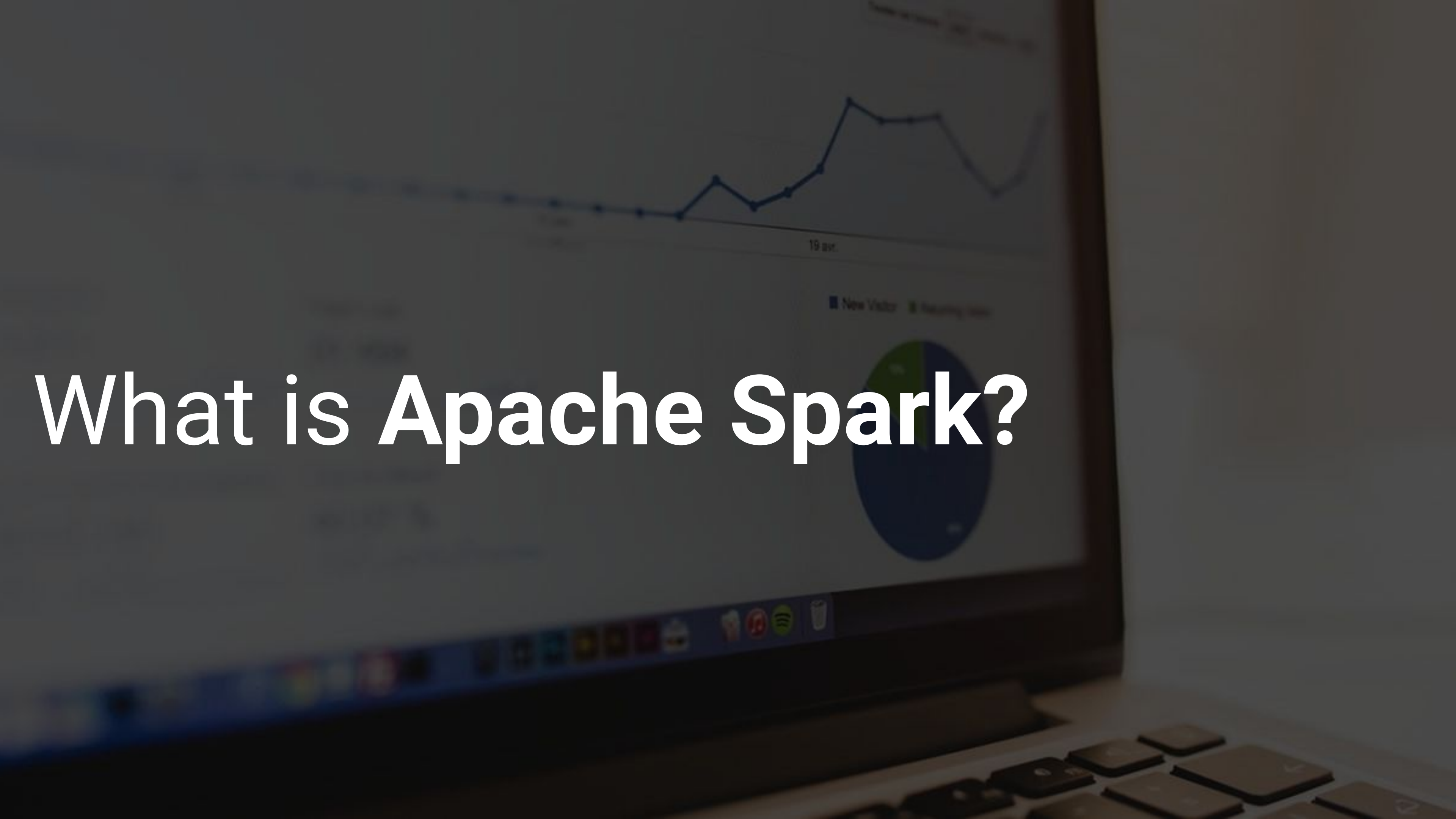
SHOW BUSINESS
NETWORK
MUSIC
CINEMA
BUSINESS/FINANCE
WORLD NEWS
ANALYSIS
EUROPE
AMERICA
ASIA
AFRICA
SEARCH
SCANNING
ANALYSIS
Spark



What is **Apache Spark?**

Apache Spark is an open-source cluster computing system that provides high-level API in Java, Scala, Python and R. It can access data from HDFS, Cassandra, HBase, Hive, Tachyon, and any Hadoop data source. And run in Standalone, YARN and Mesos cluster manager.

Apache Spark is a general-purpose & lightning fast cluster computing system.

It provides a high-level API. For example, Java, Scala, Python, and R. Apache Spark is a tool for Running Spark Applications.

Spark is 100 times faster than Big Data Hadoop and 10 times faster than accessing data from disk.

Spark is written in Scala but provides rich APIs in Scala, Java, Python, and R.

It can be integrated with Hadoop and can process existing Hadoop HDFS data.

Why **Spark?**

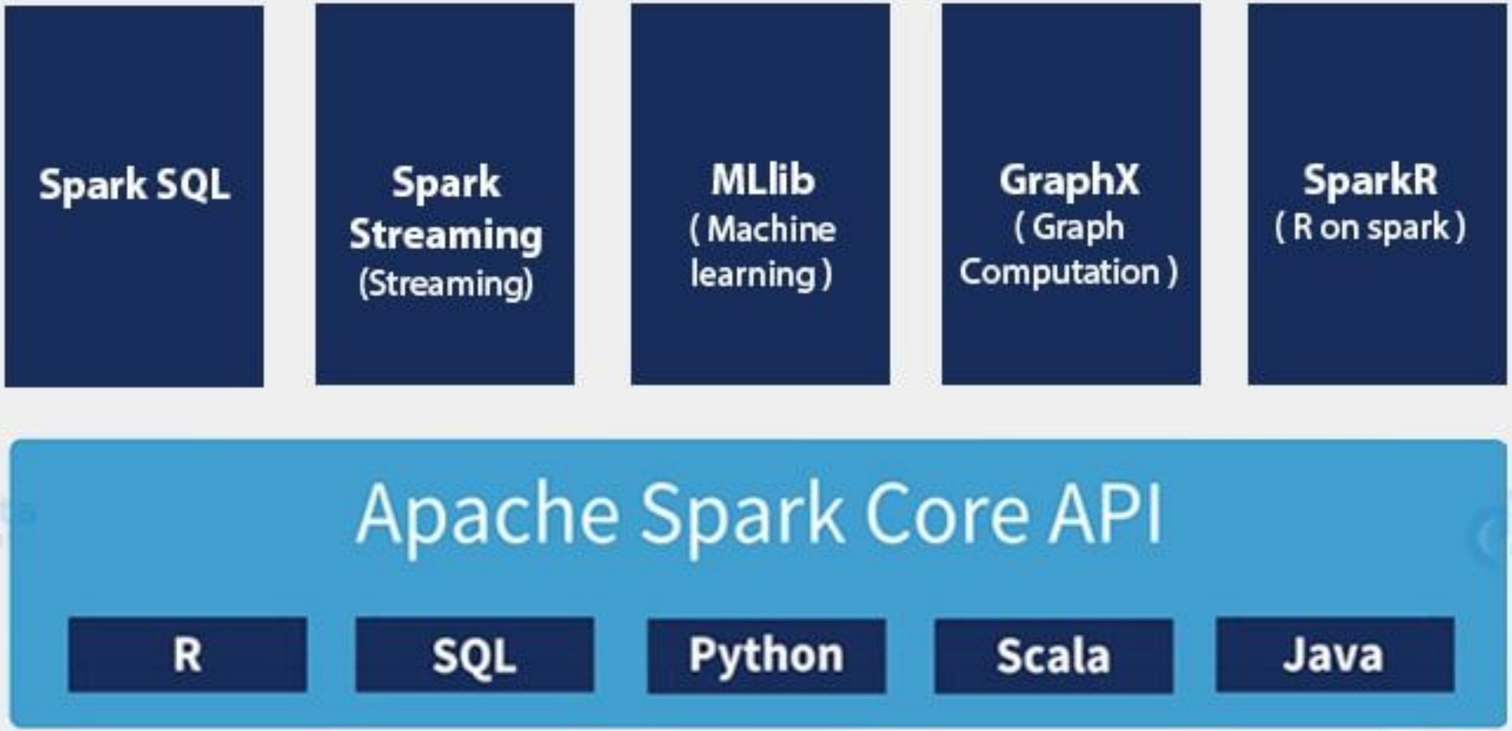
- Hadoop MapReduce can only perform batch processing.
- Apache Storm / S4 can only perform stream processing.
- Apache Impala / Apache Tez can only perform interactive processing
- Neo4j / Apache Giraph can only perform graph processing

Hence in the industry, there is a big demand for a powerful engine that can process the data in real-time (streaming) as well as in batch mode. There is a need for an engine that can respond in sub-second and perform in-memory processing.

Apache Spark Definition says it is a powerful open-source engine that provides real-time stream processing, interactive processing, graph processing, in-memory processing as well as batch processing with very fast speed, ease of use and standard interface. This creates the difference between Hadoop vs Spark and also makes a huge comparison between Spark vs Storm.

Spark Components

APACHE SPARK ECOSYSTEM



Spark SQL

Spark Streaming (Streaming)

MLlib (Machine learning)

GraphX (Graph Computation)

SparkR (R on spark)

Apache Spark Core API

R

SQL

Python

Scala

Java

Spark Core

It is the kernel of Spark, which provides an execution platform for all the Spark applications.

It is a generalized platform to support a wide array of applications.

Spark SQL

It enables users to run SQL/HQL queries on the top of Spark.

Using Apache Spark SQL, we can process structured as well as semi-structured data.

It also provides an engine for Hive to run unmodified queries up to 100 times faster on existing deployments.

Spark Streaming

Apache Spark Streaming enables powerful interactive and data analytics application across live streaming data.

The live streams are converted into micro-batches which are executed on top of spark core.

Spark MLlib

It is the scalable machine learning library which delivers both efficiencies as well as the high-quality algorithm.

Apache Spark MLlib is one of the hottest choices for Data Scientist due to its capability of in-memory data processing, which improves the performance of iterative algorithm drastically.

Spark GraphX

Apache Spark GraphX is the graph computation engine built on top of spark that enables to process graph data at scale.

SparkR

It is R package that gives light-weight frontend to use Apache Spark from R.

It allows data scientists to analyze large datasets and interactively run jobs on them from the R shell.

The main idea behind SparkR was to explore different techniques to integrate the usability of R with the scalability of Spark.

Spark RDD

Resilient Distributed Dataset (RDD) is the fundamental unit of data in Apache Spark, which is a distributed collection of elements across cluster nodes and can perform parallel operations. Spark **RDDs** are immutable but can generate new **RDD** by transforming existing **RDD**.

There are three ways to create **RDDs** in Spark:

- Parallelized collections – We can create parallelized collections by invoking parallelize method in the driver program.
- External datasets – By calling a textFile method one can create RDDs. This method takes URL of the file and reads it as a collection of lines.
- Existing RDDs – By applying transformation operation on existing RDDs we can create new RDD.

Spark RDDs Support two types of operations

- **Transformation** – Creates a new RDD from the existing one. It passes the dataset to the function and returns new dataset.

- **Action** – Spark Action returns final result to driver program or write it to the external data store.

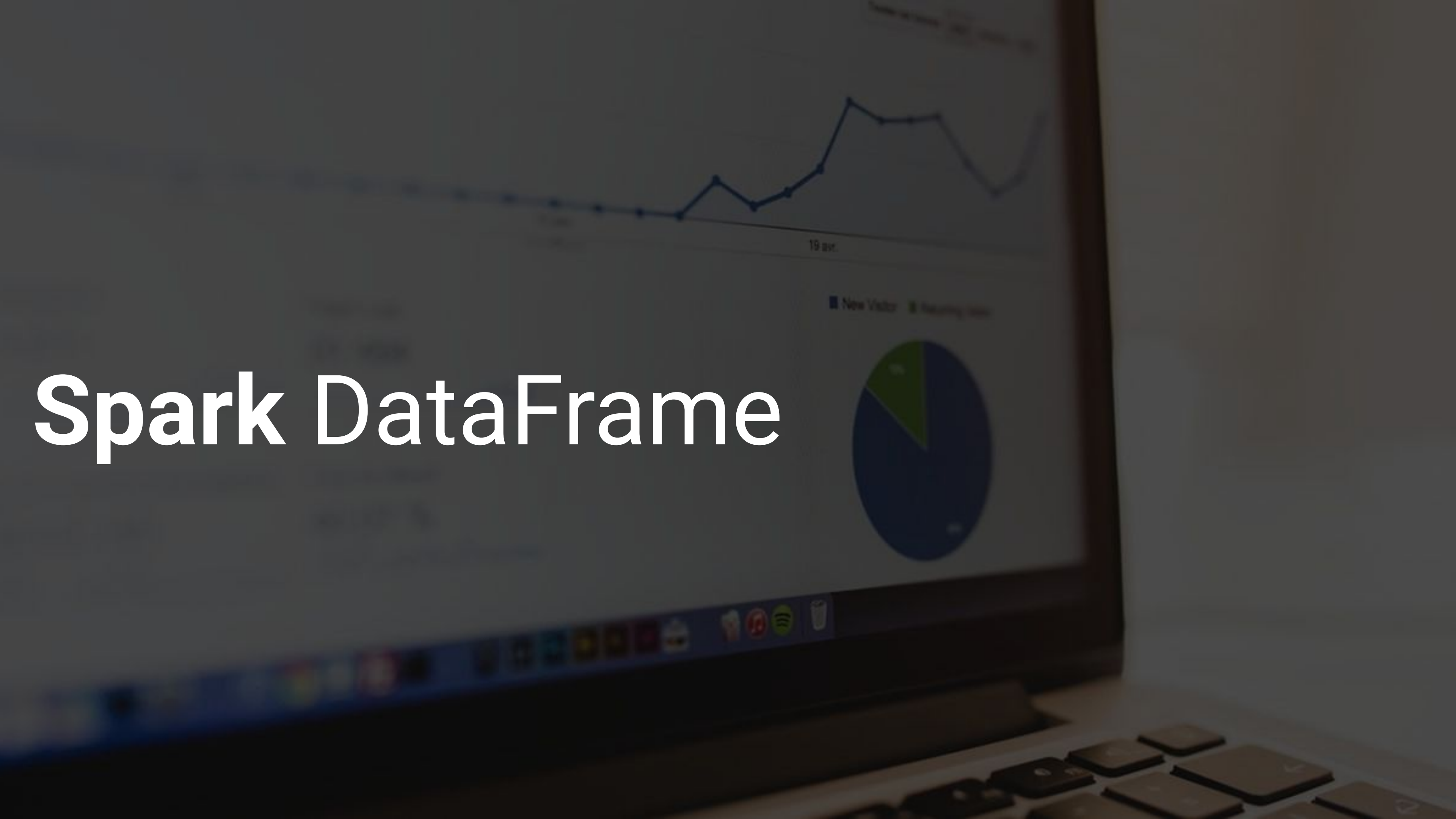
Spark Dataset

A Dataset is a distributed collection of data.

A Dataset can be constructed from JVM objects and then manipulated using functional transformations (map, flatMap, filter, etc.).

The Dataset API is available in Scala and Java.

Spark DataFrame



A DataFrame is a *Dataset* organized into named columns.

It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood.

DataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases or existing RDDs.

Spark Shell

Apache Spark provides an interactive **spark-shell.**

It helps Spark applications to easily run on the command line of the system.

Using the **Spark shell** we can run/test our application code interactively.

Spark can read from many types of data sources so that it can access and process a large amount of data.

What is **PySpark?**

PySpark is the collaboration of Apache Spark and Python.

Apache Spark is an open-source cluster-computing framework, built around speed, ease of use, and streaming analytics whereas Python is a general-purpose, high-level programming language.

Questions?