Hadoop i Spark

Mariusz Rafało
mrafalo@sgh.waw.pl
http://mariuszrafalo.pl

PRZETWARZANIE DANYCH

Map Reduce

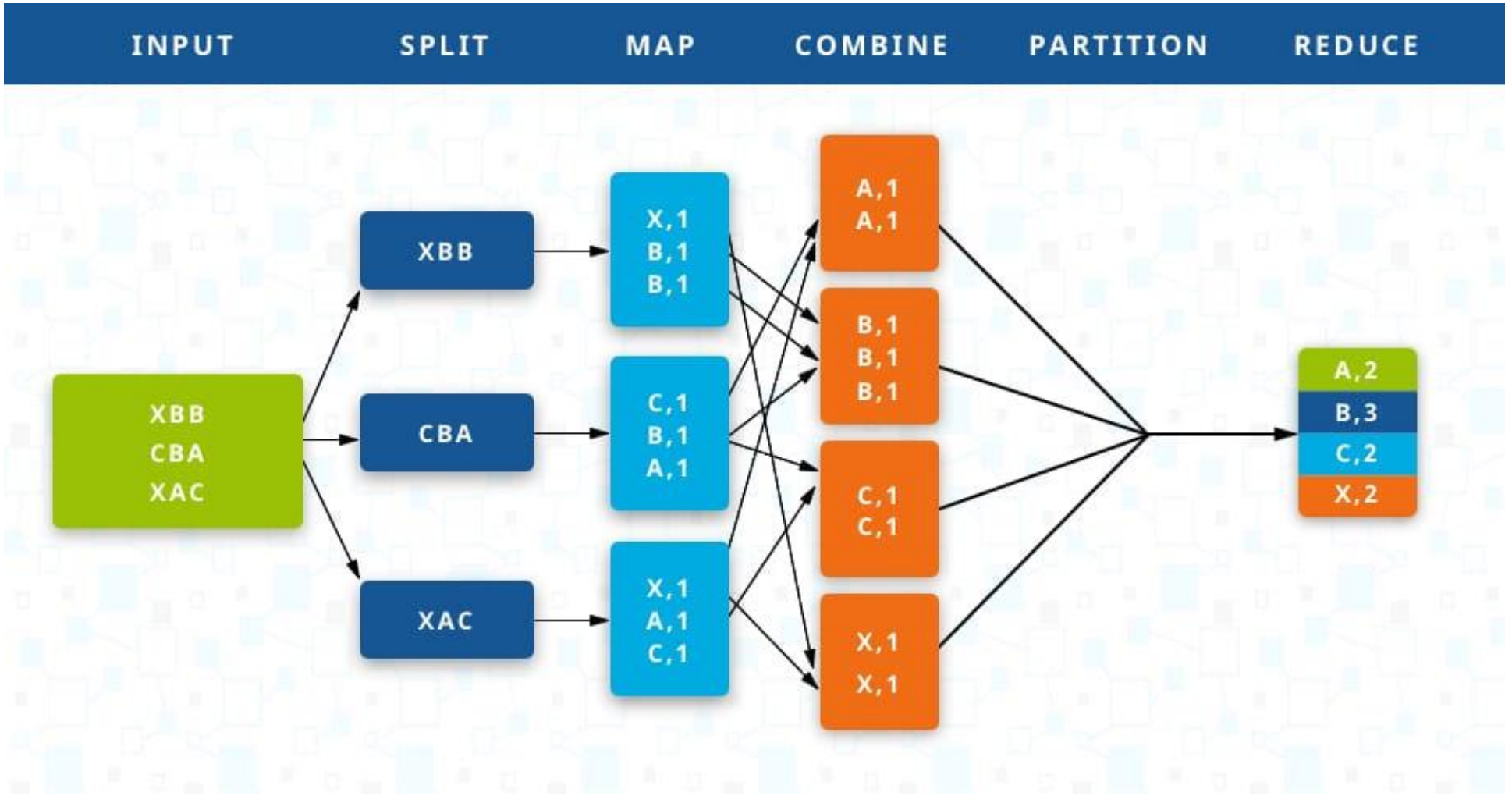
Algorytm służący przetwarzaniu równoległemu dużych zbiorów danych w rozproszonym środowisku. Podejście opracowane przez firmę Google.

Algorytm składa się z dwóch głównych kroków:

- ***map*** – pobranie danych z wejścia i ich podział na podzbiory. Dekompozycja problemu na podproblemy.

- ***reduce*** – zgromadzenie odpowiedzi, ich połączenie i przekazanie wyniku

Map reduce



Źródło: https://www.talend.com/resources/what-is-mapreduce/

Map reduce

```
public void map(Object key, Text value, Context context) throws IOException, InterruptedException
    {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens())
      {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
```

```
public void reduce(Text key, Iterable<IntWritable> values,Context context) throws IOException, InterruptedException
    {
      int sum = 0;
      for (IntWritable val : values)
      {
        sum += val.get();
      }
      result.set(sum);
      context.write(key, result);
    }
```

Pig

- Platforma służąca do analizy i przetwarzania dużych zbiorów danych
- Udostępnia język programowania, pozwalających na zrównoleglanie i rozpraszanie przetwarzania
- Język Pig stanowi warstwę działającą na HDFS. Kod źródłowy Pig jest przetwarzany przez platformę na zadania MapReduce
- Optymalizacja kodu jest wykonywana automatycznie przez platformę
- Możliwe jest dołączanie/programowanie dodatkowych funkcji, rozszerzających standardowe

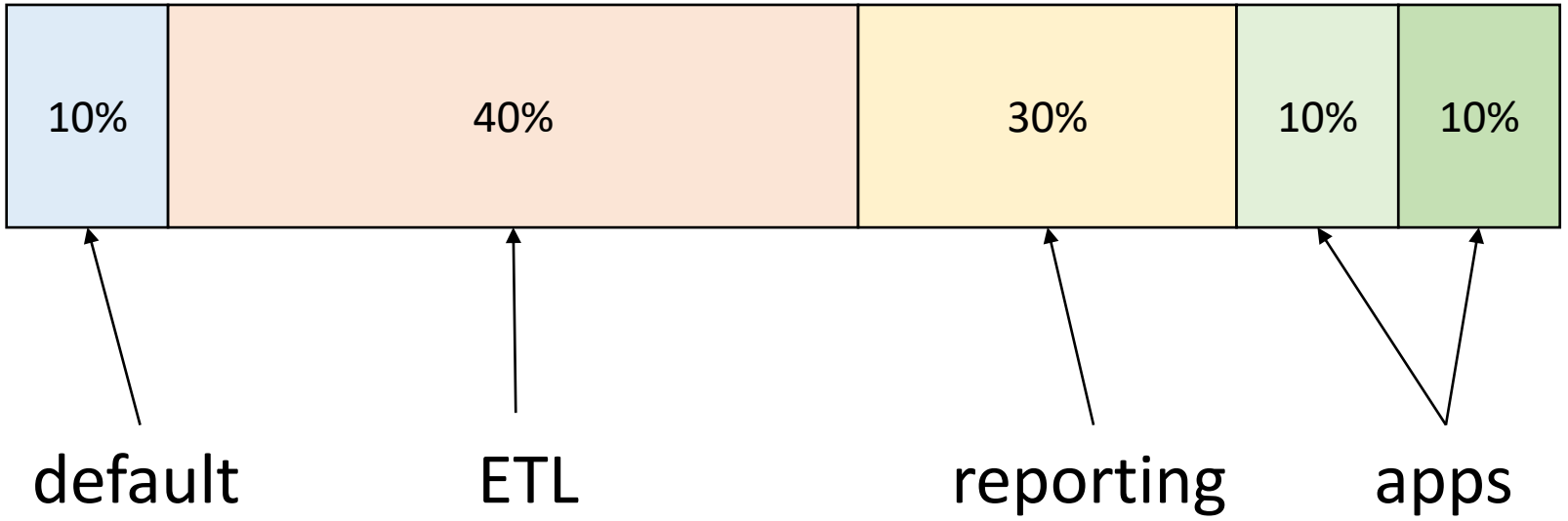
```
    STOCK_A = LOAD '/user/maria_dev/NYSE_daily_prices_A.csv' USING
PigStorage(',')
    AS (exchange:chararray, symbol:chararray, date:chararray,
    open:float, high:float, low:float, close:float, volume:int,
adj_close:float);
    DESCRIBE STOCK_A;
```

INFRASTRUKTURA

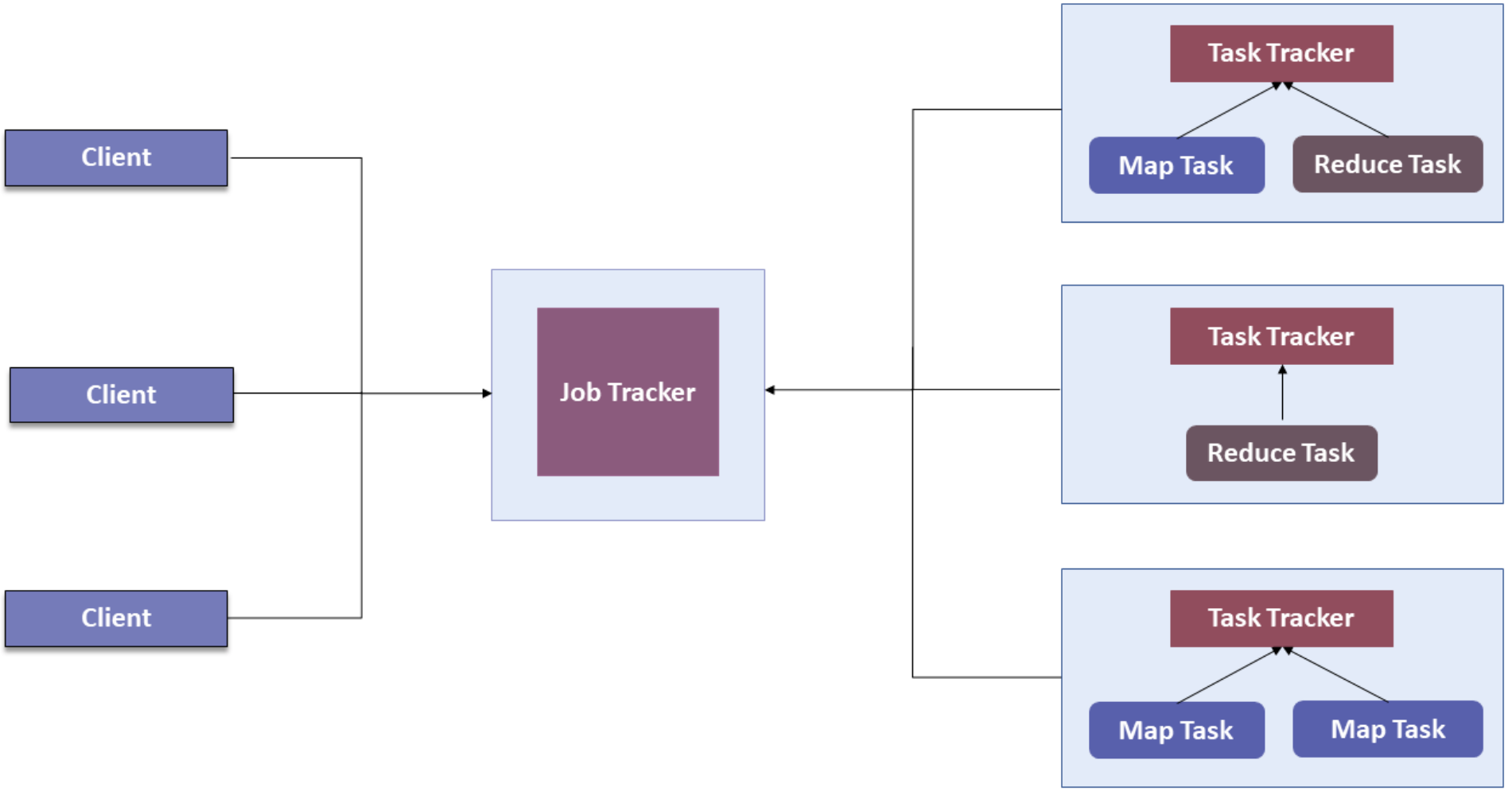
Yarn

- Manager (negocjator) zasobów klastra
- Każde zadanie realizowane przez klaster (zapytanie o dane, przetwarzanie danych, ładowanie danych, itp.) wymaga określonych zasobów. Zasoby te przydziela Yarn
- Yarn zarządza mocą procesorów (CPU), pamięcią RAM, przestrzenią dyskową oraz zasobami sieciowymi
- Współpracuje z silnikami przetwarzania (Spark, Tez) oraz wieloma technologiami składowania danych (np. Hive, HBase)

Yarn – kolejki (queues)

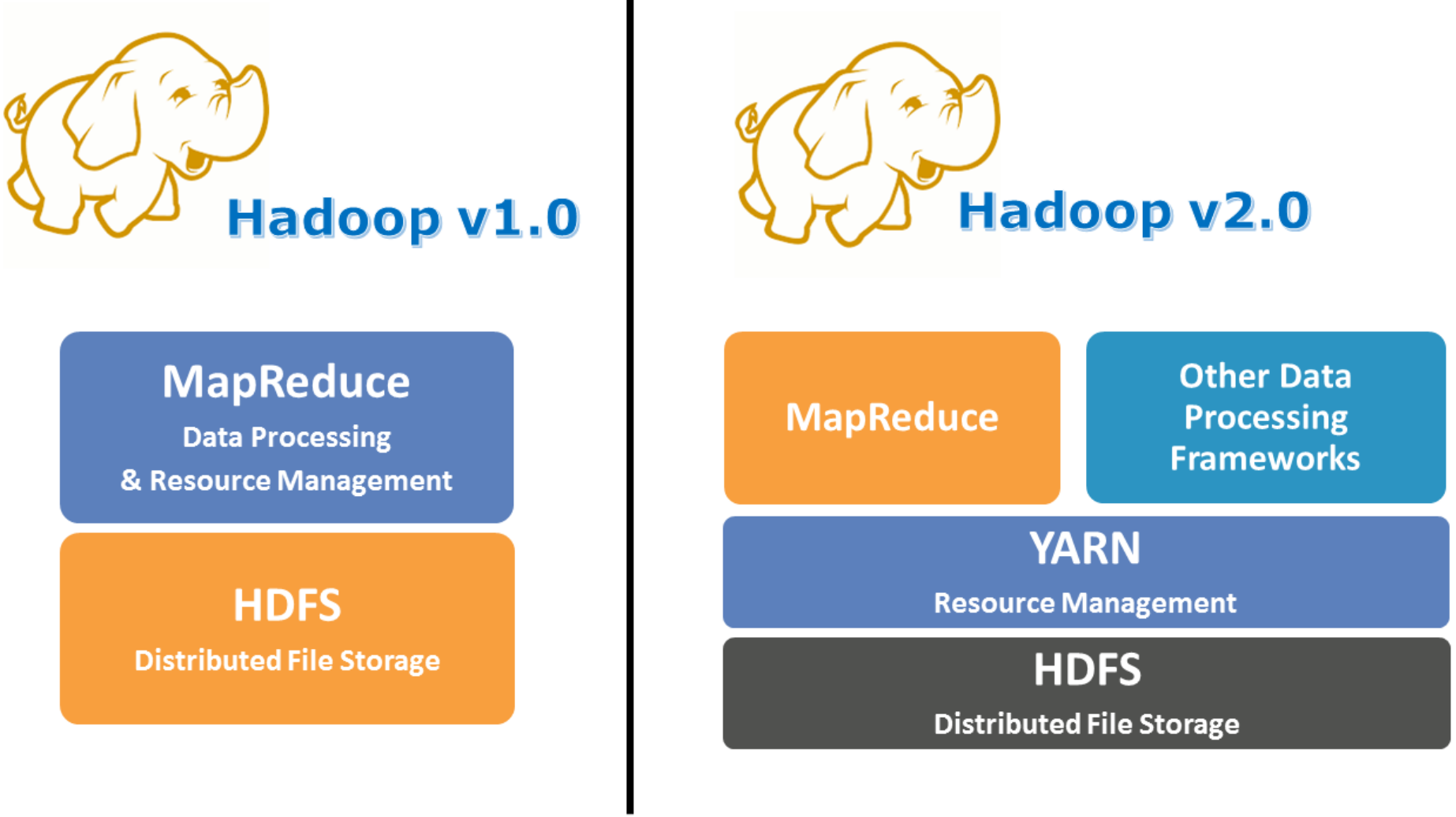


Yarn



Źródło: https://www.edureka.co/blog/hadoop-yarn-tutorial/

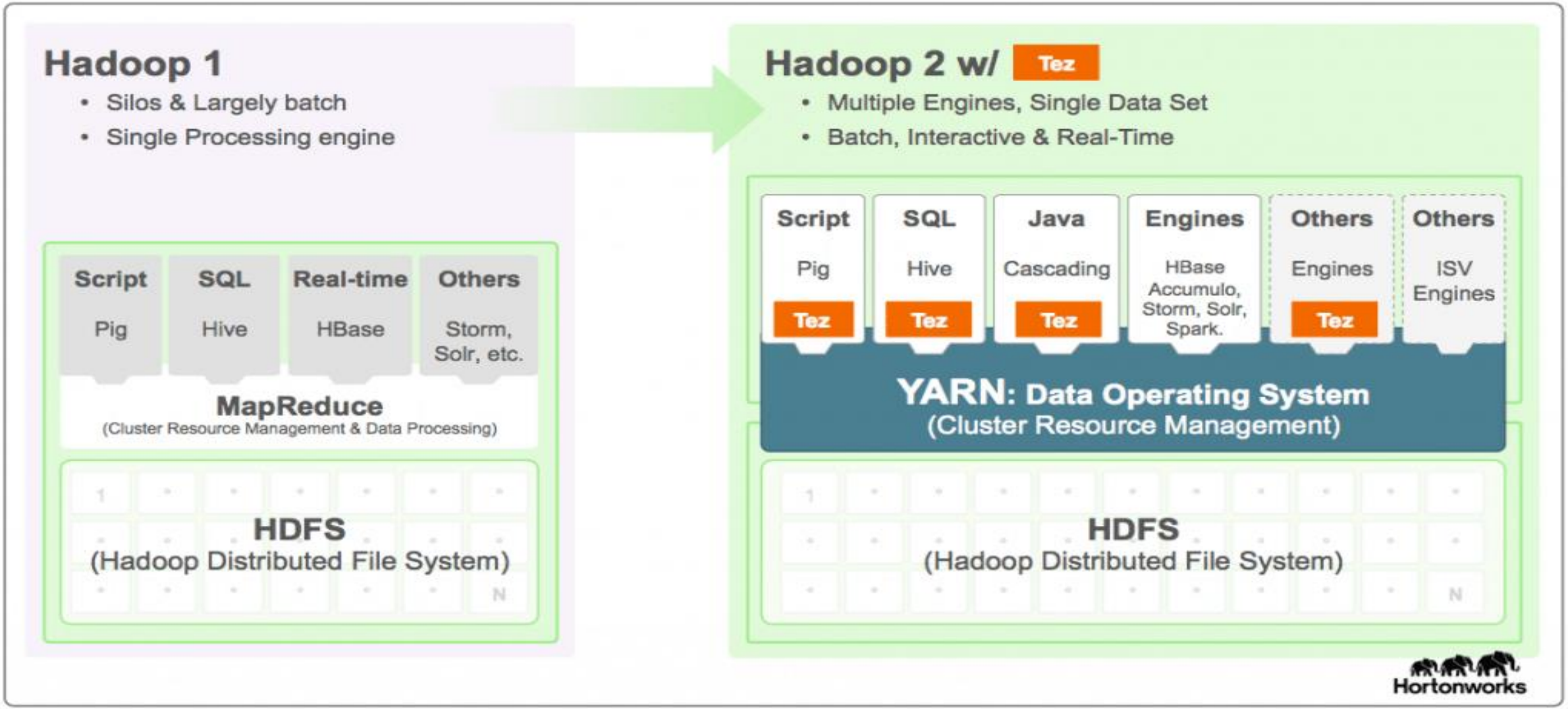
Yarn



Źródło: https://www.edureka.co/blog/hadoop-yarn-tutorial/

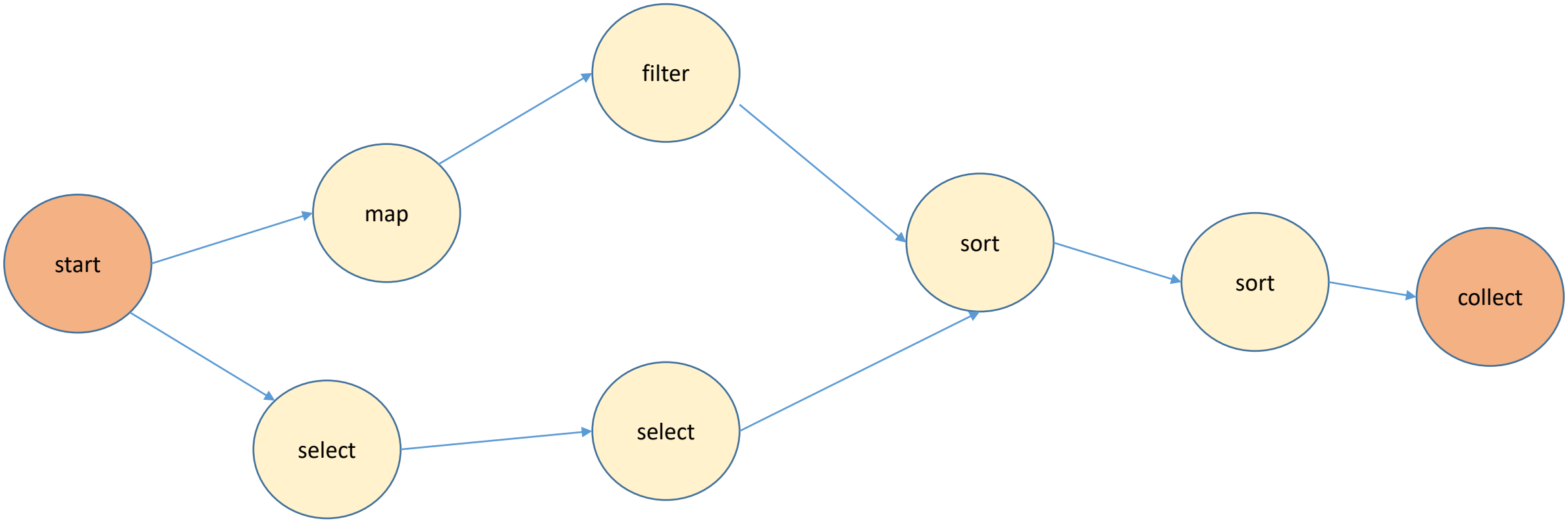
COMPUTING ENGINES

Tez

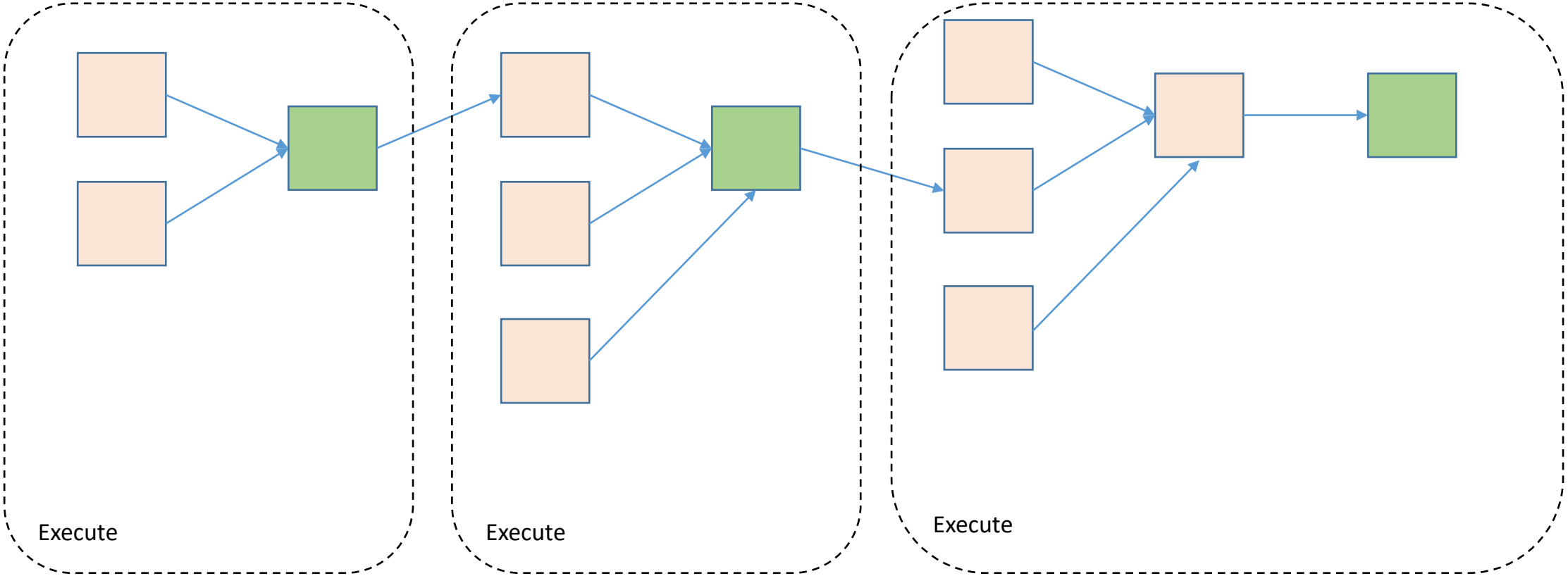


Źródło: hortonworks.com/apache/tez/

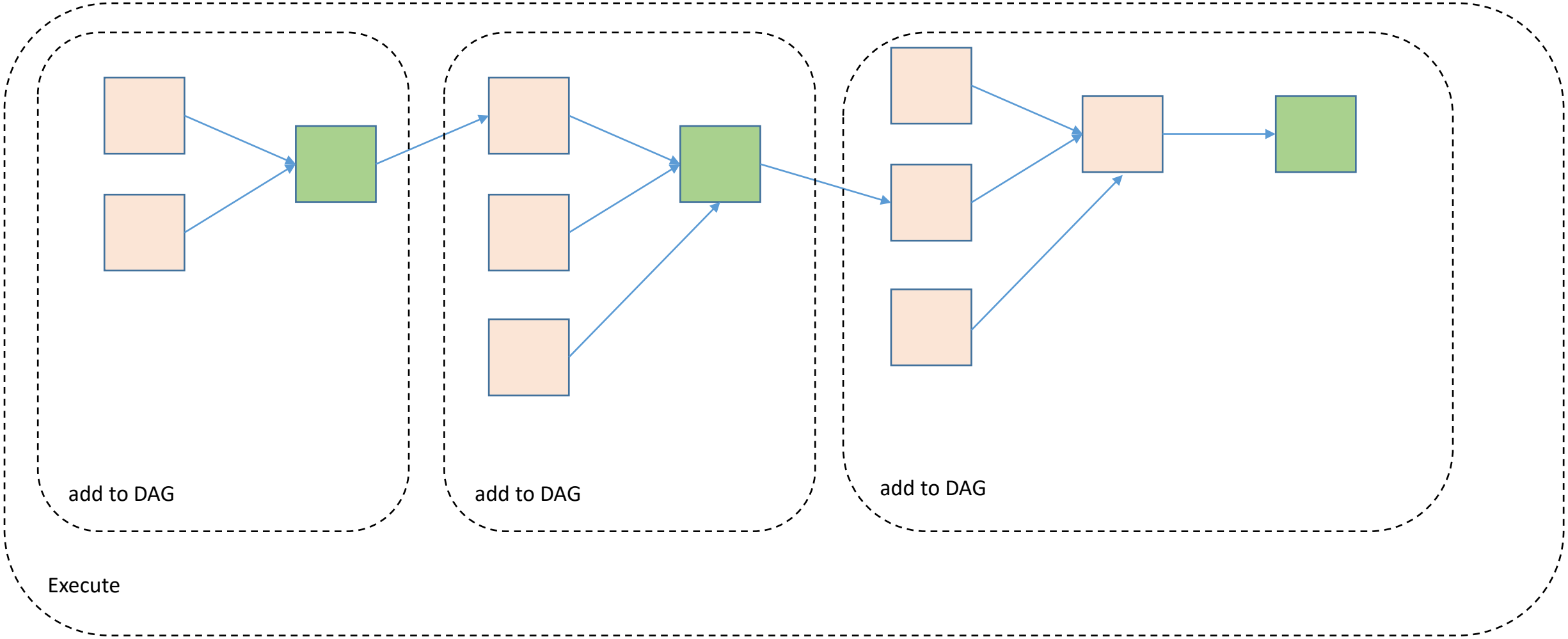
DAG (Directed Acyclic Graph)



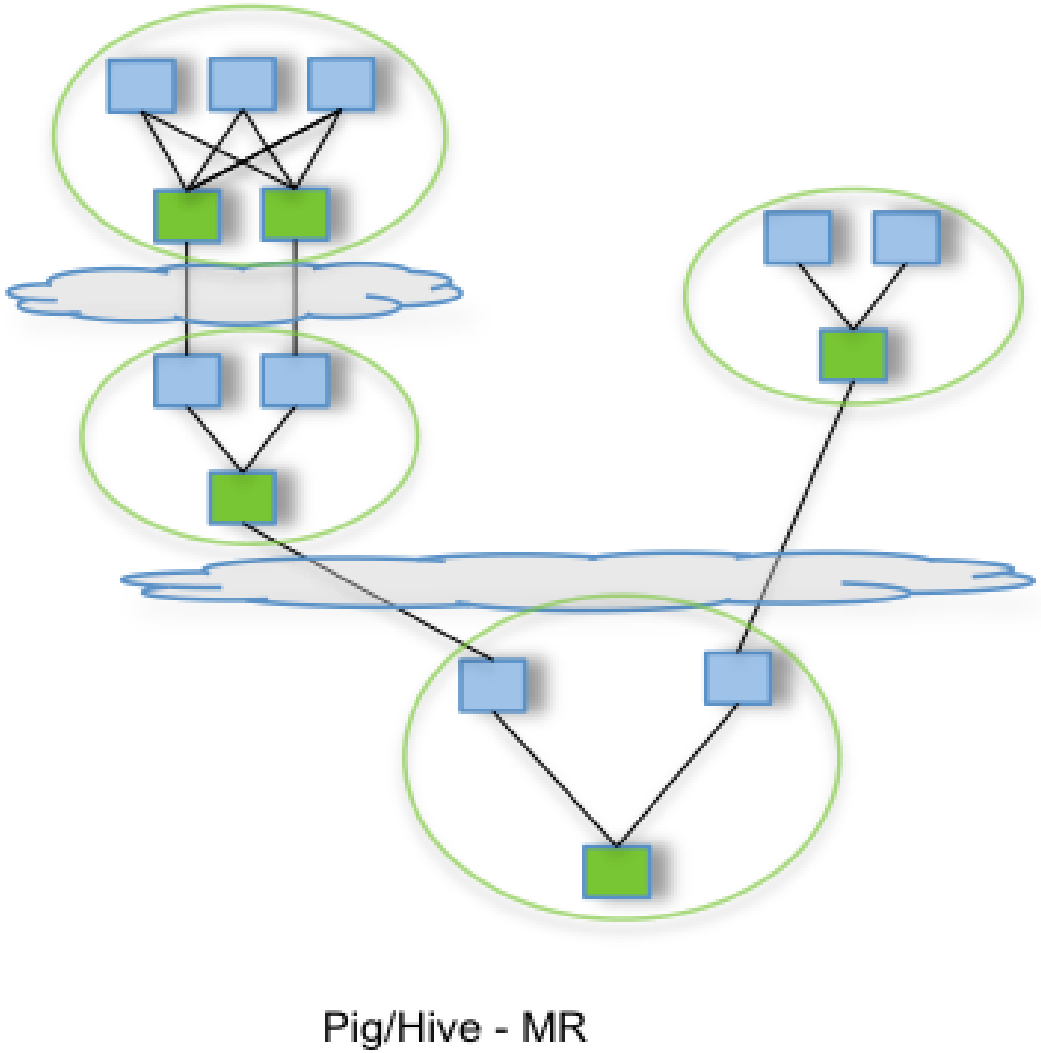
Lazy evaluation concept



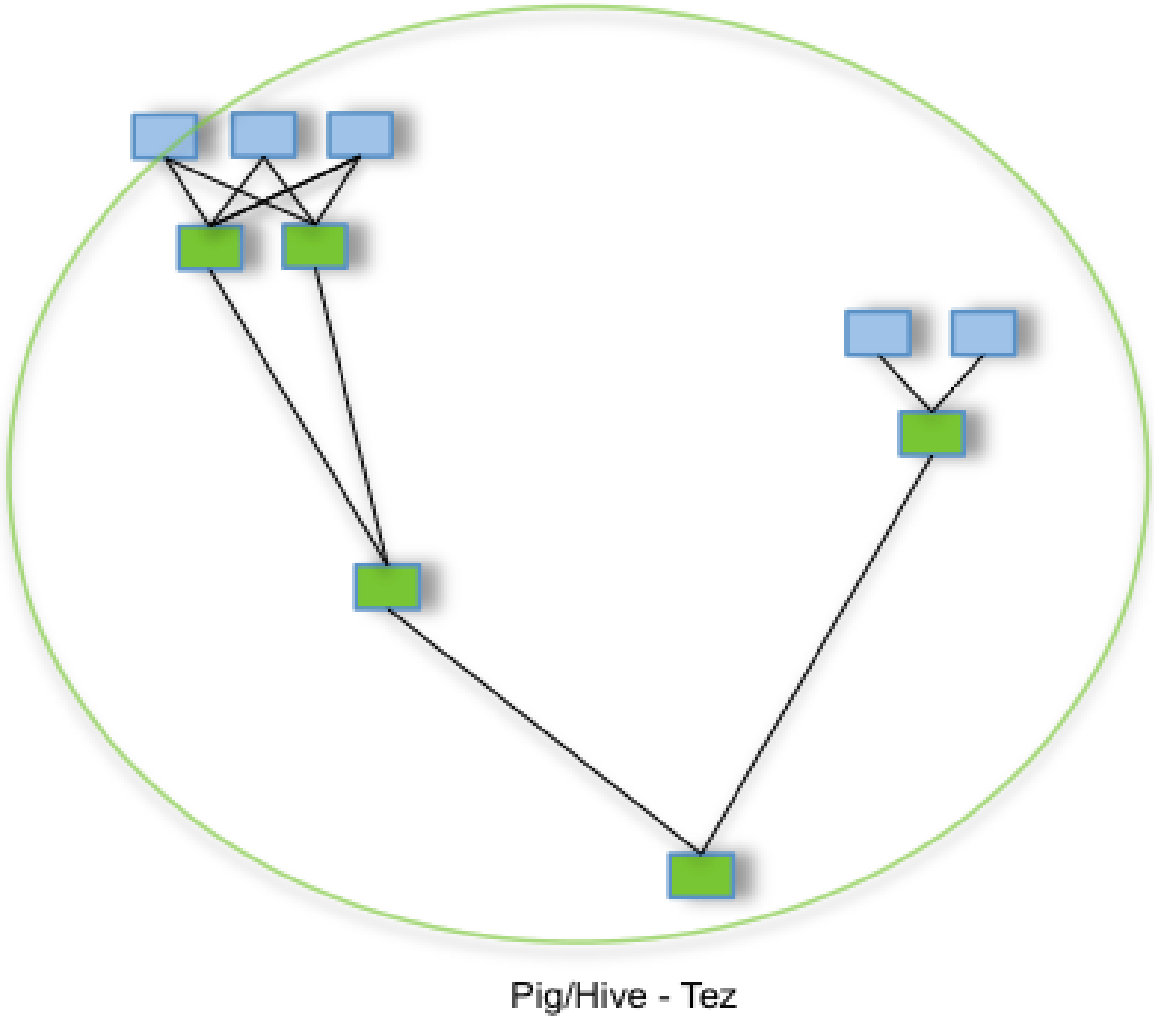
Lazy evaluation concept



Tez vs MapReduce



Pig/Hive - MR



Pig/Hive - Tez

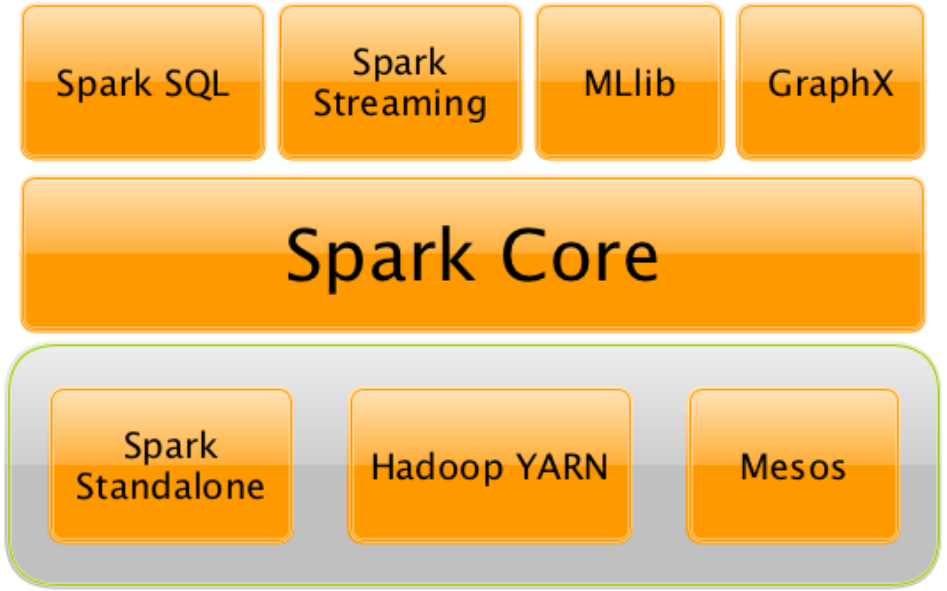
Źródło: tez.apache.org

Spark

- Transformacje
- Akcje

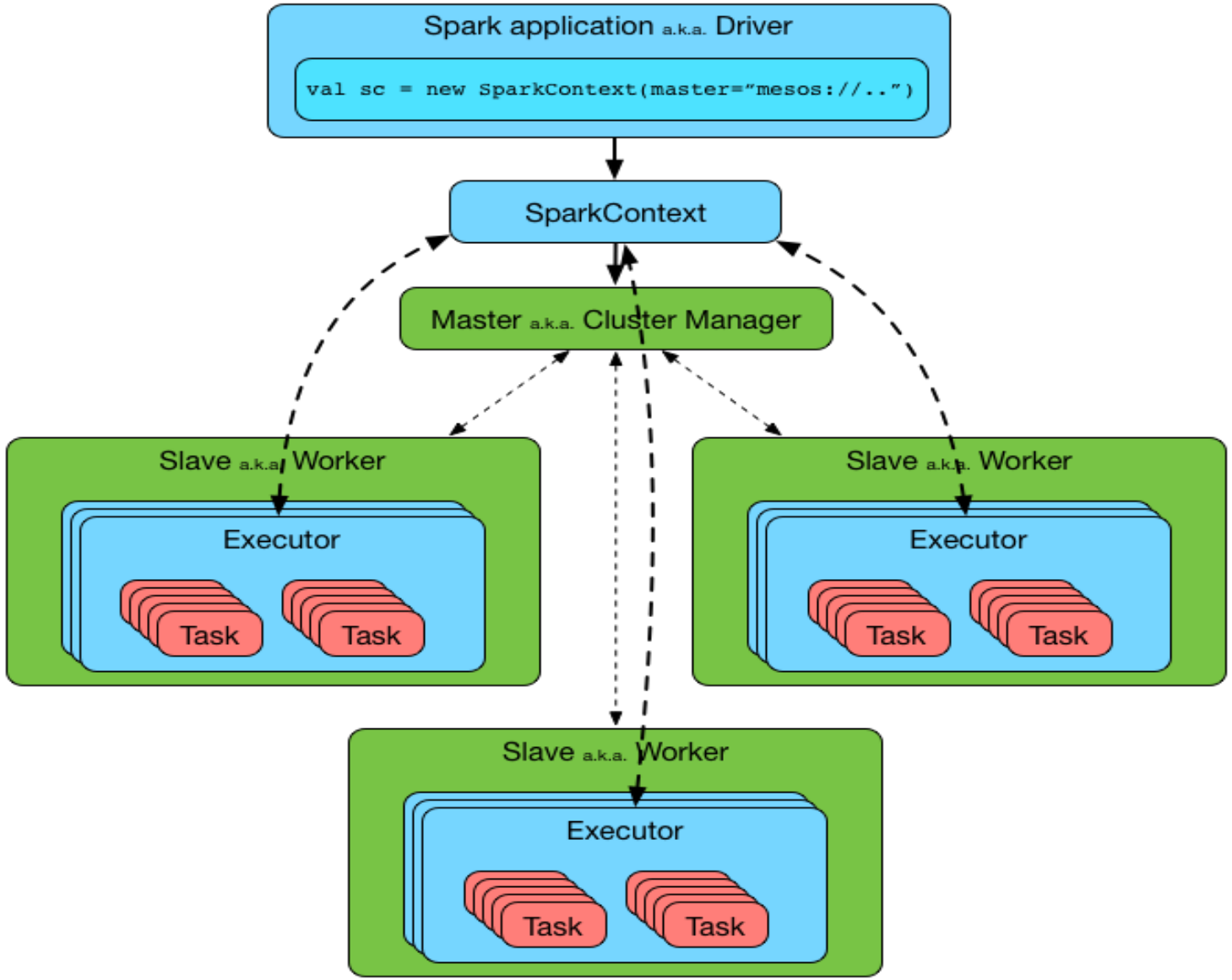
Spark

- Platforma (*engine*) do przetwarzania danych w dużej skali
- Obsługuje języki programowania: Java, Scala, Python, R, SQL
- Może pracować w trybie *batch* lub *stream*
- Może realizować zadania na jednej maszynie oraz na klastrze
- *Lazy evaluation*
- *Immutable structures*



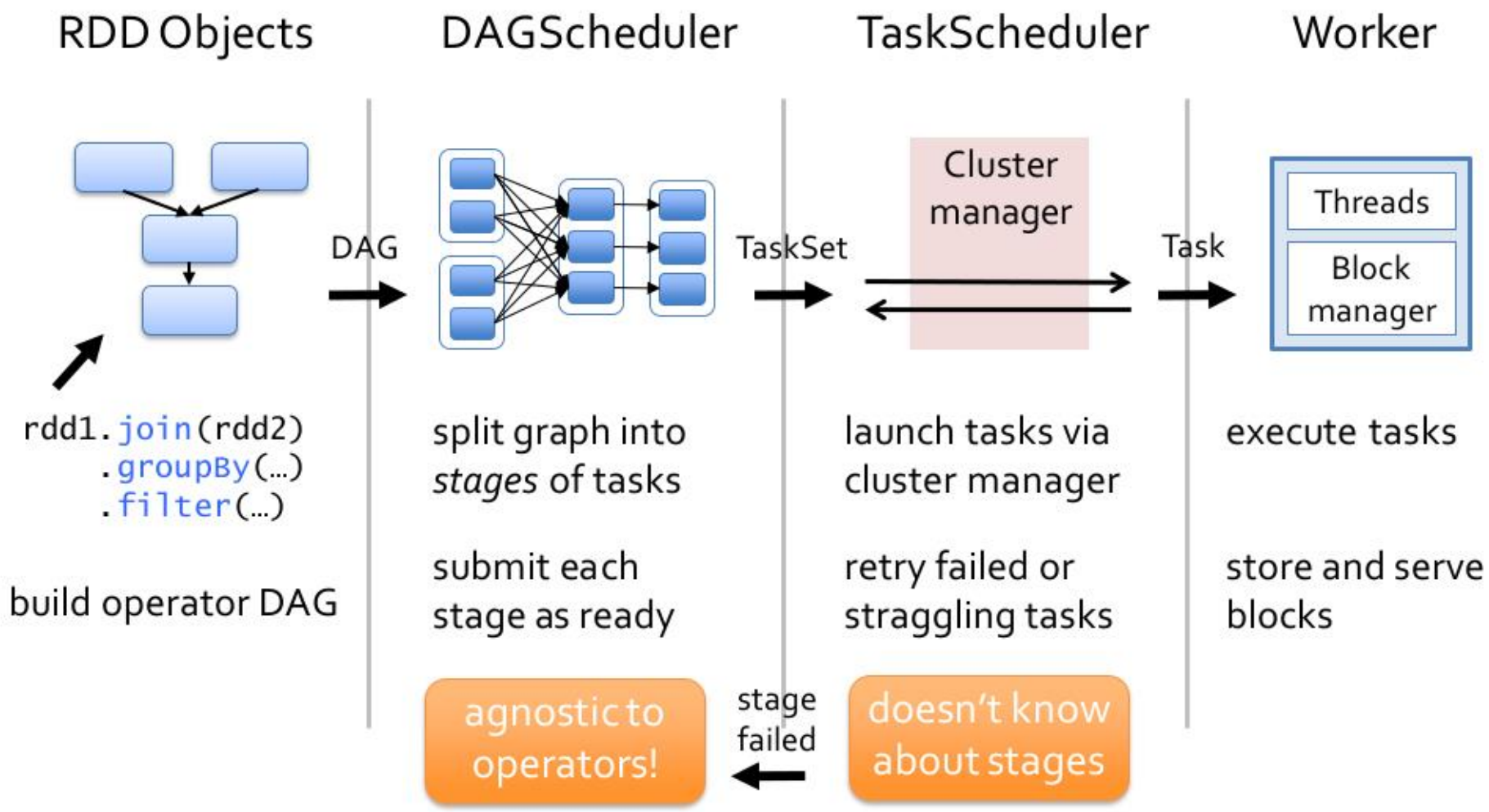
Źródło: https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-overview.html

Spark



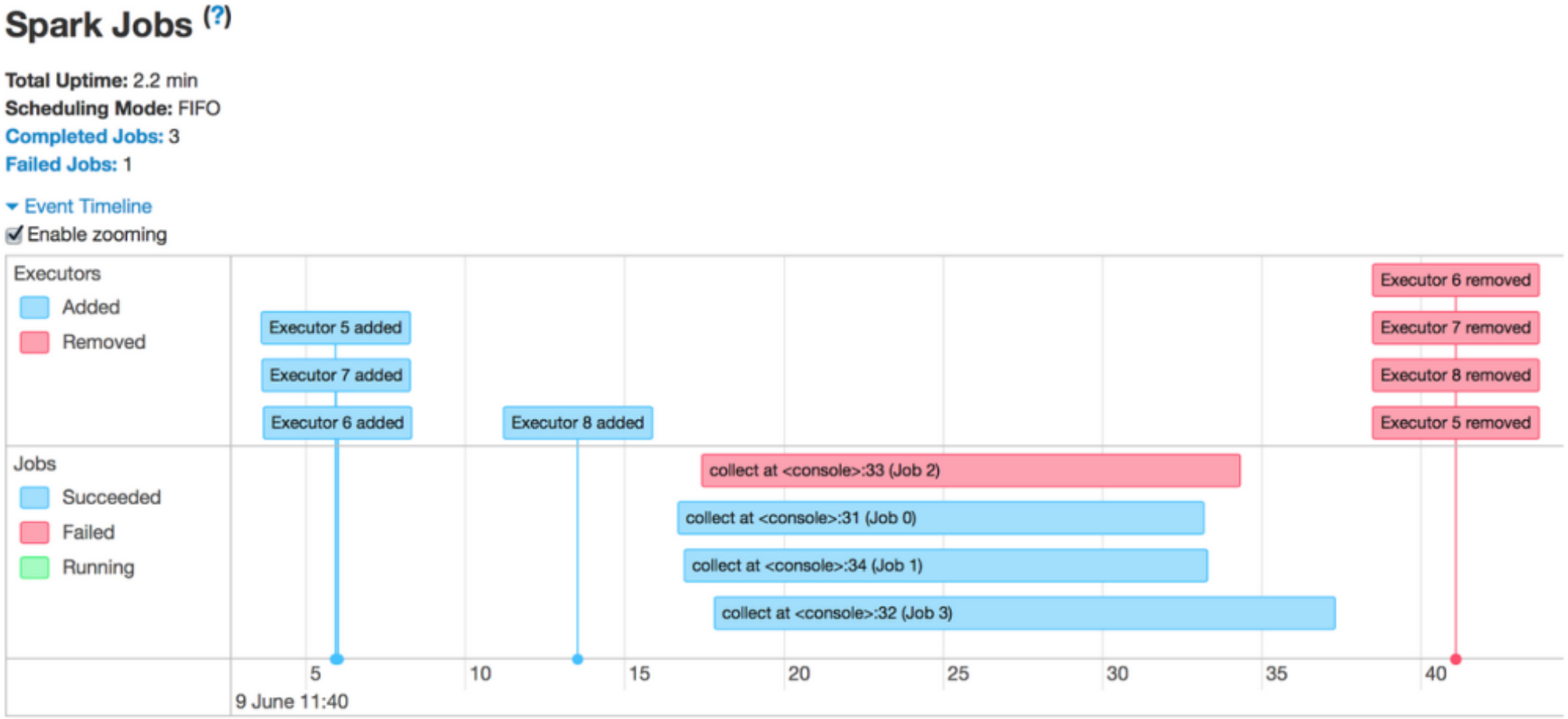
Źródło: https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-architecture.html

Spark: DAG



Źródło: 1.bp.blogspot.com

Spark: UI



Źródło: databricks.com

Spark context

```
from pyspark import SparkContext

spark = SparkContext("134.216.10.212", „raport 17")
```

Dziękuję za uwagę