Why use Apache Spark?

Objectives

After watching this presentation, you will be able to:

- Describe Apache Spark attributes
- Describe distributed computing
- List the benefits of Apache Spark and distributed computing
- Compare and contrast Apache Spark to MapReduce

Apache Spark attributes

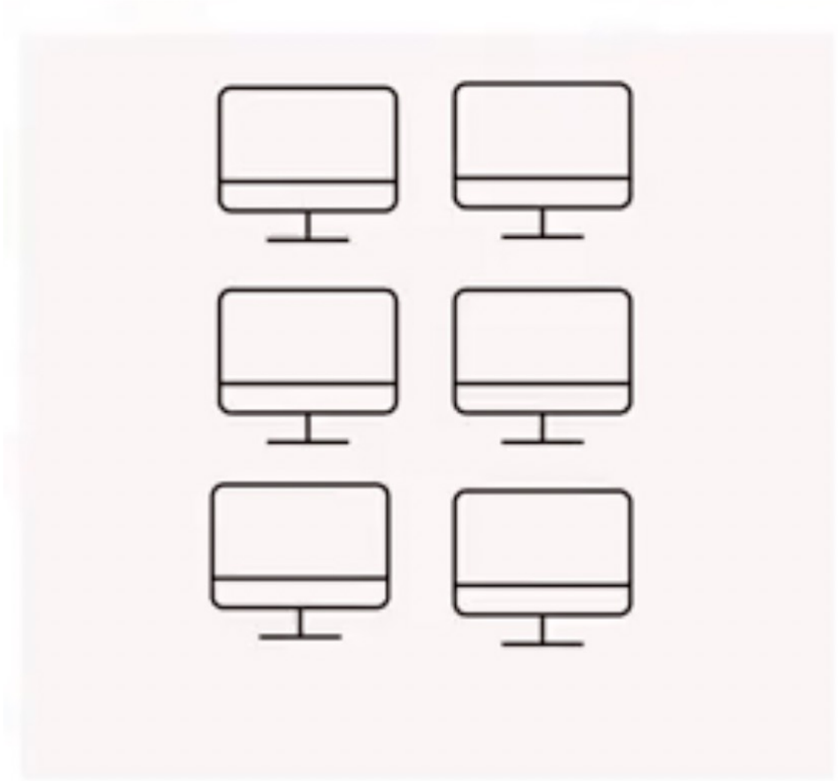
Spark is an **open** source **in-memory** **application** framework for **distributed** data processing and **iterative** analysis on **massive** data volumes

Apache Spark Attributes

Spark is written predominantly in **Scala** and runs on **Java virtual machines (JVMs)**

What is Distributed Computing?

- A group, or **cluster**, of computers working together to appear as one system to the end user

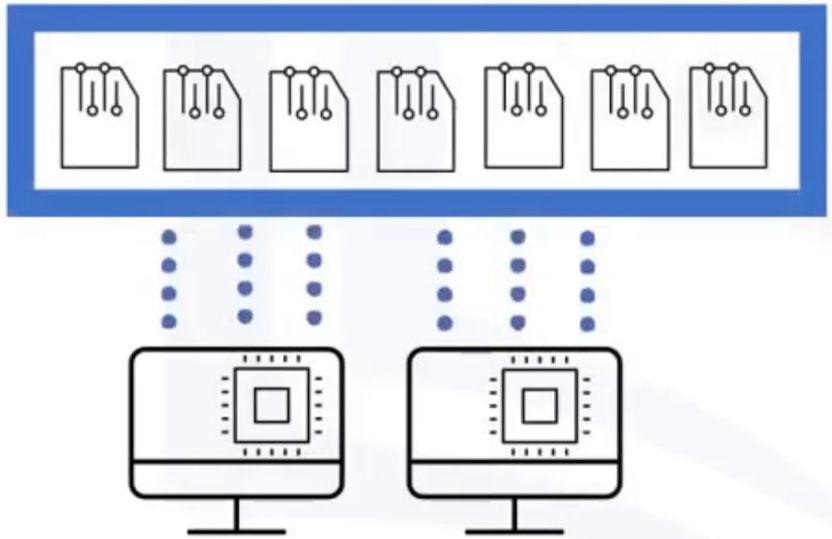


What is Distributed Computing?

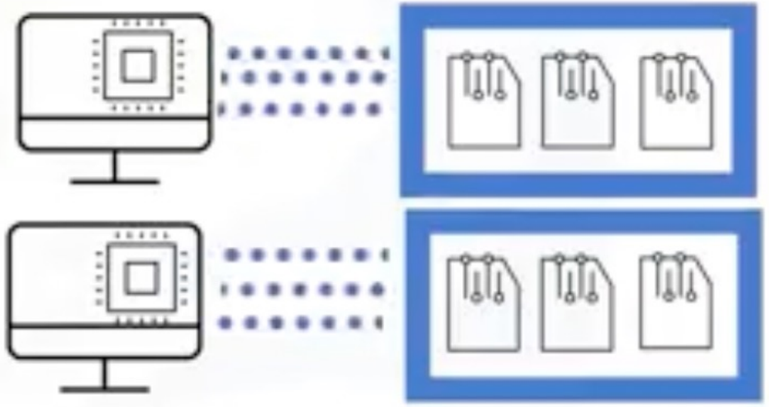
- The term distributed computing often used interchangeably with parallel computing as both are similar.

Parallel versus Distributed Computing

- Parallel computing processors access shared memory

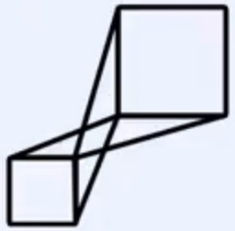


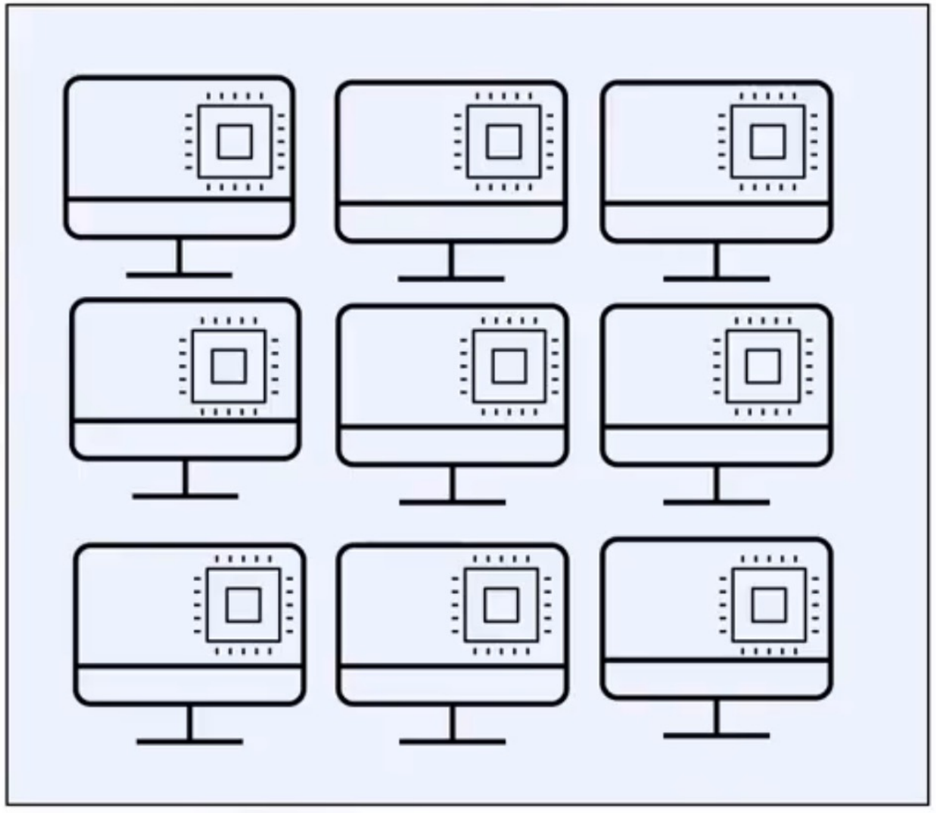
- Distributed computing processors usually have their own private or distributed memory



Distributed computing benefits

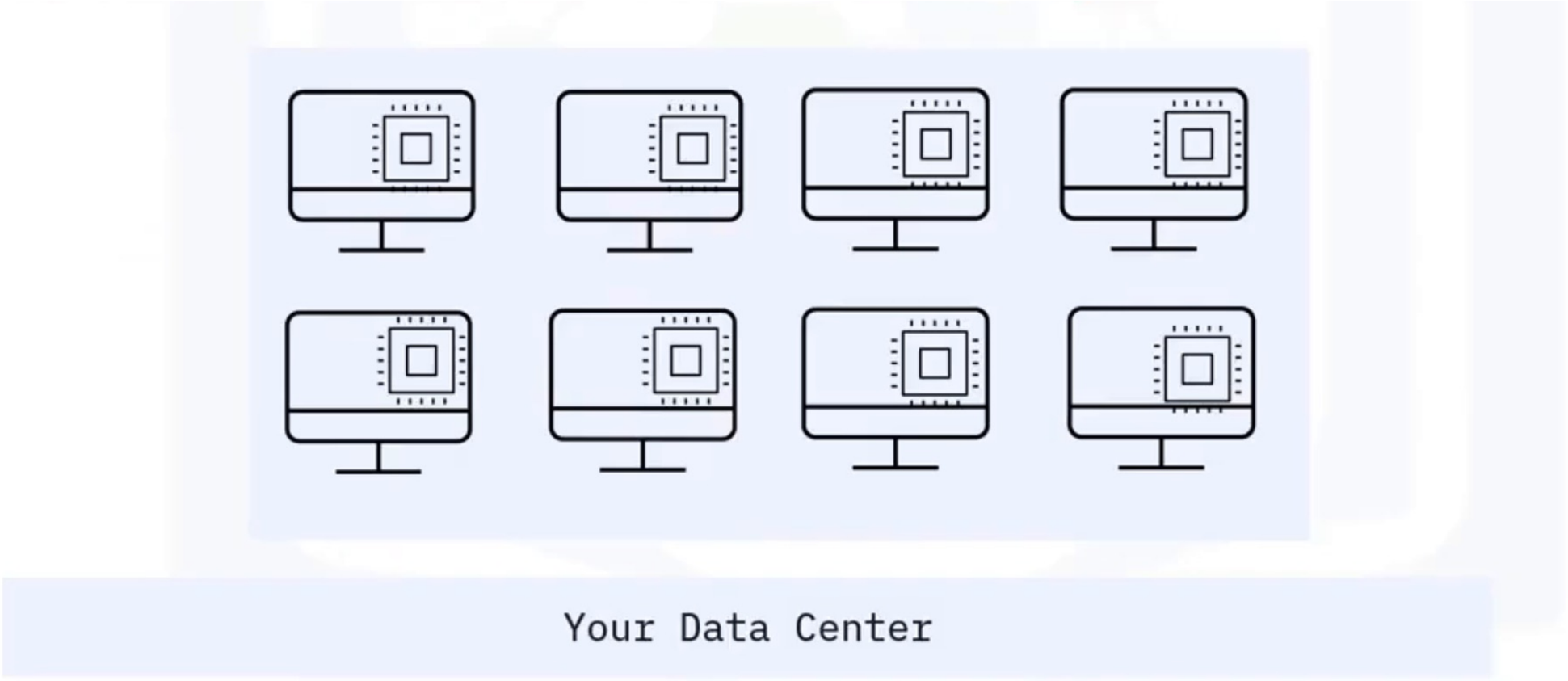
- Scalability and modular growth





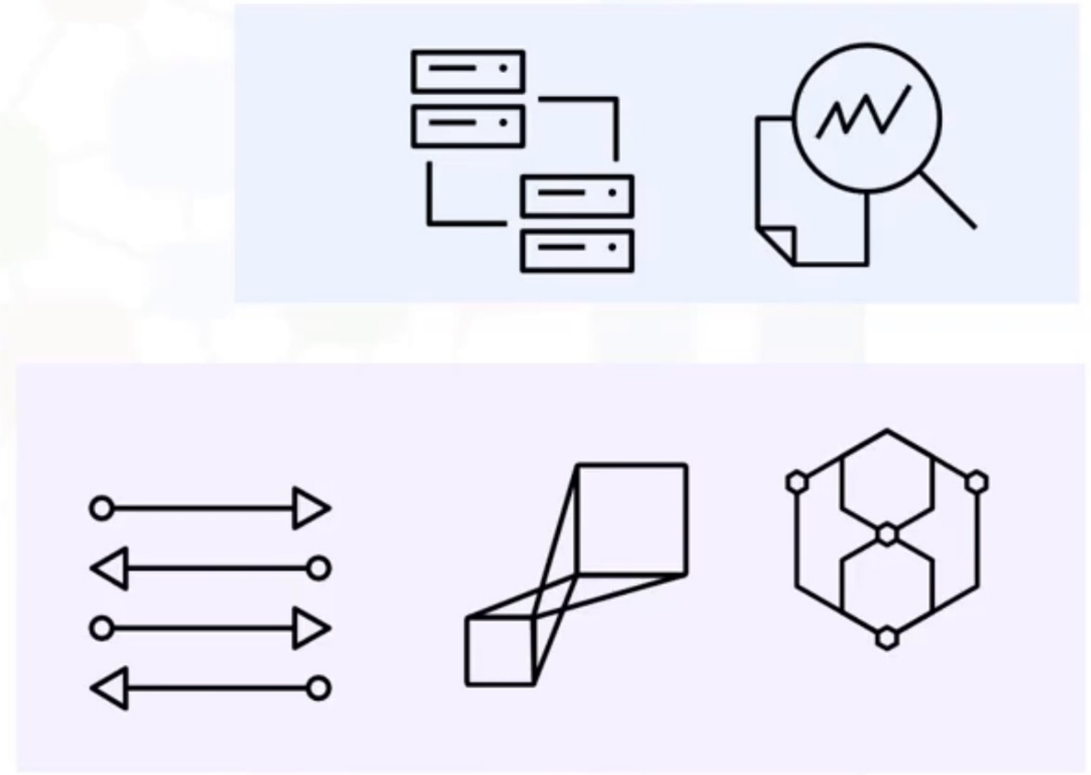
Distributed computing benefits

- Fault tolerance and redundancy



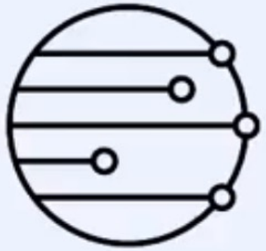
Spark Benefits

- Supports a computing framework for large scale data processing and analysis
- Provides parallel and distributed processing, scalability and fault tolerance on commodity hardware



Apache Spark Benefits

- Provides speed due to in-memory processing
- Creates a comprehensive, unified framework to manage big data processing
- Enables programming flexibility with easy-to-use Python, Scala, and Java APIs



Apache Spark & MapReduce Compared

Traditional Approach:

- Create MapReduce jobs for complex jobs, interactive query, and online event-hub processing involves lots of (slow) disk I/O

Apache Spark & MapReduce Compared

Solution:

- Keep more data in-memory with a new distributed execution engine

Spark and Big Data

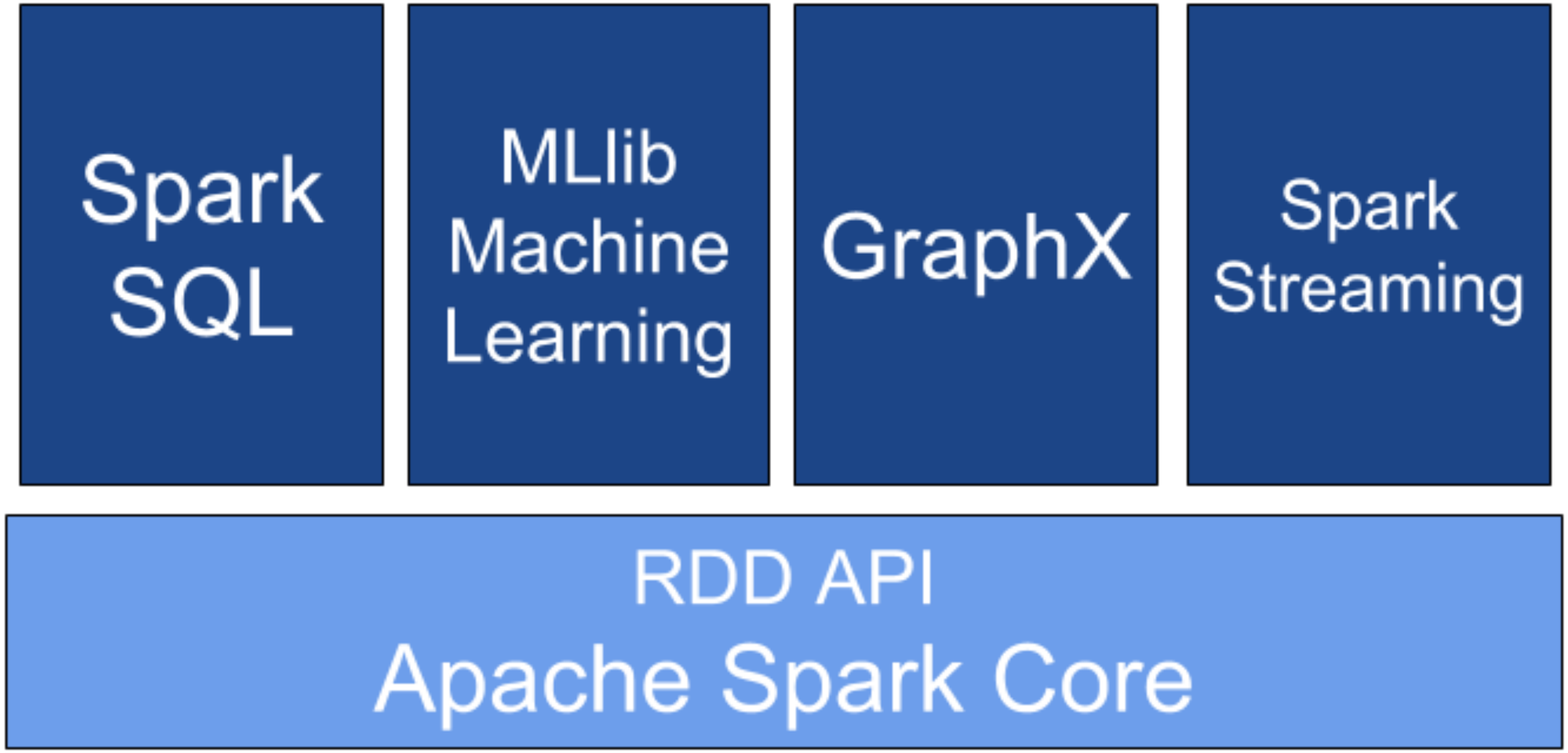
Data engineering

- Core Spark engine
- Clusters and executors
- Cluster management
- SparkSQL
- Catalyst
  Tungsten DataFrames

Data science and Machine learning

- SparkML
- DataFrames
- Streaming

Apache Spark Components



Summary

- Spark is an open source in-memory application framework for distributed data processing and iterative analysis on massive data volumes
- Distributed computing is a group of computers or processors working together behind the scenes
- Both distributed systems and Apache Spark are inherently scalable and fault tolerant
- Apache Spark a large portion of the data required in memory and avoids expensive and time-consuming disk I/O