# Spark – *Pandas API on Spark*

Programujesz w Pythonie? Znasz bibliotekę Pandas? Używasz jej na co dzień? Wiesz jak ważna jest dla osób uznających się za analityków danych (*data scientist*)? Doskwiera Ci fakt, że rozmiar Twoich danych przekracza możliwości Twojej maszyny, Pythona i Pandas? Jest na to sposób – *Pandas API on Spark*, czyli biblioteka odwzorowująca bibliotekę Pandas na platformie Apache Spark.

## Uruchomienie środowiska

Na początku przygotujemy środowisko przetwarzania danych.

Wykorzystamy w tym celu z *Dataproc* – klastra *Hadoop* z dodatkowymi komponentami dostępny w ramach *Google Cloud Platform*.

1. Korzystając z poniższego polecenia i konsoli *Cloud Shell* utwórz klaster.

```
gcloud dataproc clusters create ${CLUSTER_NAME} \
  --enable-component-gateway --region ${REGION} \
  --master-machine-type n1-standard-4 --master-boot-disk-size 50 \
  --num-workers 2 --worker-machine-type n1-standard-2 --worker-boot-disk-size 50 \
  --image-version 2.1-debian11 \
  --optional-components JUPYTER \
  --project ${PROJECT_ID} --max-age=3h
```

2. Czekając na uruchomienie klastra otwórz, mogą się przydać
   - dokumentację - `https://spark.apache.org/docs/latest/api/python/user_guide/pandas_on_spark/index.html` `https://spark.apache.org/docs/latest/api/python/reference/pyspark.pandas/index.html`
   - prezentację - *pandas API on Spark in 10 minutes* `https://docs.databricks.com/en/pandas/pandas-on-spark.html`

## Notatnik

3. Otwórz interfejs sieciowy środowiska *JupyterLab*

4. Pobierz, a następnie zaimportuj notatnik `SP08_w1_23-Spark-PandasAPI-zadania.ipynb`. Postępuj zgodnie ze znalezionymi tam instrukcjami.