



# APACHE SPARK

---

The introduction  
By Vasyl Nakvasiuk

An abstract geometric pattern consisting of white lines and dots (nodes) connected to form a network of triangles and polygons, set against a teal background.

# 01

## WHAT IS SPARK?

---

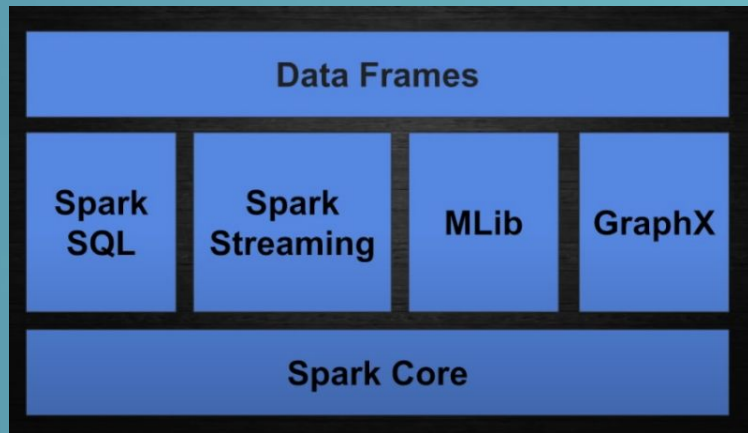
# WHAT IS SPARK

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

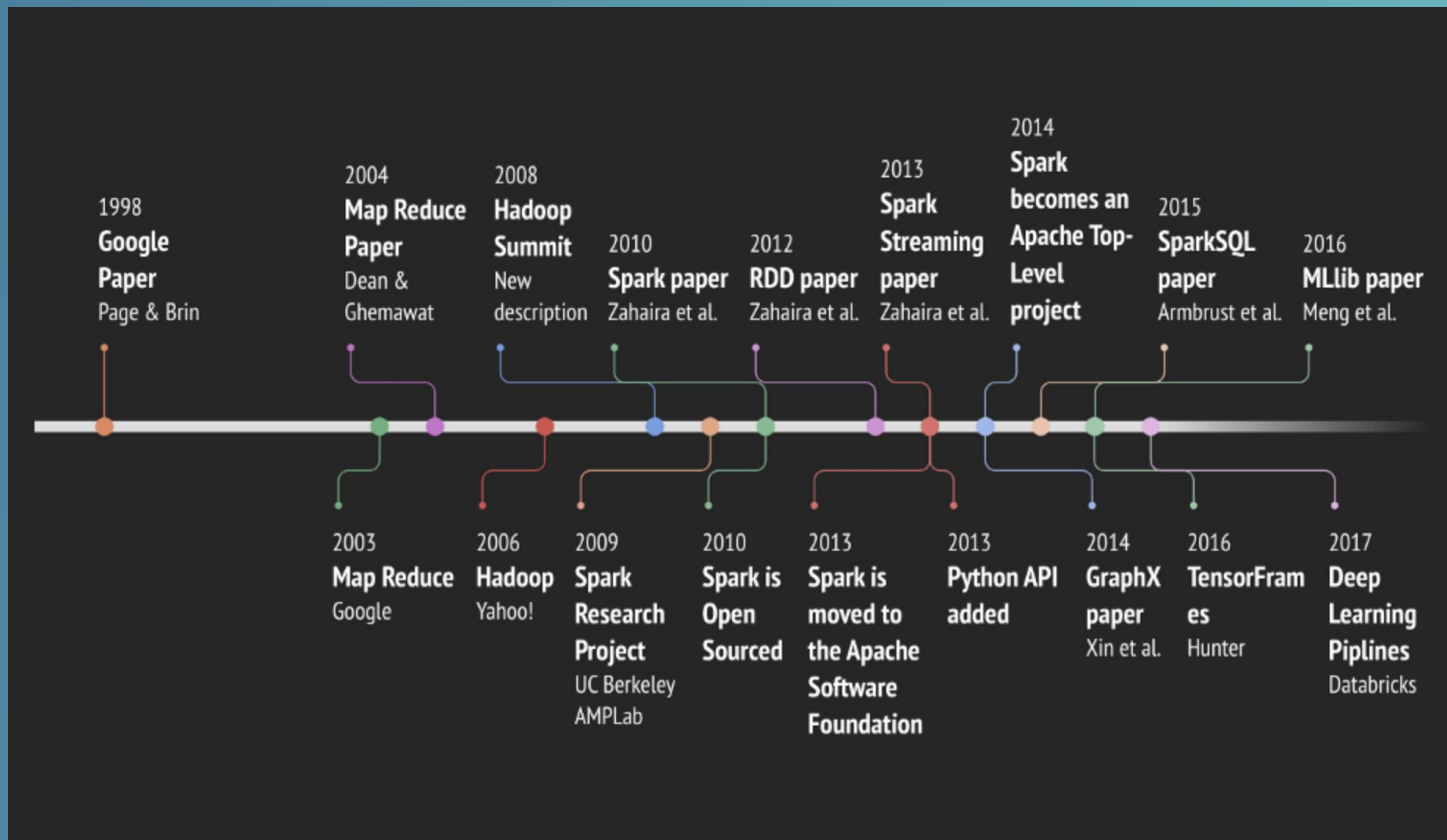


# APACHE SPARK ECOSYSTEM

- 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

resilient  
distributed  
dataset



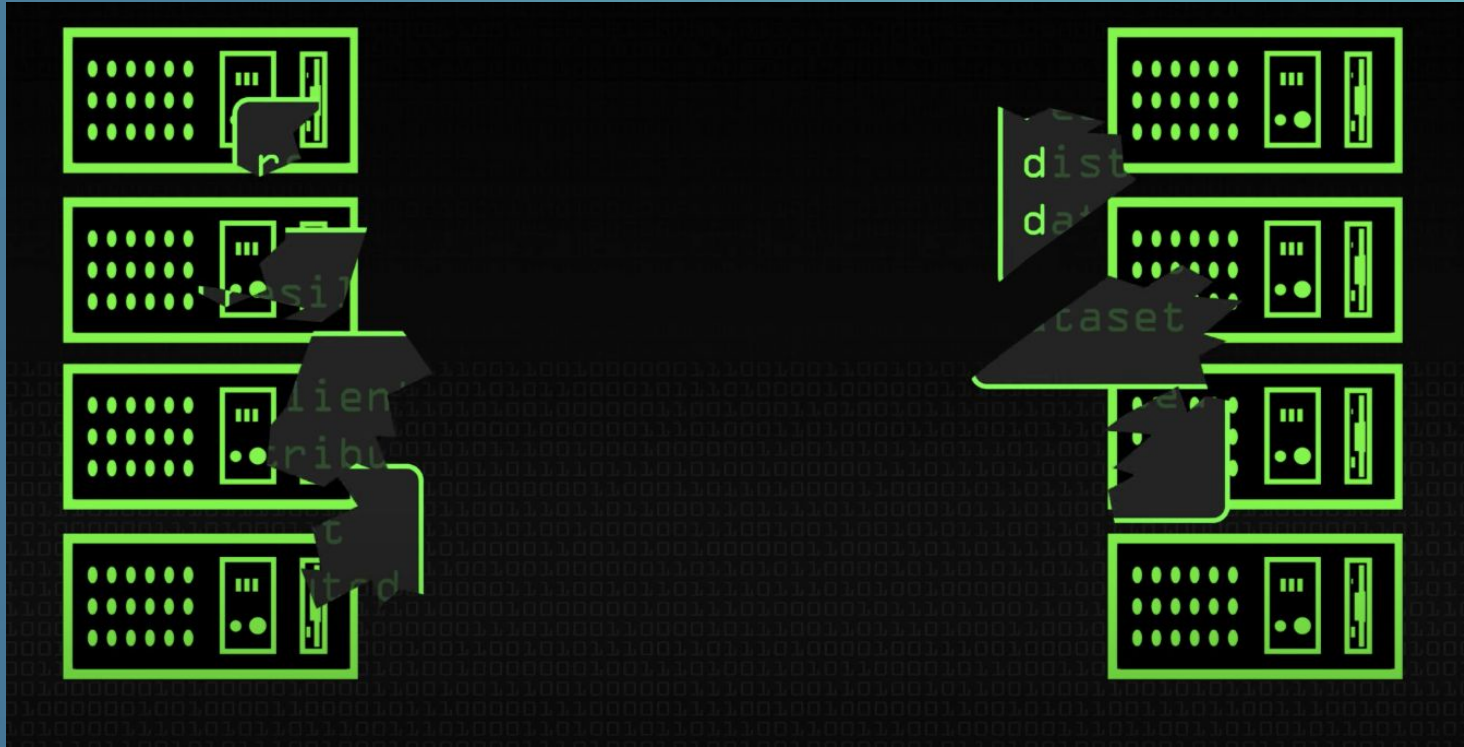
# RDD



The diagram illustrates a distributed system. In the center is a dark gray rounded rectangle with a red border containing the text "resilient distributed dataset". Surrounding this central box are eight server racks, arranged in two columns of four. Each rack is a light gray rectangle containing a 4x4 grid of small red circles, a small red square with three horizontal lines, and a small red vertical bar. The background is a dark gray field with a faint, repeating pattern of binary code (0s and 1s). On the right side of the image, there is a network diagram consisting of white dots connected by thin white lines, representing a distributed network topology.

resilient  
distributed  
dataset

# RDD



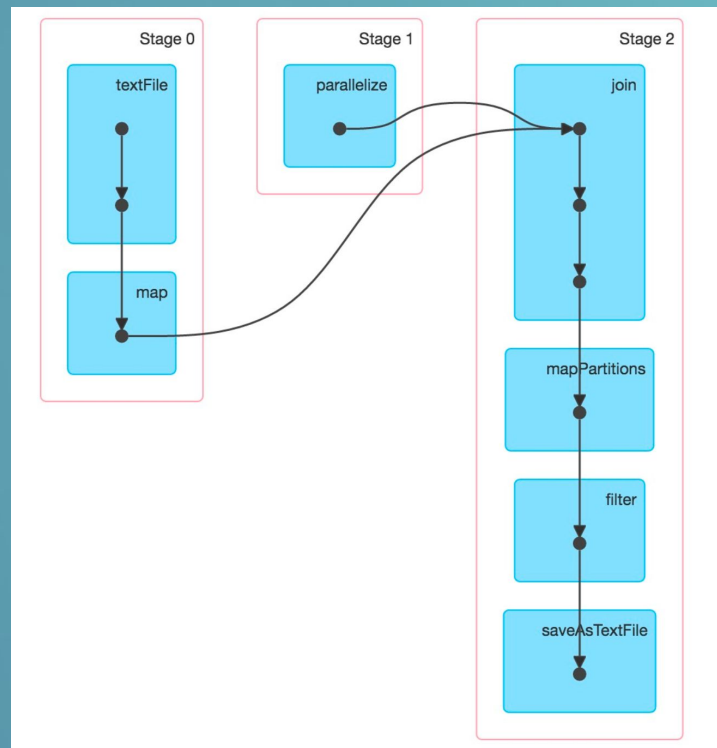
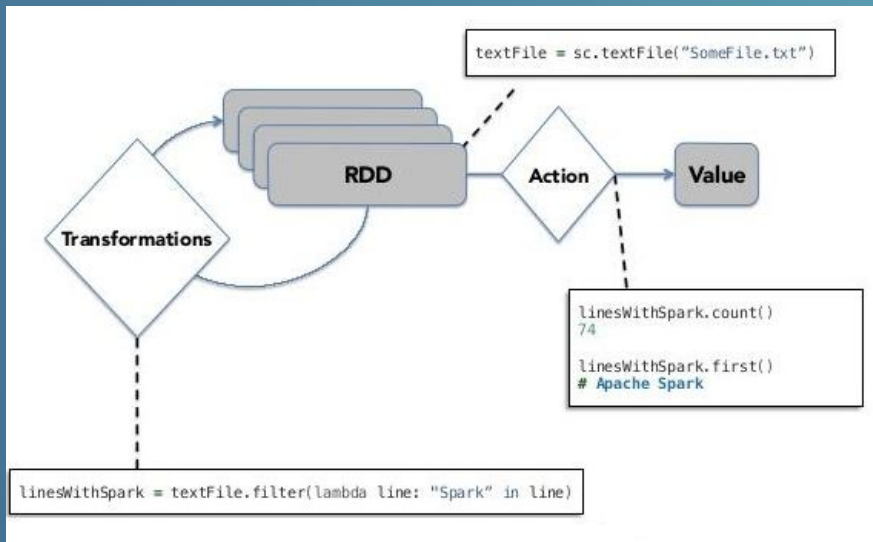
# DISTRIBUTED SYSTEM



# DISTRIBUTED SYSTEM



# WORKING WITH RDDs





The background is a solid teal color. Overlaid on this are several abstract geometric patterns. These consist of white dots (nodes) connected by thin white lines (edges). The connections form various polygonal shapes, some of which are triangles. The patterns are scattered across the slide, with a larger, more complex cluster in the upper left and another in the lower left. There are also some isolated dots and small triangles in the center and right areas.

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](mailto:v.nakvasiuk@ukma.edu.ua)

