Wprowadzenie do Apache Spark

Justyna Jankowiak

Koło Naukowe Data Science, MiNI

19.04.2016

Co to jest Apache Spark?

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

Historia rozwoju Sparka

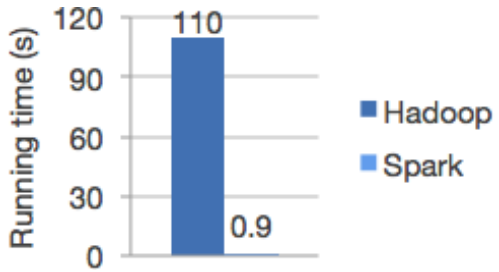
- **2009** - opracowanie oprogramowania na Uniwersytecie Kalifornijskim w Berkeley
- **2010** - opublikowanie Sparka jako wolne oprogramowanie zgodnie z licencją BSD (Berkeley Software Distribution License)
- **2013** - przekazanie Sparka do fundacji Apache Software
- **od 2014** - Spark jest jednym z czołowych projektów Apache

Berkeley
UNIVERSITY OF CALIFORNIA

THE APACHE® SOFTWARE FOUNDATION

Cechy Sparka

- **Szybkość** - działa do 100 razy szybciej niż MapReduce z wykorzystaniem pamięci operacyjnej i do 10 razy szybciej z wykorzystaniem operacji dyskowych



Rysunek : Źródło: http://spark.apache.org/

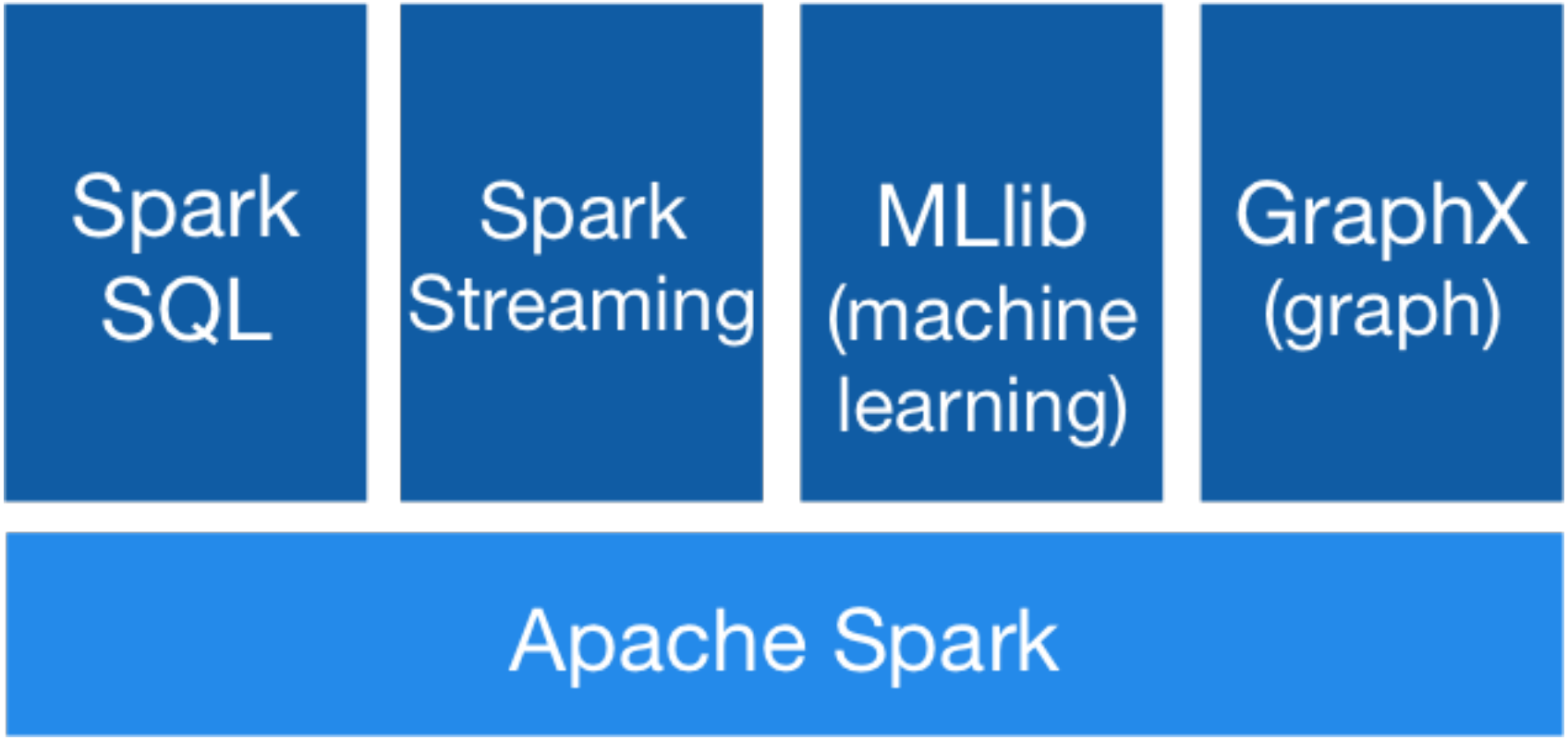
- **Łatwy w użyciu** - możliwość budowania aplikacji z wykorzystaniem języków Java, Scala, Python lub R
- **Zaawansowana analityka** - nie tylko operacje ‘map-reduce‘ ale również SQL, strumienie danych, machine learning oraz algorytmy grafowe

Resilient Distributed Datasets (leniwe kolekcje rozproszone :))
to podstawowa struktura danych w Sparku.

- Możliwość wykonywania równoległych operacji
- Sposoby tworzenia RDD
  - operacja *parallelize* na "zwykłych" obiektach (np. liście)
  - odwołanie do zewnętrznych źródeł danych (np. lokalny system plików lub HDFS)
- Wykonujemy na nich dwie operacje
  - transformacje (np. *map*, *filter*)
  - akcje (np. *collect*, *reduce*)

  Wszystkie transforacje są leniwe, tzn. nie są wykonywane dopóki nie jest to konieczne. Wykonywane są dopiero, gdy następujące po nich akcje wymagają zwrócenia wyniku.

Rozszerzenia Spark Core



Rysunek : Źródło: http://spark.apache.org/

Co dalej?

Zachęcam do uczestnictwa w bezpłatnych kursach na platformie e-learningowej edX:
https://www.edx.org/xseries/data-science-engineering-spark