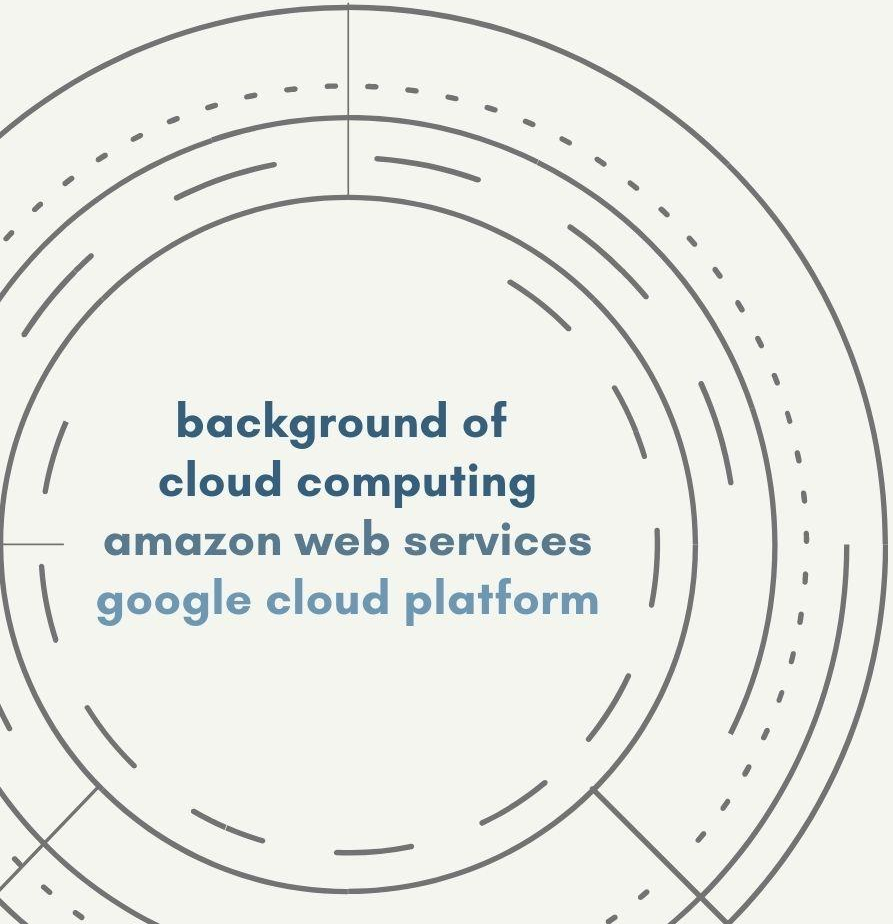
team jsa ○ jingchao zhou + simran regmi

INTRO TO CLOUD COMPUTING + MACHINE LEARNING

WEEK EIGHT

COVERED TODAY



RELEVANT CONCEPTS

- learn how to use PyTorch & Keras on cloud platforms
- implement and train a CNN model on the MNIST dataset
- set up a Jupyter notebook on GCP
- use SageMaker on AWS
- import data from & store data to S3

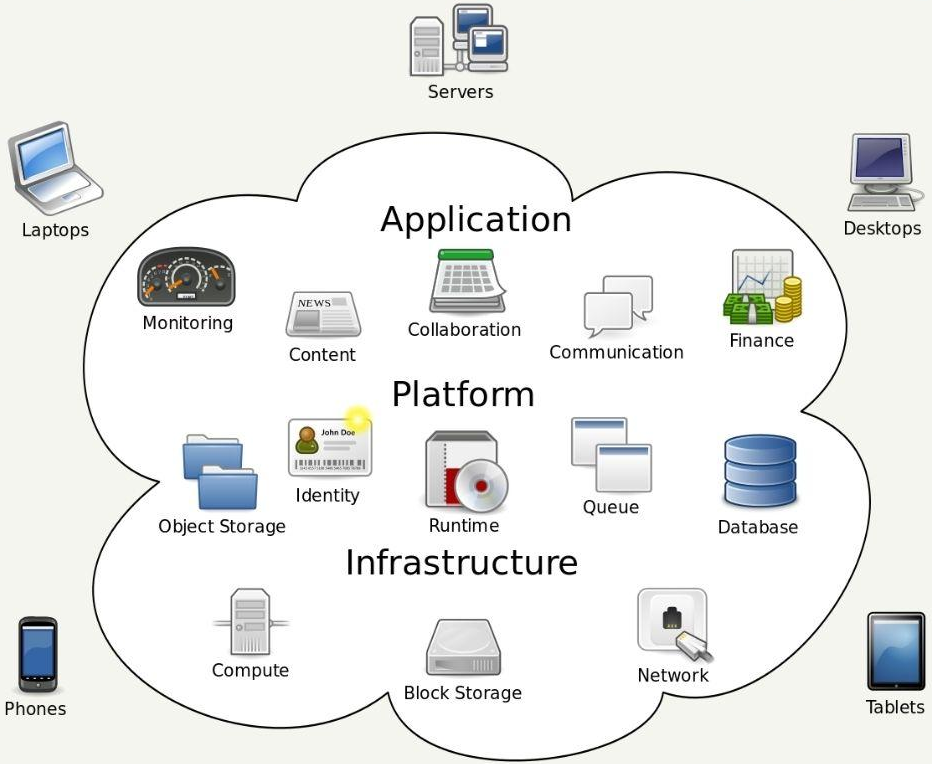
PREREQUISITES

1. cs61a or equivalent programming knowledge
2. module 23: hyperparameter tuning
3. module 24: cnns

BACKGROUND

BACKGROUND: WHAT IS CLOUD COMPUTING

on-demand delivery of computing services, such as servers, databases, intelligence, over the Internet



BACKGROUND: CLOUD SERVICE PROVIDERS

AMAZON WEB SERVICES

- pay as you go basis pricing structure
- current services offered: EC2, Aurora, VPC, etc
- more than a million active customers
- holds 32% of the cloud market share

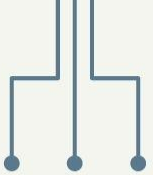
GOOGLE CLOUD PLATFORM

- variety of pricing options (usage, combined, etc)
- current services offered: dialogflow, automl tables, video/vision ai
- runs on the same infrastructure that Google uses internally for its end-user products

MICROSOFT AZURE

- pay as you go basis pricing structure
- current services offered: data science virtual machines, anomaly detector
- offers hybrid cloud solutions

BACKGROUND: CLOUD COMPUTING ADVANTAGES

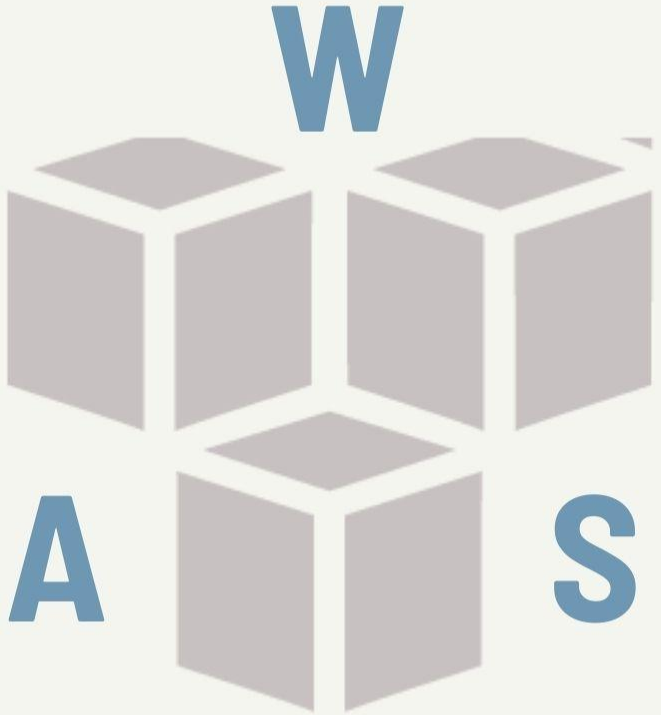


1. the cloud's pay-per-use model is good for bursty AI or machine learning workloads
2. the cloud makes it easy for enterprises to experiment w ML capabilities & scale up as projects go into production & demand increases
3. the cloud makes intelligent capabilities accessible w/o requiring advanced skills in AI or data science
4. AWS, Microsoft Azure, and GCP offer many ML options that don't require deep knowledge of AI, ML theory, or a team of data scientists

ASSESSMENT

complete part 1 of the quiz

AMAZON WEB SERVICES



AWS: MACHINE LEARNING TOOLS

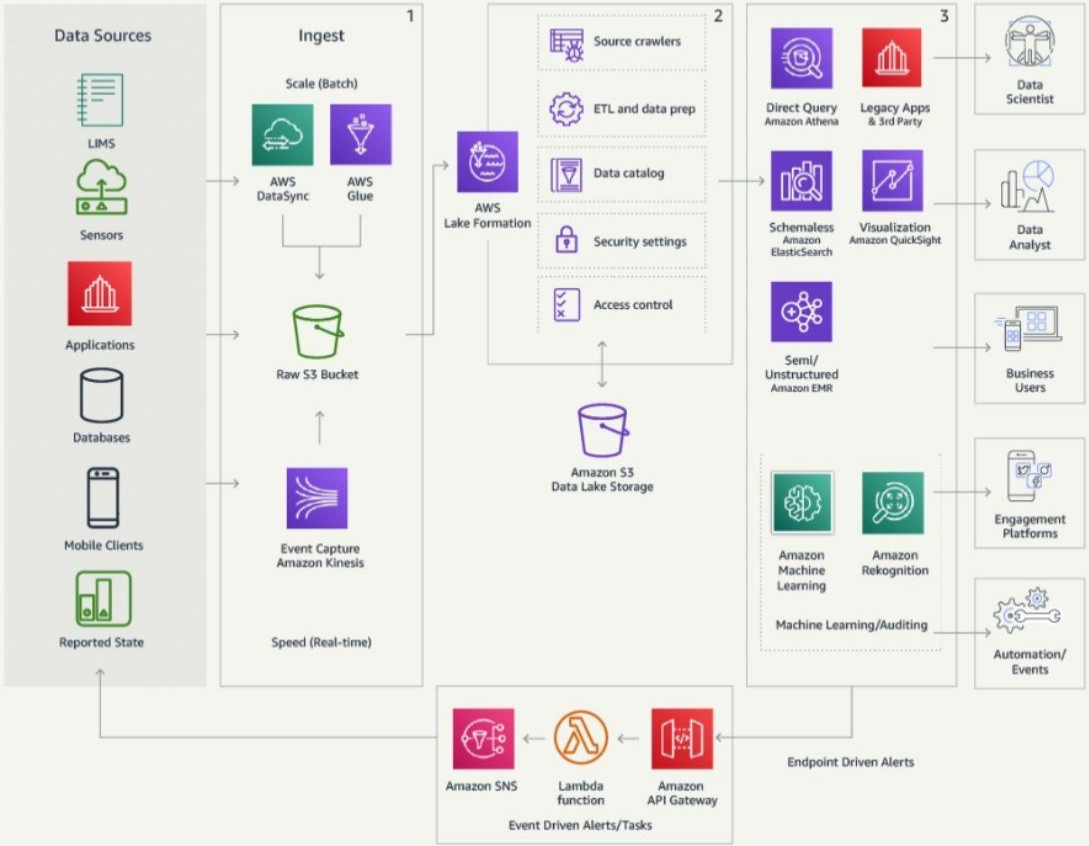
sagemaker: a fully managed machine learning service that allows oneself to build and train machine learning models

rekognition: a cv software that allows one to identify a person in a photo or video using private repositories of face images

s3: a data storage service that has the capability to store the dataset that powers our model training process

emr: a big data tool that processes vast amounts of data using open source tools such as Apache Spark

AWS: SAMPLE MACHINE LEARNING WORKFLOW



AWS SAGEMAKER SETUP VIDEO

MANDATORY

https://youtu.be/s1M8P9X6j_8

OPTIONAL

https://www.youtube.com/watch?v=pfjhNe1M2t4

P Y T O R C H

open source machine learning
library based on the Torch library

relevant packages + functions

- **torch.optim** – implementing various optimization algorithms
- **torch.distributed** – provides PyTorch support and communication primitives for multiprocess parallelism across several computation nodes running on one or more machines
- **torch.nn.functional** – applies a 1D convolution over an input signal composed of several input planes

WORK ON ASSIGNMENT ONE

assignment length: four hours

ASSESSMENT

complete part 2 of the quiz

GOOGLE CLOUD PLATFORM

GCP: MACHINE LEARNING TOOLS

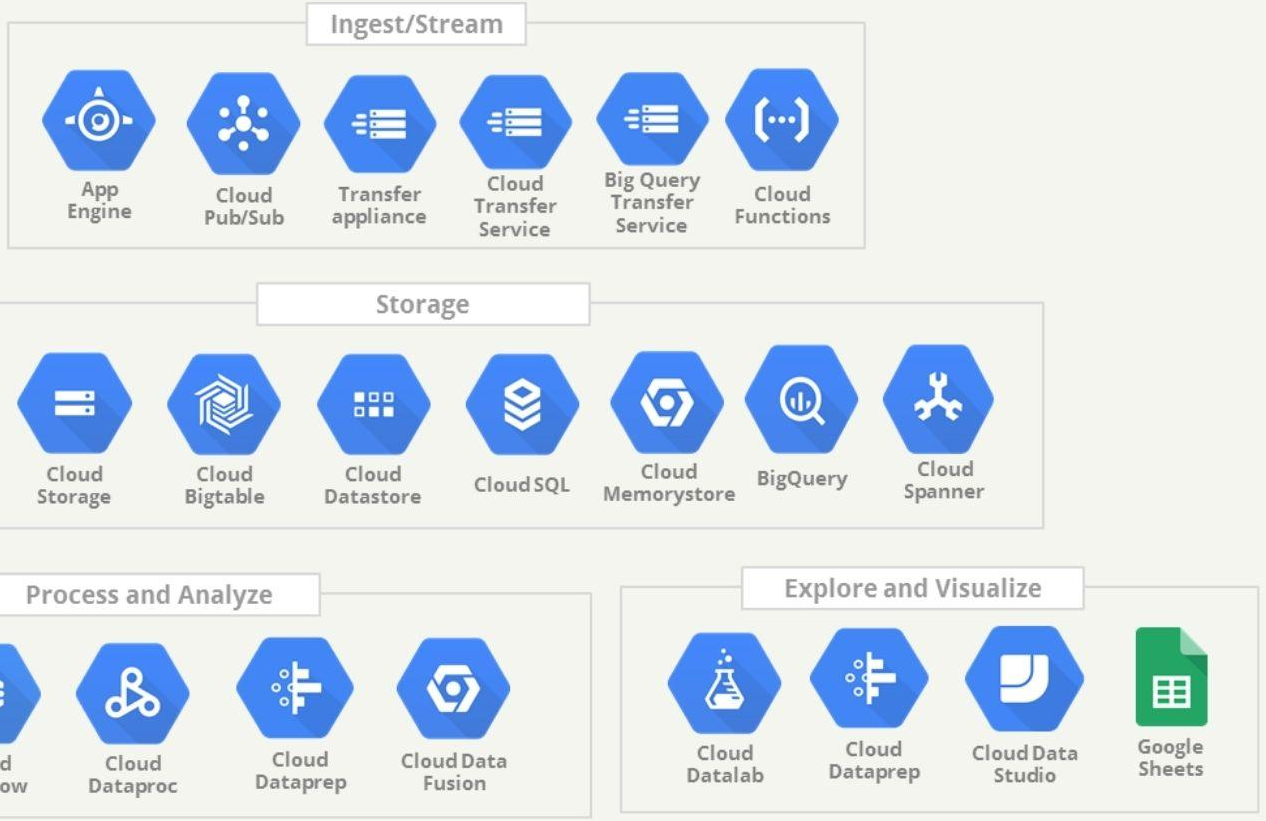
data labeling service: get highly accurate labels from human labelers for better machine learning models

deep learning vm image: instantiate a vm image containing the most popular AI frameworks on a Compute Engine instance

ai explanations: understand how each feature in your input data contributed to model's outputs

compute engine: Infrastructure as a Service component that enables users to launch virtual machines on demand

GCP: SAMPLE MACHINE LEARNING WORKFLOW



K E R A S

open-source library that provides a Python interface for artificial neural networks

relevant packages + functions

- **sequential model** - appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor
- **tf.keras.losses.SparseCategoricalCrossentropy** - computes the crossentropy loss between the labels and predictions
- **tf.keras.layers** - a layer encapsulates both a state & a transformation from inputs to outputs

WORK ON ASSIGNMENT TWO

assignment length: four hours

ASSESSMENT

complete part 3 of the quiz