DATA SCIENCE W PRAKTYCE

Analiza dużych zbiorów danych

CZĘŚĆ II: APACHE SPARK

Tematy

- Architektura platformy Apache Spark
- Spark Core (RDD)
- Spark SQL (DataFrames)
- Spark Streaming
- Spark Machine Learning Library (MLlib)
- Spark GraphX

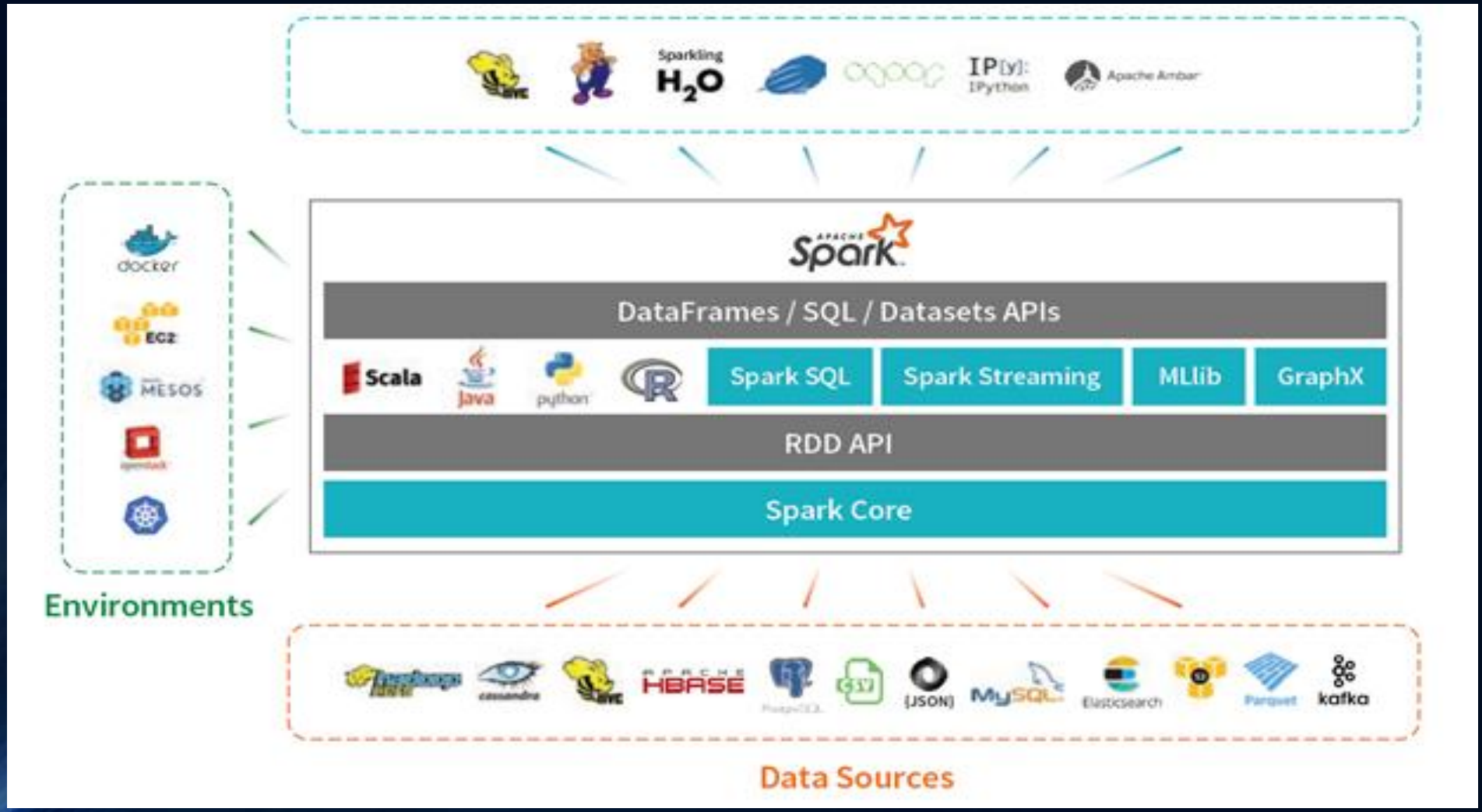
Apache Spark

PLATFORMA

Cechy Apache Spark

- Szybki silnik do ogólnego przetwarzania dużych zbiorów danych
- Zintegrowana platforma uruchamiana w klastrze maszyn
- Najpopularniejszy projekt Open Source rozwijany przez Apache Foundation, uniwersytet UC Berkeley oraz firmę DataBricks
  - http://spark.apache.org, http://databricks.com
  - Początek w 2009, publicznie 2013
- Przetwarzanie w modelu DAG (Directed Acyclic Graph) optymalizujące przepływ
- In-memory processing, optymalizacja wykonania na poziomie procesora
- Spark napisany jest w języku Scala
- Programy uruchamiane w Sparku pisane są w języku Scala, Java, Python oraz R

Architektura Apache Spark



Apache Spark vs Hadoop

- Framework
  - Pojedyncza platforma vs zestaw różnych komponentów
- Szybkość
  - Spark 100x szybszy w pamięci oraz 10x szybszy na dysku
  - Batch processing vs ~ RealTime processing / Iterative processing
- Model DAG vs MapReduce
- Łatwość użycia? Koszt?
- Odporność na błędy
- Security (Kerberos vs Shared secret, ACL na HDFS)
- Wsparcie SQL?
- Dane przechowywane w HDFS, S3, RDBMS, HBase, Cassandra, ...

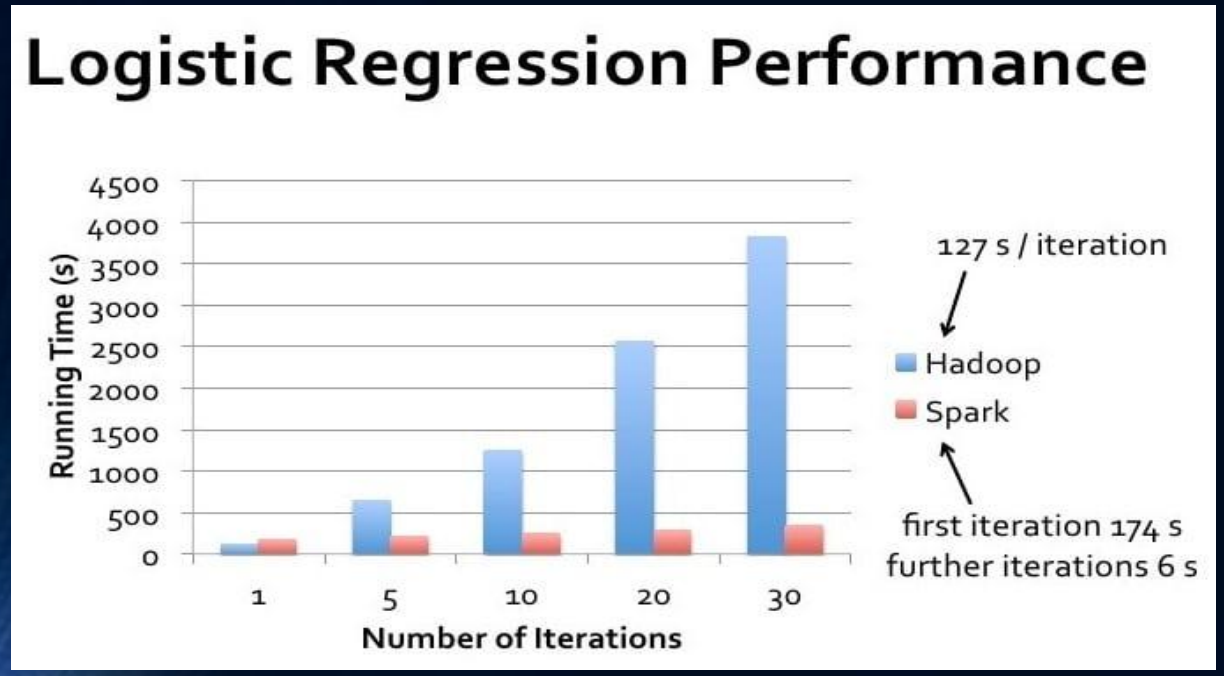
Hadoop	Spark
Hive	SparkSQL
Apache Graph	Graphax
Impala	SparkSQL
Apache Storm	Spark Streaming
Apache Mahout	MLlib

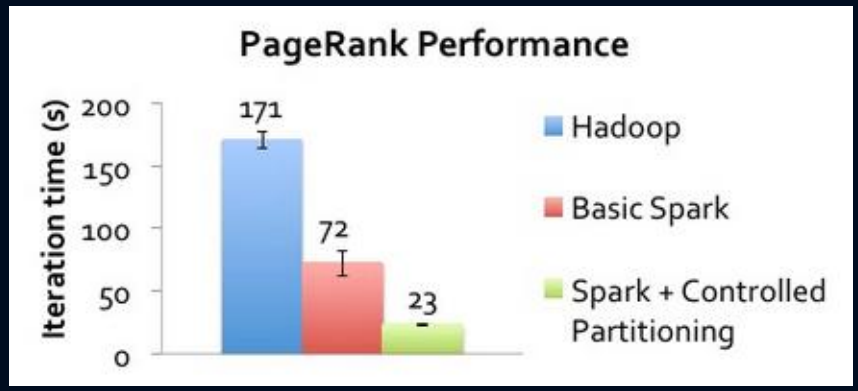
Apache Spark vs MapReduce

	MapReduce on Hadoop	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

Źródło: https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html

Apache Spark vs MapReduce





Spark Core

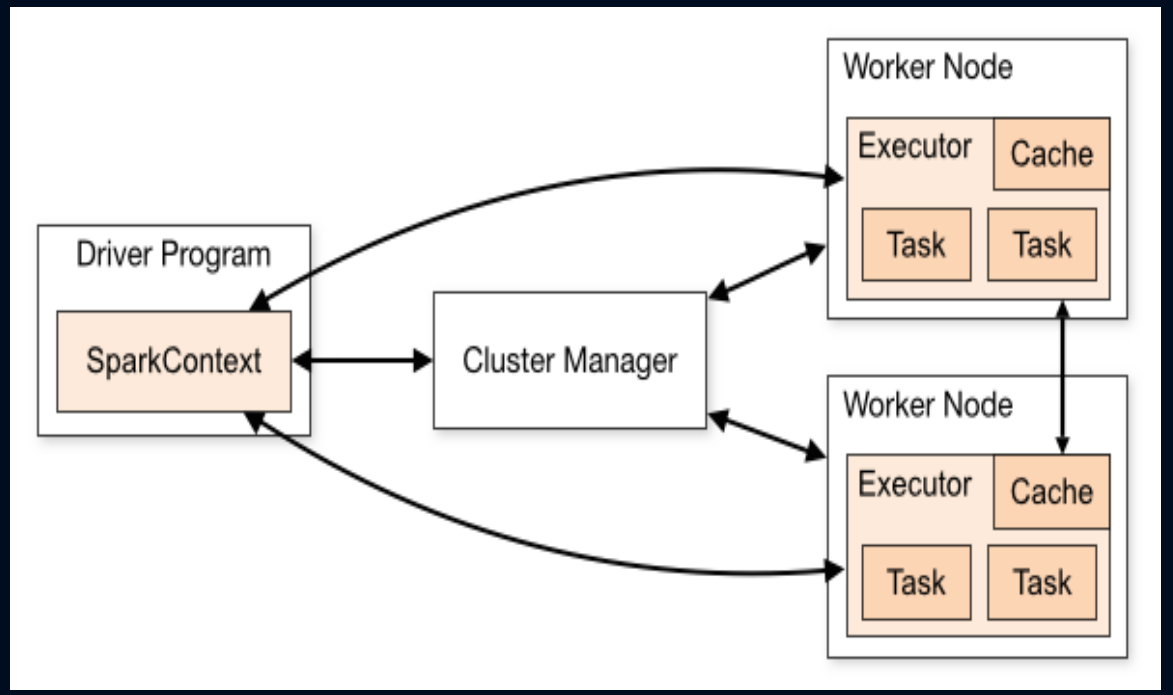
RDZEŃ PLATFORMY

Właściwości Spark Core

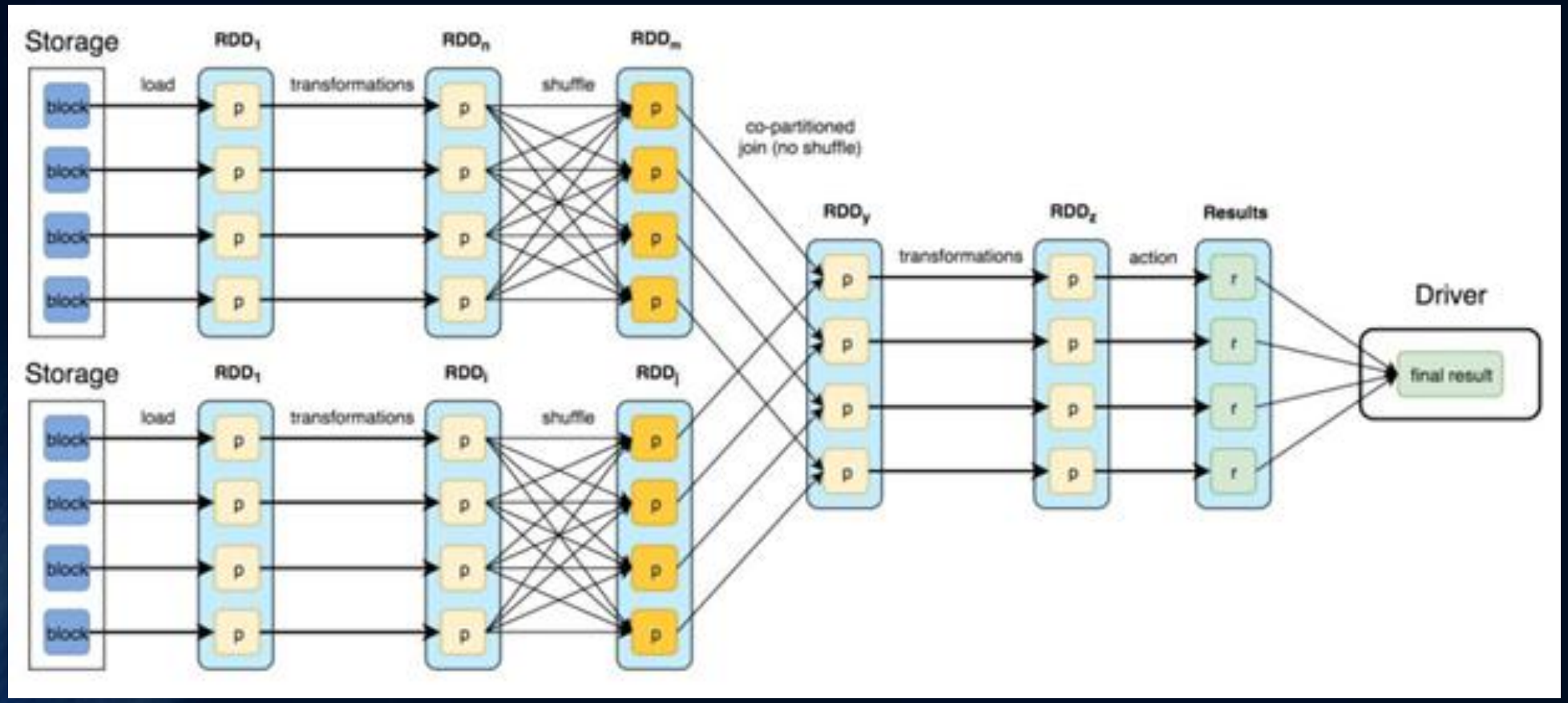
- Model grafu DAG
- Rozsyła rozproszone zadania
- Planuje wykonywanie zadań
- Zarządza podstawowymi funkcjonalnościami wejścia / wyjścia
- Interfejs API w języku Scala, Java, Python oraz R
- Opiera się na abstrakcji rozproszonych obiektów RDD
  - Leniwa ewaluacja, transformacje vs akcje
  - Partycje, Cache, Akumulatory, Broadcasting

Architektura klastra Spark

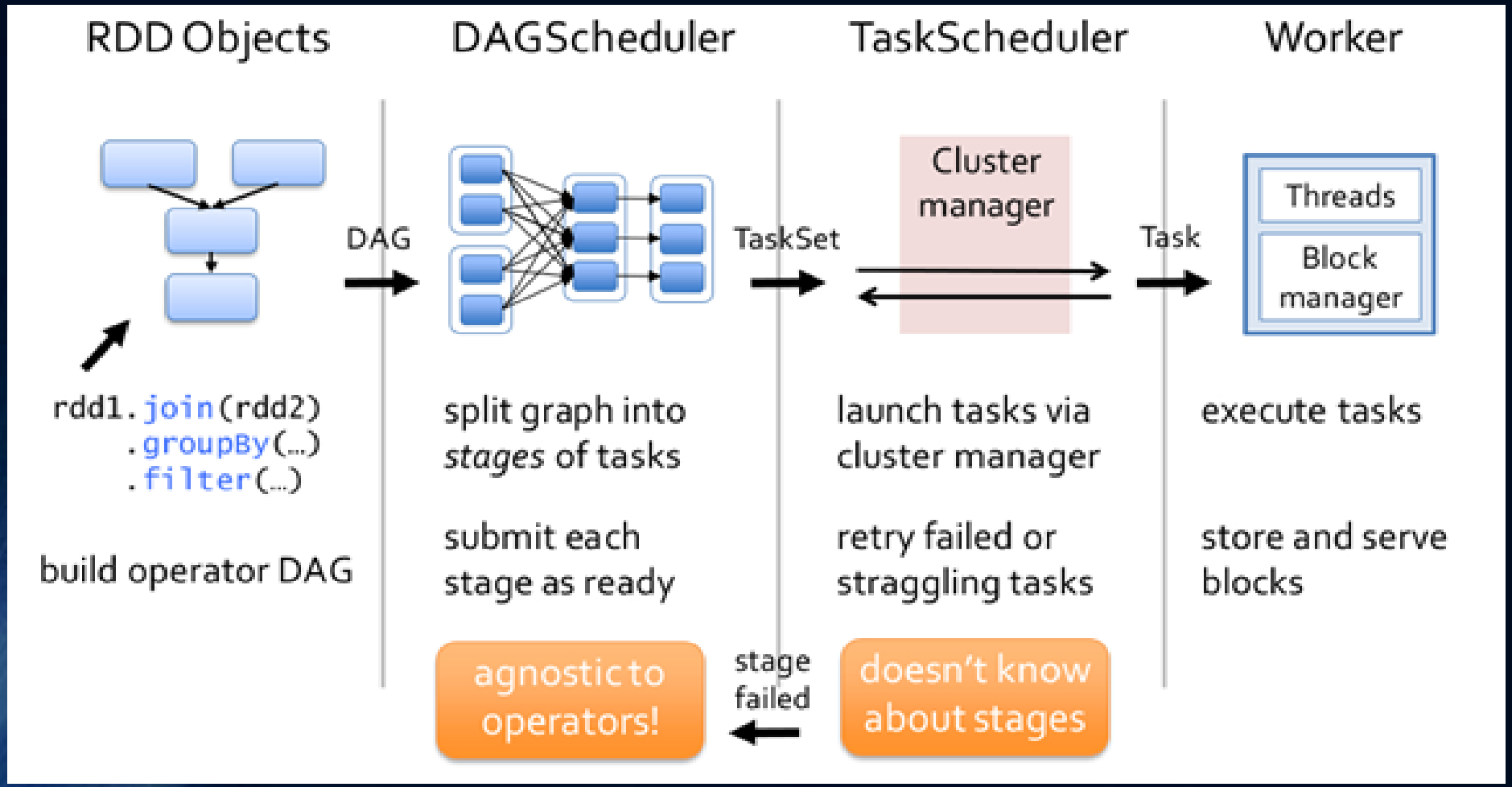
- Driver Program
  - Inicjuje przepływ pracy
  - Kontroluje przebieg działania
  - Zbiera wyniki
- Cluster Manager
  - Zarządza zasobami
    - Pamięć
    - Procesory
    - Dyski
  - Implementacje Managera
    - Natywna Sparka
    - YARN
    - Mesos
- Worker Node
  - Core Node / Data Node
  - Task Node



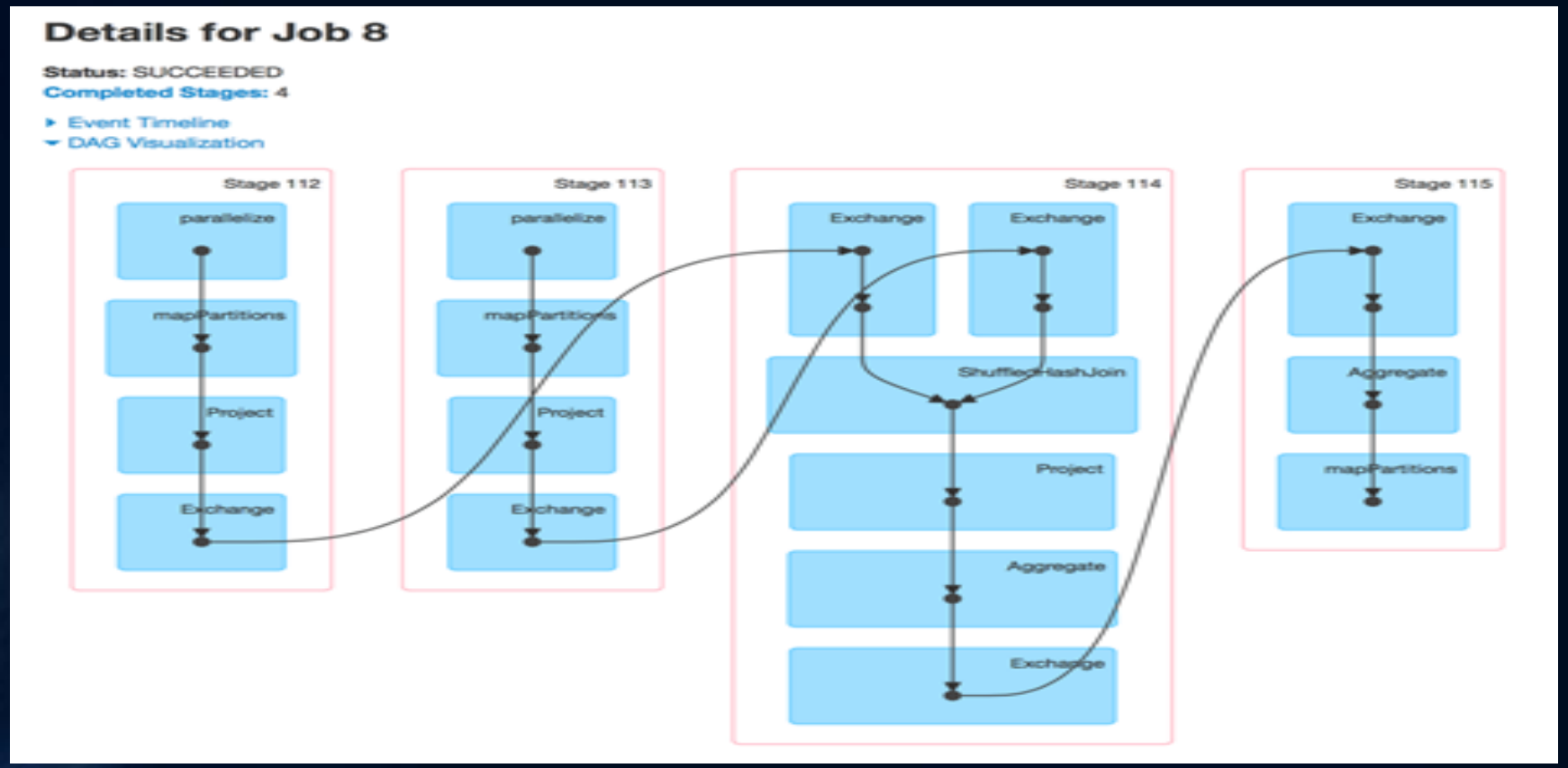
DAG – Directed Acyclic Graph



DAG – cykl życia

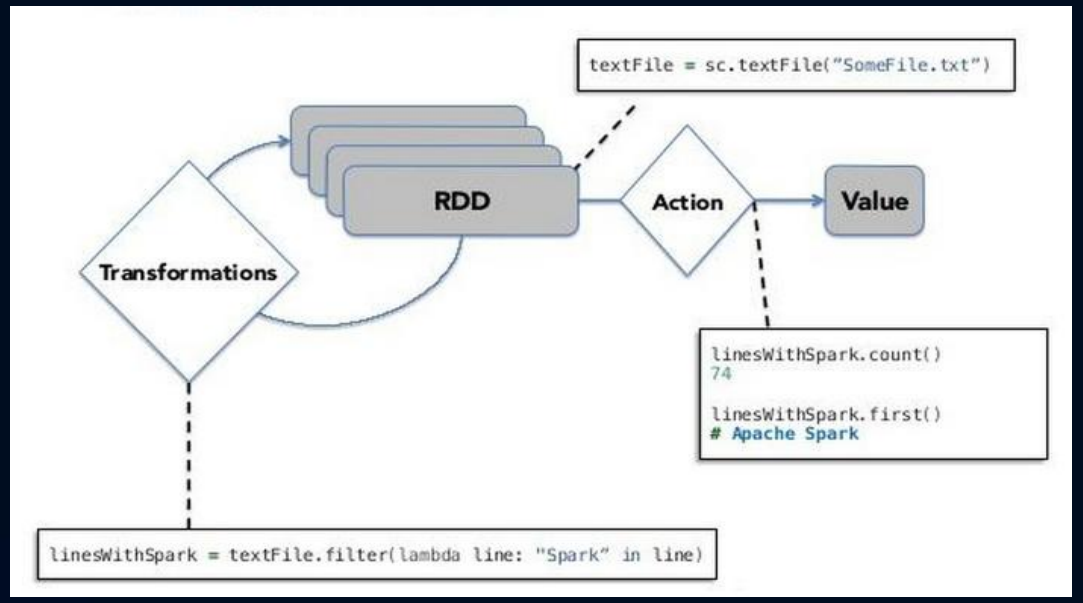


DAG – przykład grafu w Spark UI



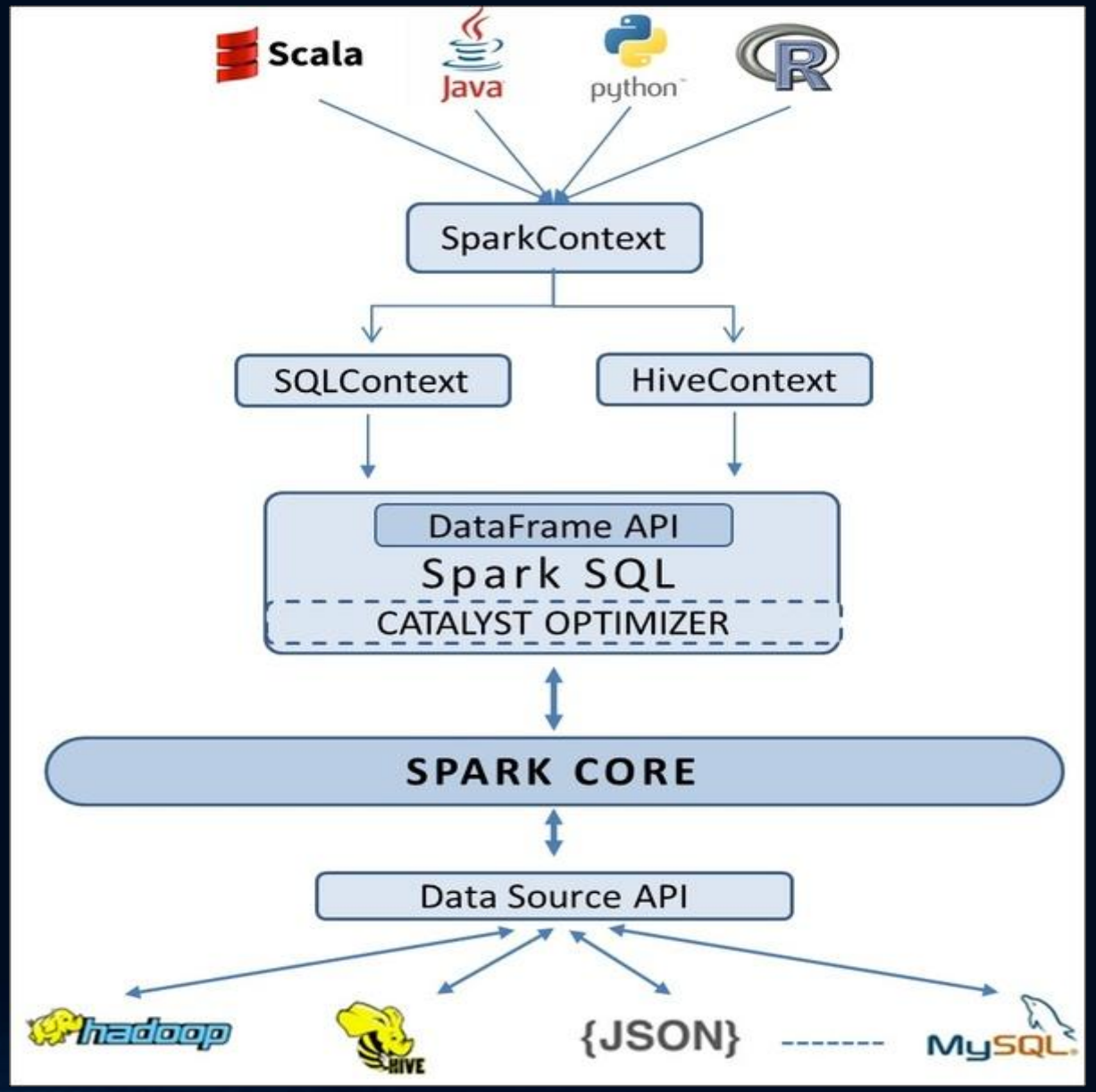
RDD – Resilient Distributed Datasets

- RDD = Odporne (na awarie) rozproszone zbiory danych
- Podstawowy typ danych w Sparku
- Źródła danych:
  - Kolekcja danych: sc.parallelize([1,2,3,4])
  - Local FS, S_3, HDFS, Hive, JDBC, Cassandra, HBase, Elasticsearch, …
  - JSON, CSV, sequence file, object files, parquet, …
- Transformacje
  - map, flatMap, filter, distinct, sample…
- Akcje
  - collect, first, count, countByValue, take, top, reduce…
- Leniwa ewaluacja (ang. lazy evaluation)
  - Bez określenia akcji Spark nie wykona fizycznie żadnych transformacji

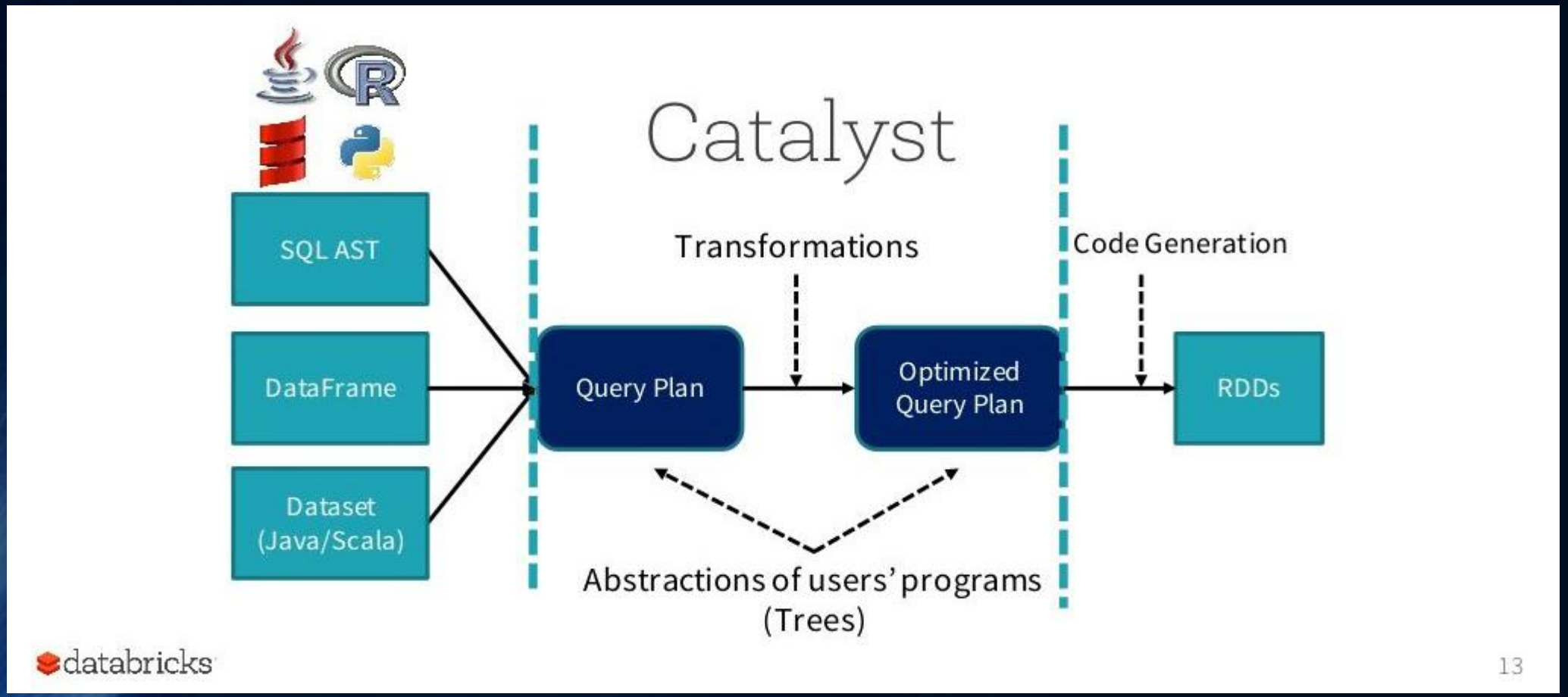


Projekt Catalyst

- Optymalizator transformacji
- Plany logiczne
- Kilka planów wykonania, z których wybierany jest najbardziej optymalny
- Optymalizacja na poziomie kodu źródłowego

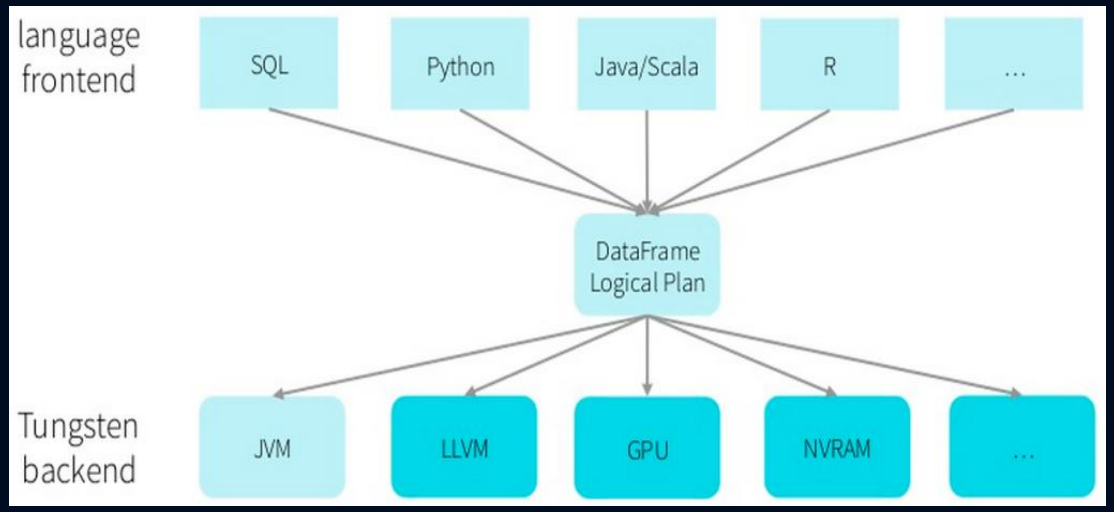


Projekt Catalyst



Projekt Tungsten

- Optymalizator fizycznego wykonania
- Translacja planu logicznego na program wykonujący
- Generowanie w locie kodu źródłowego w języku Scala
- JVM = Java Virtual Machine
- Wykorzystanie nieulotnej pamięci NVRAM
- Wykorzystanie precesorów graficznych (GPU)

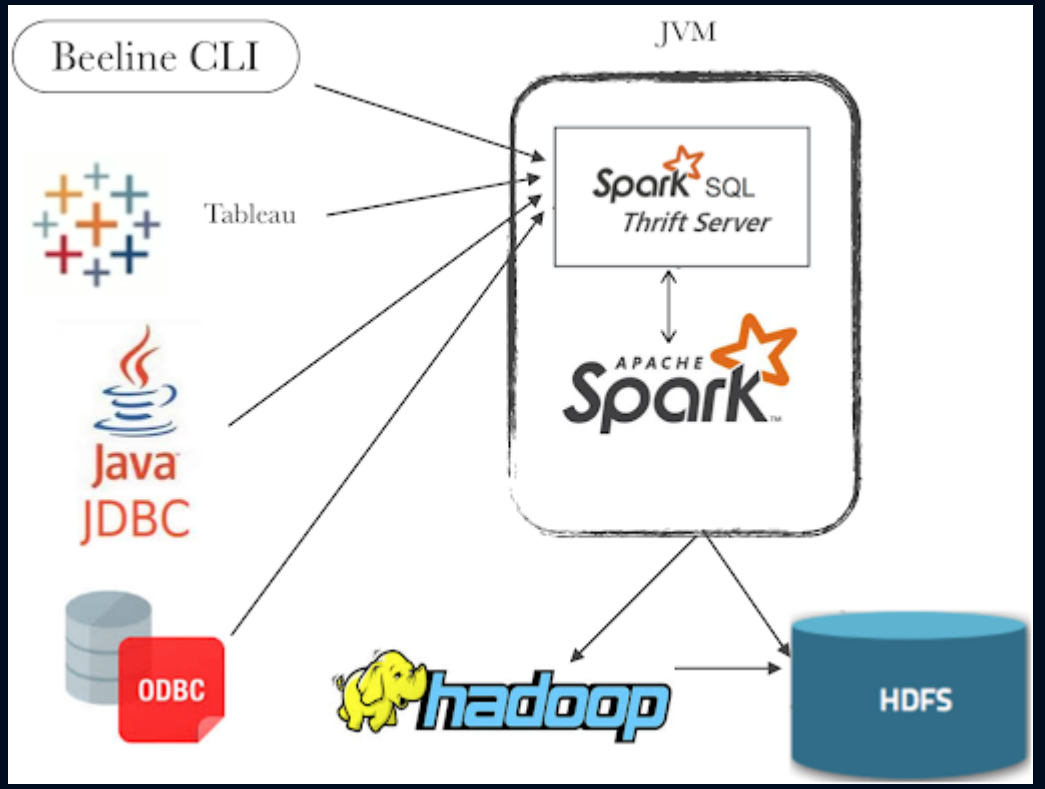


Spark SQL

DATA FRAMES, DATA SETS & SQL

DataFrame

- Rozszerzenie RDD
- Praca z danymi strukturalnymi
- Posiada schemat (+ na wydajności)
- Ramka to Dataset z obiektami typu Row
- Może wykonywać zapytania SQL
  - Korzysta z Hive Catalog Metastore
  - Na AWS możliwość użycia AWS Glue Data Catalog
- Odczyt i zapis w formacie JSON, Hive, parquet, …
- Wsparcie dla Pandas: df.toPandas()
- Funkcje UDF (ang. User Defined Functions)



DataSets

- Bezpieczne i optymalne transformacje na danych
- Silne typowanie danych możliwe w Java oraz Scala
- Python jest dynamicznie typowany
- DataFrame to DataSet z obiektami typu Row

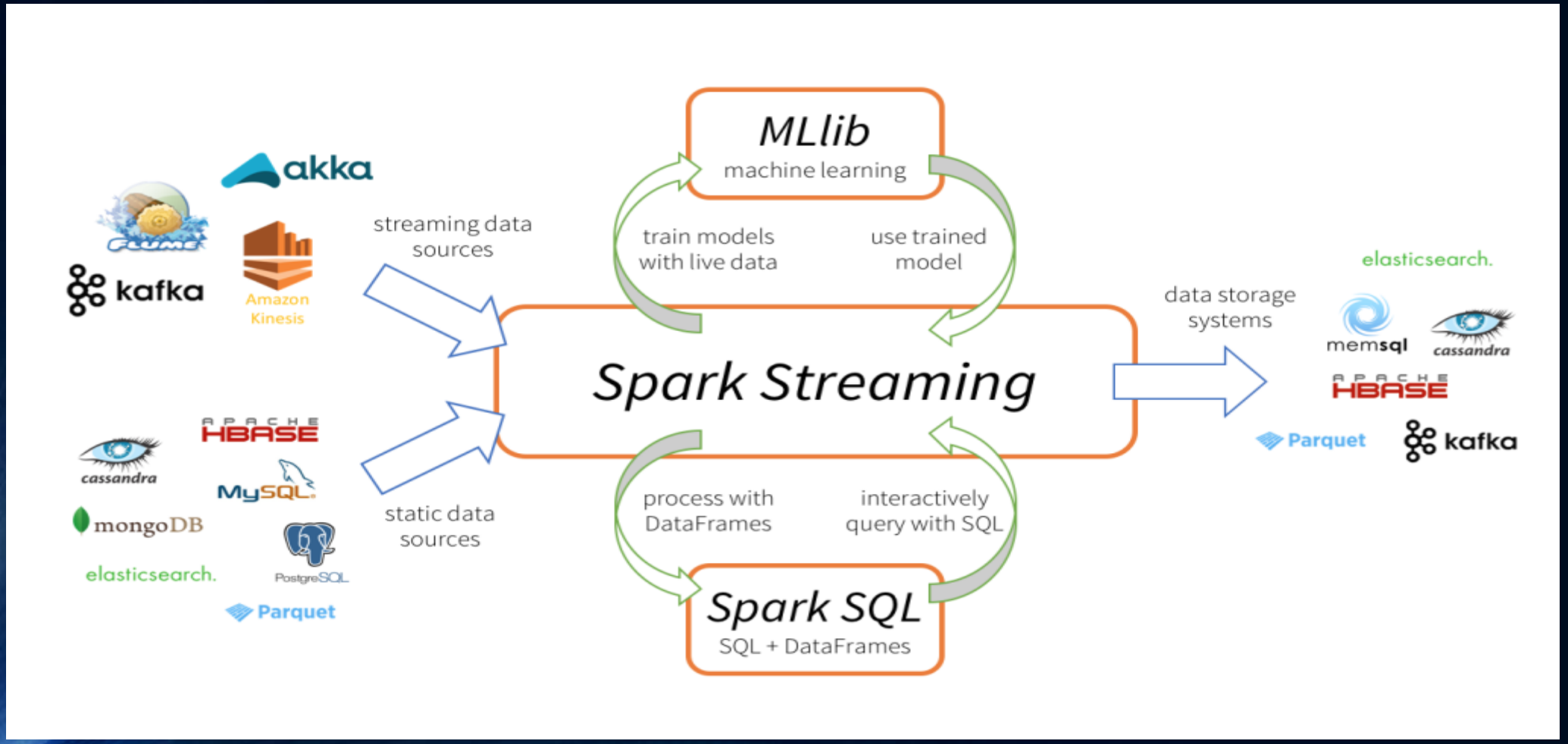
Spark Streaming

PRZETWARZANIE STRUMIENI DANYCH

Spark Streaming

- Potrzeba:
  - Analiza danych w czasie rzeczywistym zamiast dużych porcji danych dziennie (wsadów, ang. data batches).
  - Analiza strumieni logów sieciowych, by móc szybko zareagować na działanie użytkowników.
  - Analiza strumieni danych pochodzących z sensorów IoT.
- Discretized Streams (DStreams) vs Structured Streaming (na bazie DataSets)
- Micro-batch
- Transformacje stateless vs stateful
- Przetwarzanie z użyciem okien czasowych (ang. windowing)
  - Najwyższa sprzedaż z ostatniej godziny
  - Okno się przesuwa z upływem czasu

Spark Streaming



Spark MLLib

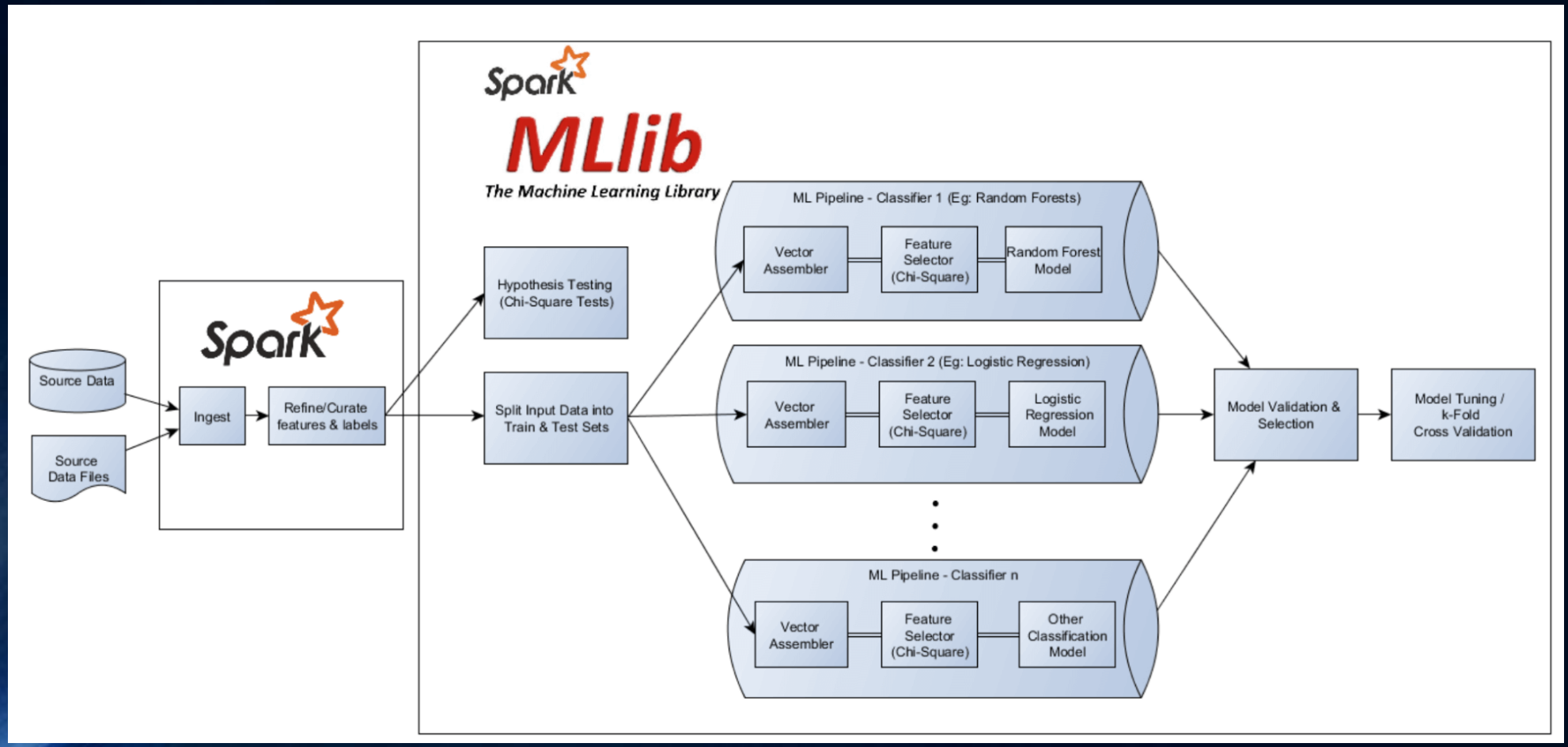
MACHINE LEARNING LIBRARY

Spark MLLib - możliwości

- Feature extraction
  - Term Frequency / Inverse Document Frequency useful for search
- Basic statistics
  - Chi-squared test, Pearson or Spearman correlation, min, max, mean, variance
- Linear regression, logistic regression

  Support Vector Machines
- Naïve Bayes classifier
- Decision trees
- K-Means clustering
- Principal component analysis, singular value decomposition
- Recommendations using Alternating Least Squares

Spark Machine Learning Pipeline



Spark MLLib – specjalne typy danych

- Dense Vector, Sparse Vector
- Dense Matrix, Sparse Matrix
- LabeledPoint
- Rating

Spark GraphX

PRZETWARZANIE DANYCH GRAFOWYCH

Spark GraphX

- Abstrakcja grafu oparta o obiekty RDD
- Graf reprezentowany przez krawędzie oraz węzły z danymi
- Operatory grafowe
- Optymalizacja grafu
- Algorytmy
  - PageRank
  - Connected Components
  - Zliczanie trójkątów