Computing & Information Sciences

Cloud Computing

Introduction

Camilo Valdes
cvalde03@fiu.edu

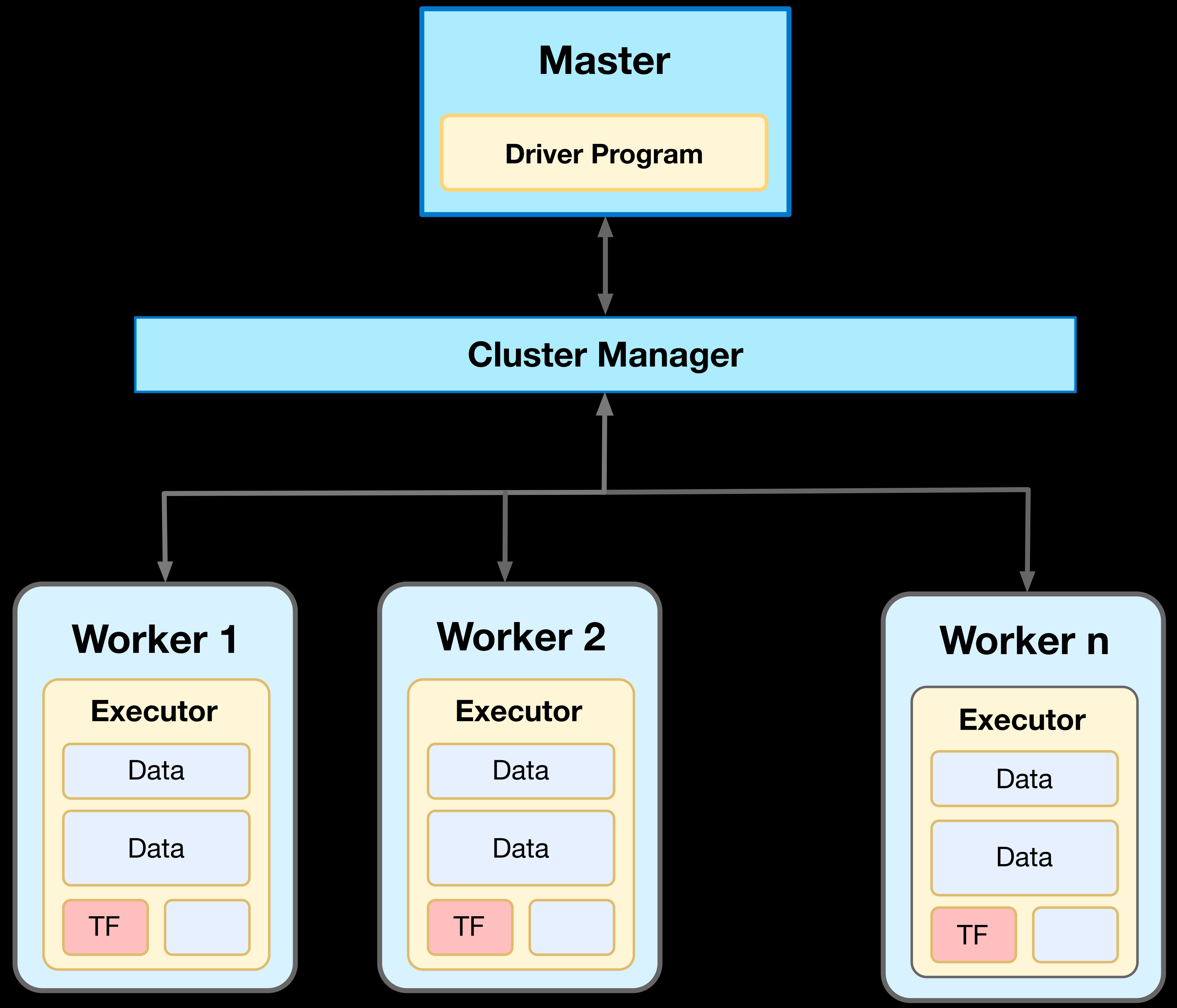
Bioinformatics Research Group
Florida International University. Miami, FL, USA

Agenda

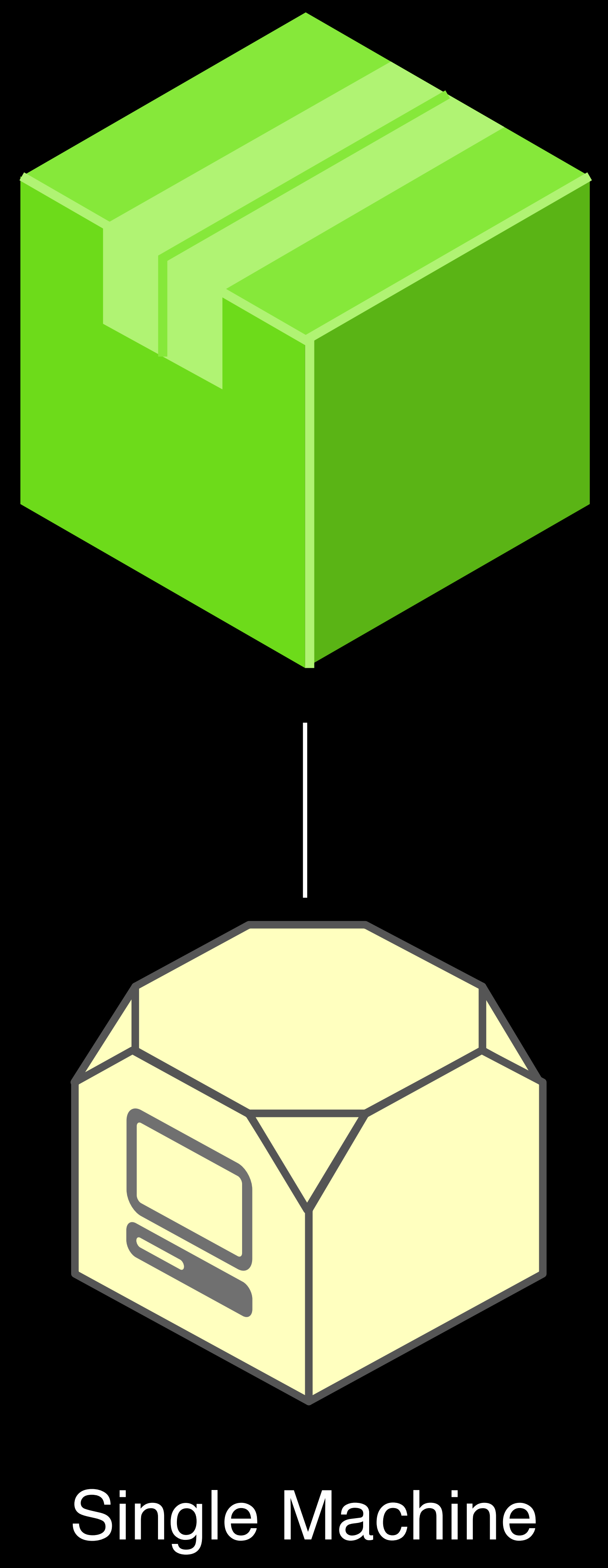
- Overview
- Amazon Web Services (AWS)
  - AWS Console
  - AWS Educate
- Spark
  - MapReduce
  - Clusters
- Tensorflow
  - Introduction
  - Examples

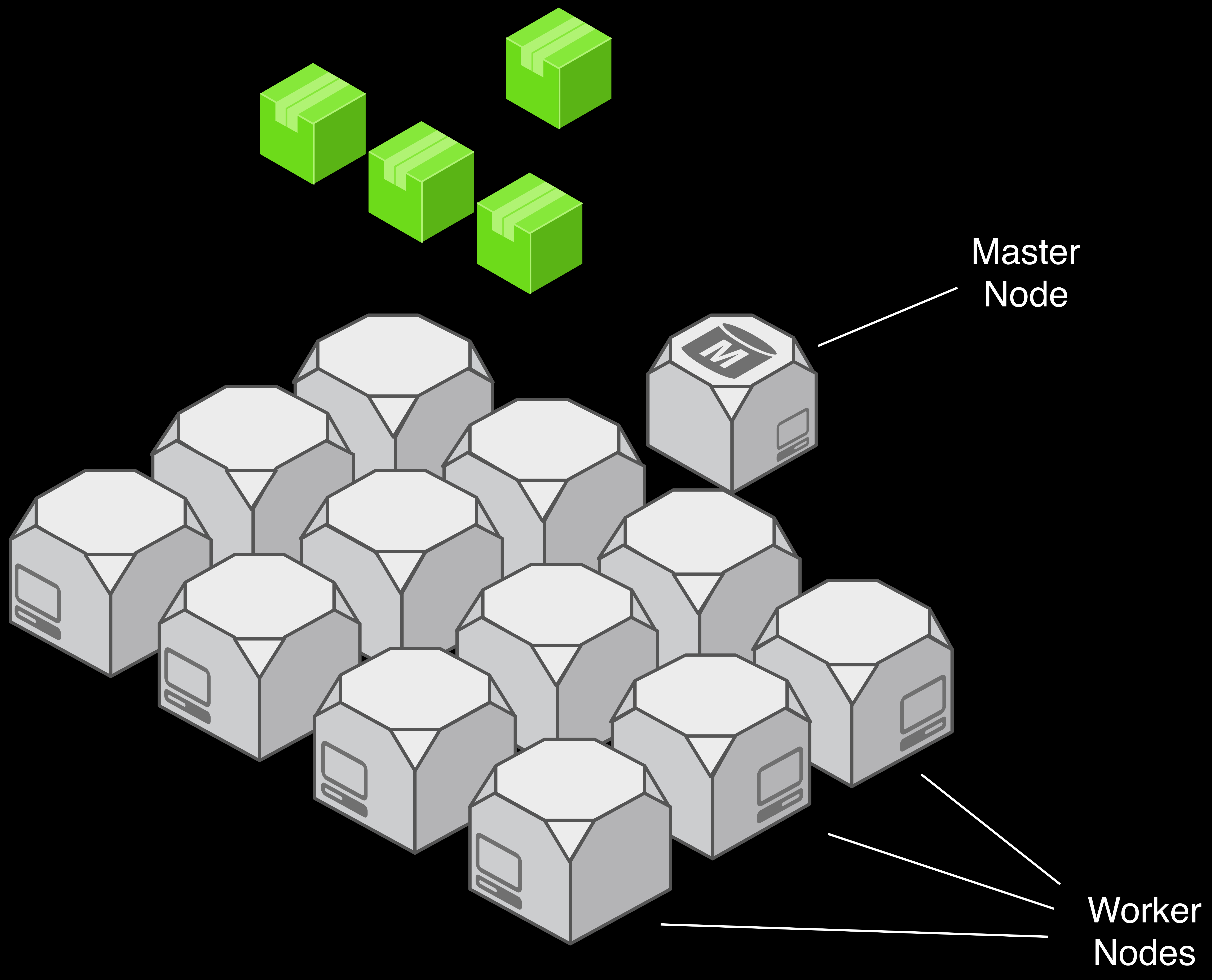
Overview

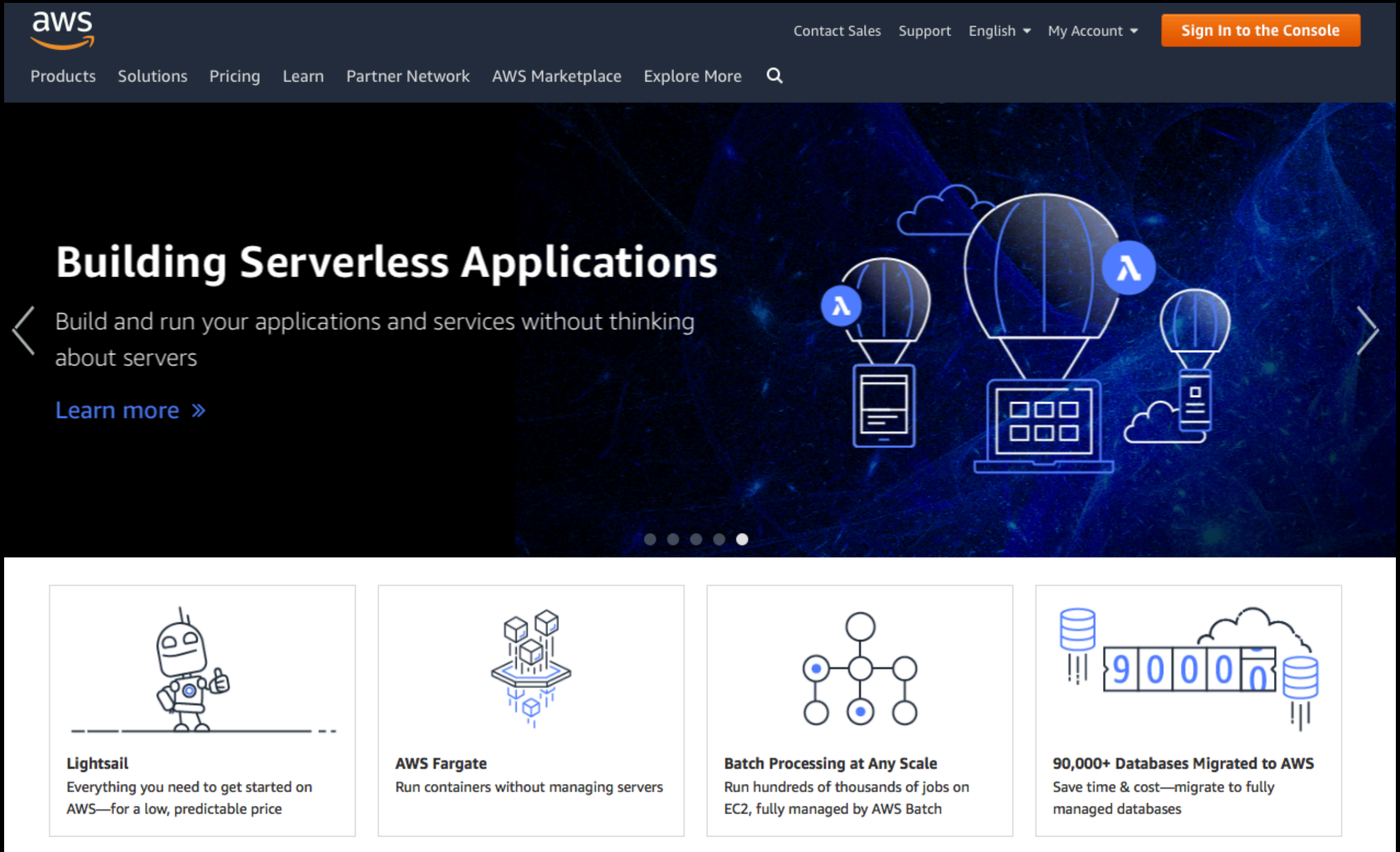
- Big Data & Computing
- Sometimes, a single computer cannot process all the data, or it would take too long
- Rather than use a single powerful machine, we could use many commodity ones
- Process data in parallel, in small chunks, and aggregate the results



MapReduce



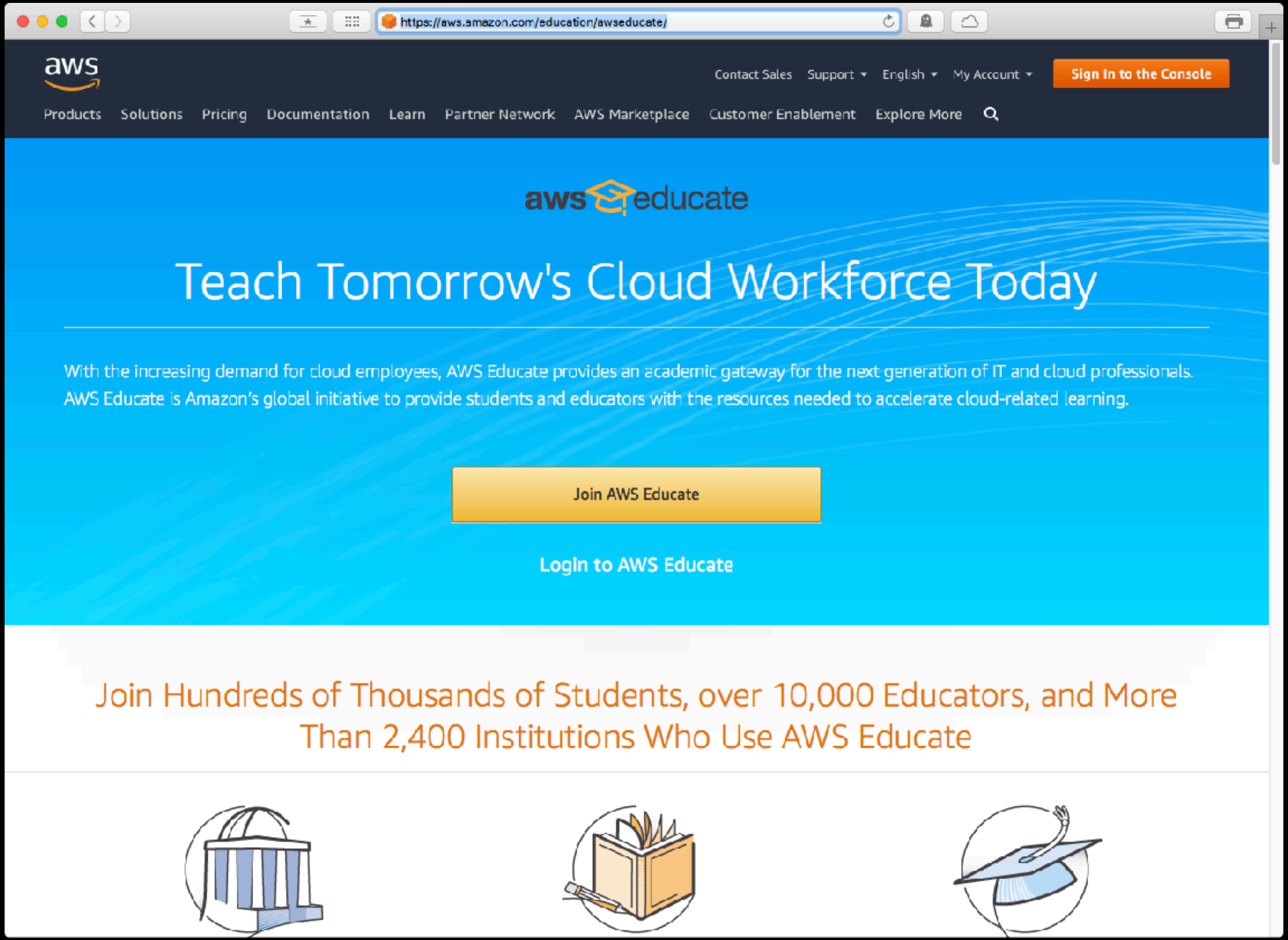




https://aws.amazon.com

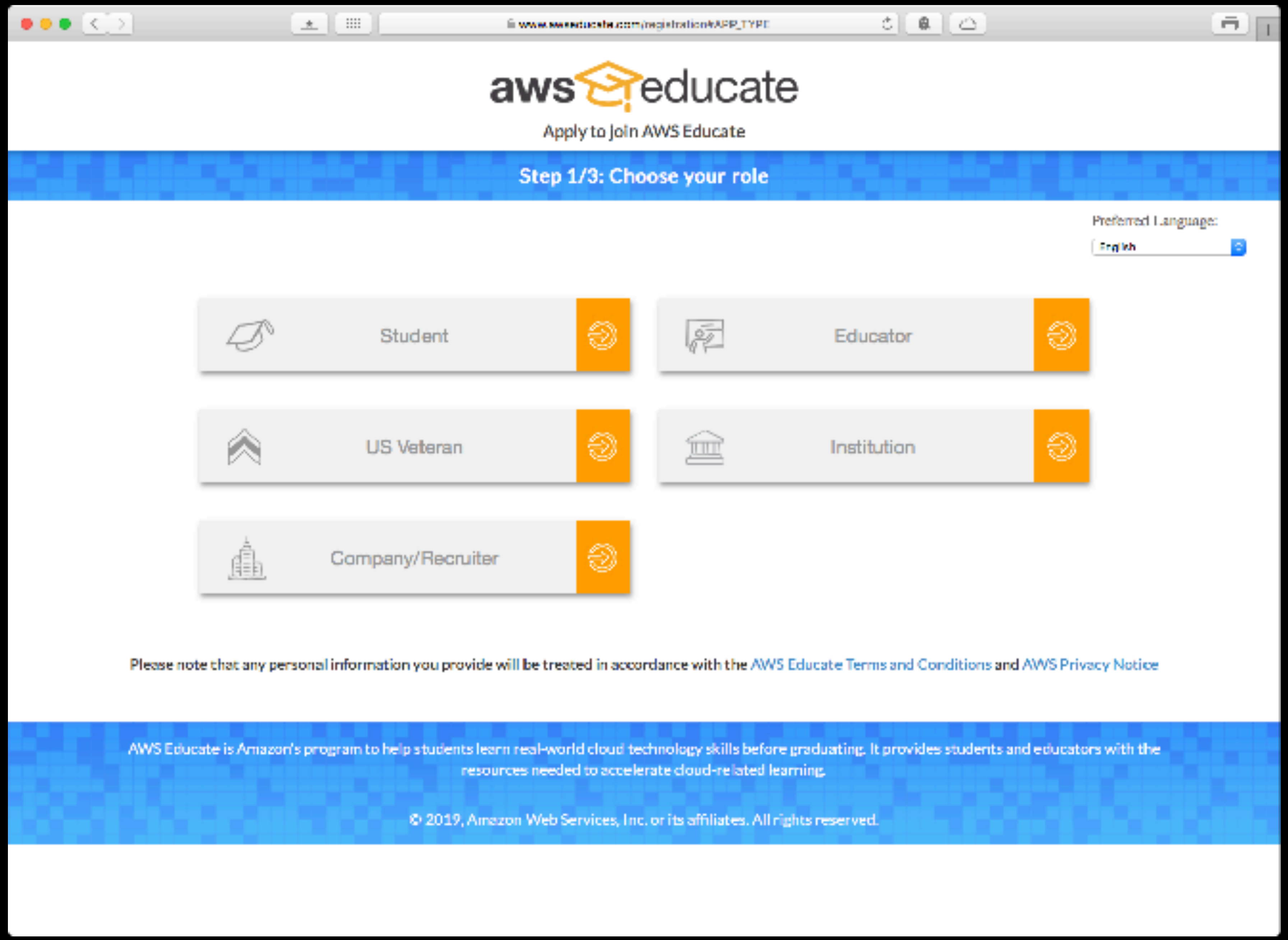
AWS Educate

https://aws.amazon.com/education/awseducate/



AWS Educate

- Students
  - Get $100 AWS credit for signing up
    - FIU email
  - Resources for Learning
    - Tutorials, videos, etc.
  - Specialization Paths
  - AWS Certifications
  - Job Board



https://aws.amazon.com/education/awseducate/

www.awseducate.com/student/s/pathways

Explore Cloud Career Pathways

Explore AWS Educate's Cloud Career Pathways to start building the key cloud skills you'll need to be successful in leading technology careers. Earn a completion credential for each pathway and share with prospective employers to show what you've learned.

Check out the roles below to learn more about each pathway and get started!

Cloud Computing 101

Take a crash course on the cloud, its history, solutions, and why companies across the globe are looking for employees with AWS cloud expertise.

START ▸ LEARN MORE ▸

Application Developer

Curious how App Developers design, test, and improve engaging web and mobile applications in the cloud? Learn more about the skills you'll need.

START ▸ LEARN MORE ▸

Cloud Support Associate

If you're excited by the future of cloud computing and enjoy working directly with customers, learn more about becoming a Cloud Support Associate.

START ▸ LEARN MORE ▸

Cloud Support Engineer

Interested in multiple technologies and working with companies to support AWS cloud solutions? Learn more about becoming a Cloud Support Engineer.

START ▸ LEARN MORE ▸

Cybersecurity Specialist

Cybersecurity Specialists use expertise in networking, programming, and coding to protect customer data every day. Learn more about the skills they use.

START ▸ LEARN MORE ▸

Data Integration Specialist

Excited about bringing data sources together to tell the story of a product's performance? Discover ways to build and improve products through data.

START ▸ LEARN MORE ▸

Data Scientist

Curious how discovering patterns in large data sets can translate into new business strategies? Learn more about how Data Scientists do this every day.

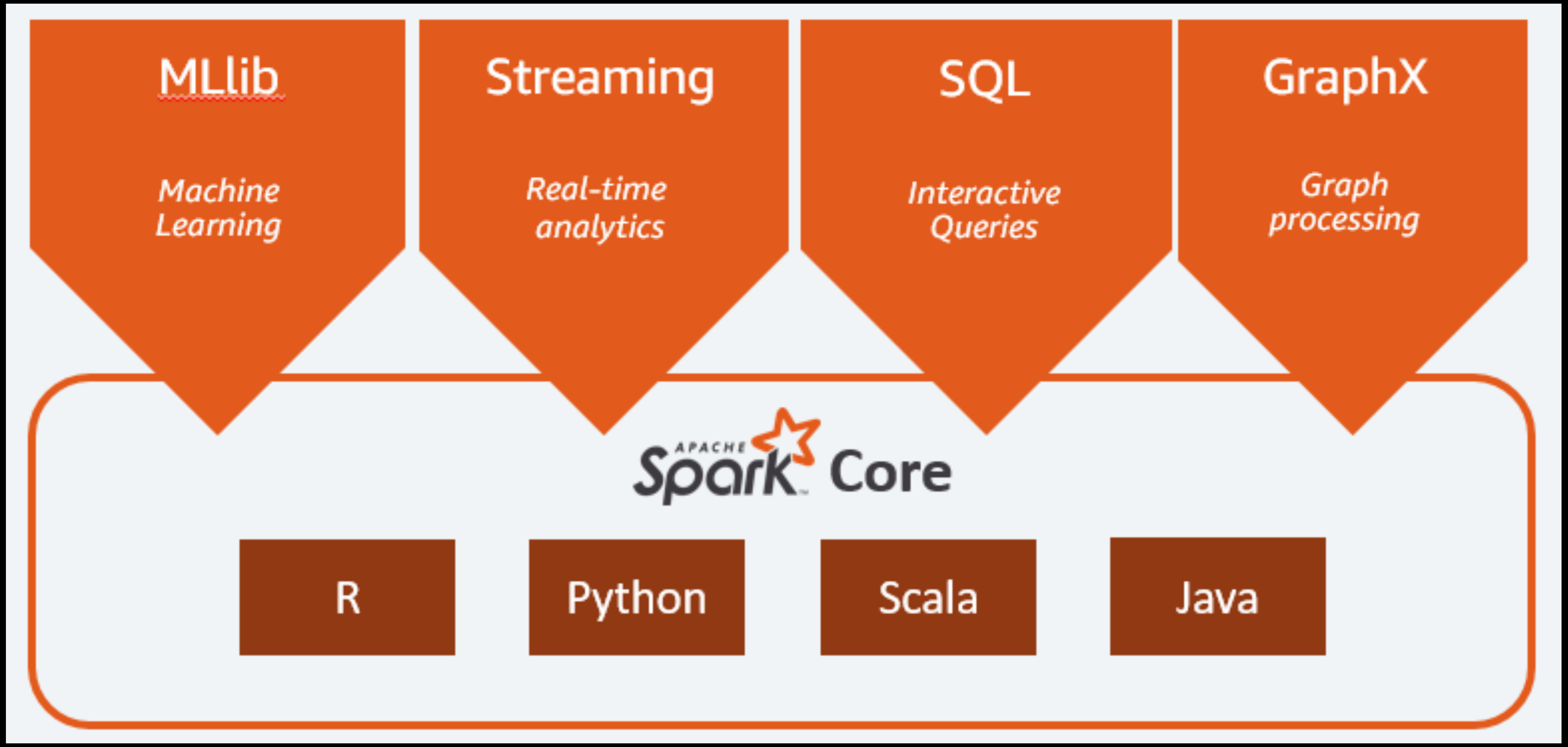
START ▸ LEARN MORE ▸

DevOps Engineer

If you like working behind the scenes to tackle challenges and are curious about skills like scripting and coding, learn more about becoming a DevOps

START ▸ LEARN MORE ▸

Apache Spark

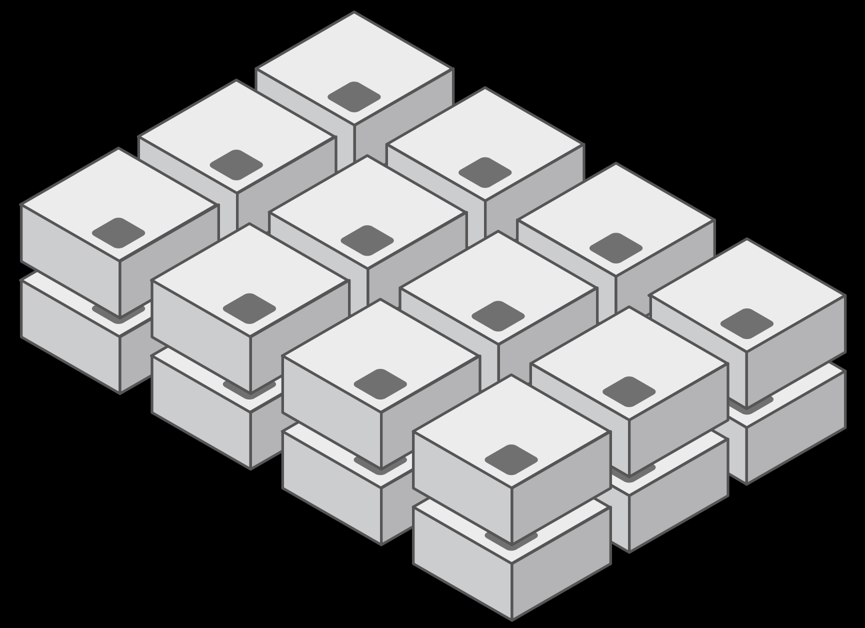


Apache Spark

- Spark is a Big Data Processing Engine — a Fast, General-Purpose, Cluster-computing Platform.
- Handles the Scheduling, Distribution, and Monitoring of applications spanning many worker machines.
- Has a Rich API to distribute data across the cluster, and process it in parallel.
- Supports a variety of workloads such as Machine Learning (MLlib), Streaming, interactive queries, graph programming and SQL.
- Execution Frameworks have language support for Python, R, Java, and Scala.

Spark — Unified Stack

- The Spark project contains multiple high-level specialized components (MLlib, Streaming, etc.).
- Spark's main programming abstraction are **Resilient Distributed Datasets (RDDs)**, a data structure distributed across nodes that can be worked on in parallel.
- Spark's multiple components operate on RDDs, which allows for close interoperability and tight integration.
- Applications that use **multiple processing models** can be written without high maintenance and development costs.



Spark — Main Benefits

Solve problems faster, and on a much larger scale

- Ease of Use — Rich, high level APIs
- Speed — Fast parallel execution
- General Engine — Combine processing models
- Open Source — Freely Available

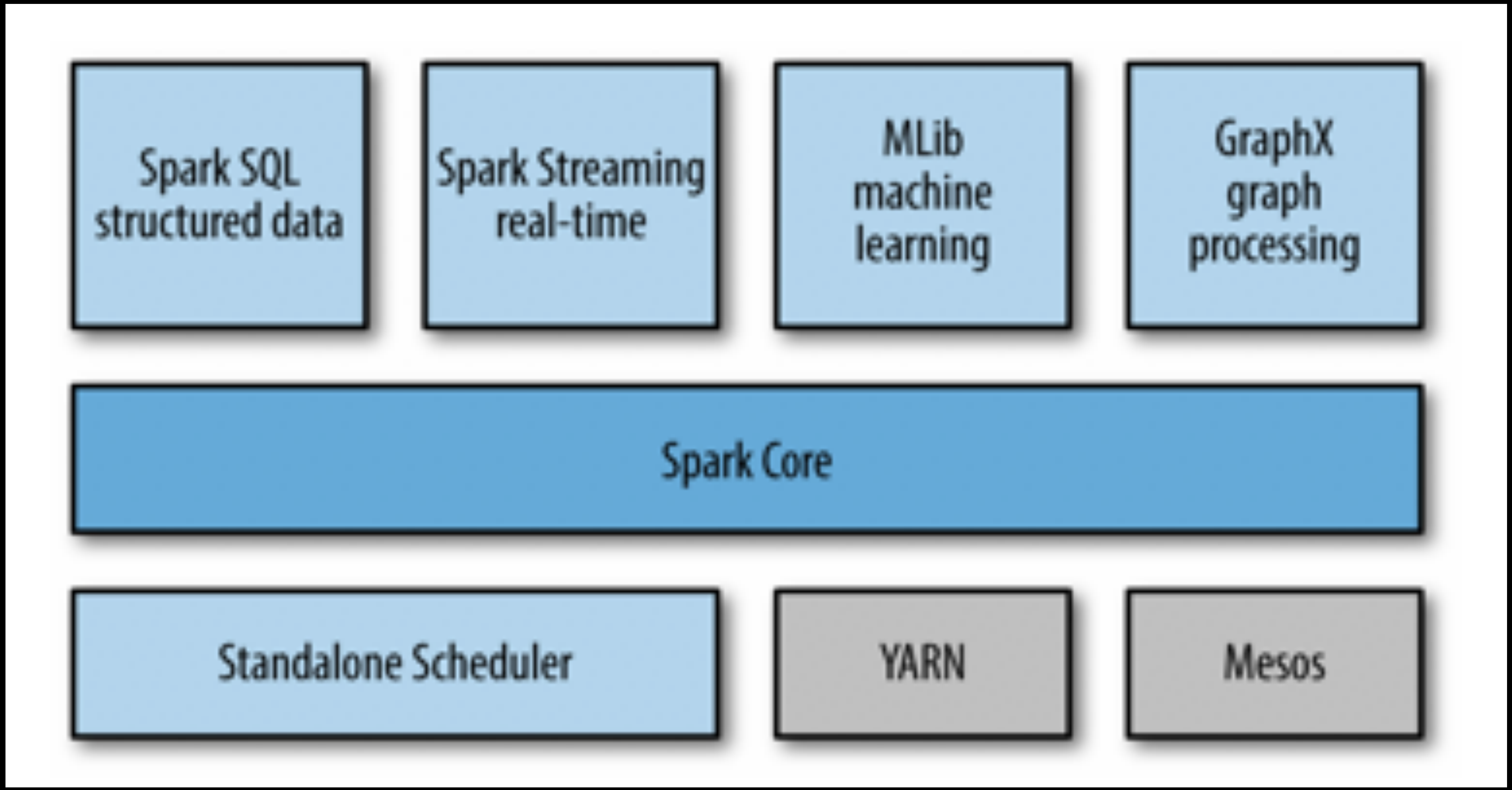
• Makes developing General Purpose Distributed programs easier, less painful.

• Reduces the management burden of maintaining separate tools.

• Allows the close Interoperability of high-level components

Spark Core

- Spark Core contains the basic functionality of Spark, including components for task scheduling, memory management, fault recovery, interacting with storage systems, and more.



- Spark Core is also home to the API that defines resilient distributed datasets (RDDs), which are Spark's main programming abstraction.
- RDDs represent a collection of items distributed across many compute nodes that can be manipulated in parallel.

Spark — Data Processing

- Spark provides a simple way to parallelize applications across clusters, and hides the complexity of distributed systems programming, network communication, and fault tolerance.
- The system gives control to monitor, inspect, and tune applications while allowing implementation of common tasks quickly.
- The modular nature of the API (based on passing distributed collections of objects) makes it easy to factor work into reusable libraries and test it locally.

Storage Layers for Spark

- Spark can create resilient distributed datasets, RDDs, from any file stored in the Hadoop distributed filesystem (HDFS).
- Spark also support other storage systems supported by the Hadoop APIs (including your local filesystem, Amazon S3, Cassandra, Hive, HBase, etc.).
- It’s important to remember that Spark does not require Hadoop.
- It simply has support for storage systems implementing the Hadoop APIs.

Spark REPL

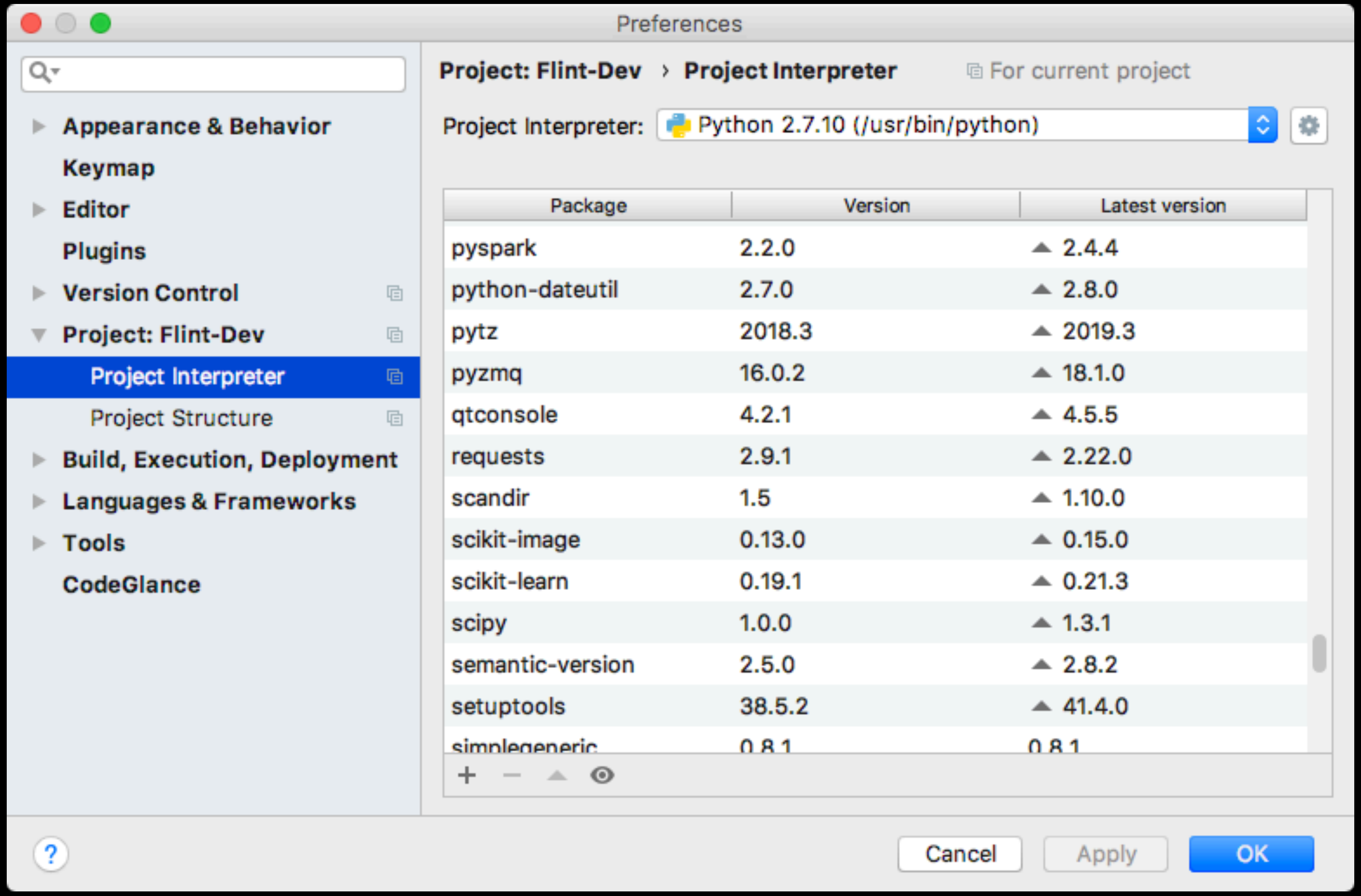
- Spark can be used from Python, R, Java, or Scala.
- Spark itself is written in Scala, and runs on the Java Virtual Machine (JVM).
- To run Spark on either your laptop or a cluster, all you need is an installation of Java 6 or newer.
- If you wish to use the Python API you will also need a Python interpreter (version 2.6 or newer).
- You don’t need to have Hadoop.
- Spark comes with interactive shells that enable ad hoc data analysis.
- Spark’s shells will feel familiar if you have used other shells such as those in R, Python, and Scala,

Installing Spark

- Spark is a framework
- Language bindings for
  - Python, Scala, Java
- Install with a package manager
  - Homebrew in macOS
  - Pycharm Repository for Windows

```python
#   Spark Modules
from pyspark import SparkConf, SparkContext
from pyspark.streaming import StreamingContext

#   Python Modules
import io, os, sys
import argparse
import time
import json
import csv
import boto3
import pandas as pd
from datetime import timedelta
import operator
```



pyspark

- Python version of the Spark Shell.

```
Last login: Sat Oct 27 16:23:14 on ttys003
Trajan.>_ pyspark
Python 2.7.14 (default, Mar 10 2018, 00:01:04)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.39.2)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/30 18:07:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
18/10/30 18:07:48 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 2.7.14 (default, Mar 10 2018 00:01:04)
SparkSession available as 'spark'.
>>>
```

pyspark

- In Spark, we express our computation through operations on distributed collections that are automatically parallelized across the cluster.
- These collections are called resilient distributed datasets, or RDDs.
- RDDs are Spark's fundamental abstraction for distributed data and computation.

```
Last login: Sat Oct 27 16:23:14 on ttys003
Trajan.>_ pyspark
Python 2.7.14 (default, Mar 10 2018, 00:01:04)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.39.2)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/30 18:07:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
18/10/30 18:07:48 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 2.7.14 (default, Mar 10 2018 00:01:04)
SparkSession available as 'spark'.
>>>
```

RDDs

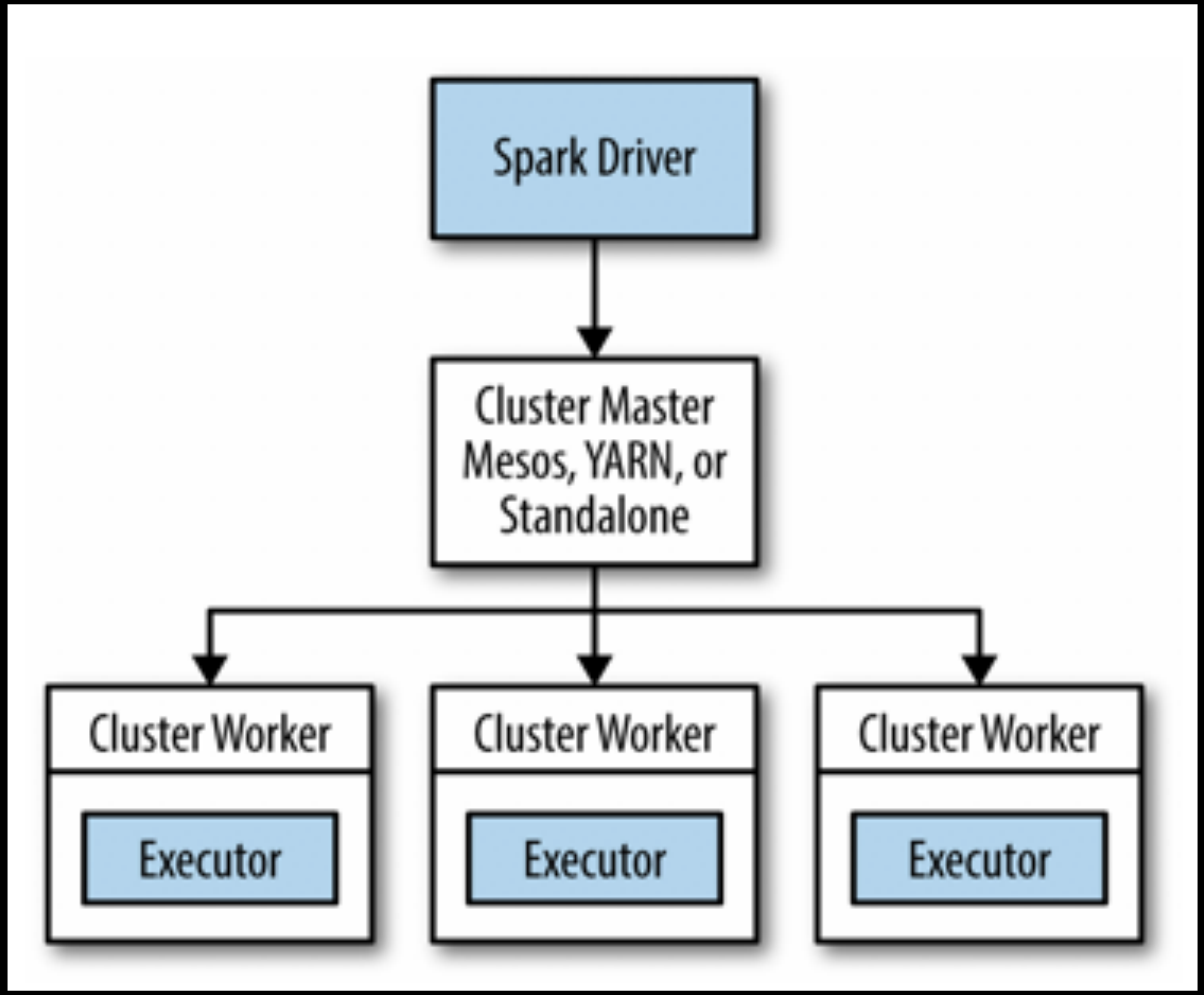
- An RDD is simply a distributed collection of elements.
- In Spark all work is expressed as either creating new RDDs, transforming existing RDDs, or calling operations on RDDs to compute a result.
- Spark automatically distributes the data contained in RDDs across your cluster and parallelizes the operations you perform on them.
- An RDD in Spark is simply an immutable distributed collection of objects.
- Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster.
- RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.
- Once created, RDDs offer two types of operations: *transformations* and *actions*.

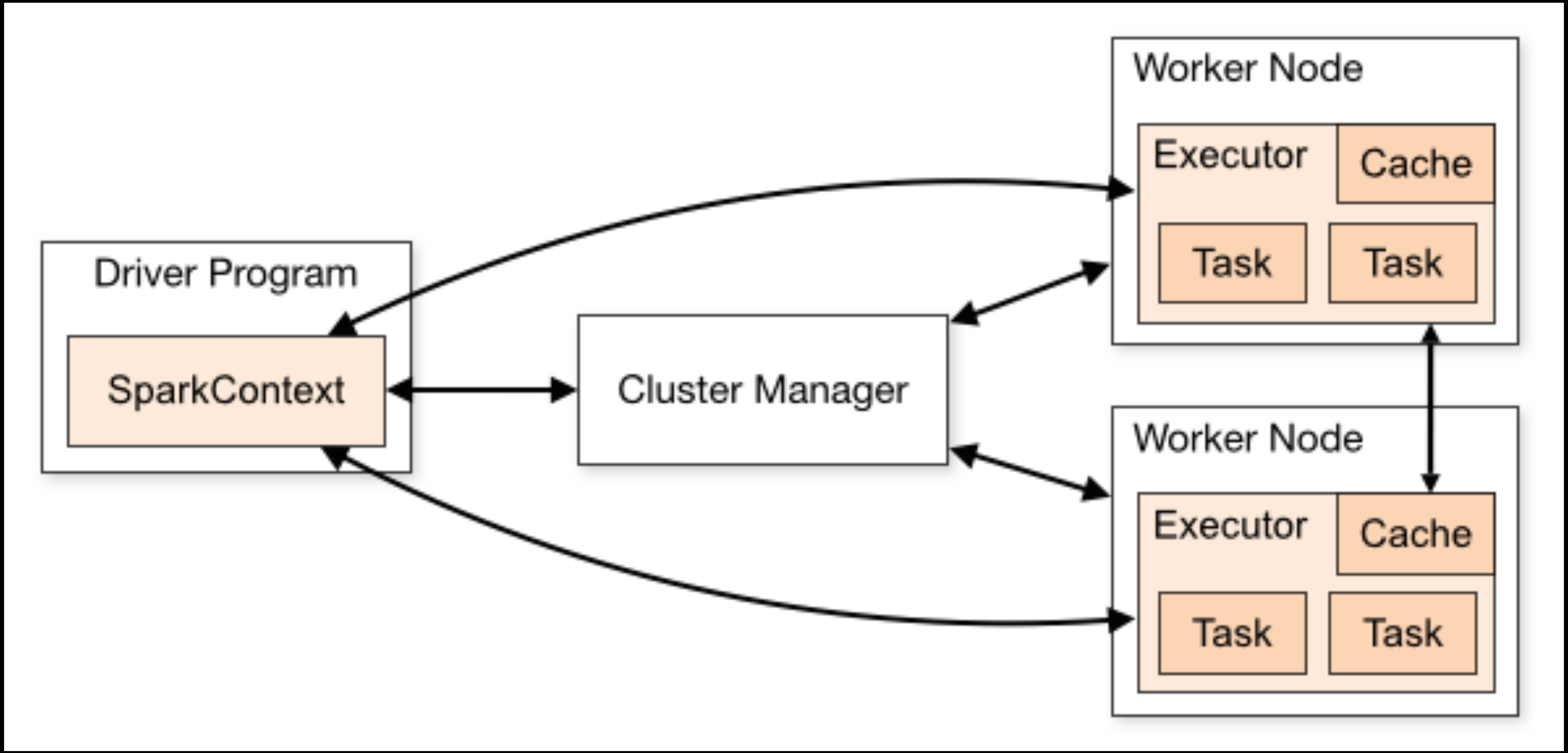
RDDs

- Transformations construct a new RDD from a previous one.
- Actions compute a result based on an RDD, and either return it to the driver program or save it to an external storage system.
- Although you can define new RDDs any time, Spark computes them only in a lazy fashion — that is, the first time they are used in an action.
- Spark provides two ways to create RDDs
  - loading an external dataset.
  - Parallelizing a collection in your driver program.

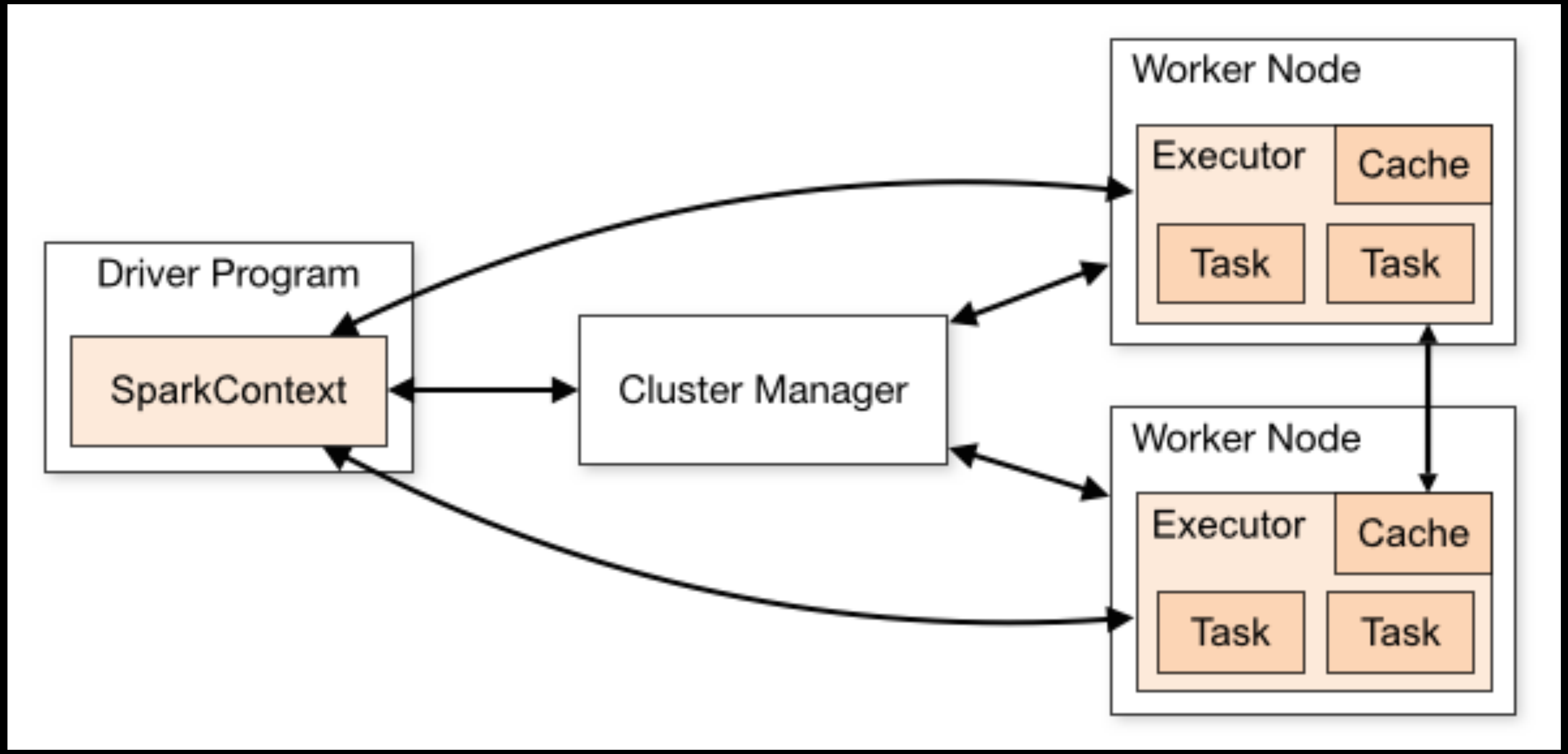
Spark Cluster

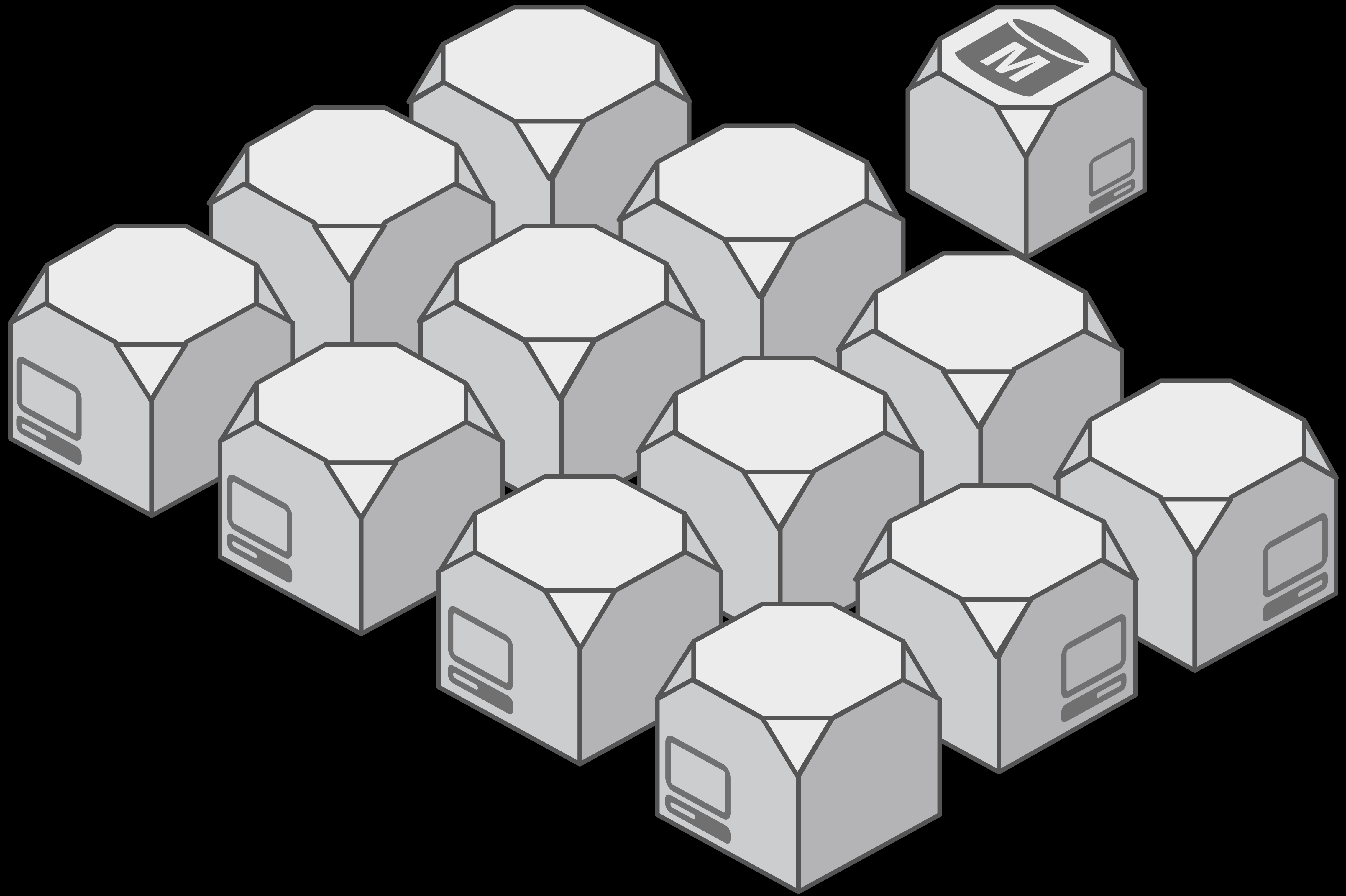
- Every Spark application consists of a driver program that launches various parallel operations on a cluster.
- The driver program contains your application's main function and defines distributed datasets on the cluster, then applies operations to them.
- The driver communicates with a potentially large number of distributed workers called executors.
- A driver and its executors are together termed a Spark application.

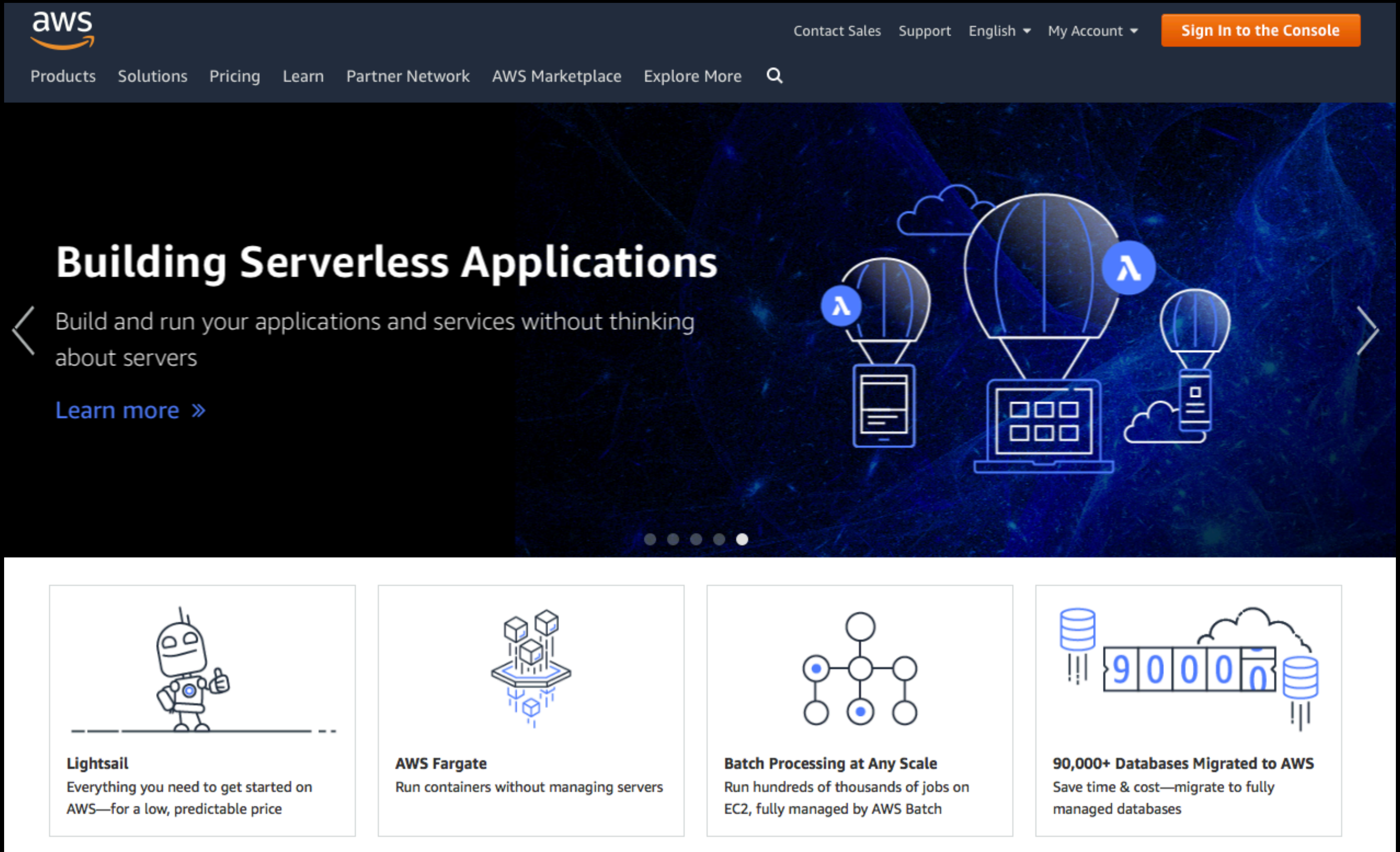




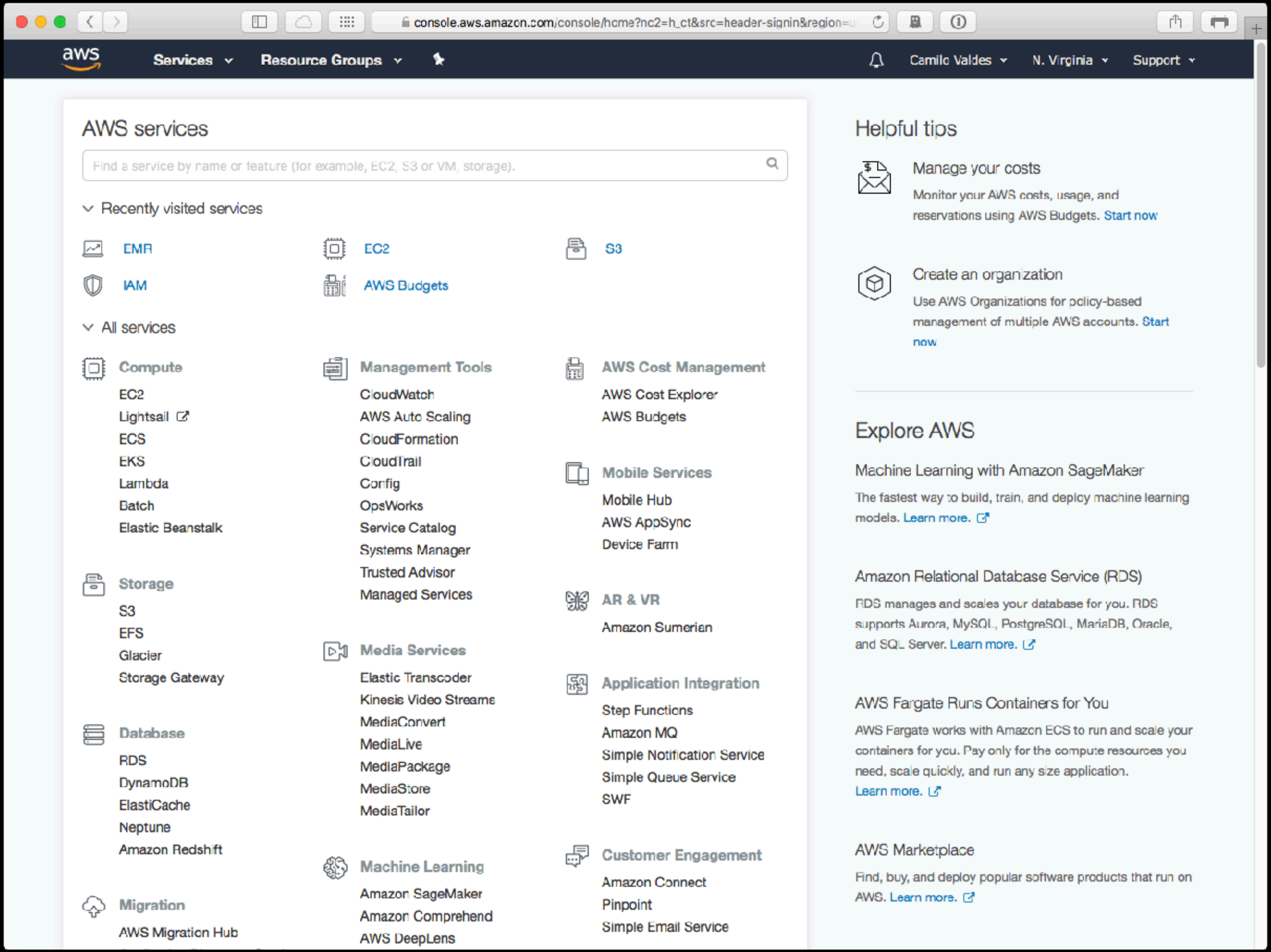
https://stackoverflow.com


Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task


M



https://aws.amazon.com

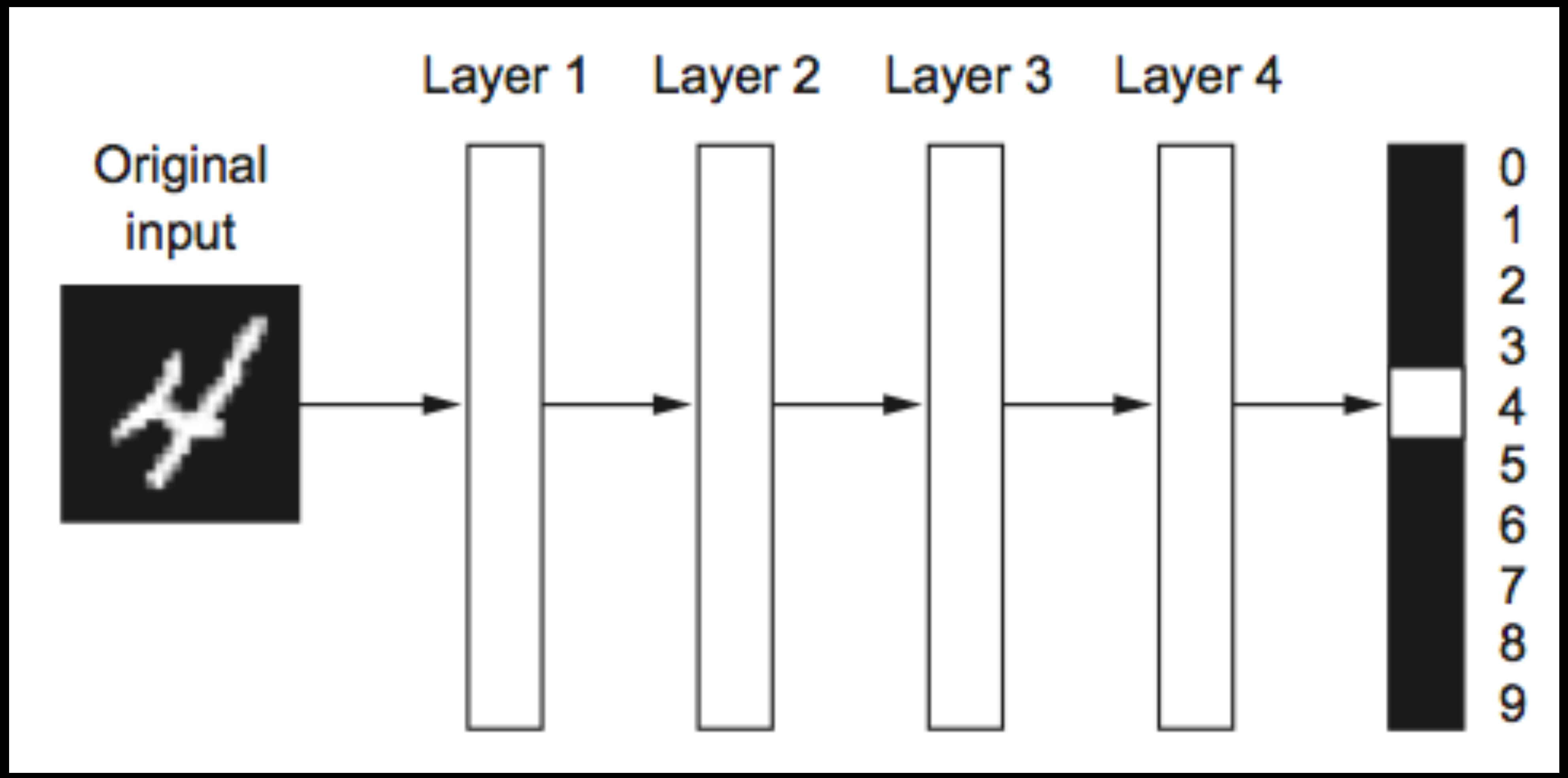

console.aws.amazon.com/console/home?nc2=h_ct&src=header-signin&region=
aws
Services
Resource Groups
Camilo Valdes
N. Virginia
Support
AWS services
Find a service by name or feature (for example, EC2, S3 or VM, storage).
Recently visited services
EMR
EC2
S3
IAM
AWS Budgets
All services
Compute
EC2
Lightsail
ECS
EKS
Lambda
Batch
Elastic Beanstalk
Storage
S3
EFS
Glacier
Storage Gateway
Database
RDS
DynamoDB
ElastiCache
Neptune
Amazon Redshift
Migration
AWS Migration Hub
Management Tools
CloudWatch
AWS Auto Scaling
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Systems Manager
Trusted Advisor
Managed Services
Media Services
Elastic Transcoder
Kinesis Video Streams
MediaConvert
MediaLive
MediaPackage
MediaStore
MediaTailor
Machine Learning
Amazon SageMaker
Amazon Comprehend
AWS DeepLens
AWS Cost Management
AWS Cost Explorer
AWS Budgets
Mobile Services
Mobile Hub
AWS AppSync
Device Farm
AR & VR
Amazon Sumerian
Application Integration
Step Functions
Amazon MQ
Simple Notification Service
Simple Queue Service
SWF
Customer Engagement
Amazon Connect
Pinpoint
Simple Email Service
Helpful tips
Manage your costs
Monitor your AWS costs, usage, and reservations using AWS Budgets. Start now
Create an organization
Use AWS Organizations for policy-based management of multiple AWS accounts. Start now
Explore AWS
Machine Learning with Amazon SageMaker
The fastest way to build, train, and deploy machine learning models. Learn more.
Amazon Relational Database Service (RDS)
RDS manages and scales your database for you. RDS supports Aurora, MySQL, PostgreSQL, MariaDB, Oracle, and SQL Server. Learn more.
AWS Fargate Runs Containers for You
AWS Fargate works with Amazon ECS to run and scale your containers for you. Pay only for the compute resources you need, scale quickly, and run any size application.
Learn more.
AWS Marketplace
Find, buy, and deploy popular software products that run on AWS. Learn more.

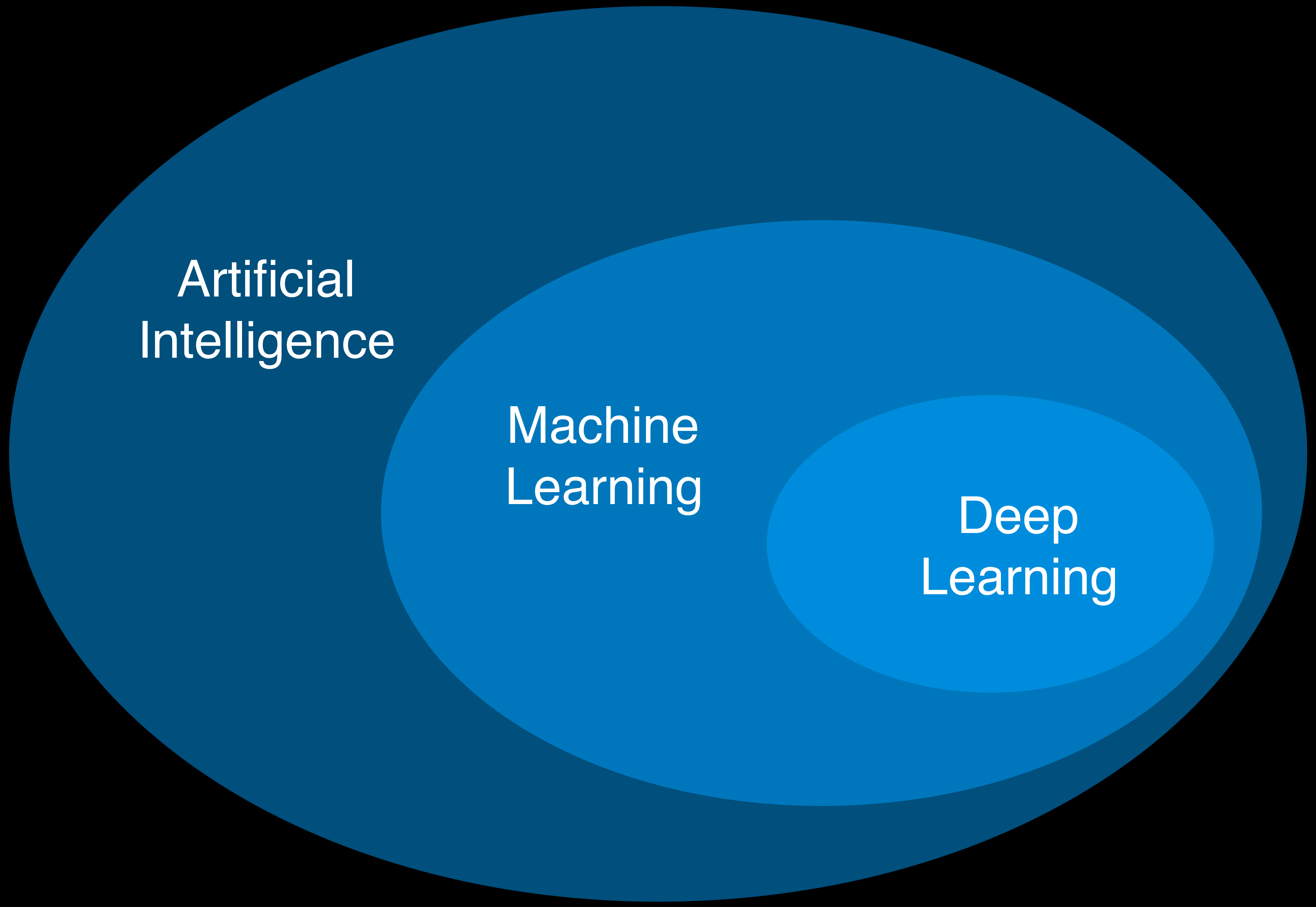
Demo

TensorFlow

TensorFlow

- Open source platform for Machine Learning
- Developed by Google
- Very popular in Deep Learning





Deep Representation

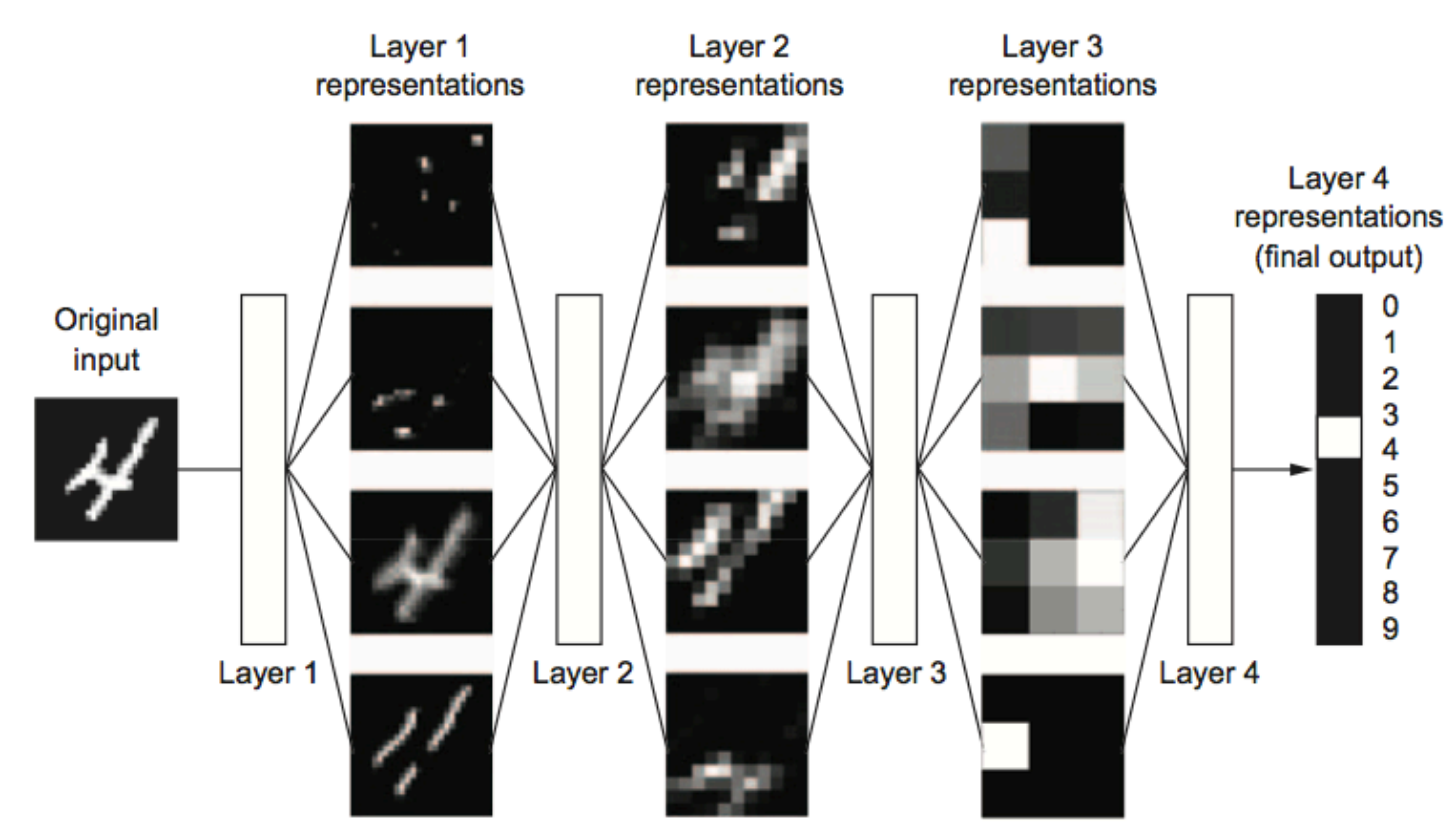


image credit: François Chollet

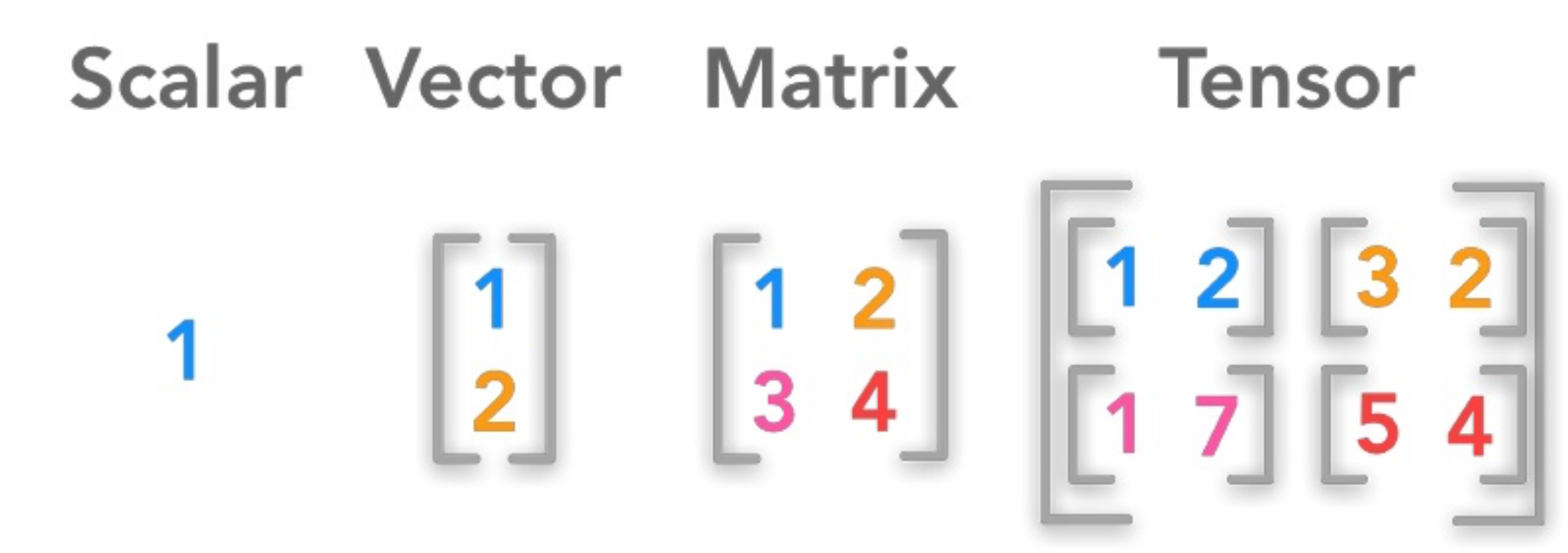
Democratization of Deep Learning

- Back in the early days, to do any deep learning, you needed a lot of experience with C++ and CUDA (NVIDIA's driver API)
- Developing deep learning models was cumbersome, and there were no tools for easy debugging
- Tensorflow was created to develop ML applications at scale an in production
    - Simplicity
    - Scalability
    - Versatility and Reusability

Tensorflow's first released to Open Source was on November 8, 2015 (version 0.5.0)

Tensor**Flow**

- Name comes from its basic data structure, a "*tensor*"
- A tensor in TensorFlow is a multidimensional data structure
  - Scalar - 0D tensor
  - Vector - 1D tensor
  - Matrix - 2D tensor
  - 3D tensor, etc.
- Tensors have attributes such as shape, data type, and the number of axes (rank)



Anatomy of a Neural Network

- Training a Neural Network involves
  - Layers (combined into a network)
  - Input Data and its Targets (labels)
  - Loss Function (feedback signal)
  - Optimizer

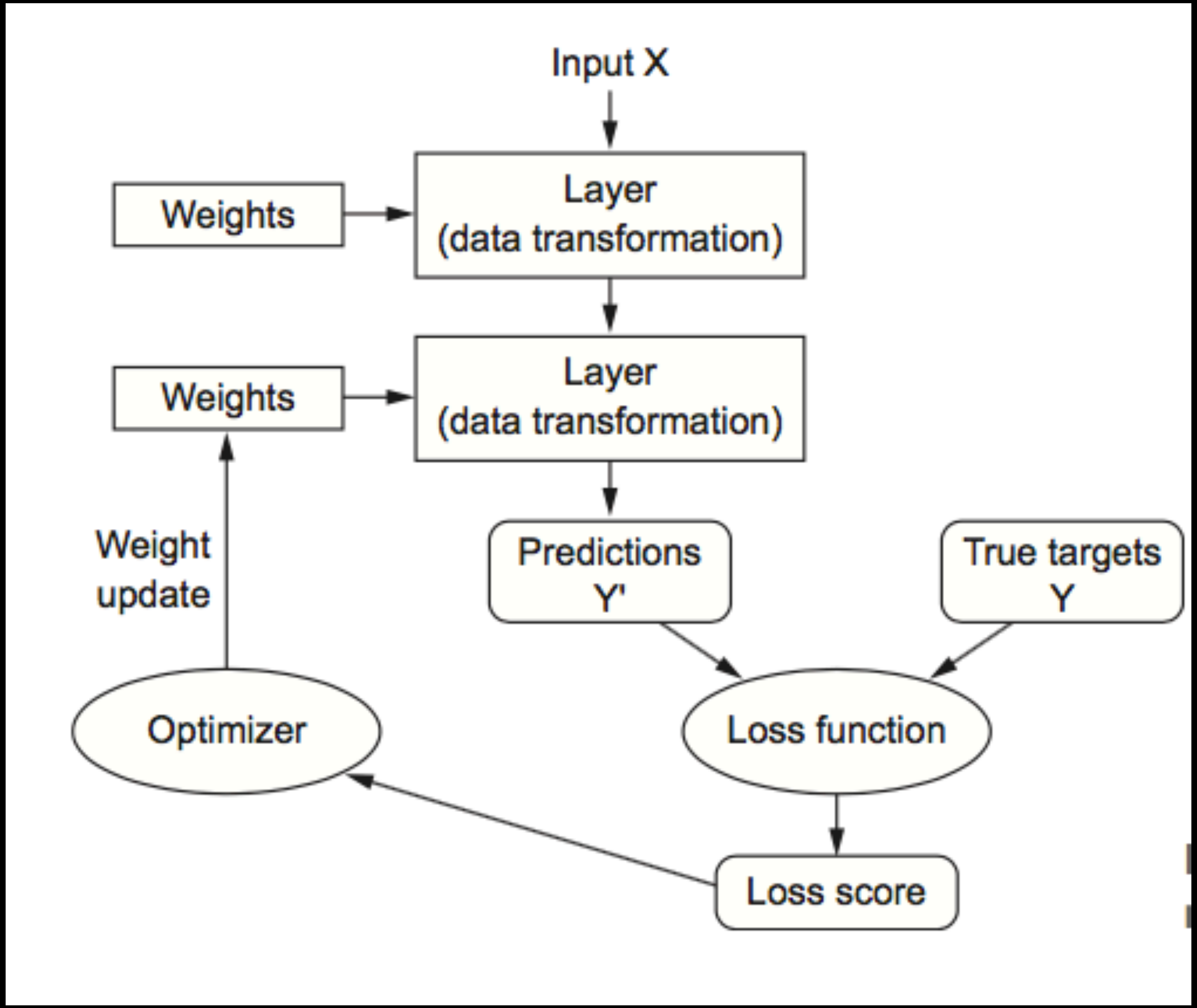


image credit: François Chollet

Demo

http://yann.lecun.com/exdb/mnist/

MNIST Dataset
60,000 Training Images
10,000 Test Images