Przetwarzanie Big Data z użyciem Apache Spark

Prowadzący

Patryk Pilarski

Data Scientist

1patryk.pilarski@gmail.com

Oliwia Wojtkowska

Data Scientist

oliwiawojtkowska@gmail.com

O czym będziemy mówić?

1. Spark – wprowadzenie
2. Spark RDD
3. Spark DataFrame i Spark SQL
4. Spark Streaming

Spark - wprowadzenie

Czym jest Spark?

- Silnik do obliczeń rozproszonych
- Projekt open source (Apache Spark)
- Napisany w Scali
- Główna abstrakcja: RDD

“Motto” Sparka:

- Łatwość użytkowania
- Szybkość obliczeń

Głównie stosowany do obliczeń iteracyjnych i interaktywnych.

Historia

2009 Projekt naukowy na UC Berkley (AmpLabs)

2010 Upublicznienie kodu źródłowego – Open Source na BSDL

2011 Shark – pierwszy silnik do przetwarzania SQL na Sparku

2013 Projekt Spark ma już ponad 100 kontrybutorow, AmpLabs oddaje go w ręce Apache Software Foundation

2013 pierwszy Spark Summit w San Francisco

2014 Spark 1.0 (pojawia się SparkSQL)

2016 Spark 2.0

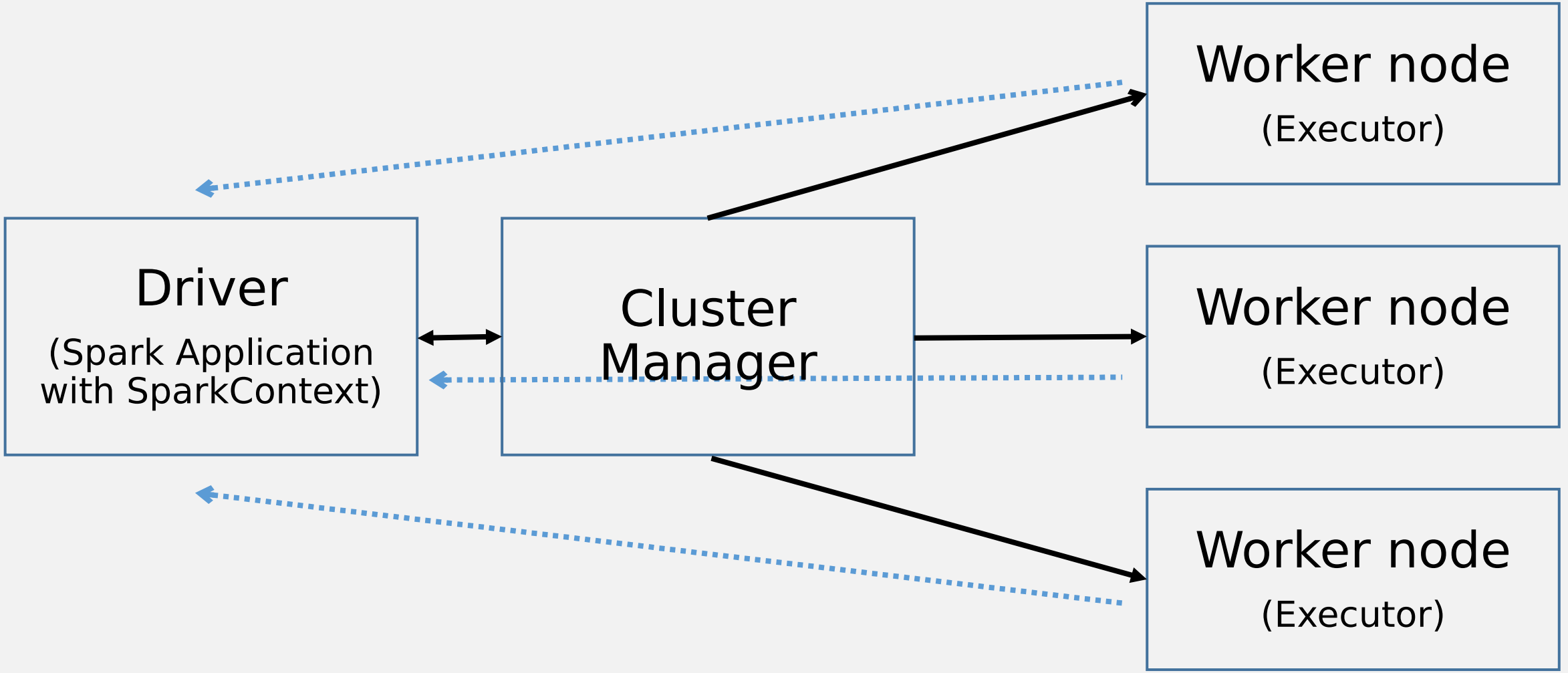
SQL & DataFrame	Streaming & Structured Streaming	MLlib & ML	Graph Tools	Community Packages
Language APIs (Scala, Python, R, Java)				

CLUSTER MANAGER

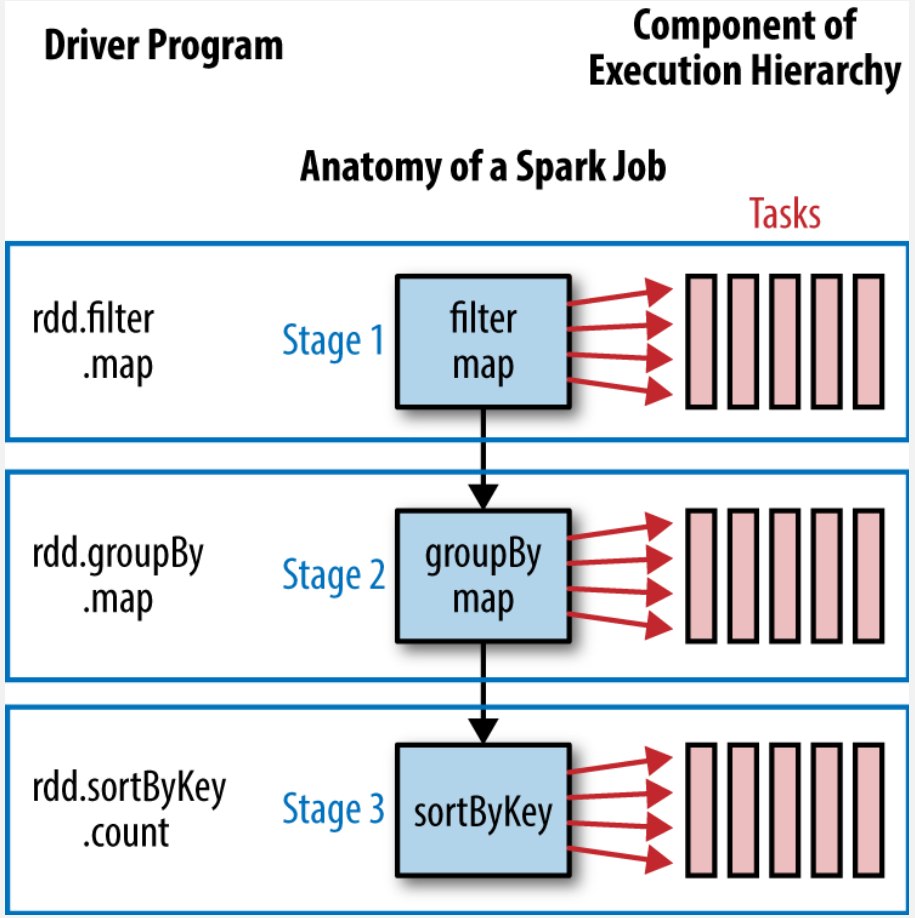
(Apache Mesos or Hadoop YARN or Standalone or Kubernetes*)

DISTRIBUTED STORAGE

Architektura Spark Core



Spark Job przykład



Źródło: High Performance Spark, Holden Karau & Rachel Warren

Spark Core Architecture. Driver

Proces JVM ze SparkContextem.

- Przez cały czas życia aplikacji utrzymuje informacje o niej
- Reaguje na input ze strony użytkownika
- Analizuje / rozdystrybuuje / planuje pracę dla poszczególnych executorów

Spark Core Architecture. Executor

Proces JVM.

- Wykonuje zlecone taski
- Wysyła komunikację zwrotną do drivera

SparkContext

SparkContext – główny, podstawowy obiekt w Sparku.
Punkt wejścia do Spark Core.

Przez niego tworzymy RDD, akumulatory, broadcast variables.

```
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("My App")
sc = SparkContext(conf = conf)
```

Intuicja: SparkContext == aplikacja Sparka

SparkSession

Wprowadzony od Spark 2.0

Zawiera m.in. SparkContext, ale również SqlContext, HiveContext.

Pierwszy obiekt, ktory tworzymy w aplikacji.

Spark Application Configuration

Spark Properties

Odpowiedzialne za większość ustawień aplikacji.

Można na ustawić:

1. Z poziomu obiektu SparkConf. Od Sparka 2.0 jest częścią SparkSession.
2. Przez parametry wywołania spark-submit

Dostępne opcje (properties) wraz z ich opisem, można znaleźć:

https://spark.apache.org/docs/latest/configuration.html#available-properties

Spark Properties

```
spark = SparkSession
    .builder()
    .appName('YourAppName')
    .master('local[2]')
    .getOrCreate()
```

```
Spark-submit --name 'YourAppName' --master 'local[2]'
```

Spark < 2.0 :

```
sparkConf = SparkConf().setAppName('YourAppName').setMaster('local[2]')
```