The cloud and why it is important

“There is no cloud, it’s just someone else’s computer”

Cole Clifford

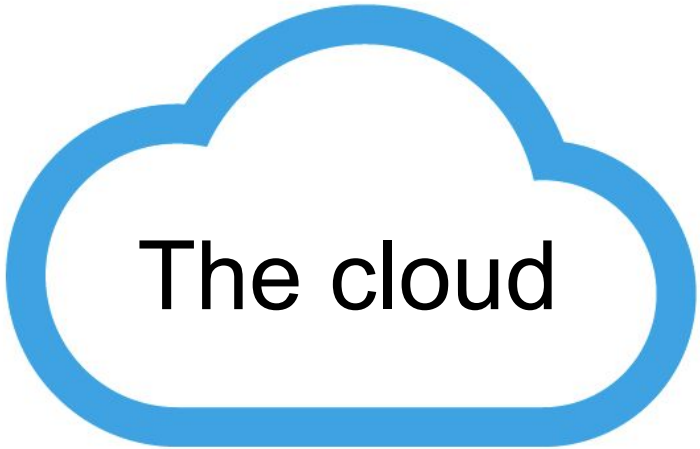
- Technical Product Manager at Dessa
- 3 years of building enterprise ML/DL systems from development to production
- Lead our Production Level Deep Learning meetup
- Worked on space2vec — a record beating supernova detection model

Preface

- Use this as an information resource
- NOT homework to know this by heart!
- There is A LOT to think about when talking about infrastructure
- This may seem overwhelming at first
- Different content will be important in different scenarios
  - Doing a quick test
  - Running a personal experiment
  - Running a hyperparameter search
  - Building a model for a production application
  - Running models for a class project

Agenda

- The cloud
- Why is it important?
- Comparison factors


The cloud

Compute environments

- Hardware and software that your code runs within
- Hardware to care about
  - Storage (SSD, Spinny Disk)
  - RAM
  - CPU
  - GPU ***
  - Network
- Software
  - Operating system
  - Packages/libraries
  - Tools

The GPU

- Graphics Processing Unit
- Originally used for gaming
- Shapes in a game are matrices
- Movement of shapes are matrix math
- Neural networks are matrix math
- GPUs are AMAZING for ***training*** neural networks
- Not necessarily needed for inference

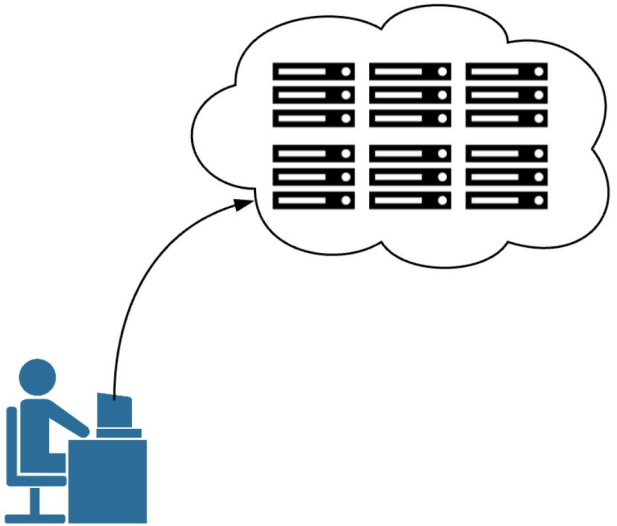
Local vs remote machines

- Local refers to a compute environment that you are sitting in front of
  - Laptop or desktop
  - Usually not that powerful
- Remote refers to a compute environment you are ***not*** in front of
  - Server or cloud
  - Usually more powerful resources
  - Usually doesn’t have GPUs… until recently

What is the cloud?

Local Compute Environment

Remote Compute Environment



Enterprise cloud (on premises)

- Hosted in a server farm that your organization owns
- Internal team supporting the machines
- Shared development environment for team members
- Have to request for changes/upgrades
  - Installation of new packages, libraries, and tools
  - More compute or storage
- Software and hardware lag behind industry standard
- Enterprise security is fighting against you in every way

Commercial cloud

- Hosted in a server farm that the cloud provider owns
- Machine support is a given
- Shared development environment for team members
- Do not have to request for changes/upgrades
  - Installation of new packages, libraries, and tools
  - More compute or storage
- Can easily use new hardware and software
- Enterprise security is fighting against you in every way

The big players

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure
- Digital Ocean

Why is it important?

Scenario - training a large network

- Want to train StyleGAN
- 41 days on a decent laptop
- 3 days on a DGX
  - 4 Tesla V100 GPUs
  - >$100,000 upfront cost
- 3 days on AWS p3.8xlarge instance
  - 4 Tesla V100 GPUs
  - ~$881.28/hour

Scenario - batch model in production

- Inference runs on the first day of the month
- Takes 3 hours to run using a GPU
- ~$3/month on AWS
- >$1000 upfront for equivalent machine
- ~27 years of production on AWS

What do you get?

- Really powerful machines
- On demand compute
- Peace of mind on hardware
- Free support
- Managed services that provide extra functionality

Mass compute

- Quickly use some of the world's most powerful machines
- Only use when and what you need
- Code can easily be moved between machines
- Code can move onto the cloud when ready
- Many different hardware configurations

System support

- Network security is a lot easier
- Instance security is a lot easier
- Instances that break will be fixed for you

Pricing

- Can be tricky
- Can be cheap for short bursts
- Instance pricing depends on
  - Hardware in the instance (CPUs, RAM, GPUs, storage, networking cards)
- Other cloud services will have different pricing
  - External, distributed storage (AWS S3, GCP Storage Buckets)
  - Managed Kubernetes clusters (AWS EKS, GCP GKE)
  - Managed log stores (AWS CloudTrail)

Google Cloud Platform Pricing Calculator

Prices are up to date. Last update: 19-July-2019

COMPUTE ENGINE | APP ENGINE | KUBERNETES ENGINE | CLOUD RUN | CLOUD STORAGE | NETWORKING | BIGQUERY

Search for a product you are interested in.

Instances

Number of instances *

What are these instances for?

Operating System / Software

Free: Debian, CentOS, CoreOS, Ubuntu, or other User Provided OS

Machine Class

Regular

Machine Family

General purpose

Machine type

Estimate [1]

Compute Engine

1 x

730 total hours per month

VM class: regular

Instance type: n1-standard-1

Region: Montréal, Canada

Total available local SSD space 1x375 GB

Sustained Use Discount: 30%

Effective Hourly Rate: USD 0.081

Estimated Component Cost: USD 59.29 per 1 month

Total Estimated Cost: USD 59.29 per 1 month

Estimate Currency

USD - US Dollars

Managed services

- Every cloud has services to make things easier
  - Distributed data buckets
  - Hosted databases
  - Messaging services
  - Monitoring services
  - Identity management
  - Queueing
  - Serverless functions
  - Machine learning
  - Virtual reality
  - *So much more… they just want you on their cloud*

Watch out for “lock in”!

- All of these services are amazing
- When used right, the cloud won’t lock you in
- Relying on services will lock you in
- Treat things as APIs
- Write code that is seperate from the services used
- Only interface with these services

Comparison factors


TO CLOUD OR
NOT TO CLOUD?
THAT ISN'T THE QUESTION...

Is the cloud right for this project?

- It isn’t always the best solution
- Important to note the pros and cons
- But what are the factors to compare against?

Factors - organization support

- Previous usage
  - Does my organization support the cloud?
  - If no proper support, will I get in trouble if I start using it?
  - Who needs to know that the cloud is being used?
  - How easy is it to use the cloud that my organization has?
  - How easy is it for me to get access to this cloud?

Factors - data storage

- Size of data
  - How much data do I need to store?
  - Do I have enough room to store the data in the organization?
  - How much will it cost to store and move my data?
- Sensitivity of data
  - Are there private features in the dataset?
  - Are there legal repercussions to storing data in certain locations?
  - How do I anonymize my data (e.g. hipaa compliance)?
  - Do I have to worry about access restrictions to the data?
- Location
  - Where will the data sit?
  - Where will I be using the data?
  - Do I have to worry about other country regulations (e.g. GDPR in the EU)

Factors - security

- Access to the system
  - Do I need to setup identity management?
  - Do I need to setup user accounts?
  - Do I need to setup networking for my machines?
- Organization compliance policies
  - Does my organization have any security compliance policies?
  - Do I need to run any security scans against my system?
- Data security
  - How do I get data in and out of the system?
  - *All questions from the previous slide*

Factors - budget

- Budget for the project
  - How much money do I have for the project?
  - Do the required services allow me to stay within the budget?

Factors - compute resources

- Required hardware
  - Do I really know what hardware I need?
  - Do I need the same amount of hardware at all times or just in bursts?
  - Does my organization have the hardware that I need? (e.g. GPUs)
  - How long would it take to get the hardware internally?

Factors - networking

- Speed
  - Is the organization network fast enough for me to do my work?
  - Does the data storage location and data consumption location have a direct connection?
- Security
  - Do I have to worry about external access to the network?
  - Do I have to worry about the security of data moving through the network?

Factors - available services

- Speed of setup
  - Are there any services that I could use that will get this done quicker?
  - Can I use services to ensure the resilience and longevity of the project?
- Lock in
  - Can I use these services in a way that allow me to easily leave later?

Factors - system reliability

- System support
  - How important is this system?
  - If it breaks, how quickly does it need to be up and running again?
  - Who will support this system?
  - How much do I trust the machines that the system is on?

Next steps - the workshop