytics-vidhya/getting-started-spark3-0-0-with-google-colab-9796d350d78>