Lecture 14

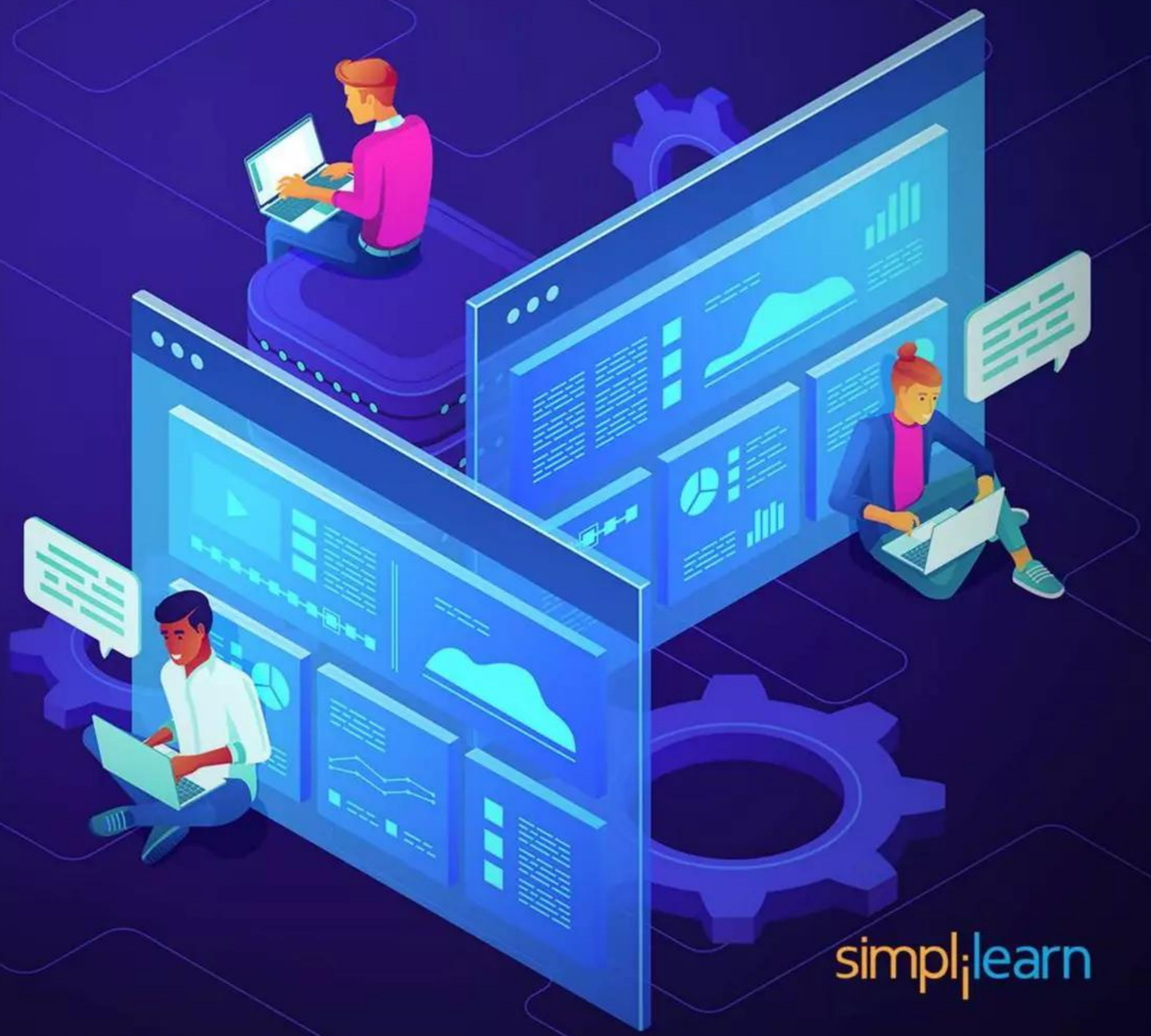
What is Apache Spark?

What's in it for you?

1. History of Spark

What's in it for you?

1. History of Spark
2. What is Spark?



What's in it for you?

1. History of Spark
2. What is Spark?
3. Hadoop vs Spark

What's in it for you?

1. History of Spark
2. What is Spark?
3. Hadoop vs Spark
4. Components of Apache Spark

Spark Core
Spark SQL
Spark Streaming
Spark MLlib
GraphX

What's in it for you?

1. History of Spark
2. What is Spark?
3. Hadoop vs Spark
4. Components of Apache Spark
5. Spark Architecture

What's in it for you?

1. History of Spark
2. What is Spark?
3. Hadoop vs Spark
4. Components of Apache Spark
5. Spark Architecture
6. Applications of Spark

What's in it for you?

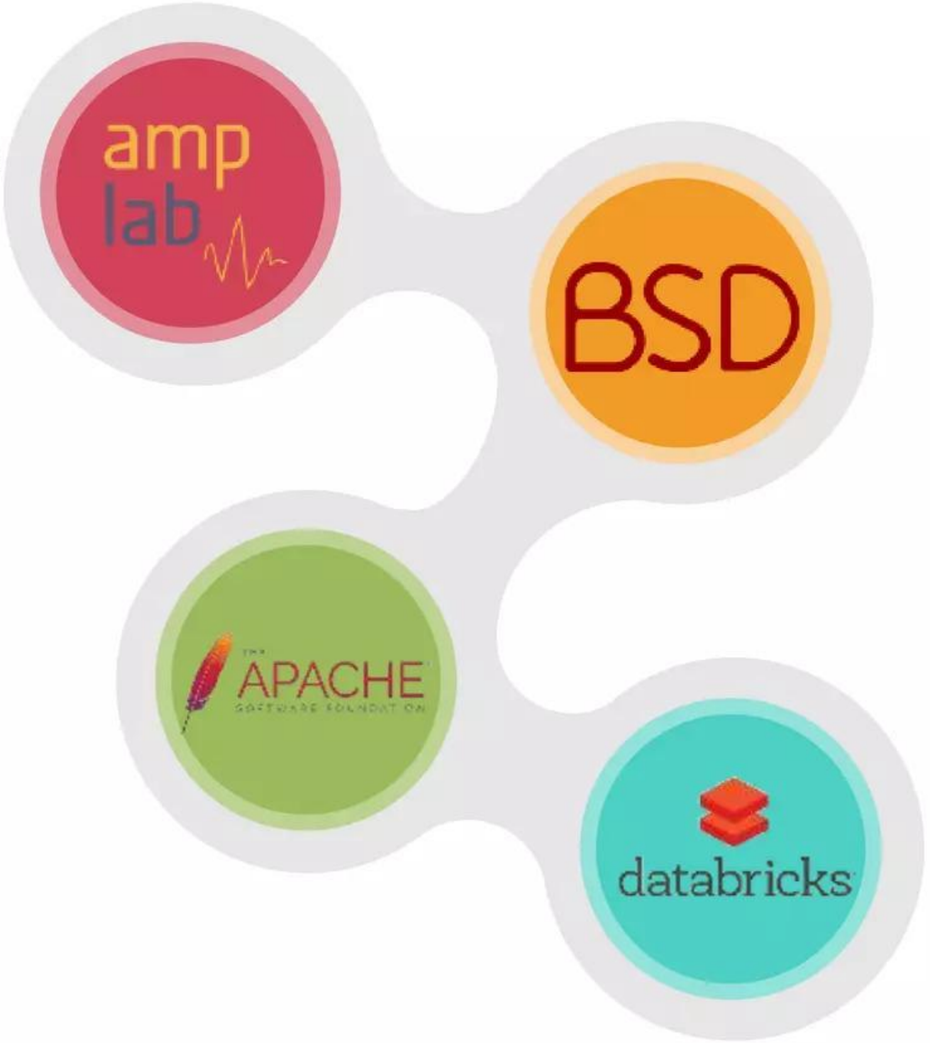
1. History of Spark
2. What is Spark?
3. Hadoop vs Spark
4. Components of Apache Spark
5. Spark Architecture
6. Applications of Spark
7. Spark Use Case

simplilearn

History of Apache Spark

2009

Started as a project at UC Berkley AMPLab

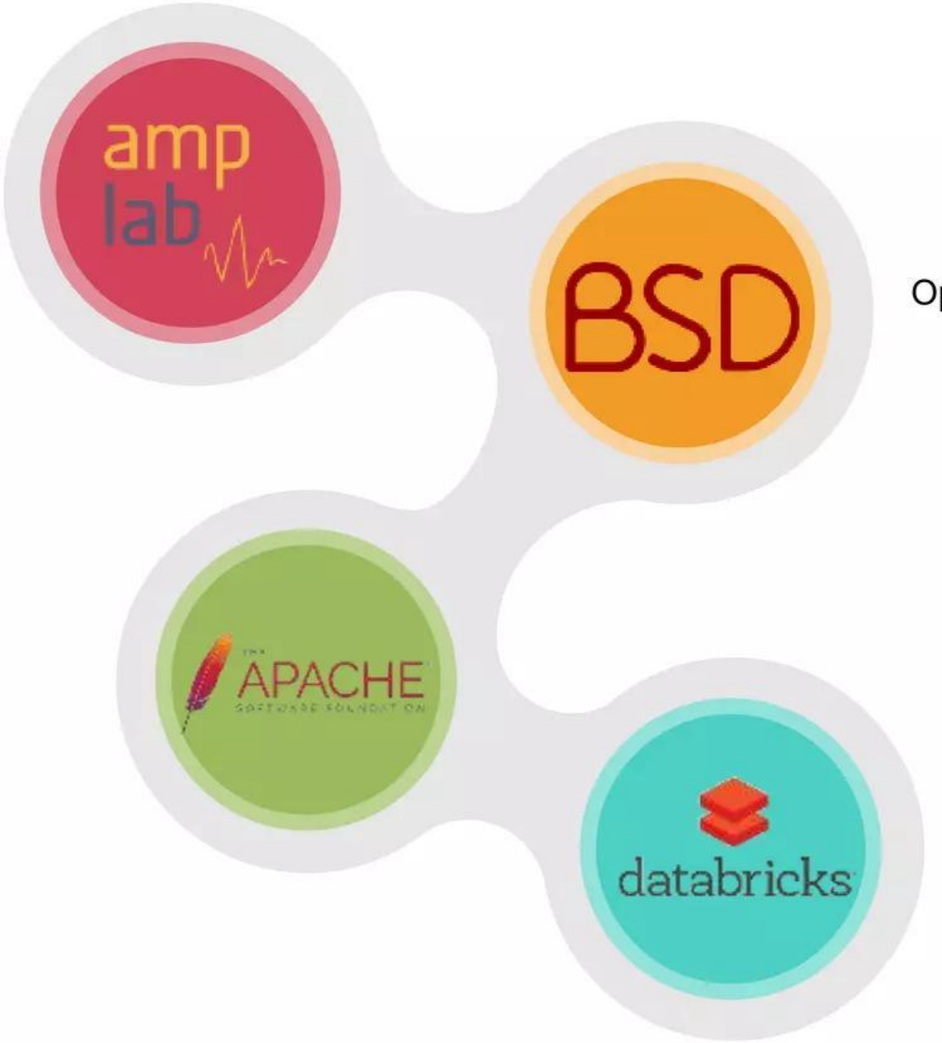


simplilearn

History of Apache Spark

2009

Started as a project at UC Berkley AMPLab



2010

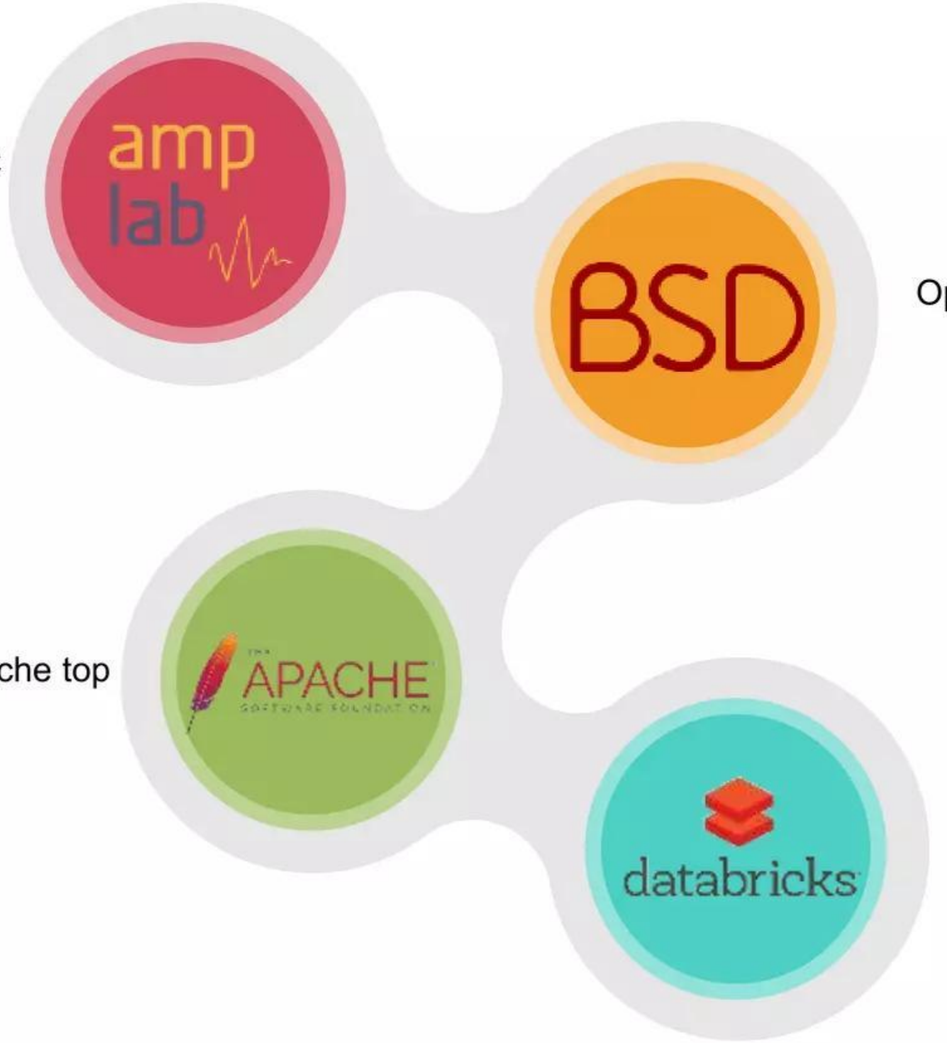
Open sourced under a BSD license

simplilearn

History of Apache Spark

2009

Started as a project at UC Berkley AMPLab



2010

Open sourced under a BSD license

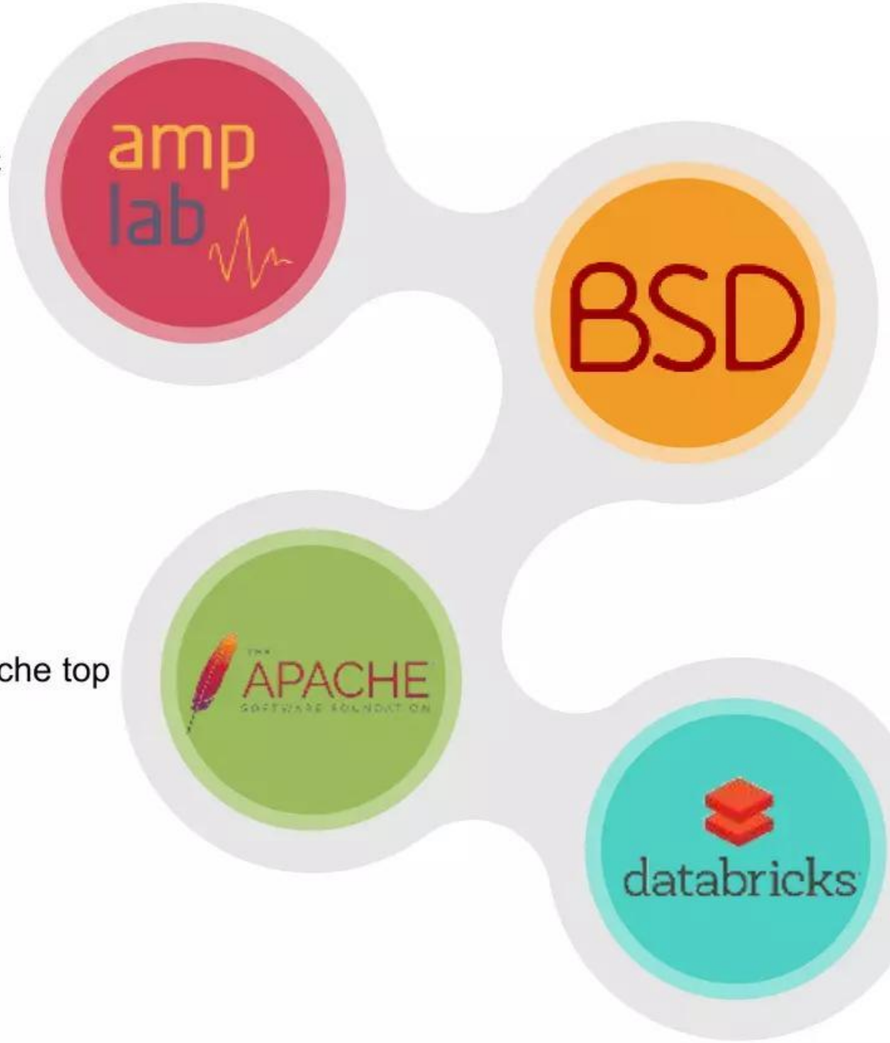
2013

Spark became an Apache top level project

History of Apache Spark

2009

Started as a project at UC Berkley AMPLab



2010

Open sourced under a BSD license

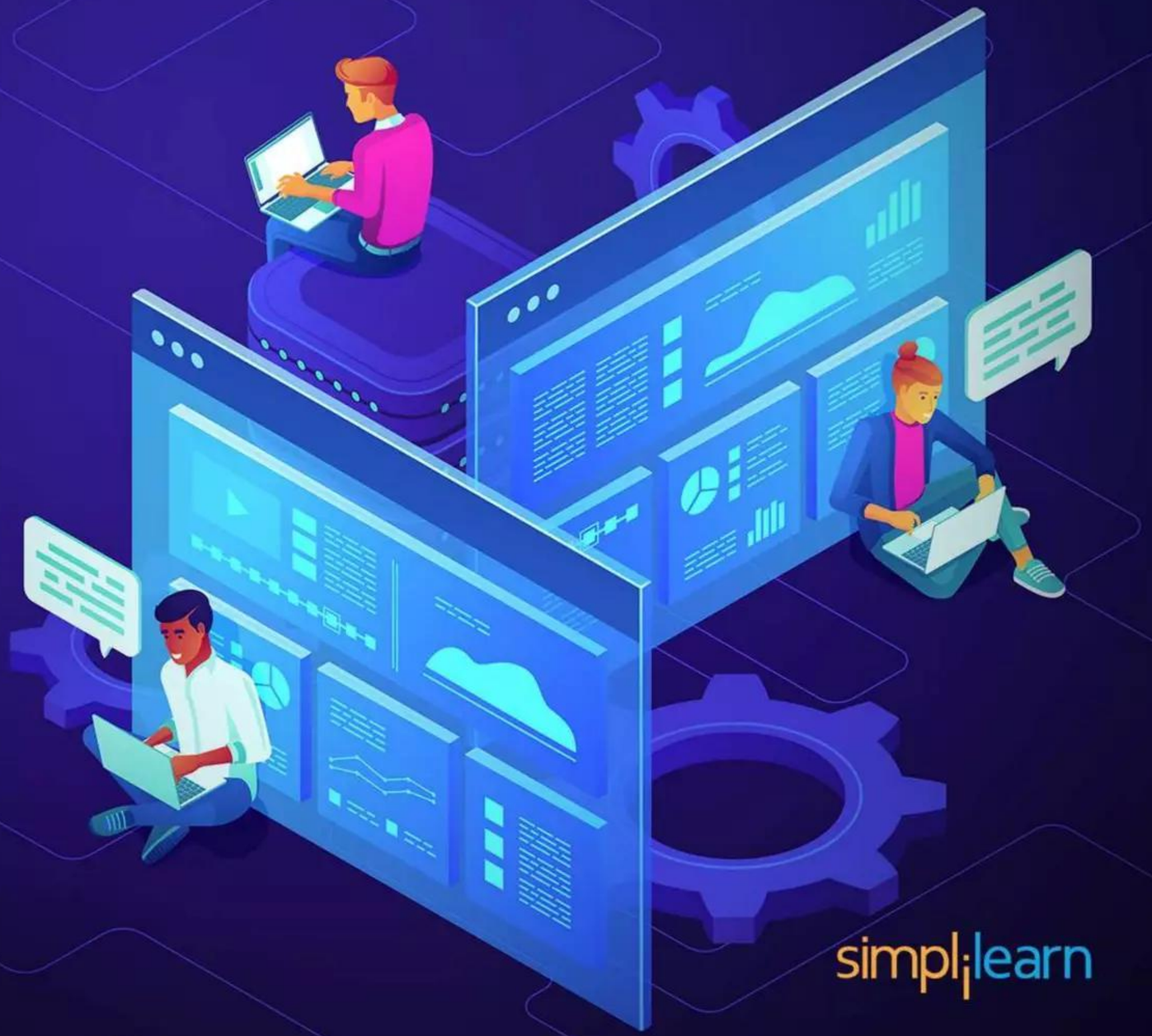
2013

Spark became an Apache top level project

2014

Used by Databricks to sort large-scale datasets and set a new world record

What is Apache Spark?



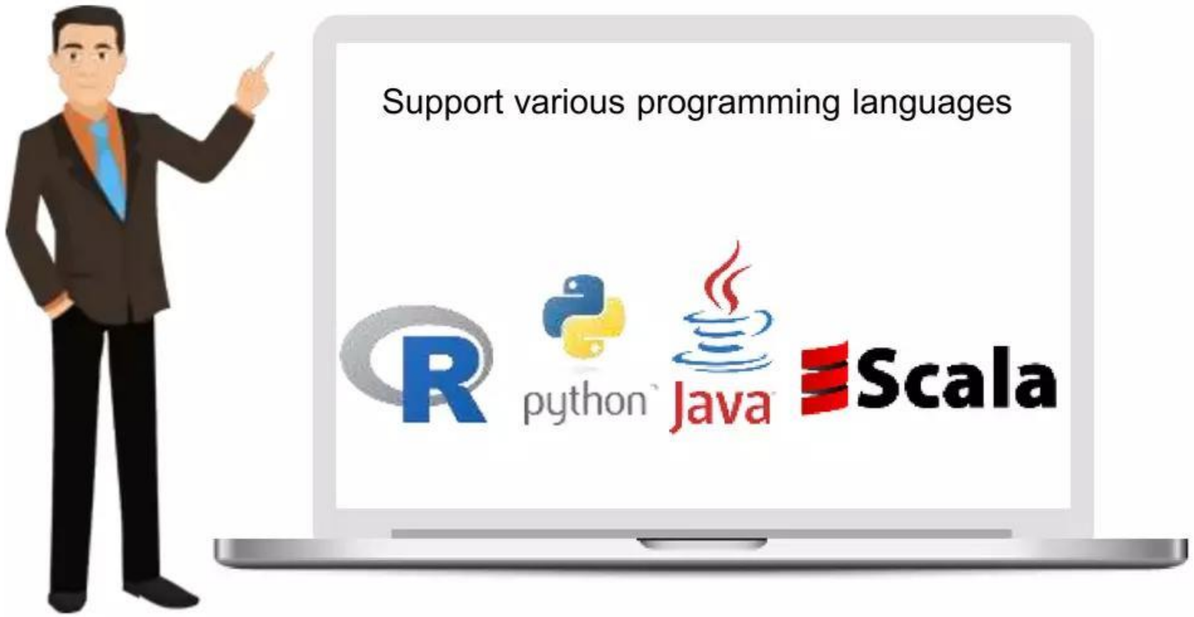
simplilearn

What is Apache Spark?

Apache Spark is an open-source data processing engine to store and process data in real-time across various clusters of computers using simple programming constructs

What is Apache Spark?

Apache Spark is an open-source data processing engine to store and process data in real-time across various clusters of computers using simple programming constructs



What is Apache Spark?

Apache Spark is an open-source data processing engine to store and process data in real-time across various clusters of computers using simple programming constructs

Hadoop vs Spark

simplilearn

Hadoop vs Spark

Processing data using MapReduce in Hadoop is slow

Spark processes data 100 times faster than MapReduce as it is done in-memory

Hadoop vs Spark

Hadoop	Spark
Processing data using MapReduce in Hadoop is slow	Spark processes data 100 times faster than MapReduce as it is done in-memory
Performs batch processing of data	Performs both batch processing and real-time processing of data

simplilearn

Hadoop vs Spark

Hadoop	Spark
Processing data using MapReduce in Hadoop is slow	Spark processes data 100 times faster than MapReduce as it is done in-memory
Performs batch processing of data	Performs both batch processing and real-time processing of data
Hadoop has more lines of code. Since it is written in Java, it takes more time to execute	Spark has fewer lines of code as it is implemented in Scala

simplilearn

Hadoop vs Spark

Hadoop	Spark
Processing data using MapReduce in Hadoop is slow	Spark processes data 100 times faster than MapReduce as it is done in-memory
Performs batch processing of data	Performs both batch processing and real-time processing of data
Hadoop has more lines of code. Since it is written in Java, it takes more time to execute	Spark has fewer lines of code as it is implemented in Scala
Hadoop supports Kerberos authentication, which is difficult to manage	Spark supports authentication via a shared secret. It can also run on YARN leveraging the capability of Kerberos

Spark Features

simplilearn

Spark Features

Fast processing

Spark contains Resilient Distributed Datasets (RDD) which saves time taken in reading, and writing operations and hence, it runs almost ten to hundred times faster than Hadoop

Spark Features

Fast processing

In-memory computing

In Spark, data is stored in the RAM, so it can access the data quickly and accelerate the speed of analytics

Spark Features

Fast processing

In-memory computing

Flexible

Spark supports multiple languages and allows the developers to write applications in Java, Scala, R, or Python

Spark Features

Fast processing

In-memory computing

Flexible

Fault tolerance

Spark contains Resilient Distributed Datasets (RDD) that are designed to handle the failure of any worker node in the cluster. Thus, it ensures that the loss of data reduces to zero

Spark Features

Fast processing

In-memory computing

Flexible

Fault tolerance

Better analytics

Spark has a rich set of SQL queries, machine learning algorithms, complex analytics, etc. With all these functionalities, analytics can be performed better

simplilearn

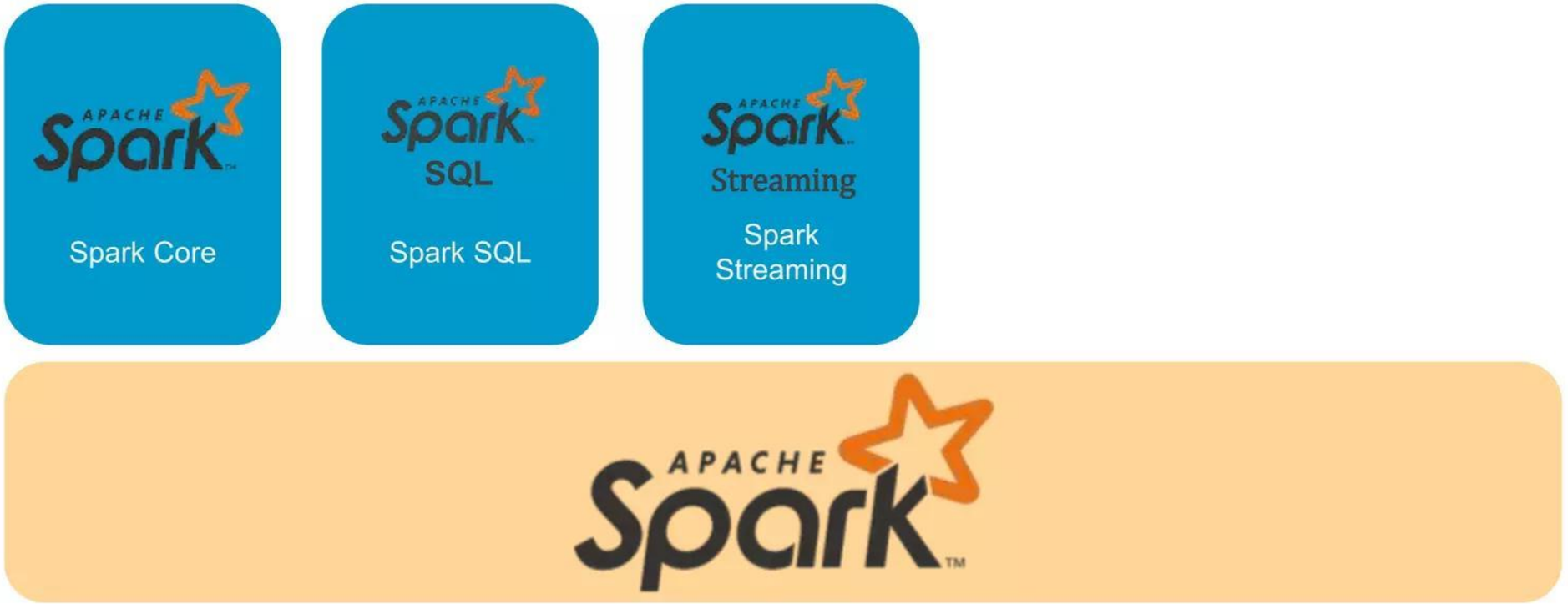
Components of Spark

simplilearn

Components of Apache Spark

Components of Apache Spark

Components of Apache Spark



Components of Apache Spark

Components of Apache Spark

Components of Spark – Spark Core

simplilearn

Spark Core

Spark Core is the base engine for large-scale parallel and distributed data processing

simplilearn

Spark Core

Spark Core is the base engine for large-scale parallel and distributed data processing

Spark Core

It is responsible for:

memory management

fault recovery

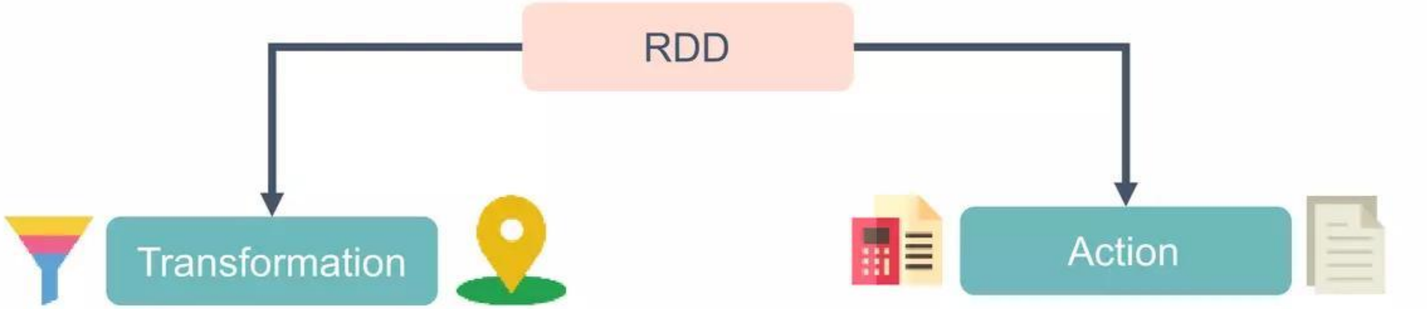
scheduling, distributing and monitoring jobs on a cluster

interacting with storage systems

Resilient Distributed Dataset

Spark Core is embedded with **RDDs** (Resilient Distributed Datasets), an immutable fault-tolerant, distributed collection of objects that can be operated on in parallel

Spark Core



These are operations (such as map, filter, join, union) that are performed on an RDD that yields a new RDD containing the result

These are operations (such as reduce, first, count) that return a value after running a computation on an RDD

Components of Spark – Spark SQL

simplilearn

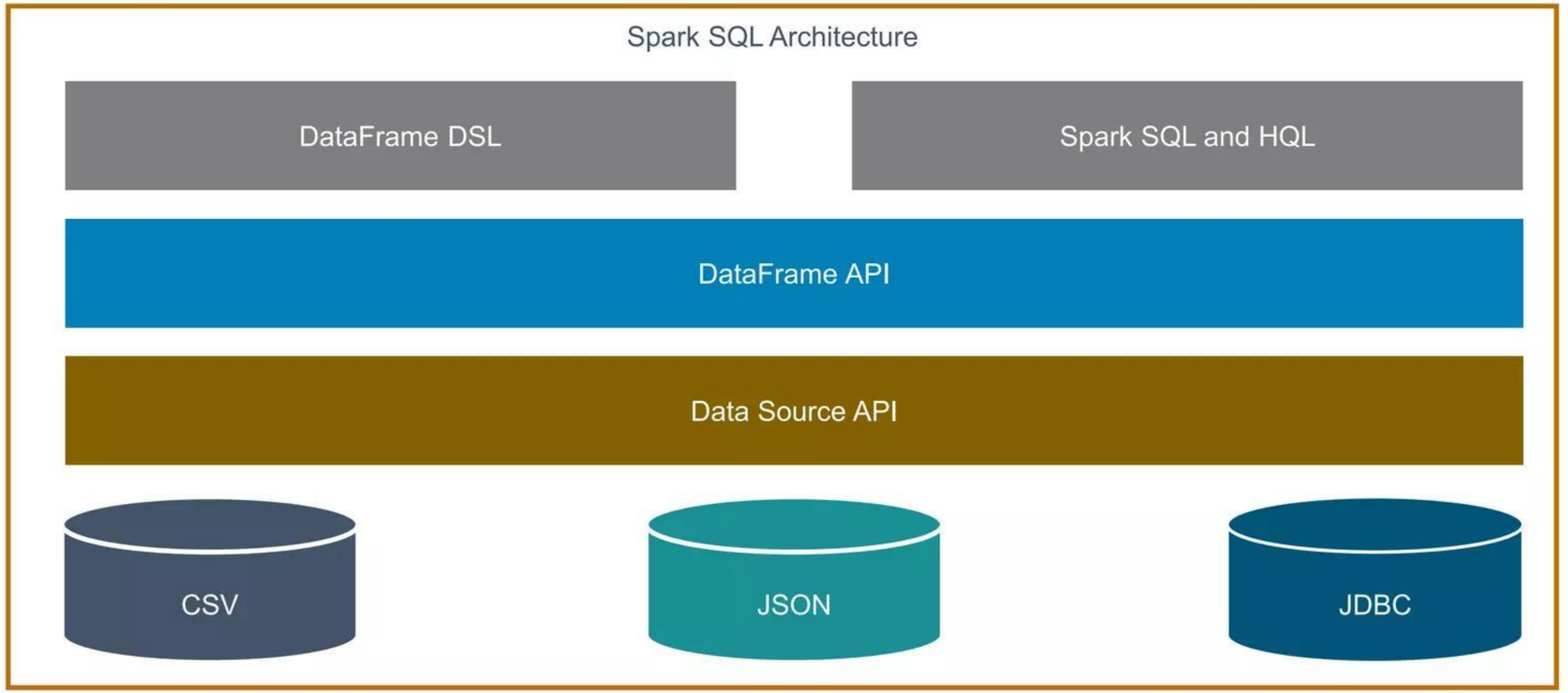
Spark SQL

Spark SQL framework component is used for structured and semi-structured data processing

simplilearn

Spark SQL

Spark SQL framework component is used for structured and semi-structured data processing



Spark SQL

Components of Spark – Spark Streaming

simplilearn

Spark Streaming

Spark Streaming is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease

Provides secure, reliable, and fast processing of live data streams

simplilearn

Spark Streaming

Spark Streaming is a lightweight API that allows developers to perform batch processing and real-time streaming of data with ease

Provides secure, reliable, and fast processing of live data streams

simplilearn

Components of Spark – Spark MLlib

simplilearn

Spark MLlib

MLlib is a low-level machine learning library that is simple to use, is scalable, and compatible with various programming languages

MLlib eases the deployment and development of scalable machine learning algorithms

simplilearn

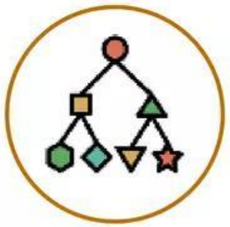
Spark MLlib

MLlib is a low-level machine learning library that is simple to use, is scalable, and compatible with various programming languages

MLlib eases the deployment and development of scalable machine learning algorithms

It contains machine learning libraries that have an implementation of various machine learning algorithms

Clustering



Classification

Collaborative Filtering

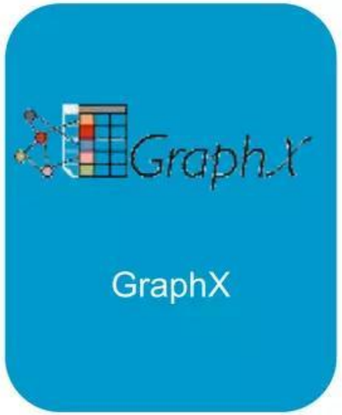
MLlib

Components of Spark – GraphX

simplilearn

GraphX

GraphX is Spark's own Graph Computation Engine and data store



GraphX

GraphX is Spark's own Graph Computation Engine and data store

GraphX

Provides a uniform tool for ETL

Exploratory data analysis

Interactive graph computations

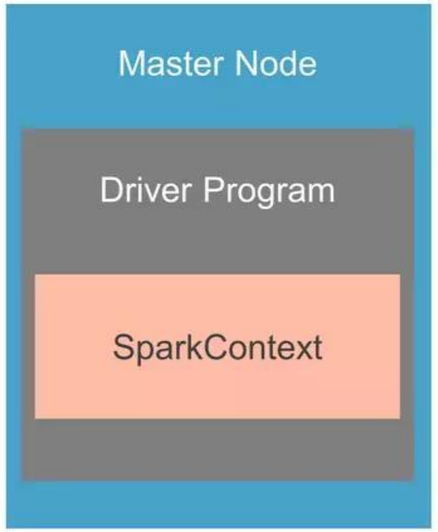
simplilearn

Spark Architecture

simplilearn

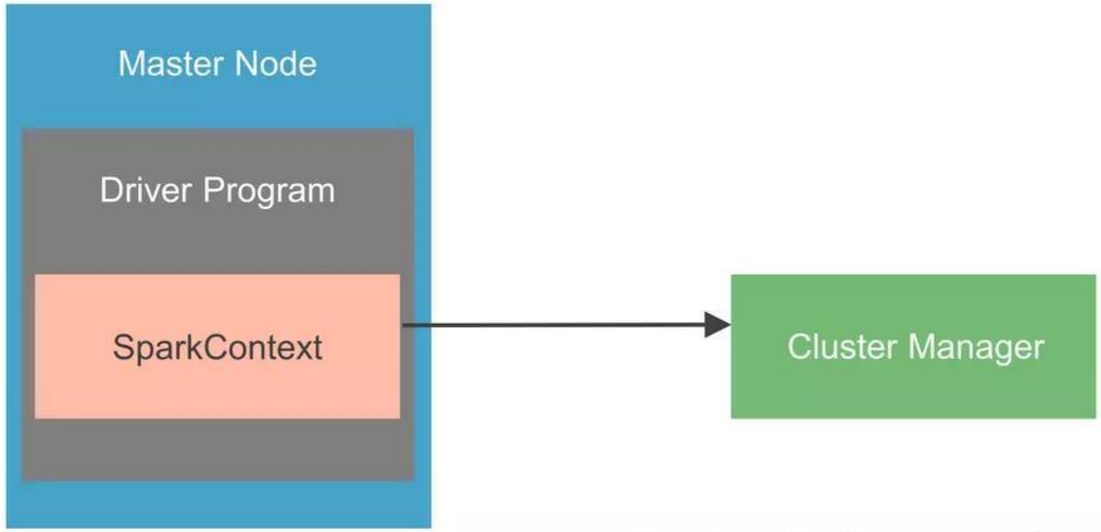
Spark Architecture

Apache Spark uses a master-slave architecture that consists of a driver, that runs on a master node, and multiple executors which run across the worker nodes in the cluster



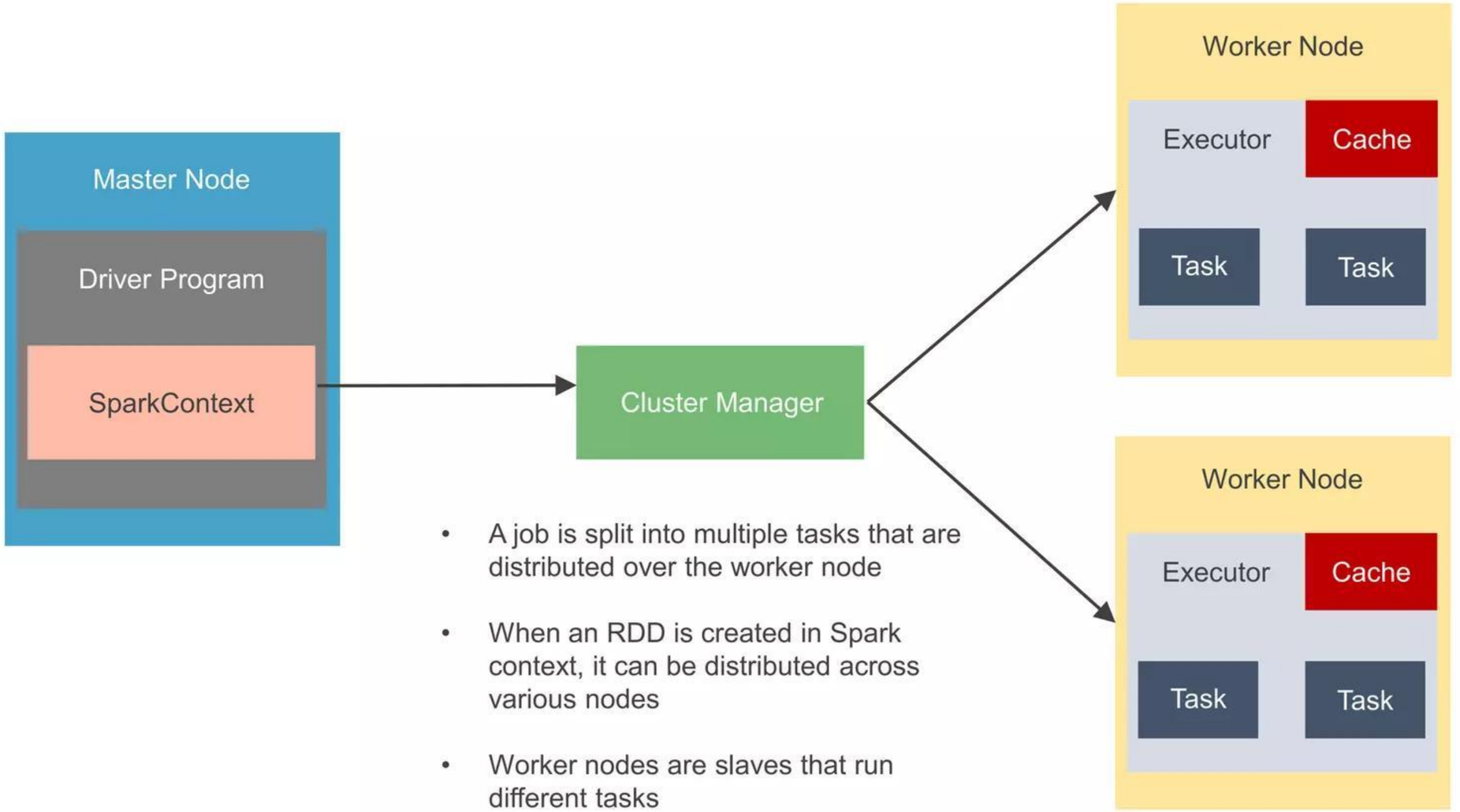
- Master Node has a Driver Program
- The Spark code behaves as a driver program and creates a SparkContext, which is a gateway to all the Spark functionalities

Spark Architecture



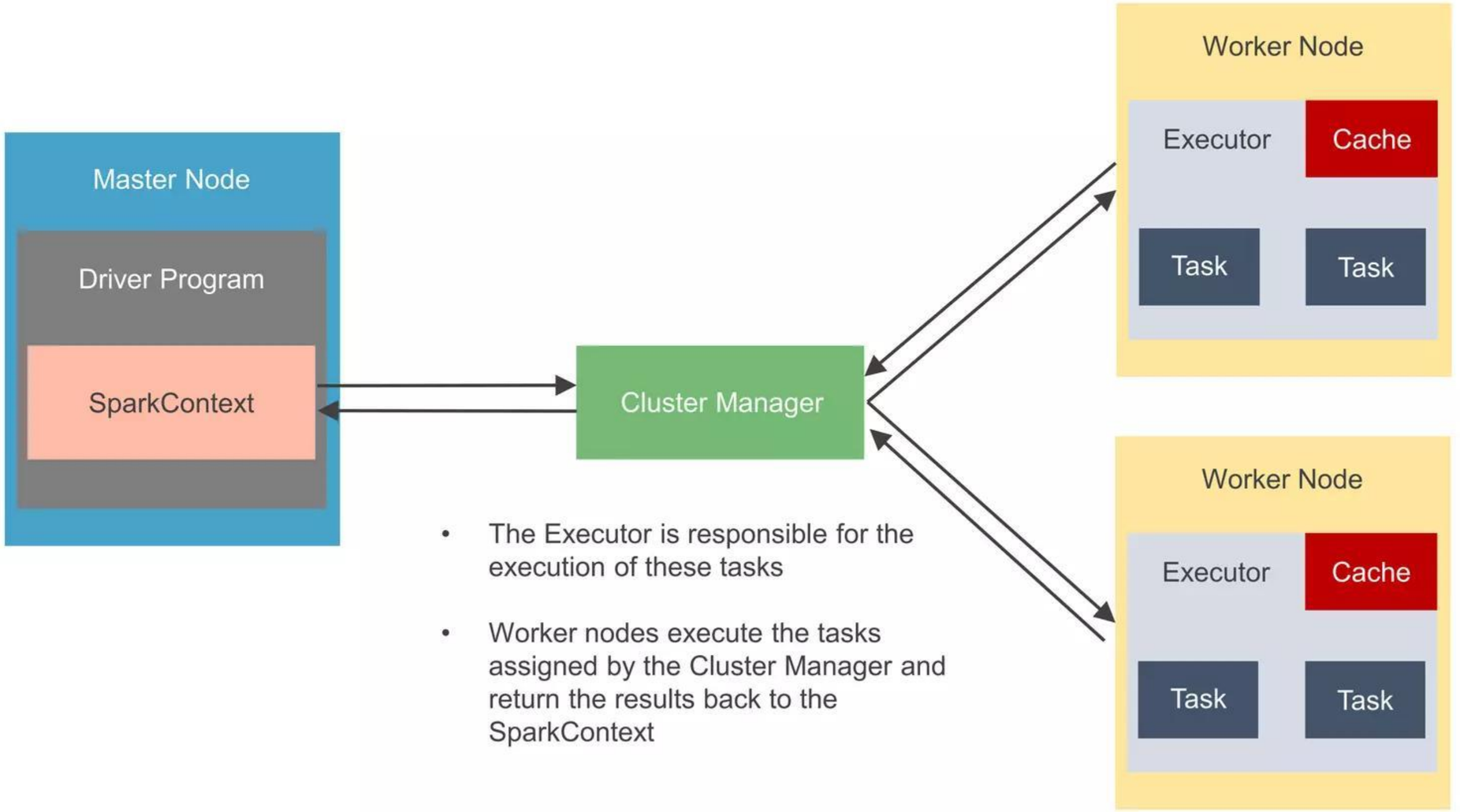
- Spark applications run as independent sets of processes on a cluster
- The driver program & Spark context takes care of the job execution within the cluster

Spark Architecture



- A job is split into multiple tasks that are distributed over the worker node
- When an RDD is created in Spark context, it can be distributed across various nodes
- Worker nodes are slaves that run different tasks

Spark Architecture



- The Executor is responsible for the execution of these tasks
- Worker nodes execute the tasks assigned by the Cluster Manager and return the results back to the SparkContext

Spark Cluster Managers

Standalone mode

1

By default, applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes

Spark Cluster Managers

Standalone mode

1

By default, applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes

2

Apache Mesos is an open-source project to manage computer clusters, and can also run Hadoop applications

simplilearn

Spark Cluster Managers

Standalone mode

1

By default, applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes

2

Apache Mesos is an open-source project to manage computer clusters, and can also run Hadoop applications

3

Apache YARN is the cluster resource manager of Hadoop 2. Spark can be run on YARN

simplilearn

Spark Cluster Managers

Standalone mode

1

By default, applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes

2

Apache Mesos is an open-source project to manage computer clusters, and can also run Hadoop applications

3

Apache YARN is the cluster resource manager of Hadoop 2. Spark can be run on YARN

4

Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications

simplilearn

Applications of Spark

simplilearn

Applications of Spark

JPMorgan Chase & Co.

JPMorgan uses Spark to detect fraudulent transactions, analyze the business spends of an individual to suggest offers, and identify patterns to decide how much to invest and where to invest

Banking

Applications of Spark

JPMorgan Chase & Co.

JPMorgan uses Spark to detect fraudulent transactions, analyze the business spends of an individual to suggest offers, and identify patterns to decide how much to invest and where to invest

Banking

Alibaba Group

Alibaba uses Spark to analyze large sets of data such as real-time transaction details, browsing history, etc. in the form of Spark jobs and provides recommendations to its users

E-Commerce

Applications of Spark

JPMorgan Chase & Co.

JPMorgan uses Spark to detect fraudulent transactions, analyze the business spends of an individual to suggest offers, and identify patterns to decide how much to invest and where to invest

Banking

Alibaba Group

Alibaba uses Spark to analyze large sets of data such as real-time transaction details, browsing history, etc. in the form of Spark jobs and provides recommendations to its users

E-Commerce

IQVIA™

IQVIA is a leading healthcare company that uses Spark to analyze patient's data, identify possible health issues, and diagnose it based on their medical history

Healthcare

Applications of Spark

JPMorgan Chase & Co.

JPMorgan uses Spark to detect fraudulent transactions, analyze the business spends of an individual to suggest offers, and identify patterns to decide how much to invest and where to invest

Banking

Alibaba Group

Alibaba uses Spark to analyze large sets of data such as real-time transaction details, browsing history, etc. in the form of Spark jobs and provides recommendations to its users

E-Commerce

IQVIA

IQVIA is a leading healthcare company that uses Spark to analyze patient's data, identify possible health issues, and diagnose it based on their medical history

Healthcare

NETFLIX

RIOT GAMES

Entertainment and gaming companies like Netflix and Riot games use Apache Spark to showcase relevant advertisements to their users based on the videos that they watch, share, and like

Entertainment

simplilearn

Spark Use Case

simplilearn

Spark Use Case

Conviva is one of the world's leading video streaming companies

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Video streaming is a challenge, especially with increasing demand for high-quality streaming experiences

simplilearn

Spark Use Case

conviva®

Conviva is one of the world's leading video streaming companies

Video streaming is a challenge, especially with increasing demand for high-quality streaming experiences

Conviva collects data about video streaming quality to give their customers visibility into the end-user experience they are delivering

simplilearn

Spark Use Case

conviva®

Conviva is one of the world's leading video streaming companies

APACHE Spark™

Using Apache Spark, Conviva delivers a better quality of service to its customers by removing the screen buffering and learning in detail about the network conditions in real-time

simplilearn

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva delivers a better quality of service to its customers by removing the screen buffering and learning in detail about the network conditions in real-time

This information is stored in the video player to manage live video traffic coming from 4 billion video feeds every month, to ensure maximum retention

simplilearn

Spark Use Case

conviva®

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva has created an auto diagnostics alert

simplilearn

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva has created an auto diagnostics alert

It automatically detects anomalies along the video streaming pipeline and diagnoses the root cause of the issue

simplilearn

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva has created an auto diagnostics alert

It automatically detects anomalies along the video streaming pipeline and diagnoses the root cause of the issue

Reduces waiting time before the video starts

simplilearn

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva has created an auto diagnostics alert

It automatically detects anomalies along the video streaming pipeline and diagnoses the root cause of the issue

Reduces waiting time before the video starts

Avoids buffering and recovers the video from a technical error

simplilearn

Spark Use Case

conviva

Conviva is one of the world's leading video streaming companies

Using Apache Spark, Conviva has created an auto diagnostics alert

It automatically detects anomalies along the video streaming pipeline and diagnoses the root cause of the issue

Reduces waiting time before the video starts

Avoids buffering and recovers the video from a technical error

Goal is to maximize the viewer engagement

simplilearn