Notatki do prezentacji wstępnej

Co to jest Spark?

Jest to szybki silnik do przetwarzania dużych danych. Główną zaletą Sparka jest możliwość wykonywania obliczeń w pamięci, co przyspiesza działanie aplikacji.

Cechy Sparka

1. SZYBKI

- Spark rozszerza funkcjonalność modelu MapReduce, oprócz obliczeń typu 'map' i 'reduce' wspiera też inne typy obliczeń, włączając interaktywane zapytania (nie musimy wykonywać całego dużego zapytania, tylko kawałek po kawałku patrząc na wyniki częściowe) i stream processing.
- szybkość jest ważna w analizie dużych zbiorów danych; jest różnica między czekaniam na wynik minutę lub godzinę.
- Spark oferuje możliwość obliczeń w pamięci co znacznie przyspiesza obliczenia (operacja czytania z dysku i pisania na dysk są bardzo kosztowne), ale nawet wykonując obliczenia na dysku dzięki złożonym aplikacjom Spark jest bardziej efektywny niż MapReduce.

2. ŁATWY W UŻYCIU

- możemy korzystać z high-levelowych API m.in. Pythona, dzięki czemu możemy się skupić na tym co chcemy zaimplementować a nie jak chcemy to zaimplementać (duże uproszczenia, ale prawdą jest, że nie musimy się zartwić i zastanawiać jak urównoleglić pracę bo Spark robi to za nas).
- Spark jest także integrowalny z innymi narzędziami Big Data, w szczególności możemy używać Spark na klastrach Hadoopa, dokonywać obliczeń na danych z Hadoopowych źródeł danych (np. Cassandra).

3. SZEROKIE ZASTOSOWANIE

- Spark jest platformą o dużych możliwościach, możemy łączyć różne typy obliczeń (zapytania SQL, text processing, machine learning). Wcześniej zaimplementowanie aplikacji, która używa tych typów obliczeń mogło wymagać kilku narzędzi, tu mamy wszystko w jednym narzędziu.
- możemy np. pobierać tweety w czasie rzeczywistym, później dokonać text processingu (czyszczenie i przygotowanie do analizy), za pomocą SQL szybko dokonać wstępnej analizy a za pomocą biblioteki Machine Learningowej dokonać predykcji jakiejś wielkości.

Rozszerzenia Spark Core

- Spark Core - zawiera podstawową funkcjonalność Sparka, włączając komponenty odpowiedzialne za rozplanowywanie zadań, zarządzanie pamięcią, interakcję z systemami gdzie przechowywane są dane itd. Spark Core definiuje również RDD.
- Spark SQL - biblioteka do pracy z ustrukturyzowanymi danymi. Oprócz SQL Spark pozwala na "mieszanie" zapytań SQL z programatycznym manipulowaniem danymi, które kest oferowane przez RDD.
- Spark Streaming - komponent umożliwiający przetwarzanie strumieni danych, np. danych o logowaniu generowane przez serwery webowe czy posty publikowane przez użytkowników serwisów społecznościwych (np. Twitter, Facebook)

- MLlib - biblioteka zawierająca szereg algorytmów machine learningowych, np. klasyfikacja, regresja, klasteryzacja. Wszystkie te metody są tak zaimplementowane, żeby obliczenia mogły być rzoproszone pomiędzy klastry.
- GraphX - bibliteka do przetwarzania grafów (np. graf sieci znajomych na Facebooku)