APACHE SPARK

The introduction

By Vasyl Nakvasiuk

01

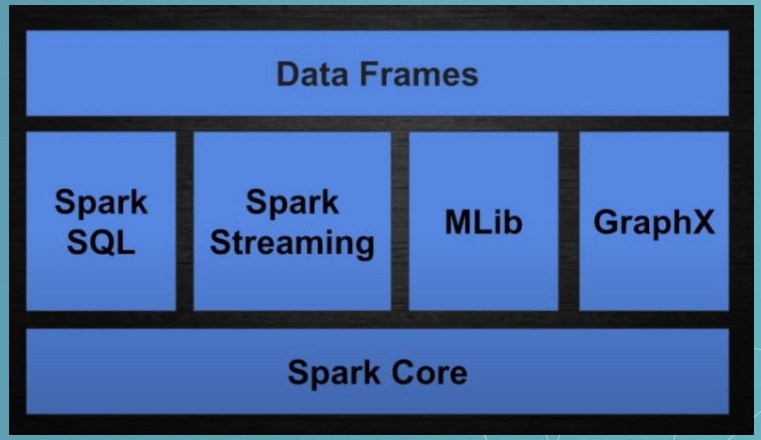
WHAT IS SPARK?

WHAT IS SPARK

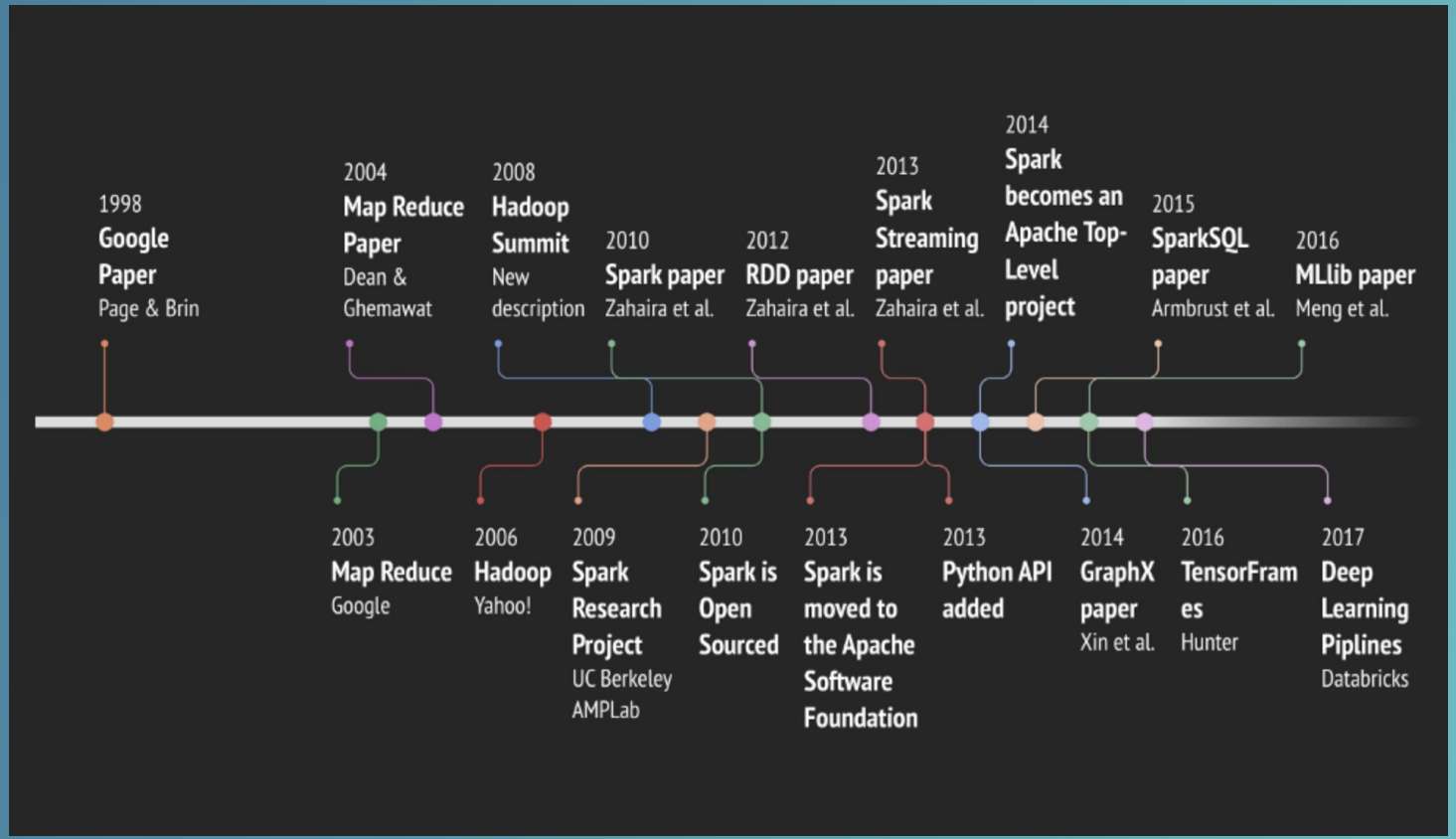
- distributed processing frameworks
- developer-friendly API
- in-memory data engine
- unified engine
- directed acyclic graph (DAG)

APACHE SPARK ECOSYSTEM

- Spark SQL + DataFrames
- Streaming
- MLlib - Machine Learning
- GraphX - Graph Computation



SPARK TIMELINE



SPARK INTEGRATIONS

APACHE SPARK ECOSYSTEM

Spark Core API

- R
- SQL
- Python
- Scala
- Java

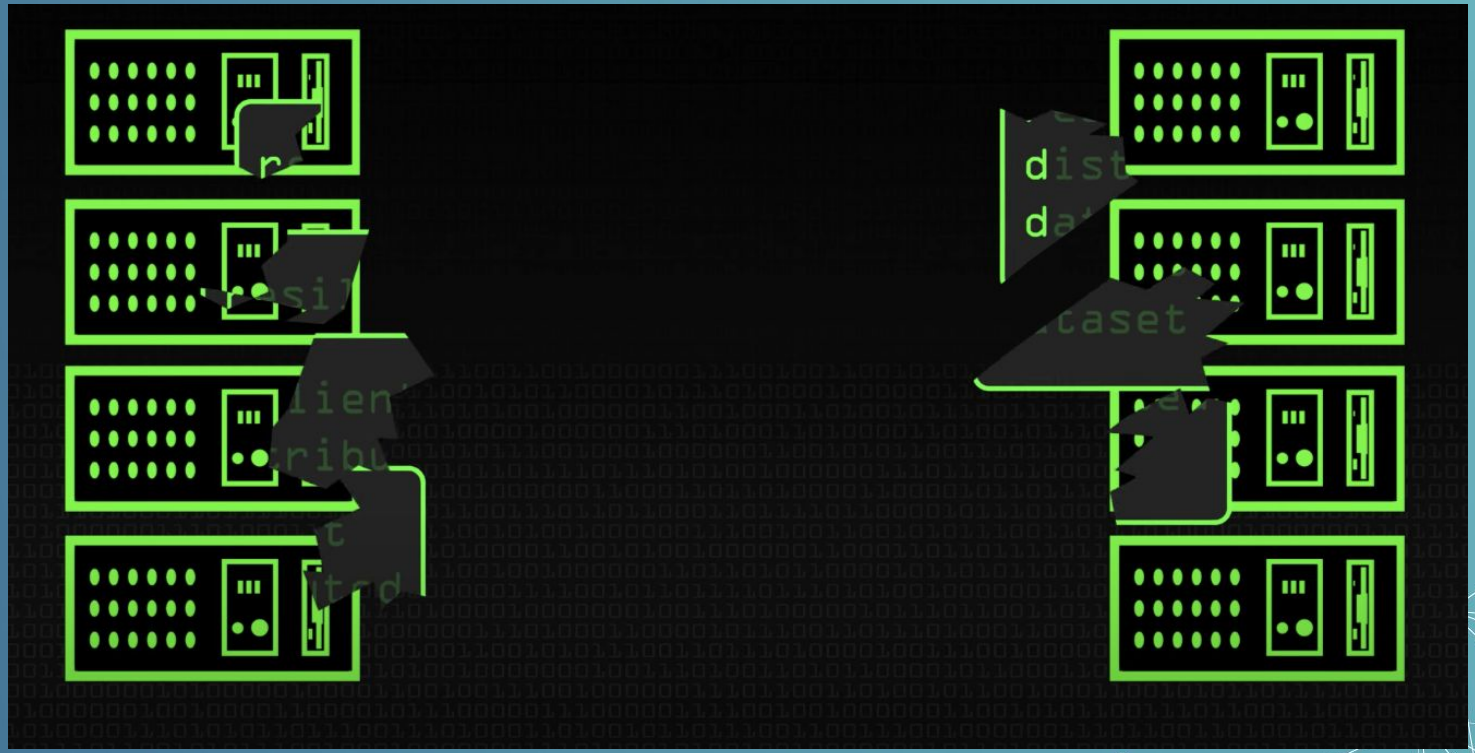
02

UNDER THE HOOD

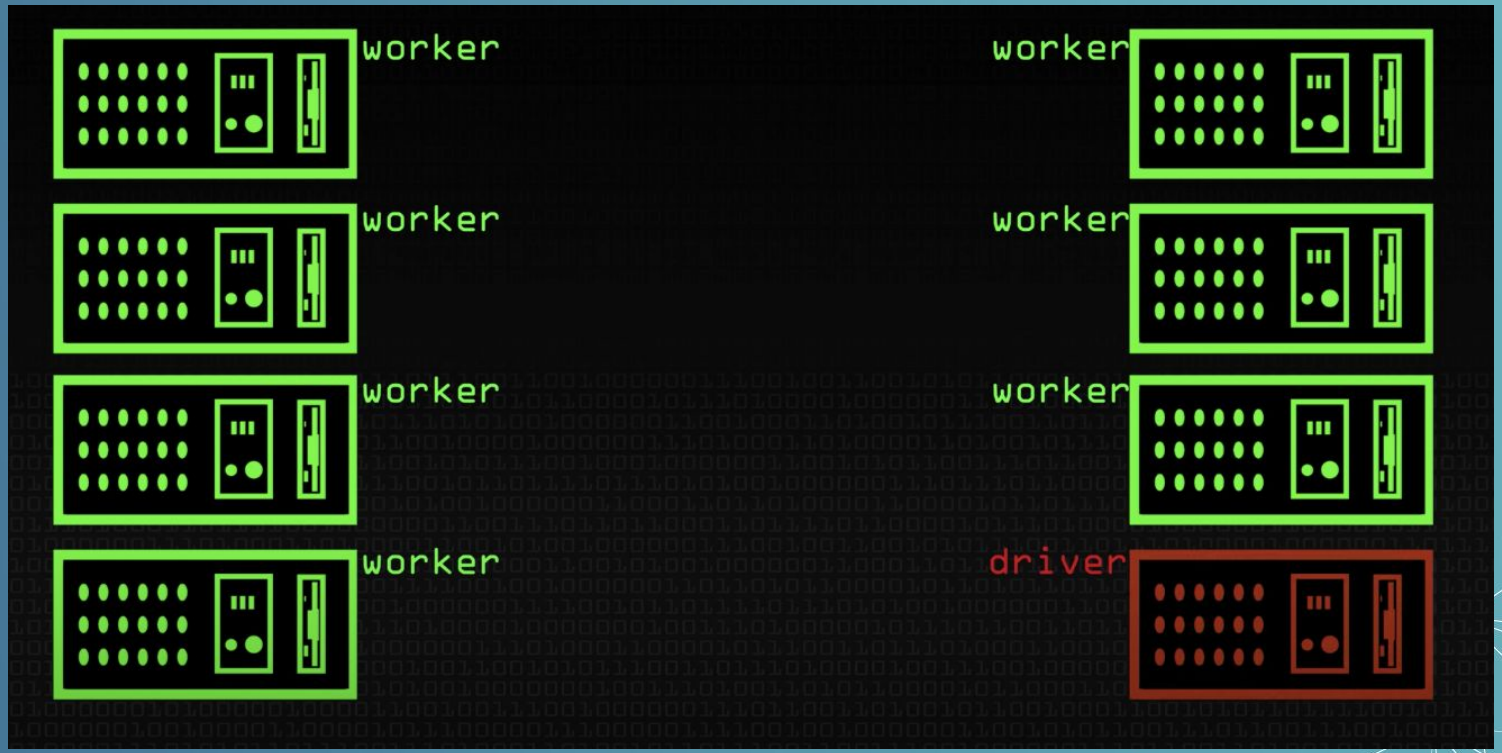
RDD

RDD

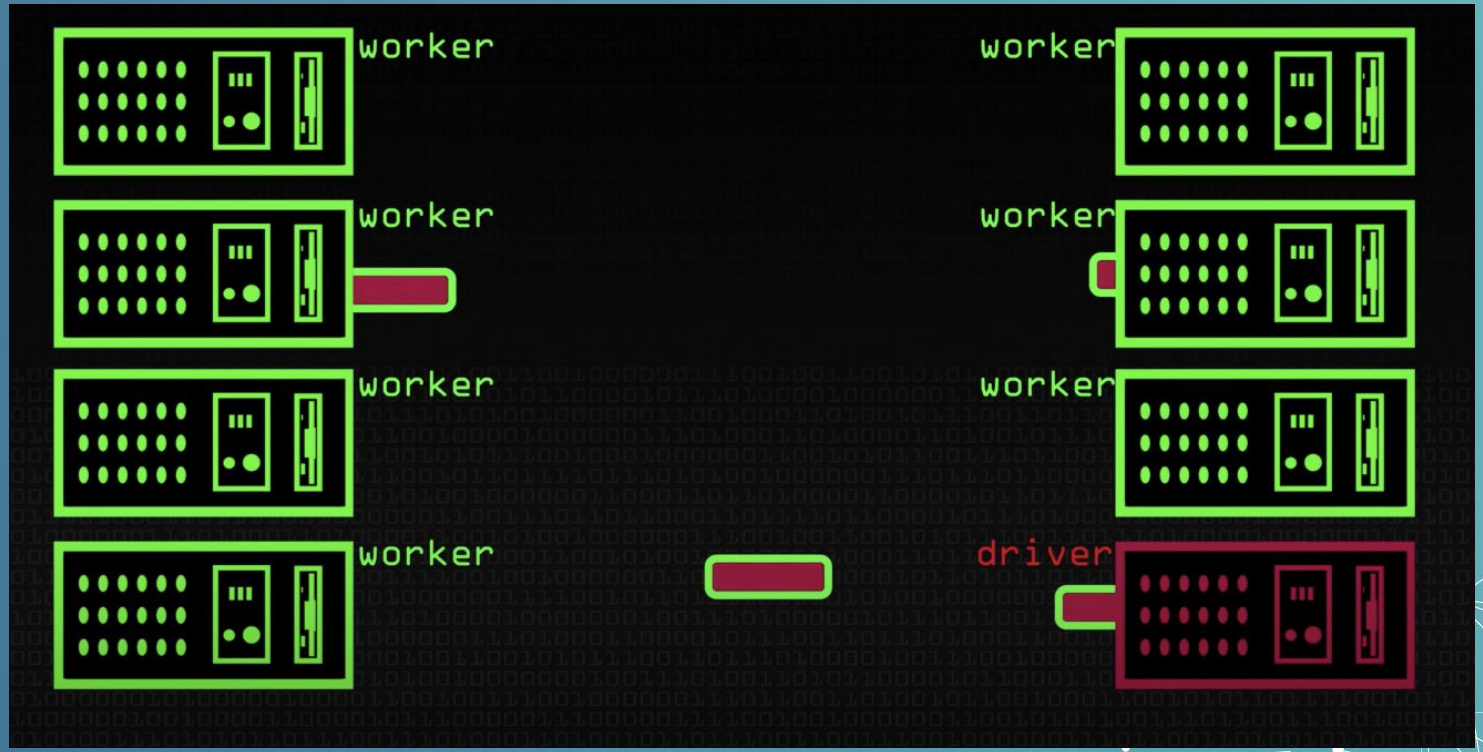
RDD



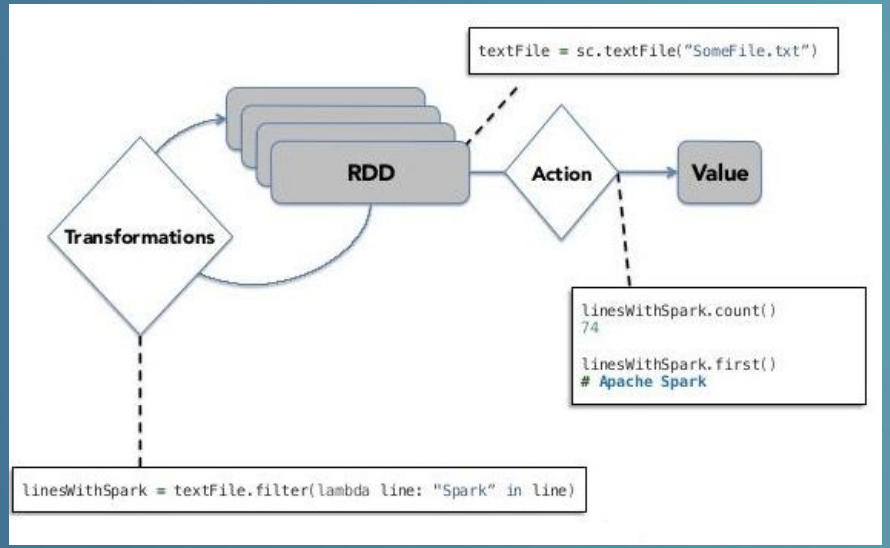
DISTRIBUTED SYSTEM

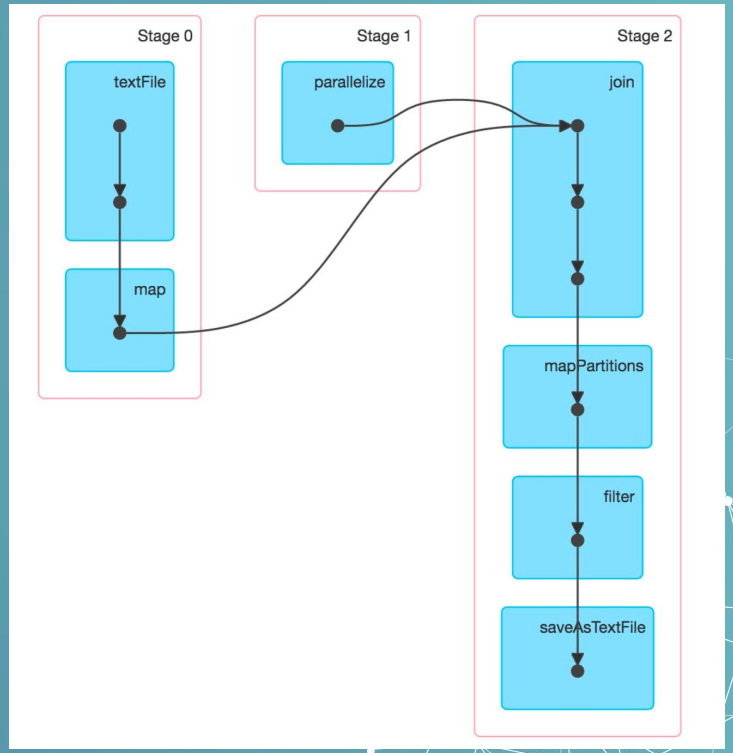


DISTRIBUTED SYSTEM



WORKING WITH RDDs





03

SPARK VS MAPREDUCE

SPARK VS MAPREDUCE

SPARK VS MAPREDUCE

Key Features	Apache Spark	Hadoop MapReduce
Speed	10–100 times faster than MapReduce	Slower
Analytics	Supports streaming, Machine Learning, complex analytics, etc.	Comprises simple Map and Reduce tasks
Suitable for	Real-time streaming	Batch processing
Coding	Lesser lines of code	More lines of code
Processing Location	In-memory	Local disk

THANKS

Does anyone have any questions?

v.nakvasiuk@ukma.edu.ua