MapReduce, Hadoop Streaming, Hive – opis projektu

Ogólny opis projektu

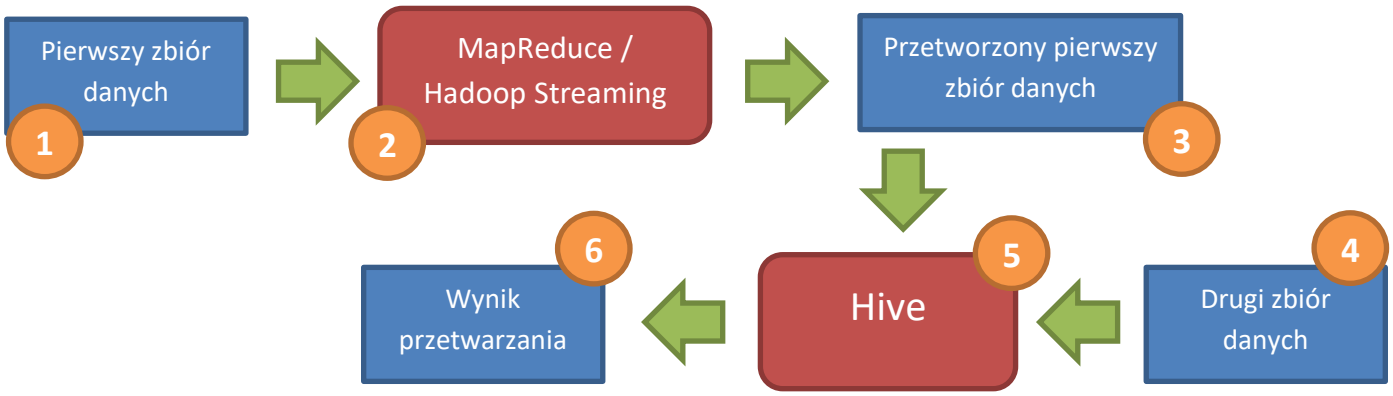
Celem projektu jest praktyczne wykorzystanie podstawowych platform przetwarzania danych stosowanych w środowiskach Big Data.

W ramach każdego z projektów będziemy przetwarzali dwa powiązane ze sobą zbiory danych.

Projekt będzie składał się z dwóch części.

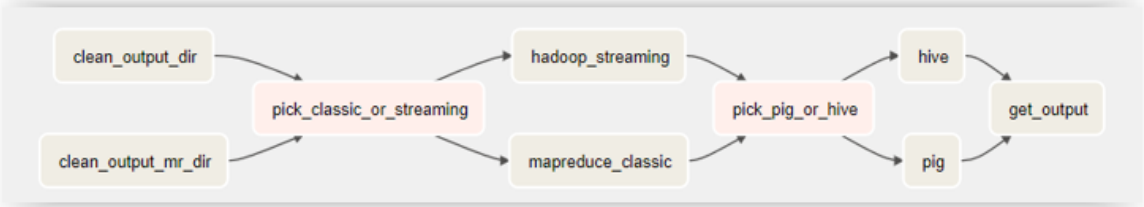
- W pierwszej części, za pomocą przetwarzania MapReduce w wariancie klasycznym (Java) lub Hadoop Streaming, będziemy przetwarzali jeden ze zbiorów danych dokonując jego filtrowania oraz agregacji.
- W drugiej części, za pomocą platformy Hive, będziemy przetwarzali wynik z pierwszej części oraz drugi ze zbiorów danych dokonując połączenia tych danych, dalszej agregacji, sortowania i filtrowania ostatecznych wyników.

Graficznie można projekt przedstawić następująco:



Technicznie projekt będzie składał się z:

1. Programu MapReduce w wariancie klasycznym (Java) lub Hadoop Streaming (2), który działając na pierwszym zbiorze danych (1) będzie generował wynik (3) umieszczając go w systemie plików HDFS.
2. Skryptu Hive (5), które działając na wyniku programu MapReduce (3) oraz drugim zbiorze danych (4), będzie generował ostateczny wynik przetwarzania (6) umieszczając go w systemie plików HDFS w formacie JSON
3. Przepływu Apache Airflow uruchamianego z poziomu jego interfejsu sieciowego, który będzie:
    a. przygotowywał system plików HDFS usuwając katalogi wynikowe z poprzednich uruchomień
    b. uruchamiał program MapReduce (2)
    c. uruchamiał program Hive (5)
    d. pobierał gotowy wynik przetwarzania (6) do lokalnego systemu plików i prezentował jego zawartość



Kilka wskazówek

Nie twórz rozwiązań bezpośrednio na GCP. Postaraj się tworzyć Twoje rozwiązania lokalnie. Oszczędzaj zasoby.

Nie uruchamiaj początkowych wersji programów na pełnych zbiorach danych. Postaraj się sprawdzić swoje rozwiązania na próbce danych, dopiero kiedy Twój program będzie gotowy, przetestuj go na pełnym wolumenie danych.

Nie ładuj danych bezpośrednio na klaster w GCP. Załaduj dane na zasobnik (*bucket*) i dopiero z zasobnika skopiuj je na klaster (`hadoop fs -copyToLocal gs://`), ewentualnie przetwarzaj je bezpośrednio z zasobnika.

Niniejszy dokument opisuje kwestie dotyczące programu MapReduce oraz skryptu Hive. Sposób przygotowania trzeciego składnika projektu - przepływu Apache Airflow – opisuje oddzielny dokument.

W przypadku wątpliwości odnośnie interpretacji danych zgłaszaj je i sprawdzaj na forum. Ustalenia, które tam będą miały miejsce są obowiązujące i mogą mieć wpływ na uznanie wyniku/przetwarzania za poprawny/poprawne.
Dużo kwestii zostało także rozwiązane w opisach poszczególnych zestawów zadań, które znajdziesz poniżej w sekcji Zestawy danych.

Hierarchia ważności ustaleń: ten dokument, forum, ustalenia z prowadzącym, inne.

Ogólne wymogi

Program MapReduce

- Program ma być parametryzowany
  - katalogiem danych źródłowych
  - katalogiem danych wynikowych
- Format plików wynikowych to `TextOutputFormat`

Skrypt Hive

- Skrypt ma być parametryzowany
  - `input_dir3` – katalogiem wejściowym dla przetworzonego pierwszego zbioru danych (wynik przetwarzania MapReduce) (3)
  - `input_dir4` – katalogiem wejściowym dla drugiego zbioru danych (4)
  - `output_dir6` – katalogiem wyjściowym, który będzie zawierał ostateczny wynik całości przetwarzania (6)
- Format plików wynikowych to pliki tekstowe z danymi w formacie JSON

Punktacja i terminy projektu

Punktacja poszczególnych składowych projektu jest opisana na stronach kursu.

Terminy oddawania projektu wynikają z aktywności obsługującej projekt. Oddajemy projekty i oceniamy je w sposób wynikający z dostępnych w tym celu aktywności Moodle.

Zestawy danych

Wszystkie zestawy danych pobieramy ze strony
`https://drive.google.com/drive/folders/1bQJoo63T-x6V27bR-K4DLRynPVNPZrCZ?usp=sharing`
niezależnie od ich oryginalnego źródła pochodzenia.

Pobieranie danych źródłowych, rozpakowanie plików i ładowanie ich do zasobnika (ewentualnie dodatkowo do systemu plików HDFS) nie należy do projektu – nie należy uzupełniać przepływu Apache Airflow o dodatkowe operatory, lub polecenia, które wykonują powyższe operacje. Operacje te należy wykonać wcześniej. W założeniu projektu dane źródłowe mają znajdować się już w zasobniku w podkatalogach:

- `projekt1/input/datasource1` – pierwszy zbiór danych
- `projekt1/input/datasource4` – drugi zbiór danych

Opis zestawów danych

Opisy poszczególnych zbiorów dostępne są w oddzielnych dokumentach.