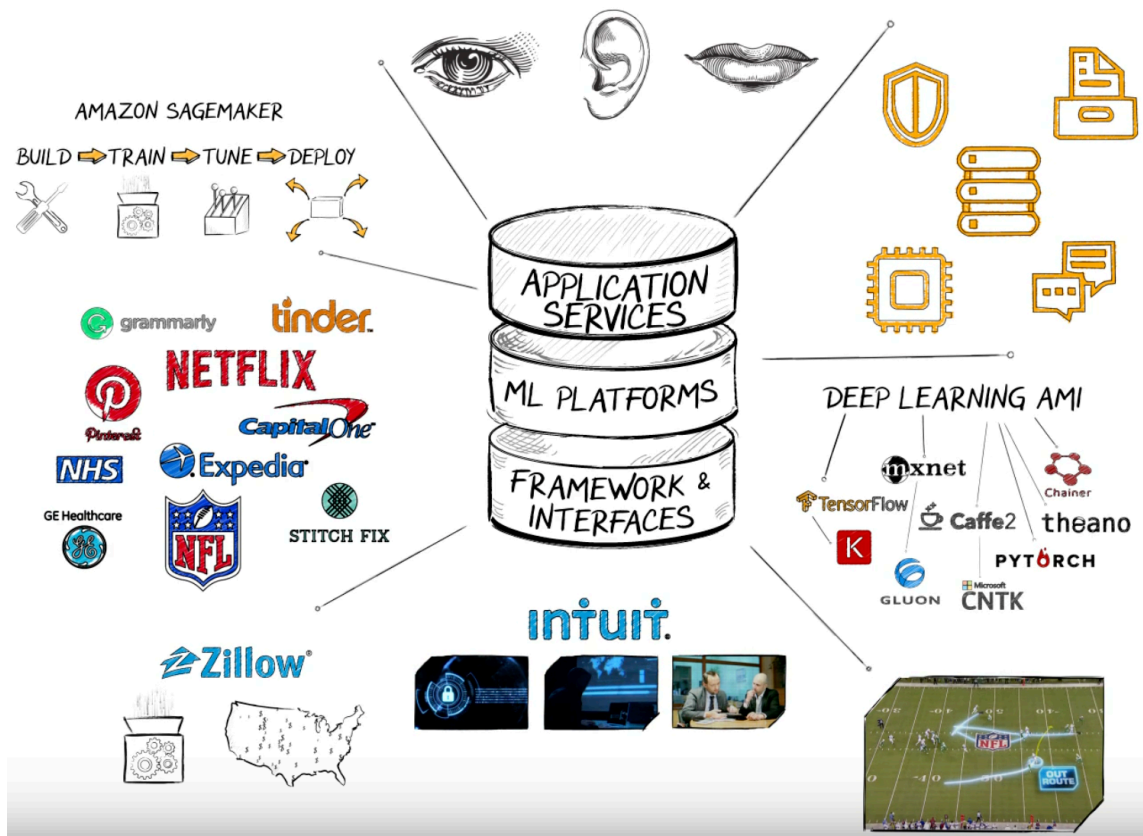# AWS Tutorial: Analyze Big Data with Hadoop

Time spent: 70min

What I learned:
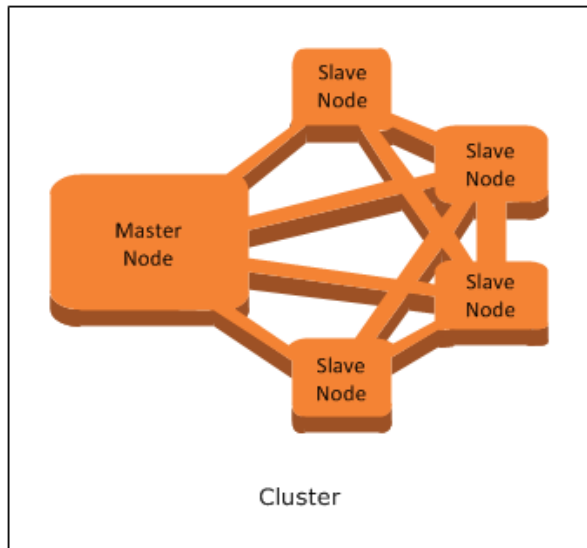In the overview video, I learned that there was three layers on the AWS ML.

AWS Machine Learning Overview

AWS EMR Overview

The central component of Amazon EMR is the *cluster*. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a *node*. Each node has a role within the cluster, referred to as the *node type*. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

Cluster

How to create my cluster

Add SSH Access

Start the script by adding steps

Clone   Terminate   AWS CLI export

Cluster: My First EMR Cluster   Waiting   Cluster ready after last step completed.

| Summary | Application history | Monitoring | Hardware | Events | Steps | Configurations | Bootstrap actions |

Add step   Clone step   Cancel step

Steps

Filter: All steps ▾   Filter steps ...   2 steps (all loaded) ⟳

ID	Name	Status	Start time (UTC-5) ▾	Elapsed time	Log files ⧉	Actions
▾ ◯ s-2WPNWJIUZUQ44	ShiboHive	Pending		0 seconds	View logs	View jobs

JAR location : command-runner.jar
Main class : None
Arguments : hive-script --run-hive-script --args -f s3://us-east-1.elasticmapreduce.samples/cloudfront/code/Hive_CloudFront.q -d INPUT=s3://us-east-1.elasticmapreduce.samples -d OUTPUT=s3://bigdata0705/HiveQueryResults/
Action on failure: Continue

Clone   Terminate   AWS CLI export

Cluster: My First EMR Cluster   Starting

| Summary | Application history | Monitoring | Hardware | Events | Steps | Configurations | Bootstrap actions |

Connections: --
Master public DNS: --
Tags: -- View All / Edit

Summary
ID: j-KVLH4L3EWSF
Creation date: 2018-12-09 08:35 (UTC-5)
Elapsed time: 0 seconds
Auto-terminate: No
Termination protection: Off Change

Configuration details
Release label: emr-5.19.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://aws-logs-238665400611-us-east-1/elasticmapreduce/ 📁
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware
Availability zone: --
Subnet ID: subnet-b64bd098 ⧉
Master: Provisioning 1 m4.large
Core: Provisioning 2 m4.large
Task: --

Security and access
Key name: shibokeypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master:
Security groups for Core & Task:

Get the result in S3

Amazon S3 > bigdata0705 / HiveQueryResults / os_requests

Overview

🔍 Type a prefix and press Enter to search. Press ESC to clear.

⬆ Upload | ✚ Create folder | Download | Actions ⌄

☐ **Name** ↑≡

☐ 📄 000000_0

○ ○ ○   📄 000000_0 (1) ⌄

```
Android855
Linux813
MacOS852
OSX799
Windows883
iOS794
```

Summary

In this tutorial, I learned the mechanism and structure of AWS EMR. With the help of it, I finished the data process by the Hadoop.