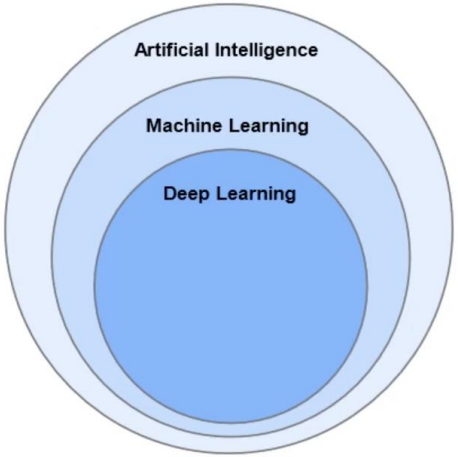
Machine Learning and AI Services



What is Artificial Intelligence (AI)?
Machines that perform jobs that mimic human behavior

What is Machine Learning (ML)?
Machines that get better at a task without explicit programming

What is Deep Learning (DL)?
Machines that have an artificial neural network inspired by the human brain to solve complex problems.

Amazon SageMaker is a fully managed service to **build, train, and deploy machine learning models** at scale

- **Apache MXNet on AWS**, open-source deep learning framework
- **TensorFlow on AWS** open-source machine intelligence library
- **PyTorch on AWS** open-source machine learning framework

Amazon SageMaker Ground Truth is **data-labeling service**. Have humans label a dataset that will be used to train machine learning models

Amazon Augmented AI human-intervention review service. When SageMaker's uses machine Learning to make a prediction is not confident it has the right answer queue up the predication for human review.

Machine Learning and AI Services

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

Amazon CodeGuru is **machine-learning code analysis service**. CodeGuru performs code-reviews and will suggest changes to improve the quality of code. It can show visual code profiles (show the internals of your code) to pinpoint performance.

Amazon Lex is a **conversion interface service.** With Lex you can build **voice and text chatbots**

Amazon Personalize is a **real-time recommendations** service. Same technology used to make product recommendations to customers shopping on the Amazon platform

Amazon Polly is a **text-to-speech** service. Upload your text and an audio file spoken by synthesized voice is generated.

Amazon Rekognition is **image and video recognition service**. Analyze images and videos to detect and label objects, people, celebrities.

Amazon Transcribe is a **speech-to-text service**. Upload your audio file and it is converted

Amazon Textract and **OCR (extract text from scanned documents) service**. When you have paper forms and you want to digitally extract the data.

Amazon Translate **neural machine learning translation service**. Uses deep learning models to deliver more accurate and natural sounding translations.

Amazon Comprehend is a **Natural Language Processor (NLP) service**. Find relationships between text to produce insights. Looks at data such as Customer emails, support tickets, social media and makes predictions.

Machine Learning and AI Services

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

Amazon Forecast is a **time-series forecasting service**. Forecast business outcomes such as product demand, resource needs or financial performance.

AWS Deep Learning AMIs Amazon EC2 instances **pre-installed with popular deep learning frameworks** and interfaces such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, and Keras

AWS Deep Learning Containers Docker images instances pre-install with popular deep learning frameworks and interfaces such as TensorFlow, PyTorch, and Apache MXNet.

AWS DeepComposer is **machine-learning enabled musical keyboard**

AWS DeepLens is a **video-camera that uses deep-learning**.

AWS DeepRacer a **toy race car** that can be powered with machine-learning to perform **autonomous driving**.

Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to EC2 instances to reduce the cost of running deep learning inference by up to 75%.

Amazon Fraud Detector is a **fully managed fraud detection a service**. identify potentially fraudulent online activities such as online payment fraud and the creation of fake accounts.

Amazon Kendra **enterprise machine learning search engine service**. Uses natural language to suggest answers to question instead of just simple keyword matching

Big Data and Analytics Services

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

What is BigData?

A term used to describe **massive volumes of structured/unstructured data** that is so large it is difficult to **move and process** using traditional database and software techniques.

Amazon Athena is a **serverless interactive query service**. It can take a bunch of CSV or JSON files in a S3 Bucket and load them into temporary SQL tables so you can run SQL queries. ***When you want to query CSV or JSON files***

Amazon CloudSearch is a fully managed **full-text search service**. ***When you want add search to your website***

Amazon Elasticsearch Service (ES) is a **managed Elasticsearch cluster**. Elasticsearch is a open-source full-text search engine. It is more robust than CloudSearch but requires more server and operational maintaince.

Amazon Elastic MapReduce (EMR) is for data processing and analysis. Its can be used for creating reports just like Redshift, but is more suited when you need to transform unstructured data into structured data on the fly.

Kinesis Data Streams is a **real-time streaming data service**. Create **Producers** which send data to a stream. **Multiple Consumers** can consume data within a stream. Use for real-time analytics, click streams, ingesting data from a fleet of IOT Devices

Kinesis Firehose is serverless and a simpler version of Data Streams, You pay-on-demand based on how much data is consumed through the stream and you don't worry about the underlying servers.

Amazon Kinesis Data Analytics allows you to run queries against data that is flowing through your real-time stream so you can create reports and analysis on emerging data.

Amazon Kinesis Video Streams allows you to analyze or apply processing on real-time streaming video.

Big Data and Analytics Services

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

Managed Kafka Service (MSK) a **fully managed Apache Kafka service.** Kafka is an open-source platform for building real-time streaming data pipelines and applications. It is similar to Kinesis but with more robust functionalities

Redshift is a **petabyte-size data-warehouse.** Data-warehouses are for Online Analytical Processing) OLAP Data-warehouses can be expensive because they are keeping data "hot". Meaning that we can run a very complex query and a large amount of data and get that data back very fast.

When you to quickly generate analytics or reports from a large amount of data.

Amazon QuickSight is **business intelligence (BI) dashboard**. You can use it to create business dashboards to power business decisions. It requires little to no programming knowledge and connect and ingest to many different types of databases

AWS Data Pipeline **automates the movement of data**. You can reliably move data between compute and storage services.

AWS Glue is an **Extract, Transform, Load (ETL) service**. Moving data from one location to another and where you need to perform transformations before the final destination. Similar to Database Migration Service (DMS) but more robust

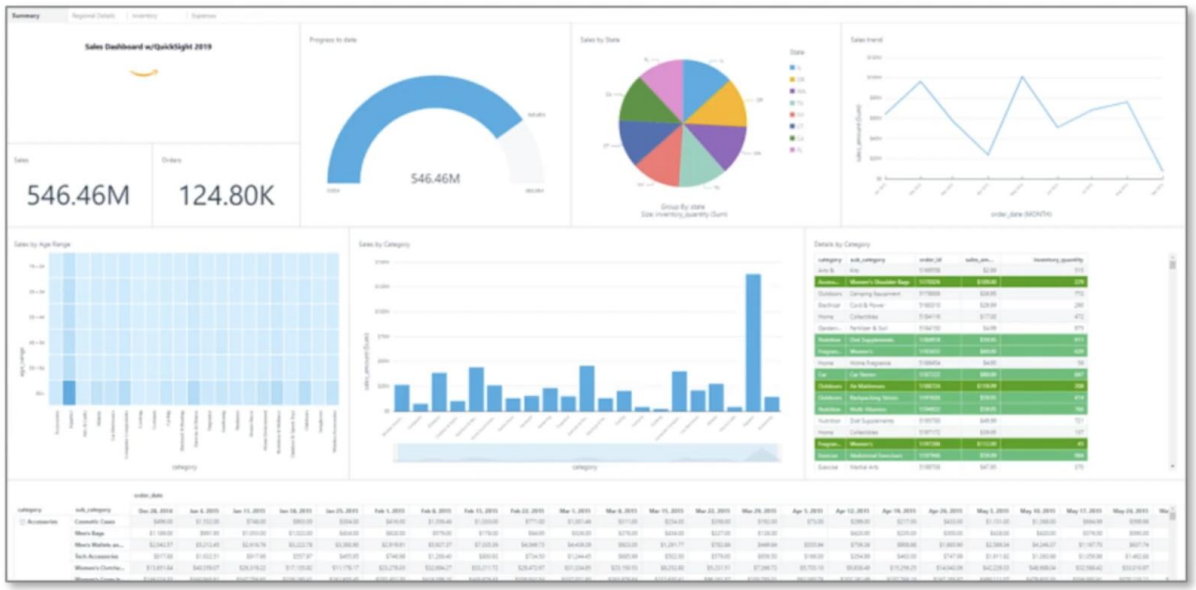
AWS Lake Formation is as a **centralized, curated, and secured repository that stores all your data.**
A **data lake** is a storage repository that holds a vast amount of raw **data** in its native format until it is needed.

AWS Data Exchange is a catalogue of third-party datasets. You can download for free subscribe or purchase datasets.
Eg. COVID-19 Foot Traffic Data, IMDB TV and Movie data, Historical Weather Data

Amazon QuickSight

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

Amazon QuickSight is a **Business Intelligence (BI) Dashboard** that allows you to ingest data from various AWS storage or database services to **quickly visualize business data** with minimal programming or data formula knowledge.



QuickSight uses **SPICE** (super-fast, parallel, in-memory, calculation engine) to achieve blazing fast performance at scale

Amazon QuickSight ML Insights – Detect Anomalies, Perform accurate forecasting, Generate Natural Language Narratives.
Amazon QuickSight Q - Ask question using natural language, on all your data, and receive answers in seconds.

Machine Learning and AI Services – Extended

Amazon Bedrock
A **Large Language Model (LLM) cloud service offering to** generate text and image responses. ***Think like ChatGPT***

Amazon CodeWhisper
An AI code generator that will predict code to meet your usecase. Think like ***Github Copilot***

Amazon DevOps Guru
Uses ML to analyze your operational data and application metrics and events to detect operational abnormalities.
Is there something wrong with our cloud operations?

Amazon Lookout for Equipment / Metrics / Vision
Uses ML models for quality control and performed automated inspections.

Amazon Monitron
Uses ML models to predict unplanned equipment downtime. Monitor has an IOT sensor that captures vibrations and sensor data

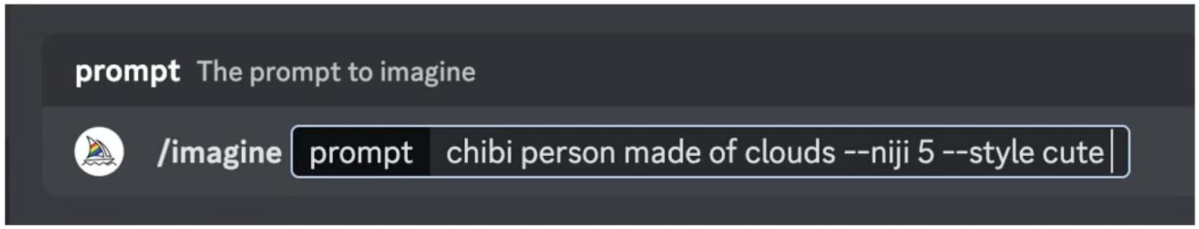
AWS Neuron
An SDK used to run deep learning workloads on AWS Inferentia and AWS Trainium based instance

Generative AI

What is Generative AI (Gen AI)?

Gen AI is a type of artificial intelligence **capable of generating new content**, such as **text**, **images**, **music**, or **other forms** of media.

Example of **Midjourney** generating a graphic



ML and DL Frameworks and Tools

Here are some common Machine Learning and Deep Learning Frameworks
Most of which can be used within SageMaker, or have direct support

Apache MXNet – adopted by AWS, supports both imperative and symbolic
PyTorch — optimized tensor library for deep learning using GPUs and CPU (created by Facebook)
TensorFlow — low-level machine learning framework (created by Google)

- **Keras** — high-level machine learning framework built on top and ships with TensorFlow

Apache Spark — unified analytics engine for large-scale data processing

- **SparkML** — uniform set of high-level APIs that help users create and tune practical machine learning pipelines

Chainer — powerful, flexible and intuitive deep learning framework, supports CUDA

Hugging Face — An AI Community of ML Models and dataset

ML frameworks which you might come across but are no longer in active development have been absorbed or abandon or the community just doesn't really use them: *Caffe2, Theano, DSSTNE, CNTK*

Apache MXNet

Apache MXNet is **a deep learning machine learning framework** which supports many different programming languages:

- Python, Java, Julia, Matlab, R, Javascript, Go, R, Scala, Perl, Wolfram langauge

They key features of Apache MXNet:

- **Scalable** — designed for distributed cloud infrastructure
- **Flexible** — supports both imperative and symbolic programming
- **Portable** — can be used on low-end or edge devices and on serverless compute
- **Multiple** Programing languages

AWS has made Apache **MXNet their ML framework of choice**
There is lots of support to use Apache MXNet within AWS SageMaker and AWS ML containers.

MXNet has two high-level interfaces:
Gluon API — **imperative** programming
Module API — **symbolic** programming

Very simple example of a Neural Next in Apache MXNet's Gluon API

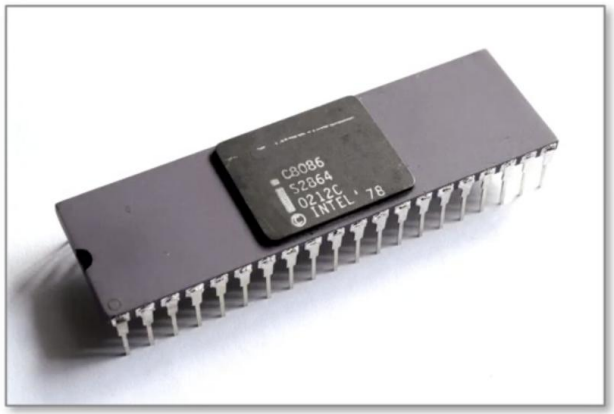
```
from mxnet import nd
from mxnet.gluon import nn

layer = nn.Dense(2) # dense layer with 2 output units
layer.initialize() # initialize its weights with the default initialization method
x = nd.random.uniform(-1,1,(3,4))
layer(x) # forward pass with random data
layer.weight.data() # access the weight after the first forward pass
```

What is Intel?

What is Intel?

Intel is a multinational corporation and is one of the world's largest semiconductor chip manufacturers. Intel is the inventor of the **x86 instruction set**

Intel 8086 chip from 1978




```
section .data
    num1 db 5 ; define byte with value 5
    num2 db 3 ; define byte with value 3

section .bss
    result resb 1 ; reserve byte for result

section .text
    global _start

_start:
    mov al, [num1] ; move num1 into al
    add al, [num2] ; add num2 to al
    mov [result], al ; move the sum into result

    ; Exit the program
    mov eax, 1 ; syscall number for exit
    mov ebx, 0 ; status
    int 0x80    ; call kernel
```

Example of **x8606 Assembly language**

There is another popular instruction set called **ARM** which uses fewer instructions and usually results better power efficiency which results in lower costs.

Intel Xeon Scalable and Intel Gaudi

Intel Xeon Scalable Processor

The Intel Xeon Scalable Processor is a high-performance CPU designed for enterprise and server applications, commonly used in AWS instances

Intel Habana Gaudi

AI training processor developed by Habana Labs, a company acquired by Intel. Gaudi is tailored for training deep learning models. Gaudi is often viewed as a competitor to NVIDIA's GPUs. While NVIDIA's GPUs are highly versatile and widely adopted, Gaudi offers a specialized alternative that's optimized specifically for AI training.

What is a GPU?

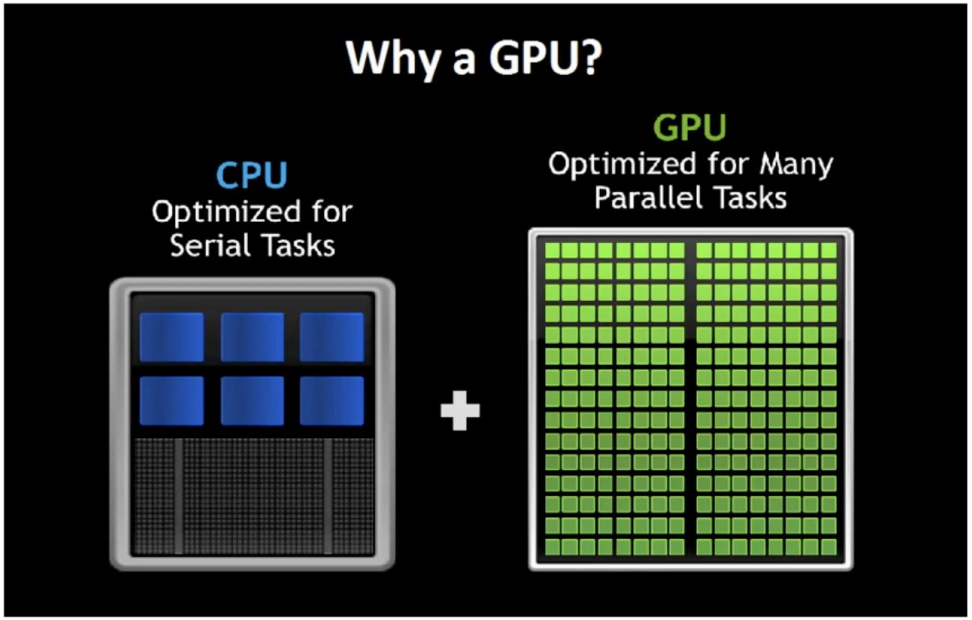
What is a GPU?
A General Processing Unit (CPU) is a processor that is specialized to quickly render high-resolution images and video **concurrently**. **Great for Video Games**

GPUs can perform parallel operations on multiple sets of data, and so they are commonly used for non-graphical tasks such as machine learning and scientific computation.

CPU can have average 4 to 16 processor cores…
GPUs can **thousands of processor cores**
4 to 8 GPUs can provide as many as 40,000 cores

GPUs are best suited for repetitive and highly-parallel computing tasks:

- Rendering graphics
- Cryptocurrency mining
- Deep Learning and ML



What is CUDA?

What is NVIDA?

NVIDA is a company that manufactures **graphical processing units (GPUs)** for gaming and professional markets

What is CUDA?

Compute Unified Device Architecture (CUDA) is a **parallel computing platform** and **API** by NVIDIA that allows developers to use **CUDA-enabled GPUs** for general-purpose computing on GPUs (GPGPU)

All major deep learning frameworks are integrated with **NVIDIA Deep Learning SDK**

EC2 P3 Instances have up to 8 NVIDIA **Tesla V100** GPUs
EC2 G3 Instances have up to 4 NVIDIA **Tesla M60** GPUs
EC2 G4 Instances have up to 4 NVIDIA **T4** GPUs
EC2 P4 Instances have up to 8 NVIDIA **Tesla A100** GPUs

The NVIDA Deep Learning SDK is a collection of NVIDIA libraries for deep learning.
One of those libraries is the **CUDA Deep Neural Network library (cuDNN)**
cuDNN provides highly tuned implementations for standard routines such as:

- forward and backward convolution
- Pooling
- Normalization
- activation layers