Spark – DataFrames

DataFrames to sposób na zapewnienie podczas przetwarzania danych za pomocą Sparka bardzo wysokiej wydajności (dzięki między innymi takim rozwiązaniom jak *Catalyst* i *Tungsten*), a zarazem możliwości wykorzystania tak dobrze znanych mechanizmów przetwarzania relacyjnego oraz języka wysokiego poziomu jakim jest SQL.

W ramach zadań, które mają na celu zapoznać nas z podstawami przetwarzania tego typu danych wykorzystamy dwa zbiory danych.

- Pierwszym zbiorem danych będzie *20 Years of Games* zawierających recenzje gier z portalu `ign.com` pochodzących z:
  `https://www.kaggle.com/egrinstein/20-years-of-games`
- `DataFrames` to jednak nie tylko dane strukturalne, to także, w swojej postaci źródłowej, semistruktury jak np. JSON. Aby się o tym przekonać i jednocześnie zapoznać się z przetwarzaniem danych o bardziej złożonej budowie wykorzystamy dane uzyskane z portalu
  `https://www.dbis.informatik.uni-goettingen.de/Mondial/`
  ograniczone do informacji o krajach i miastach świata.

W ramach zadań będziemy starali się trzymać API dostarczanego przez typ `DataFrames` (nie korzystaj z SQL jeśli nie będzie to wynikało jednoznacznie z treści polecenia).

Przygotowanie środowiska i zbioru danych

Na początku przygotujemy środowisko przetwarzania danych.

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

1. Otwórz konsolę GCP `https://console.cloud.google.com/`, następnie korzystając z poniższego polecenia i terminala *Cloud Shell* utwórz klaster.

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

2. Przejdź do strony ze szczegółami klastra, na zakładkę *VM Instances*. Otwórz terminal SSH do serwera master utworzonego klastra, a następnie załaduj do katalogu HDFS dane, które będziemy przetwarzali
    a. `ign.csv` - *20 Years of Games*
    b. `mondial.countries.json` – dane z MondialDB dotyczące państw
    c. `mondial.cities.json` – dane z MondialDB dotyczące miast

```
wget https://jankiewicz.pl/bigdata/bigdata-sp/ign.csv
wget https://jankiewicz.pl/bigdata/bigdata-sp/mondial.countries.json
wget https://jankiewicz.pl/bigdata/bigdata-sp/mondial.cities.json
hadoop fs -mkdir -p .
hadoop fs -copyFromLocal * .
hadoop fs -ls
```

```
jankiewicz_krzysztof@hadoop-intro-m:~$ hadoop fs -ls
Found 3 items
-rw-r--r--   1 jankiewicz_krzysztof hadoop    2019628 2023-11-03 15:42 ign.csv
-rw-r--r--   1 jankiewicz_krzysztof hadoop     801242 2023-11-03 15:42 mondial.cities.json
-rw-r--r--   1 jankiewicz_krzysztof hadoop     130712 2023-11-03 15:42 mondial.countries.json
```

3. Zapoznaj się z zawartością pliku `ign.csv`

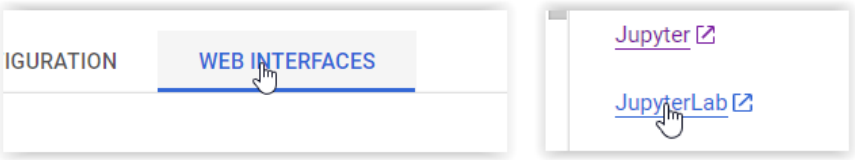
```
head ign.csv
```

```
jankiewicz_krzysztof@hadoop-intro-m:~$ head ign.csv
,score_phrase,title,url,platform,score,genre,editors_choice,release_year,release_month,release_day
0,Amazing,LittleBigPlanet PS Vita,/games/littlebigplanet-vita/vita-98907,PlayStation Vita,9.0,Platformer,Y,2012,9,
12
1,Amazing,LittleBigPlanet PS Vita -- Marvel Super Hero Edition,/games/littlebigplanet-ps-vita-marvel-super-hero-ed
ition/vita-20027059,PlayStation Vita,9.0,Platformer,Y,2012,9,12
2,Great,Splice: Tree of Life,/games/splice/ipad-141070,iPad,8.5,Puzzle,N,2012,9,12
3,Great,NHL 13,/games/nhl-13/xbox-360-128182,Xbox 360,8.5,Sports,N,2012,9,11
4,Great,NHL 13,/games/nhl-13/ps3-128181,PlayStation 3,8.5,Sports,N,2012,9,11
5,Good,Total War Battles: Shogun,/games/total-war-battles-shogun/mac-142565,Macintosh,7.0,Strategy,N,2012,9,11
6,Awful,Double Dragon: Neon,/games/double-dragon-neon/xbox-360-131320,Xbox 360,3.0,Fighting,N,2012,9,11
7,Amazing,Guild Wars 2,/games/guild-wars-2/pc-896298,PC,9.0,RPG,Y,2012,9,11
8,Awful,Double Dragon: Neon,/games/double-dragon-neon/ps3-131321,PlayStation 3,3.0,Fighting,N,2012,9,11
```

Zawartość pliku to zawierający nagłówek, rozdzielany przecinkami plik CSV. Zawiera on następujące pola:

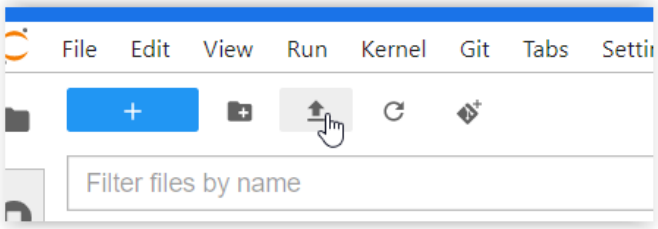
- `score_phrase` – opis oceny,
- `title` – tytuł gry,
- `url` – adres na stronie ign.com,
- `platform` – nazwa platformy dla której gra została zaprojektowana,
- `score` – wartość oceny,
- `genre` – gatunki gry,
- `editors_choice` – czy gra należy do gier wybranych przez redakcję,
- `release_year` – rok wydania
- `release_month` – miesiąc wydania
- `release_day` – dzień wydania

4. Zadania zrealizujemy korzystając z funkcjonalności środowiska notatnikowego *JupyterLab*. Przejdź zakładki *Web Interfaces*, a następnie wybierz stosowny link



5. Pobierz eksport notatnika `SP04_w1_23-DataFrames-API-SQL.ipynb`

6. Na stronie *JupyterLab* zaimportuj ten notatnik, wykorzystując opcję *Upload Files*



a następnie otwórz zaimportowany notatnik. Znajdziesz tam dalsze instrukcje.