Kafka, Stream Processing – projekt

Ogólny opis projektu

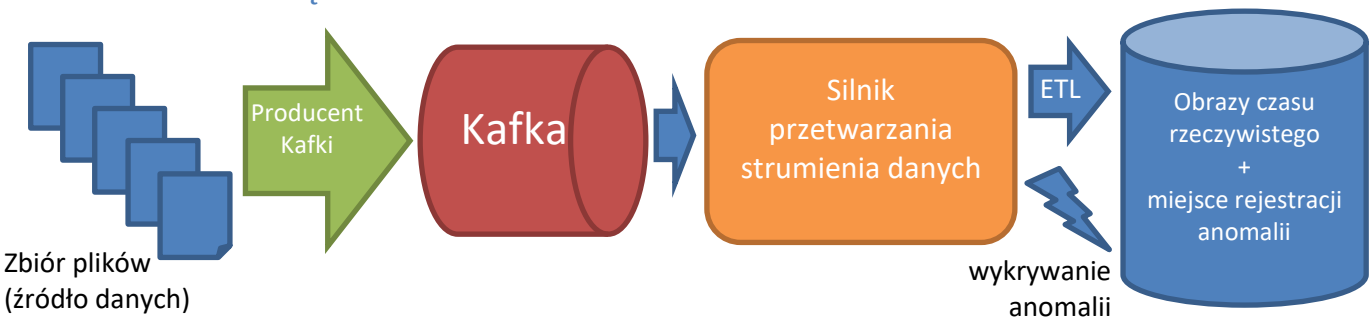
W ramach projektu należy samodzielnie zaimplementować rozwiązanie dokonujące przetwarzania strumieni danych w oparciu o:

- brokera wiadomości Kafka oraz
- określony silnik przetwarzania strumieni danych wykorzystywany w środowiskach Big Data, a także
- wybrane miejsce docelowe.

Dostępne silniki przetwarzania strumieni danych:

- Spark Structured Streaming
- Kafka Streaming
- Flink

Architektura rozwiązania



Opis

Dane źródłowe w naszym rozwiązaniu będą miały postać zbioru plików (do 100) dostępnych w jednym z katalogów.

Producent Kafki (zaimplementowany w ramach jednego z zestawów zadań) będzie odczytywał zawartość kolejnych plików z tego zbioru i wysyłał je, linia po linii, do brokera Kafki symulując w ten sposób zachodzenie zdarzeń w świecie rzeczywistym.

Twoim zadaniem będzie implementacja rozwiązania, które będzie:

- odczytywało dane z serwera Kafki
- utrzymywało na podstawie tych danych obraz czasu rzeczywistego
- reagowało na zachodzące "anomalie" rejestrując ich wystąpienia

Ponadto konieczne będzie wybranie właściwego (ze względu na własności) miejsca przechowywania obrazów czasu rzeczywistego oraz miejsca rejestracji anomalii. W obu przypadkach może to być to samo narzędzie/miejsce. Uwzględnij fakt, że na platformie Dataproc dostępna platforma Dockerowa – to daje praktycznie nieograniczone możliwości. Niestety nie każda z platform przetwarzania strumieni danych posiada konektory do każdego możliwego miejsca docelowego. Ważnym też są własności konektora.

Zbiory danych

Wszystkie zbiory danych pobieramy ze strony `http://www.cs.put.poznan.pl/kjankiewicz/bigdata/stream_project` niezależnie od ich oryginalnego źródła pochodzenia.

Kilka wskazówek

1. Nie ładuj wejściowych danych bezpośrednio na klaster. Załaduj dane jeden raz na zasobnik (bucket), a następnie, za każdym razem kiedy będzie taka potrzeba, kopiuj je z zasobnika na klaster (`hadoop fs -copyToLocal gs://`).
2. Nie twórz rozwiązań bezpośrednio na GCP. Postaraj się w miarę możliwości tworzyć Twoje rozwiązania lokalnie. Oszczędzaj zasoby.
3. Nie uruchamiaj początkowych wersji programów na pełnym zbiorze danych. Postaraj się sprawdzić swoje rozwiązania na próbce danych, dopiero kiedy Twój program będzie gotowy, przetestuj go na pełnym wolumenie danych.
4. Nie twórz początkowych wersji programów opierając się na złożonych przykładach. Rozpocznij od przepisywania danych z tematu wynikowego do ujścia. Jeśli to działa, wprowadzaj kolejno poszczególne transformacje cały czas mając wszystko pod kontrolą.
5. Rozpocznij tworzenie Twojego rozwiązania od zasilania wejściowego tematu "z konsoli", mając pod kontrolą dostarczanie każdej wiadomości, obserwując po każdej z nich to co dostajesz na wyjściu.

Punktacja projektu

Kryterium	Poziom 0 – 0%	Poziom 1 – 75%	Poziom 2 – 100%	Liczba punktów
Producent; skrypt zasilający	Brak, lub fundamentalne błędy uniemożliwiające działanie	Drobne błędy uniemożliwiające działanie, lub działanie jest możliwe ale niepoprawne	Ideał, spójny z resztą projektu	2
Silnik przetwarzania danych; program – procesy ETL	Brak lub brak spójności z tematem projektu lub fundamentalne błędy uniemożliwiające działanie	Drobne błędy uniemożliwiające działanie, lub działanie jest możliwe ale niepoprawne	Ideał, spójny z tematem i resztą projektu	8
Silnik przetwarzania danych; program – wykrywanie anomalii	Brak lub brak spójności z tematem projektu lub fundamentalne błędy uniemożliwiające działanie	Drobne błędy uniemożliwiające działanie, lub działanie jest możliwe ale niepoprawne	Ideał, spójny z tematem i resztą projektu	8
Silnik przetwarzania danych; jar	Brak	Istnieje, występują problemy z jego użyciem	Ideał, spójny z resztą projektu	4
Silnik przetwarzania danych; skrypt uruchamiający	Brak	Istnieje, występują problemy z jego użyciem	Ideał, spójny z resztą projektu	2
Konsument: skrypt odczytujący wynik przetwarzania	Brak, lub fundamentalne błędy uniemożliwiające działanie	Drobne błędy uniemożliwiające działanie, lub działanie jest możliwe ale niepoprawne	Ideał, spójny z resztą projektu	2
Miejsce utrzymywania obrazów czasu rzeczywistego; skrypt tworzący i użycie	Brak	Istnieje, występują problemy z jego użyciem	Ideał, spójny z resztą projektu	4
			Razem	30