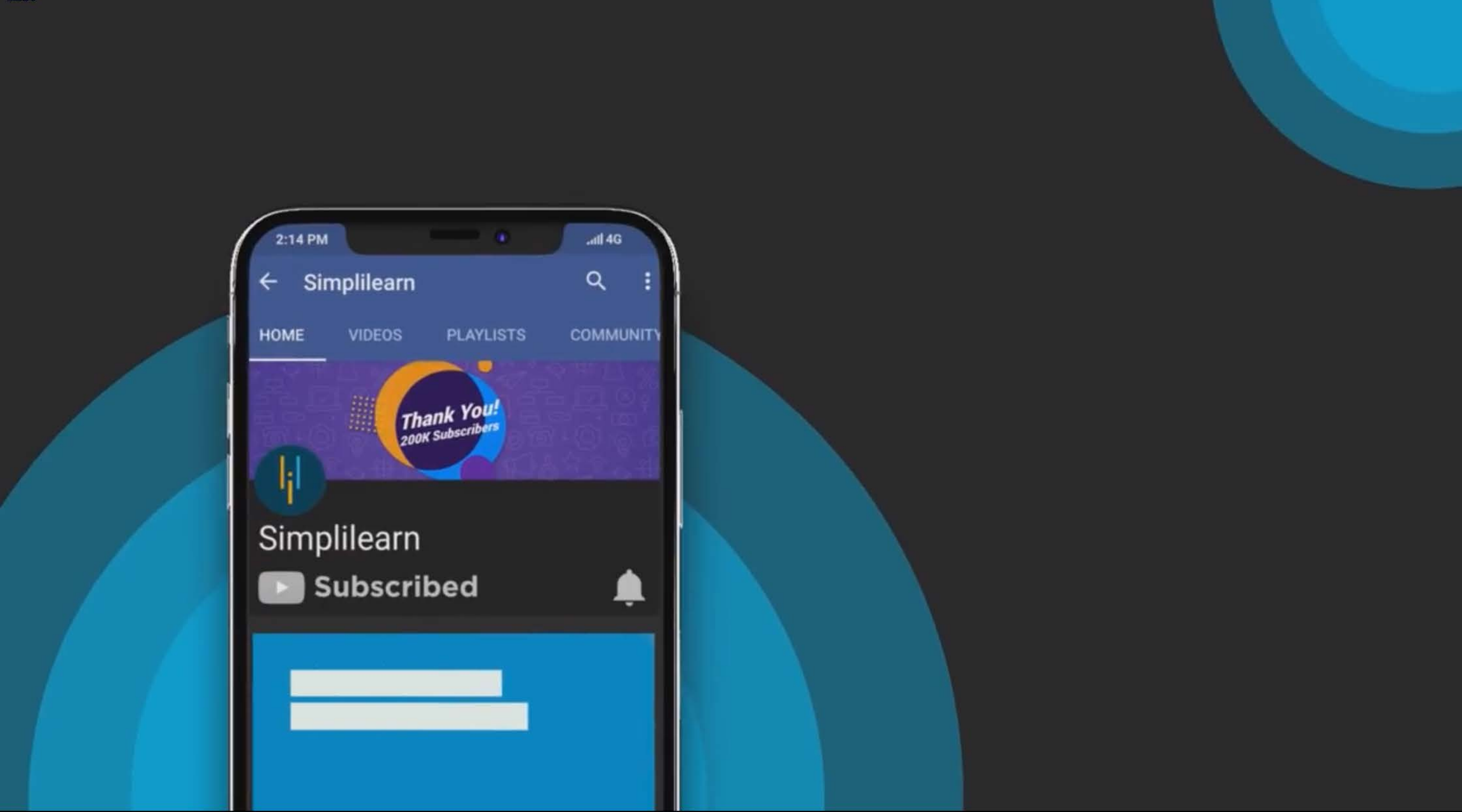
全屏


2:14 PM
4G
Simplilearn
HOME
VIDEOS
PLAYLISTS
COMMUNITY
Thank You!
200K Subscribers
Simplilearn
Subscribed

搜索：步退 5 秒(关键帧) / 00:00:15(0%)

# What is Apache Spark?

APACHE Spark

simplilearn

搜索：步进 5 秒(关键帧) / 00:00:41(1%)

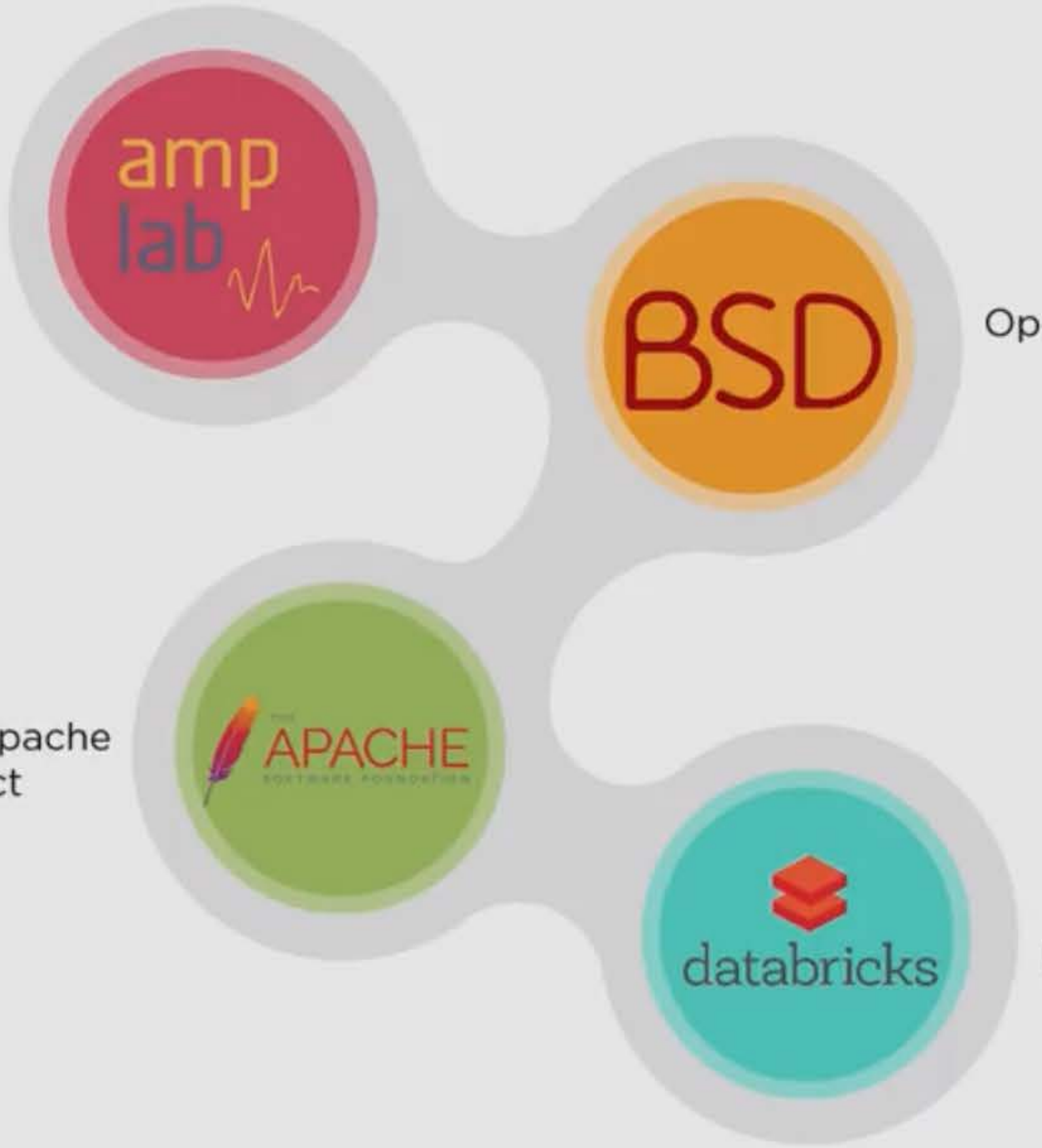
# What's in it for you?

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

搜索：步进 5 秒(关键帧) / 00:01:06(2%)

# 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?

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

Support various programming languages

Developers and data scientists incorporate Spark into their applications to rapidly query, analyze, and transform data at scale

simplilearn

搜索：步进 5 秒(关键帧) / 00:02:28(6%)

# Hadoop vs Spark

simplilearn

无字幕。

# Hadoop vs Spark

| hadoop | APACHE 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

APACHE Spark

Fast processing

In-memory computing

Flexible

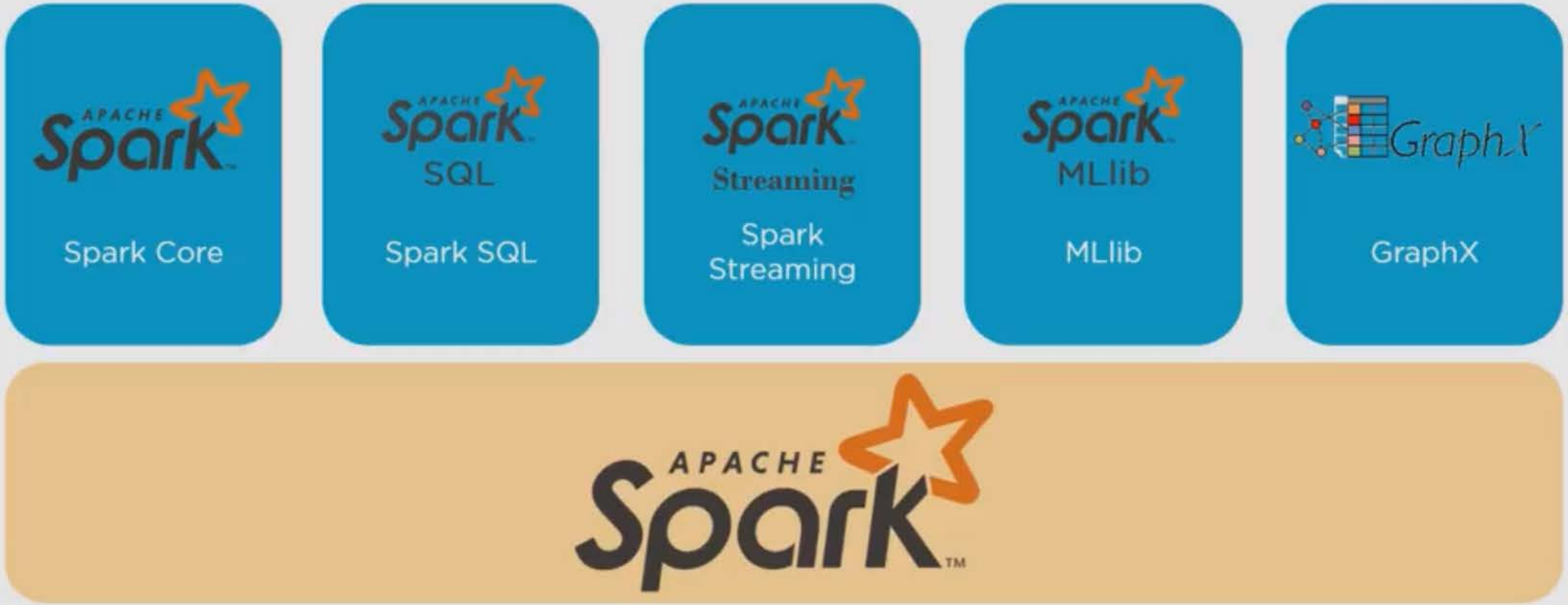
Fault tolerance

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

# Components of Apache Spark

# Components of Apache Spark



Spark Core

Spark SQL

Spark Streaming

MLlib

GraphX

搜索：步进 5 秒(关键帧) / 00:10:24(27%)

# Components of Spark – Spark Core

simplilearn

# Spark Core

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

Spark Core

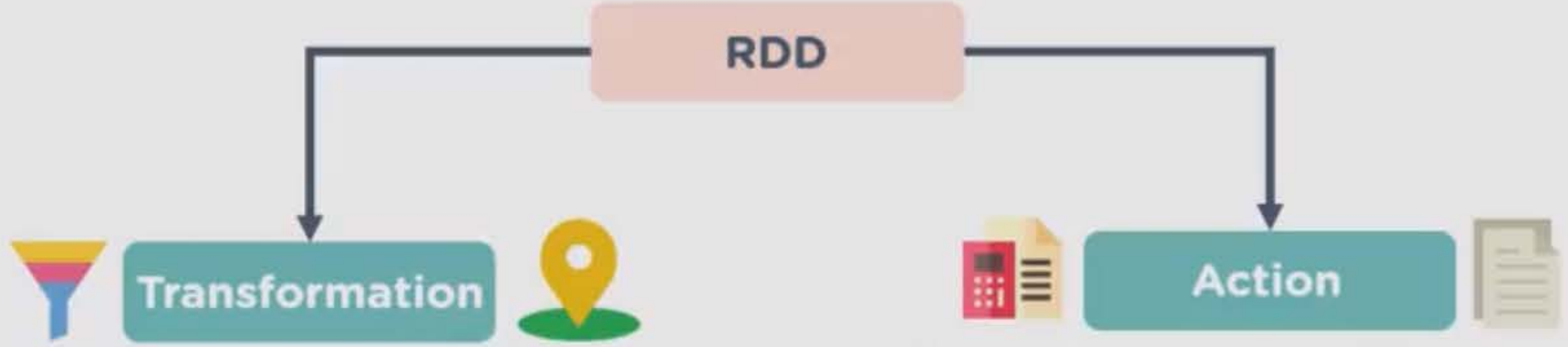
It is responsible for:

- 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



RDD

Transformation

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

Action

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

simplilearn

暂停

# Components of Spark – Spark SQL

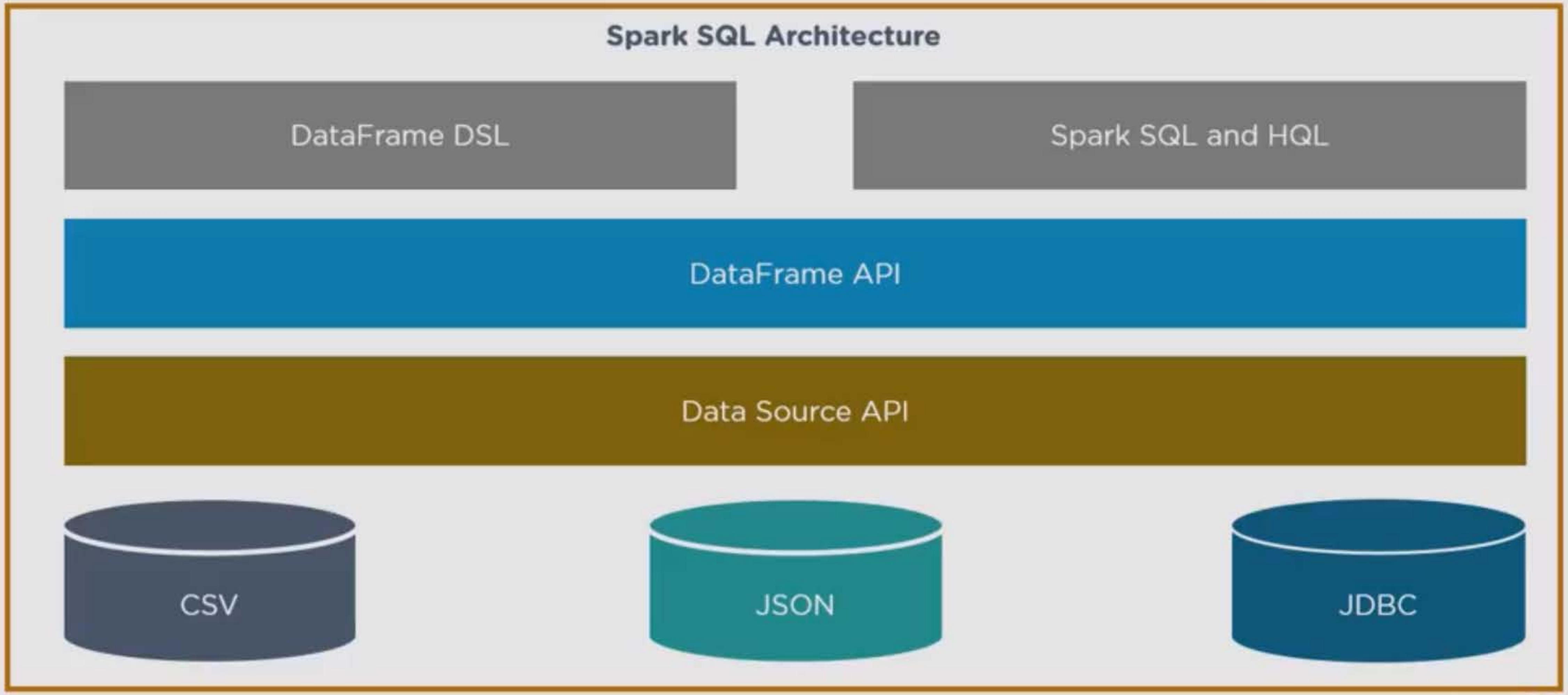
simplilearn

搜索：步进 5 秒(关键帧) / 00:19:01(49%)

# Spark SQL

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

Spark SQL



Spark SQL Architecture

DataFrame DSL

Spark SQL and HQL

DataFrame API

Data Source API

CSV

JSON

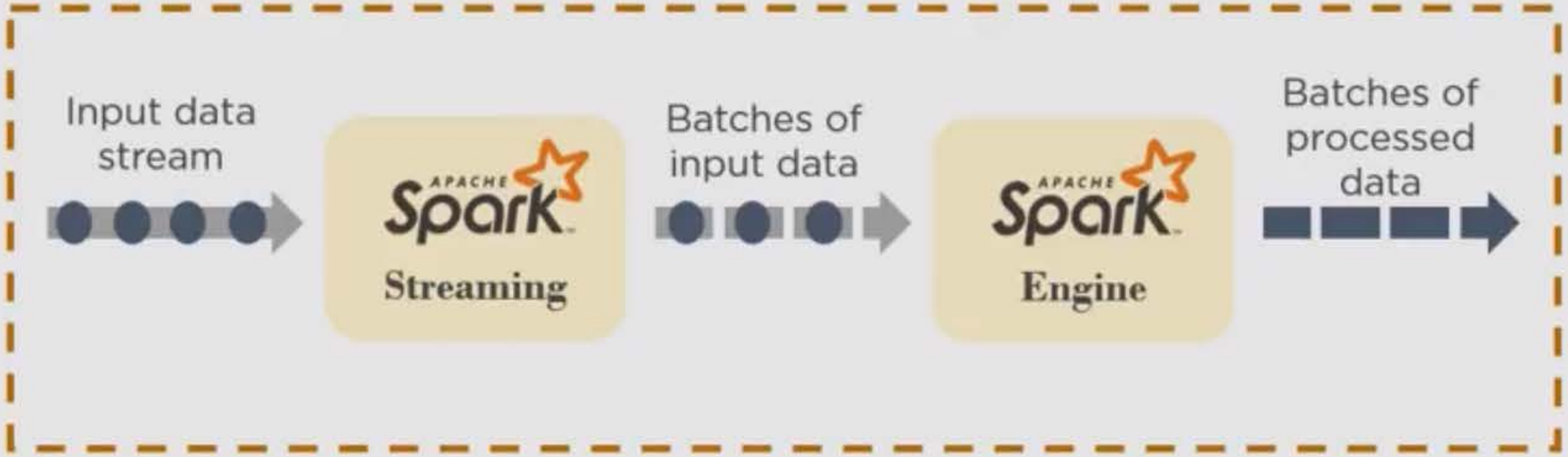
JDBC

# 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

Spark Streaming



搜索：步进 5 秒(关键帧) / 00:25:30(66%)

# 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

# GraphX

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

GraphX

Provides a uniform tool for ETL

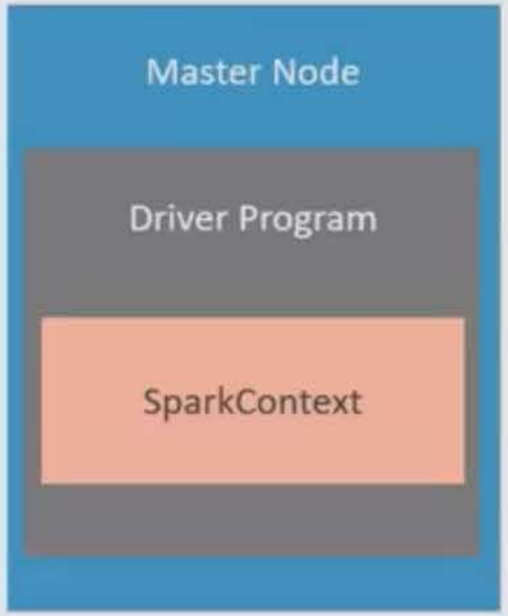
ETL
Extract
Transform
Load
DATA WAREHOUSE
DATA MART
HADOOP
FLAT FILES
XML

Exploratory data analysis

Interactive graph computations

# Spark Architecture

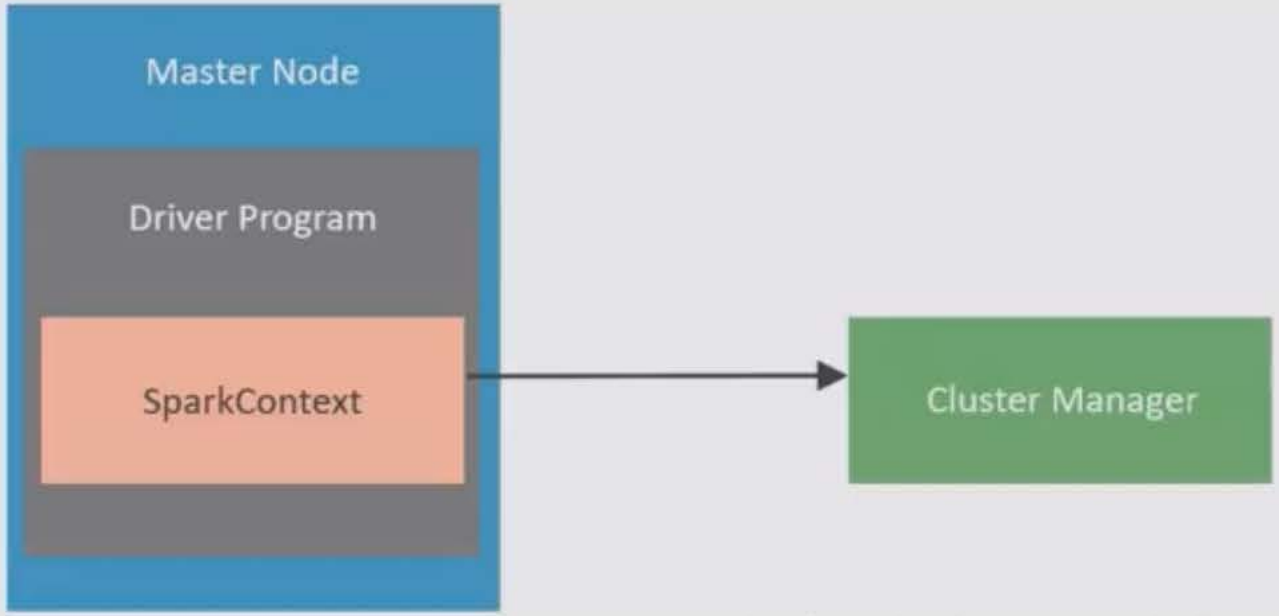
Apache Spark uses a master-slave architecture that consists of a driver, that runs on a master node, and multiple executors which run across the worker nodes in the cluster



- 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

暂停

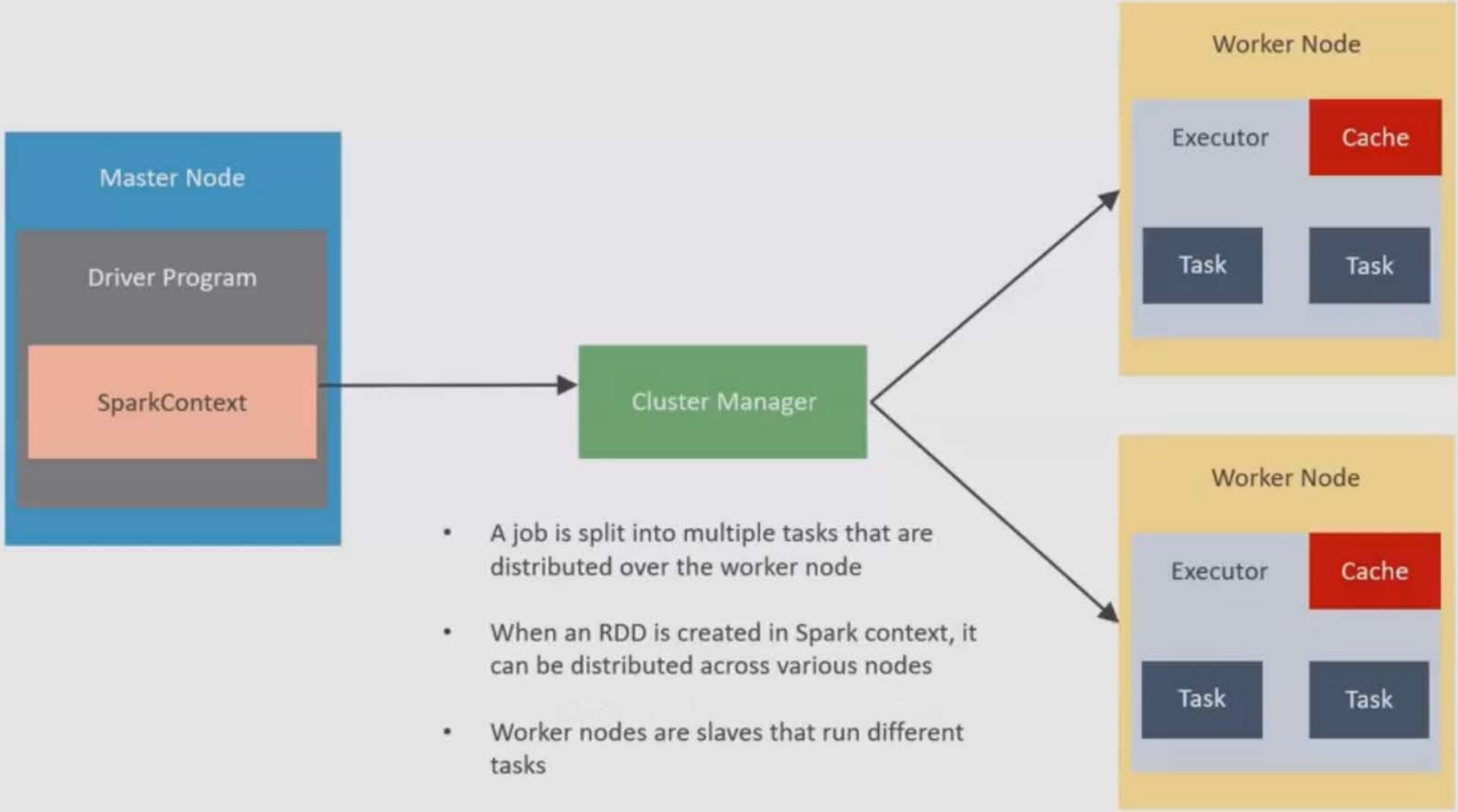
# Spark Architecture



- 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

暂停

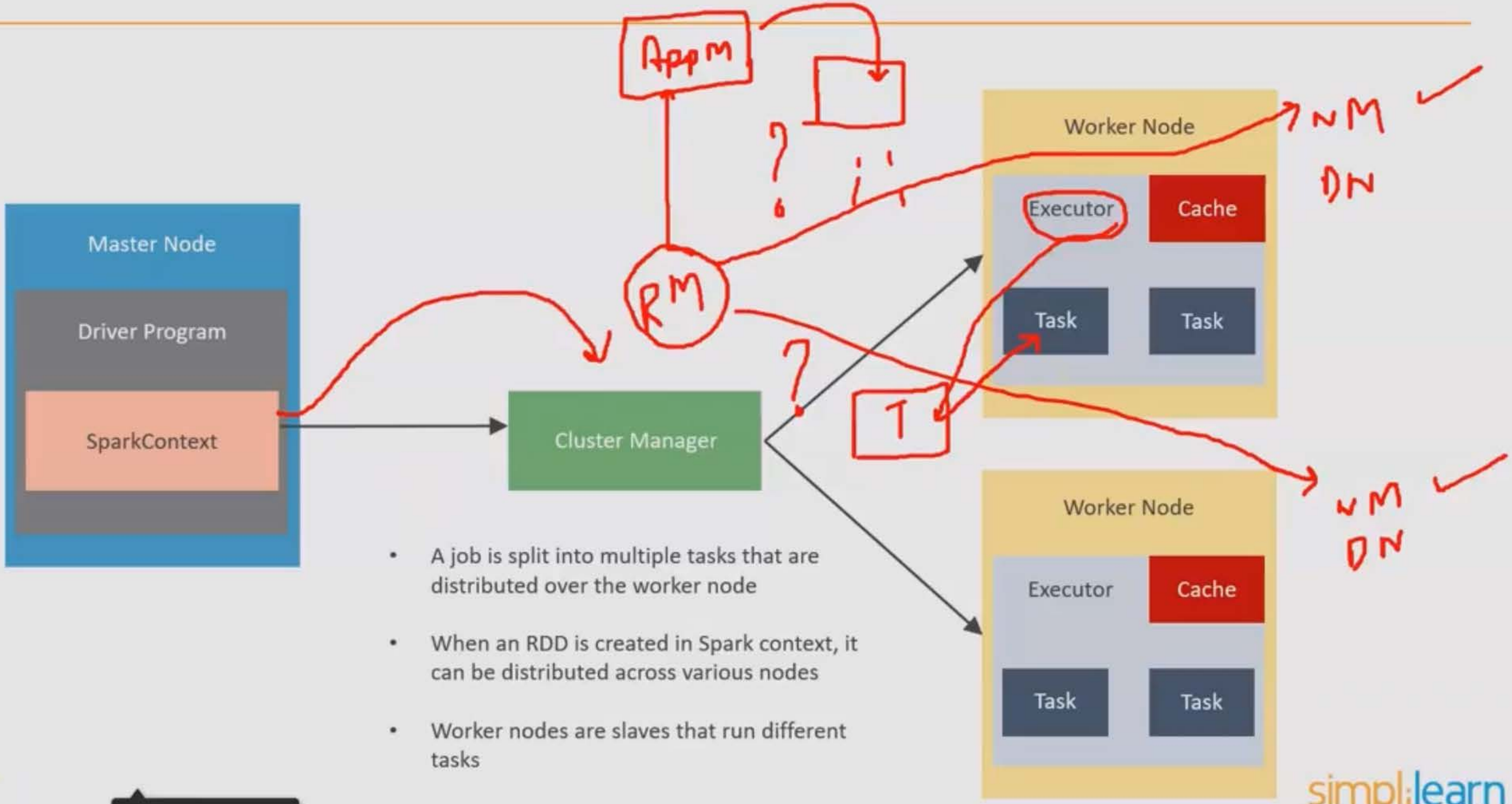
# Spark Architecture



- 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

暂停

# Spark Architecture

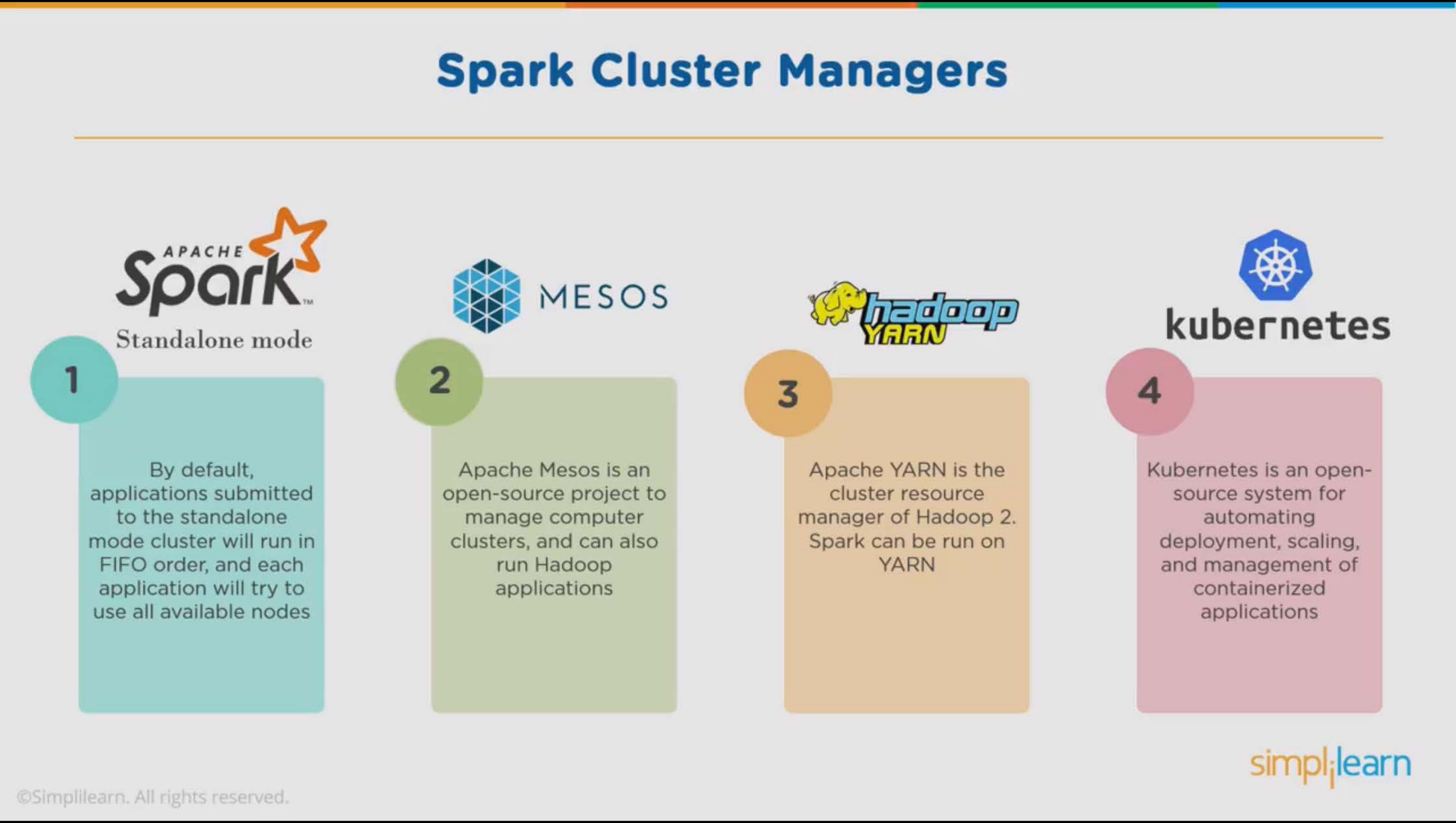


- A job is split into multiple tasks that are distributed over the worker node
- When an RDD is created in Spark context, it can be distributed across various nodes
- Worker nodes are slaves that run different tasks

simplilearn

Turn on Draw & Zoom

暂停



# Spark Cluster Managers

APACHE Spark™

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

MESOS

2

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

hadoop YARN

3

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

kubernetes

4

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

simplilearn

# Applications of Spark

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 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 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

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

暂停

# 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

暂停

# 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

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

暂停

# 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

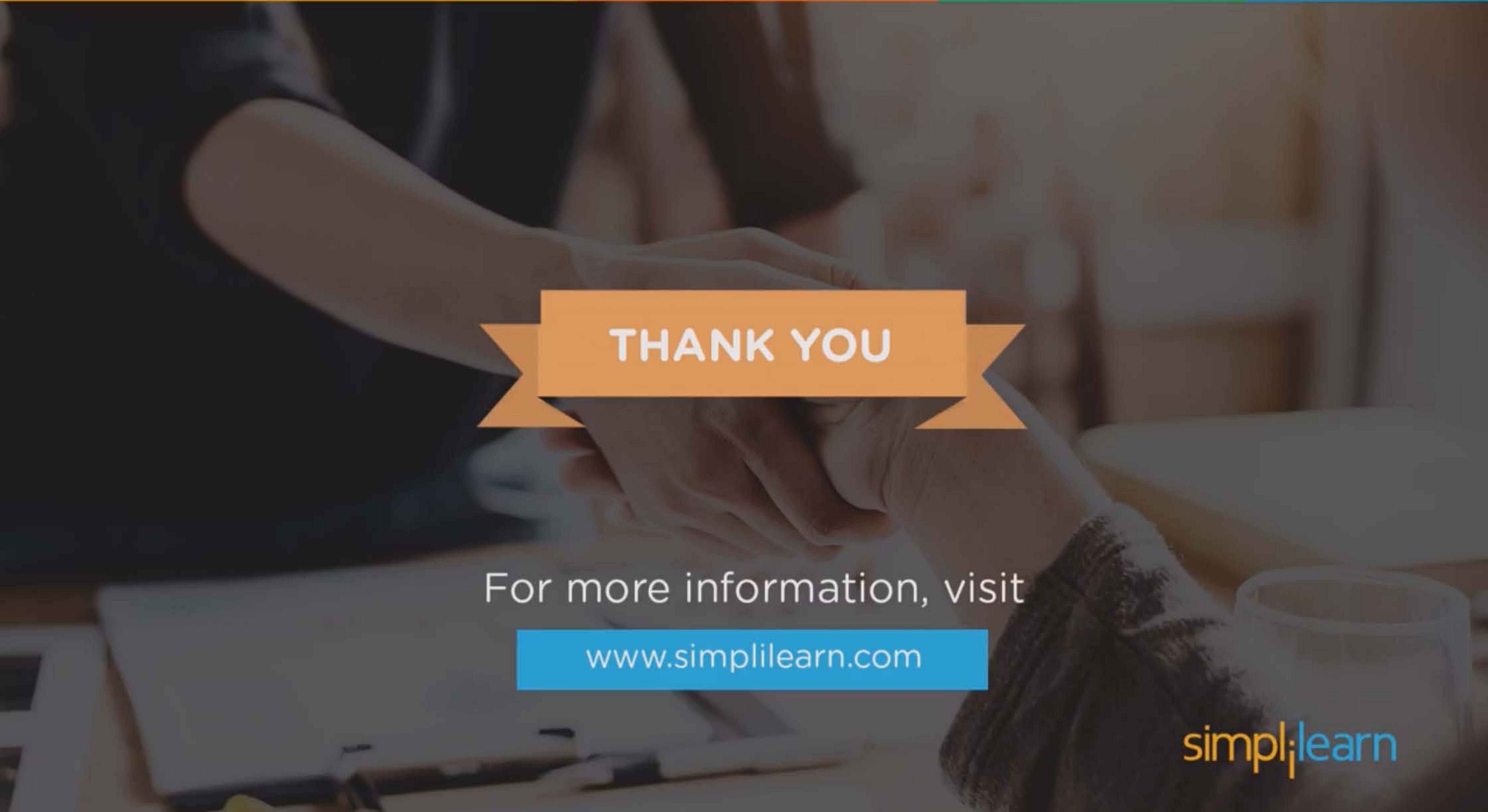
It automatically detects anomalies along the video streaming pipeline and diagnoses the root cause of the issue

Reduces waiting time before the video starts

Avoids buffering and recovers the video from a technical error

simplilearn

暂停



# THANK YOU

For more information, visit

www.simplilearn.com

simplilearn