Spark – projekt

Ogólny opis projektu

Celem projektu jest praktyczne wykorzystanie platformy Spark służącej do przetwarzania danych w systemach klasy Big Data.

Na podstawie danych źródłowych:

- posiadających różną charakterystykę (niestrukturalnych, semistrukturalnych, strukturalnych)
- składających się z wielu podzbiorów symulujących wiele źródeł danych

należy:

- dokonać analizy danych źródłowych
- zaprojektować na ich podstawie architekturę hurtowni danych w architekturze gwiazdy, składającą się z
  - tabeli faktów z miarami w liczbie 1-5
  - tabelami wymiarów w liczbie 3-5 (z obowiązkowym wymiarem czasu)
- zaprojektować 2-3 analizy danych oparte na zaprojektowanej hurtowni danych
- zaimplementować wspomnianą powyżej w hurtownię danych w Apache Hive w miejscu osiągalnym przez Sparka
- zaimplementować procesy ETL mających postać programów Sparka skompilowanych do postaci programów wykonywalnych `*.jar` rejestrowanych do powtarzalnego i bezobsługowego wykonywania, które z będą zasilały hurtownię danych bezpośrednio z danych źródłowych.
  Każdy z programów ma być wykorzystywany do zasilania pojedynczej tabeli w hurtowni danych
- zaimplementować wspomniane powyżej analizy mające postać interaktywnego notatnika Zeppelin – jeden notatnik na wszystkie analizy.

Uwagi szczegółowe:

- należy zadbać o odpowiednią parametryzację ścieżek, programów itp. w taki sposób aby było możliwe uruchamianie programów z dowolnego miejsca, w ramach dowolnego konta użytkownika, przy założeniu, że wszystkie wymagane i rozpakowane pliki źródłowe znajdują się w bieżącym katalogu
- należy zadbać o właściwy dobór abstrakcji danych przy przetwarzaniu danych o określonej charakterystyce
- podczas korzystania ze Spark SQL (przy implementacji programów ETL korzystających z danych strukturalnych) należy szczególną uwagę zwrócić na wykorzystywanie typowanych transformacji (*Typed Transformations*).
- projekty implementowane są w grupach 4-osobowych, przy czym każda osoba powinna pełnić określoną, istotną i jasno sprecyzowaną rolę w projekcie, ponadto każda osoba musi znać każdy fragment projektu (musi być gotowa do wyjaśnienia dowolnej jego części). Ponadto w każdym projekcie musi być wyznaczony leader, znający szczegóły zaangażowania oraz rolę w projekcie swoją i każdej z pozostałych osób.
- najpóźniej na 2 tygodnie przed terminem oddania projektu konieczne jest przedstawienie prowadzącemu projektu hurtowni danych oraz projektów planowanych analiz.

Kilka wskazówek

Nie twórz rozwiązań bezpośrednio na GCP. Postaraj się stworzyć fragmenty rozwiązania lokalnie. Oszczędzaj zasoby.

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

Nie ładuj danych bezpośrednio na klaster w GCP jeśli trwa to długo. Załaduj dane na zasobnik (bucket) i dopiero z zasobnika skopiuj je na klaster (`hadoop fs -copyToLocal gs://`).

Punktacja projektu

Za całość projektu można otrzymać 40 punktów

Podział punktów za poszczególne elementy składowe zostanie ustalony wkrótce i umieszczony poniżej.

Termin oddawania projektów wynika z harmonogramu zajęć. Oddawanie projektów odbywa się tylko i wyłącznie na zajęciach poszczególnych grup. Każdy tydzień spóźnienia powoduje naliczenie 10 punktów karnych.
Ostatecznym terminem oddawania projektów jest termin poprawki zaliczenia przedmiotu.

Opis zbiorów danych

Opis zbiorów danych jest dostępny w oddzielnych dokumentach na stronie kursu.