Spark

Apache Spark

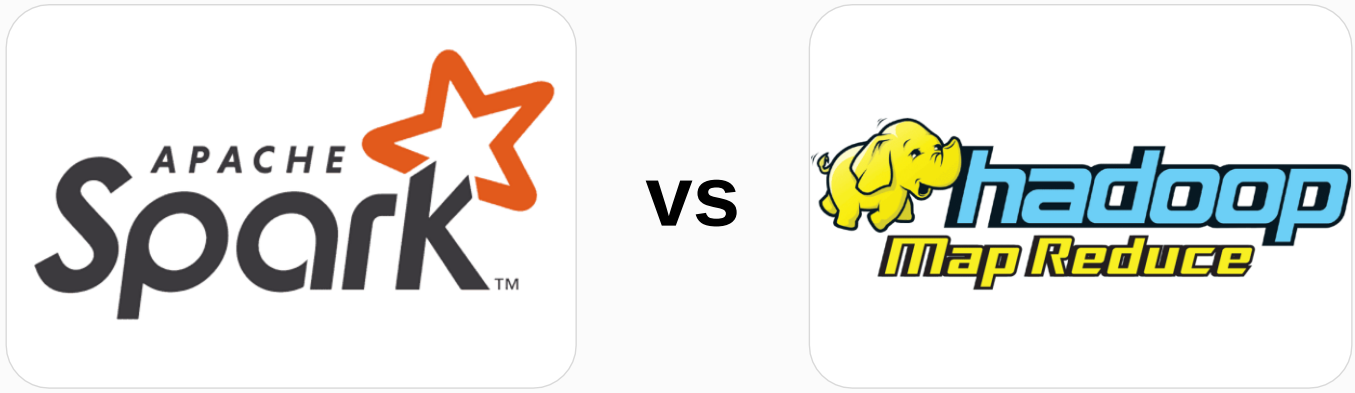
- Что такое Apache Spark
- Архитектура Spark и преимущества перед MapReduce
- Библиотека Spark SQL
- Библиотека Spark MlLib

Изображение: Apache Spark

Apache Spark

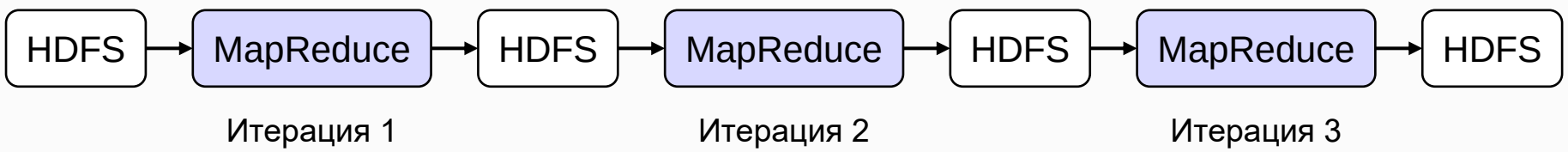
Apache Spark — фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop.

Отличия Spark от MapReduce

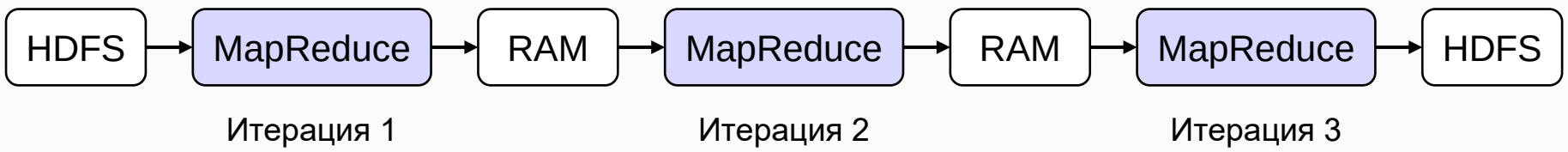


- In memory computing

MapReduce



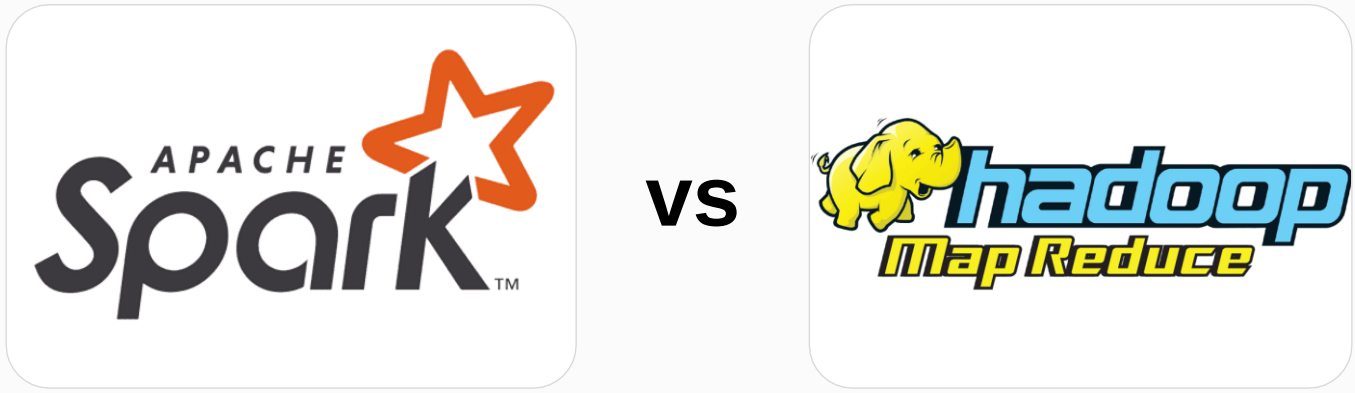
Spark



Огромный выигрыш в скорости обработки данных

Схема: Кирилл Богданов

Отличия Spark от MapReduce



- Spark Streaming — обработка потоковых данных.

Изображения: Apache Spark, Hadoop, Spark Streaming

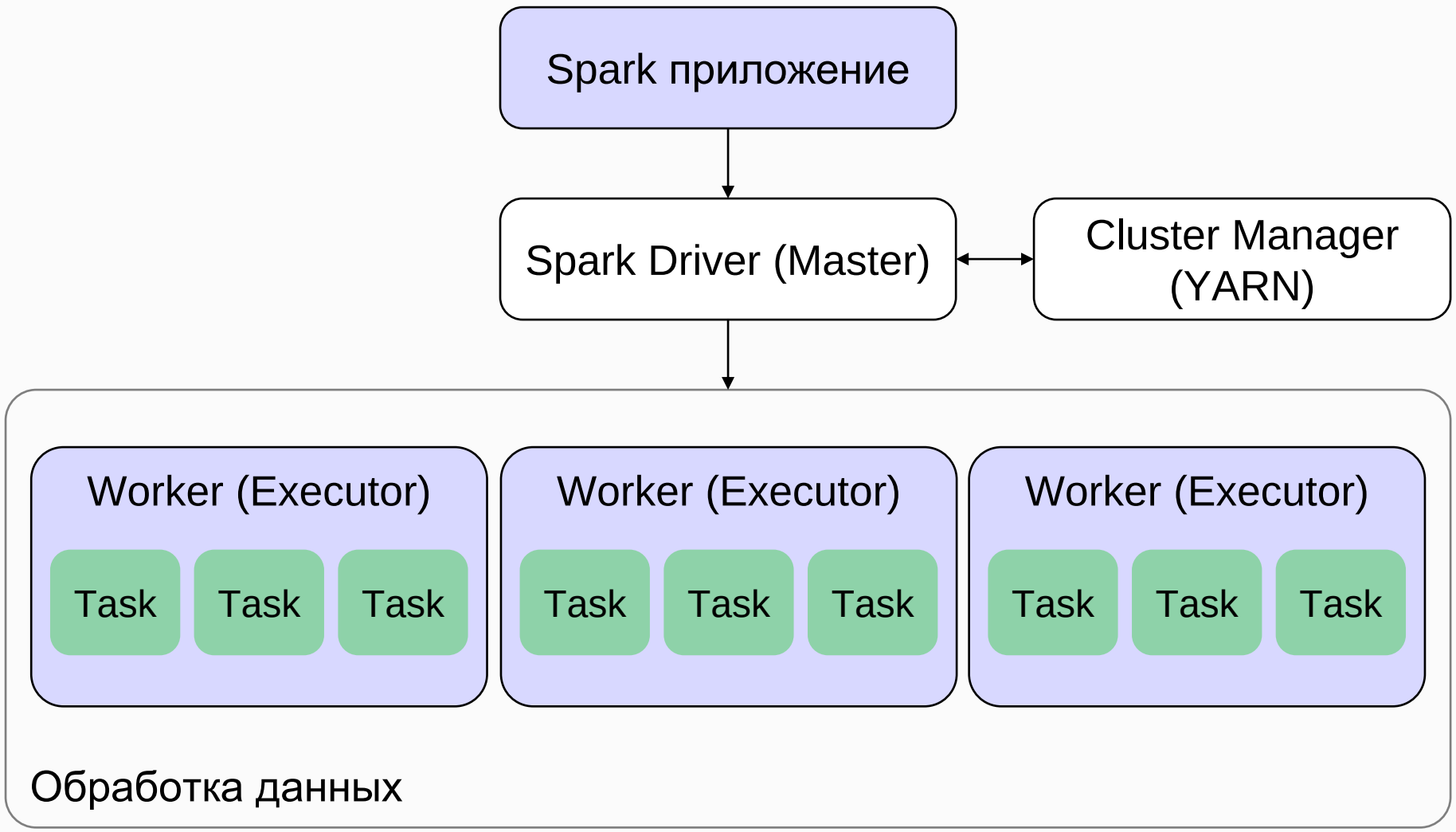
Отличия Spark от MapReduce

vs

- Программные интерфейсы для языков Java, Scala, Python, R
- Библиотеки для задач ML

Изображения: Apache Spark, Hadoop, Spark Jupyter, Spark MLlib

Архитектура Spark



1 Task = 1 Partition
(обработка каждой партиции делается
одним вычислительным ядром worker’а)

Схема: Skillbox

Структуры данных в Spark

RDD (Resilient Distributed Datasets)

- набор объектов, распределённых по нескольким узлам кластера

Spark DataFrame

- это оболочка над RDD, которая упрощает выполнение аналитики и обработки данных
- представляет собой таблицу с именами столбцов и типами данных для каждого столбца
- поддерживает Spark SQL

Spark Dataset

- DataFrame + типизация

Выводы

- Узнали, что такое Apache Spark, как и с какими сущностями он работает
- Узнали, как в Spark делать предобработку данных
- Узнали, как в Spark обучать модели