Big Data Technologies: Hadoop, MapReduce & Spark

CS 448 – Spring 2023

Trevor Bonjour

Agenda

- Motivation
- Characteristics of Big Data
- Hadoop
- Spark

Estimated data generated in 2022*

97 Zettabytes

1 zettabyte = 1 billion terabyte

*source: statista

Data generated every minute*

- Email users send 231.4M messages
- Google users conduct 5.9M searches
- Snapchat users send 2.43M snaps
- Tinder users swipe 1.1M times
- Instagram users share 66K photos

*source: Domo

Characteristics of Big Data

Volume	Velocity	Variety	Veracity
Size of data	Rate of data	Types of data	Quality of data

Core Components of Hadoop

Hadoop Distributed Filesystem (HDFS)

Distributed file system designed to run on commodity hardware.

Hadoop MapReduce

Software framework for processing large data sets in a distributed computing environment.

HDFS Design Goals

Fault Tolerant

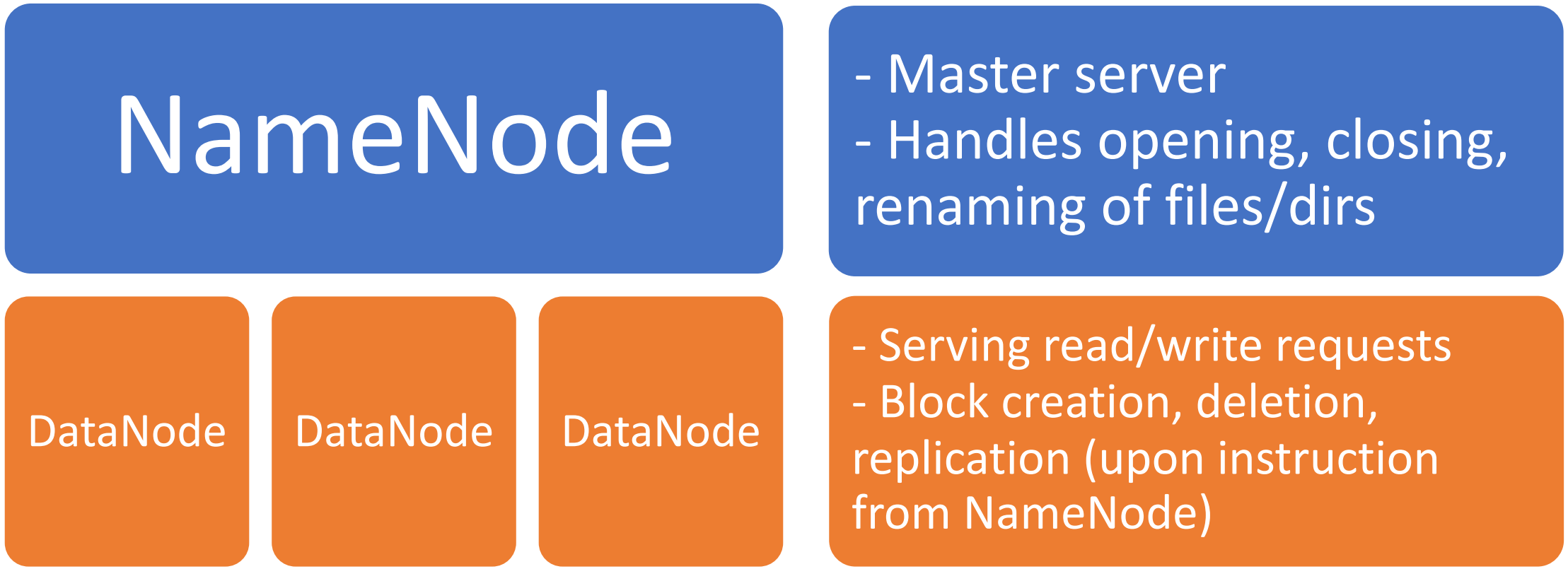
High Throughput Access

Support Large Datasets

Low-cost Hardware

Append Only Data Write Model

HDFS Architecture



MapReduce

Software framework for processing large data sets in a distributed computing environment.

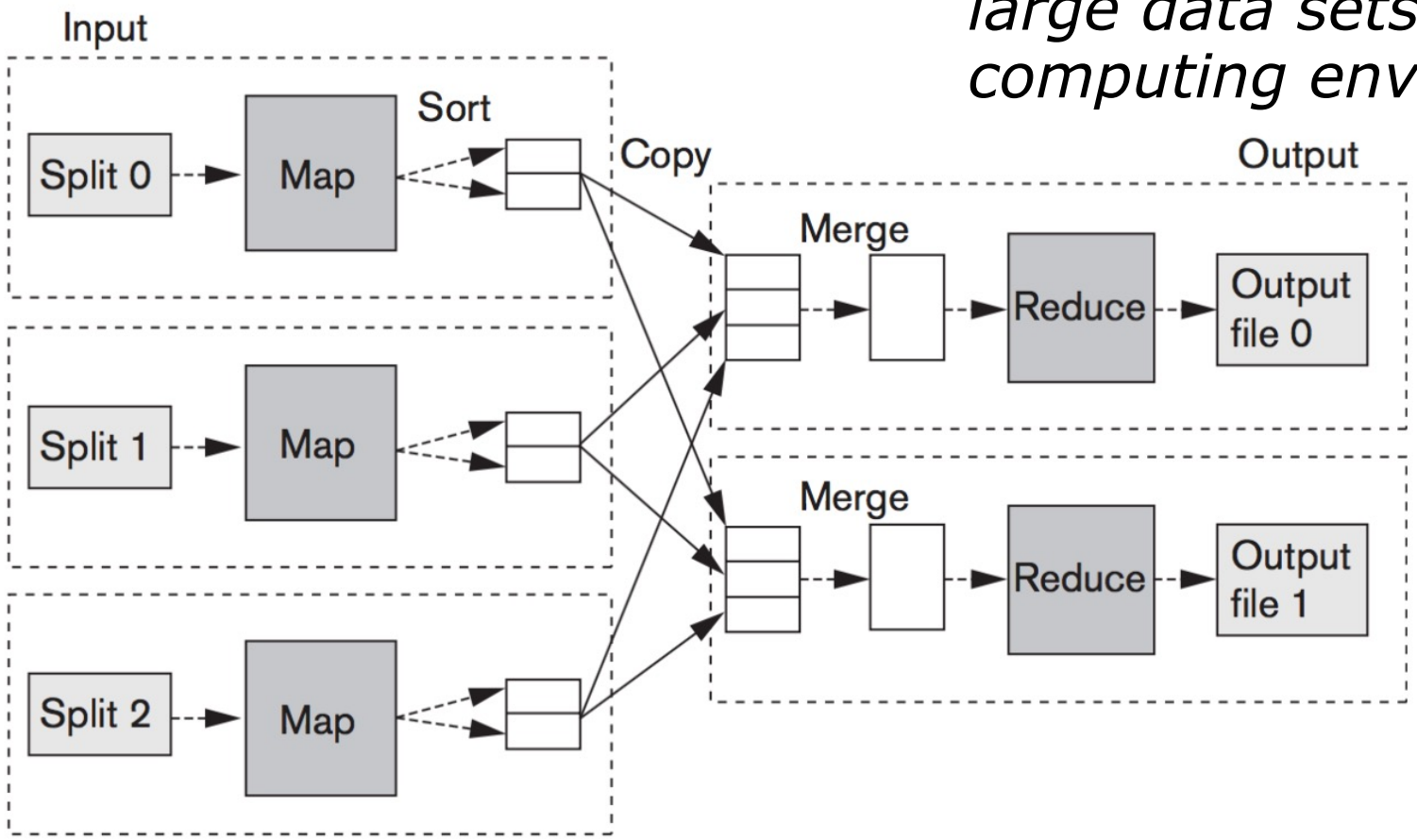


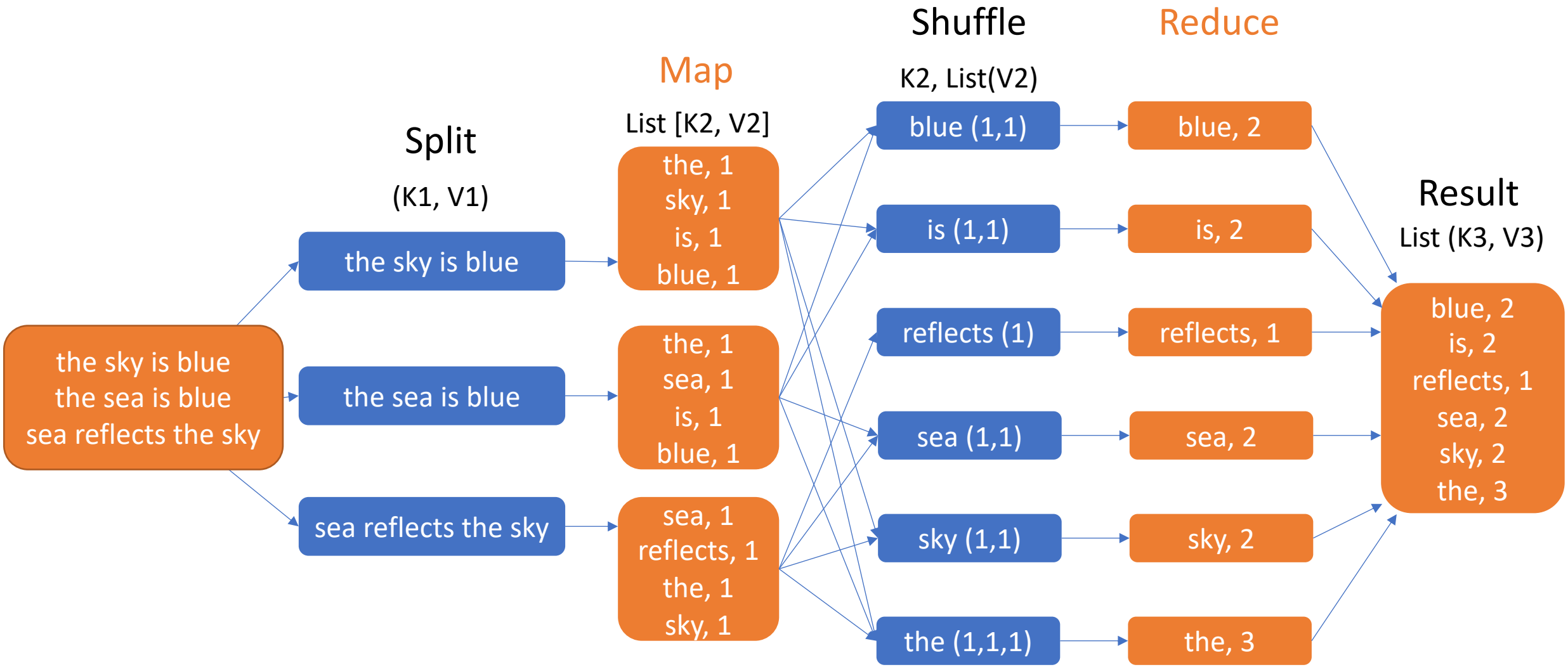
Figure 25.1
Overview of MapReduce execution. (Adapted from T. White, 2012)

MapReduce Model

map(K1,V1) : List[K2,V2]

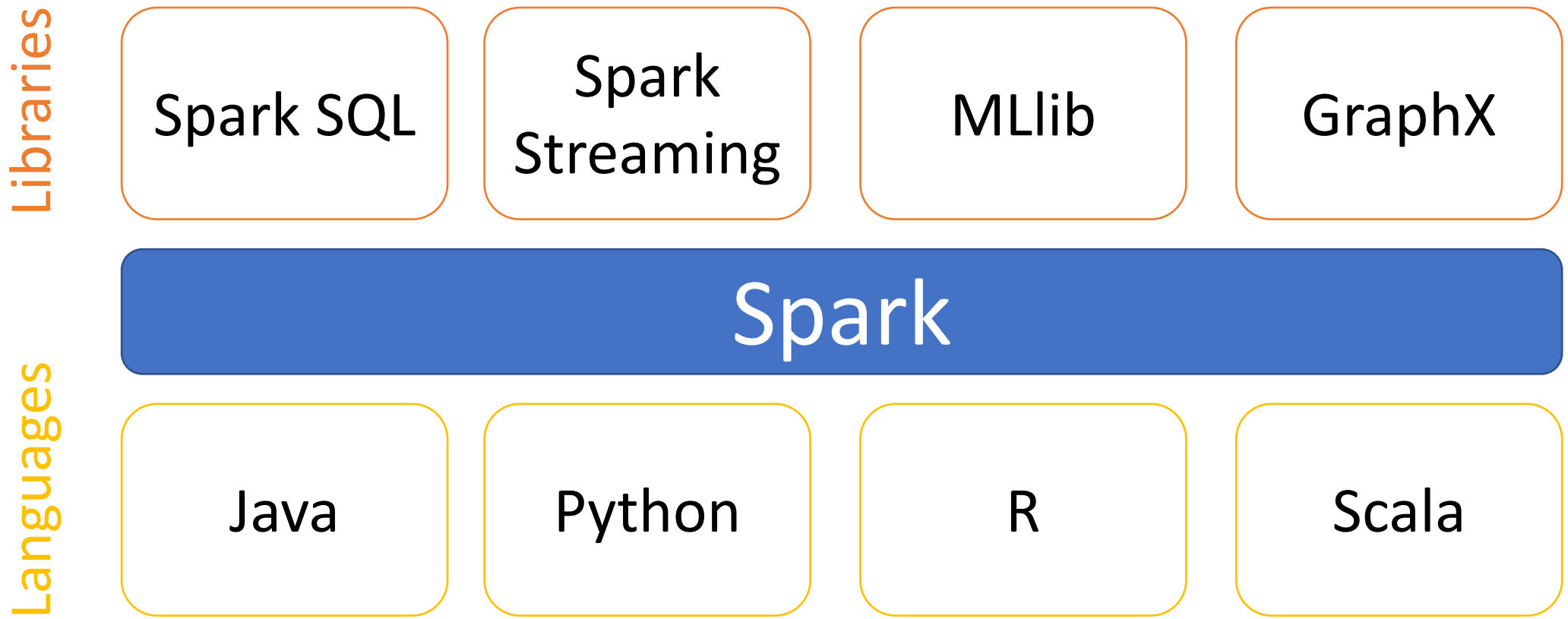
reduce(K2, List[V2]) : List[K3,V3]

MapReduce: Word Count Example



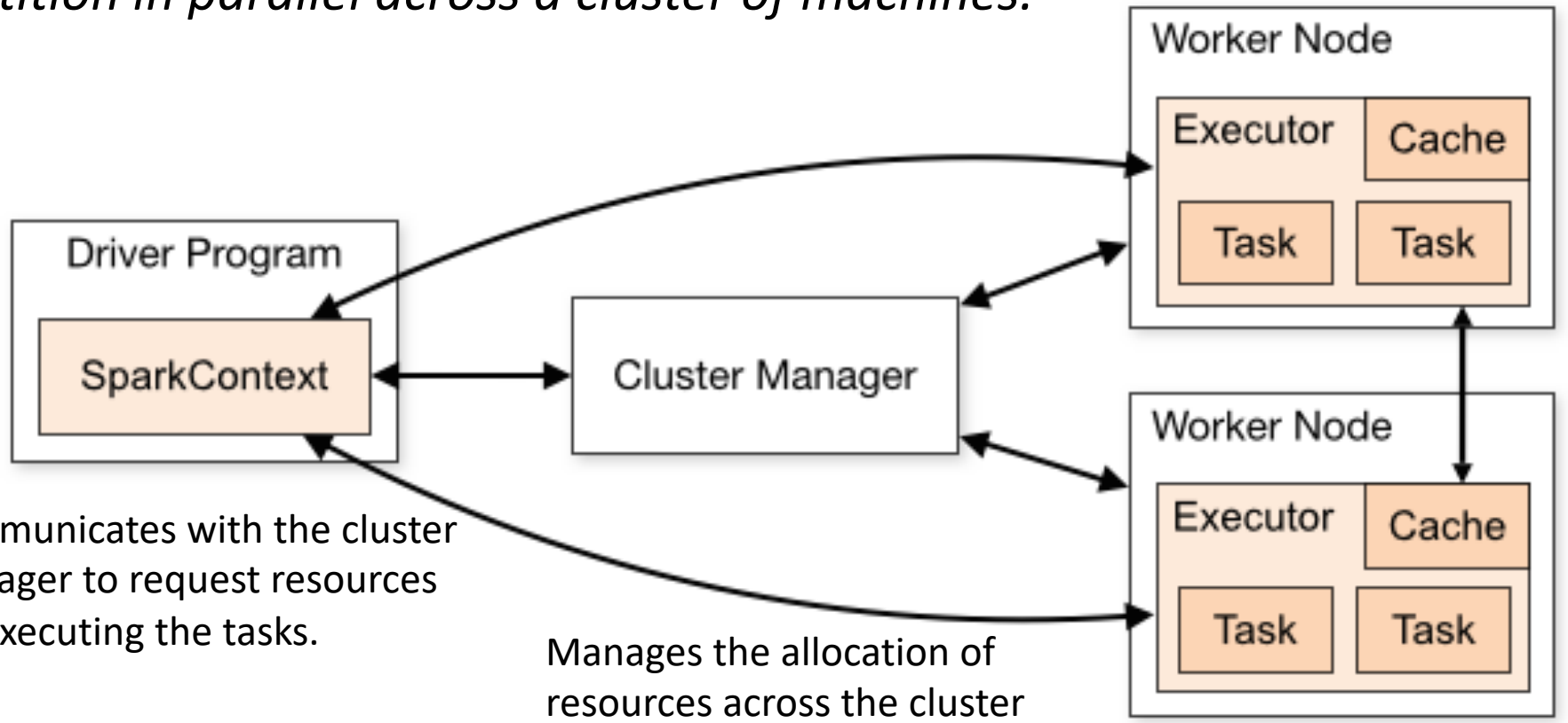
Spark

Unified analytics engine for large-scale data processing.



Spark Architecture

Works by dividing the input data into smaller partitions and processing each partition in parallel across a cluster of machines.



Machines in the cluster that perform the actual data processing operations.

Communicates with the cluster manager to request resources for executing the tasks.

Manages the allocation of resources across the cluster of machines.

Spark vs MapReduce

More efficient: 100x on smaller jobs to 3x on large jobs

- Caches data in RAM instead of disk
- Faster startup, better CPU utilization
- Richer functional programming
- Specially suited for iterative algorithms

Spark Common Use Cases

- SQL Batch Jobs Across Large Datasets
- Processing Streaming Data
- Machine Learning
- Web Analytics