Introduction to AWS

Cloud Computing

Amazon Web Services (AWS) is a *cloud computing* service offering on-demand, pay-as-you-go computing power.

Cloud computing services offer access to compute resources, such as data storage or computing power without requiring direct active management by the user.

Cloud services take advantage of economies of scale to charge very little for the services they offer

Cloud Computing

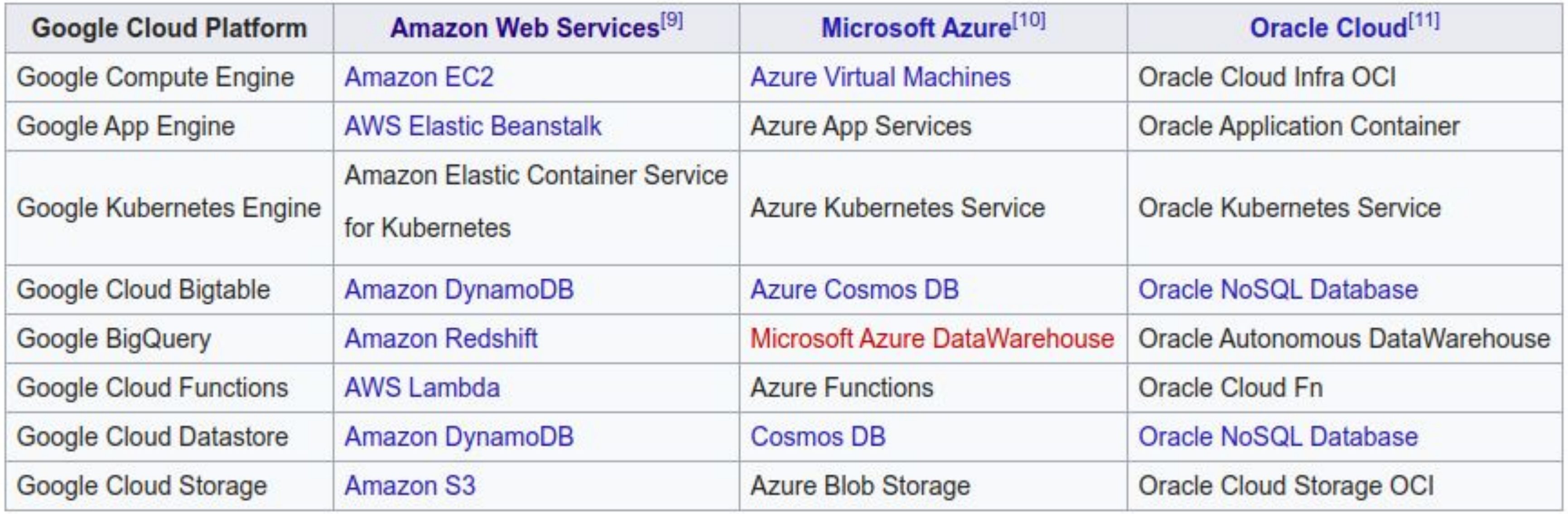
Example: Setting up a Hadoop cluster can take weeks to do.

As an alternative, amazon EMR (Elastic MapReduce) lets you start working on as large a Hadoop or Spark cluster as needed in minutes.

Cloud Computing

Biggest Cloud Computing Services:

- Amazon Web Services
- Google Cloud
- Microsoft Azure
- Oracle Cloud

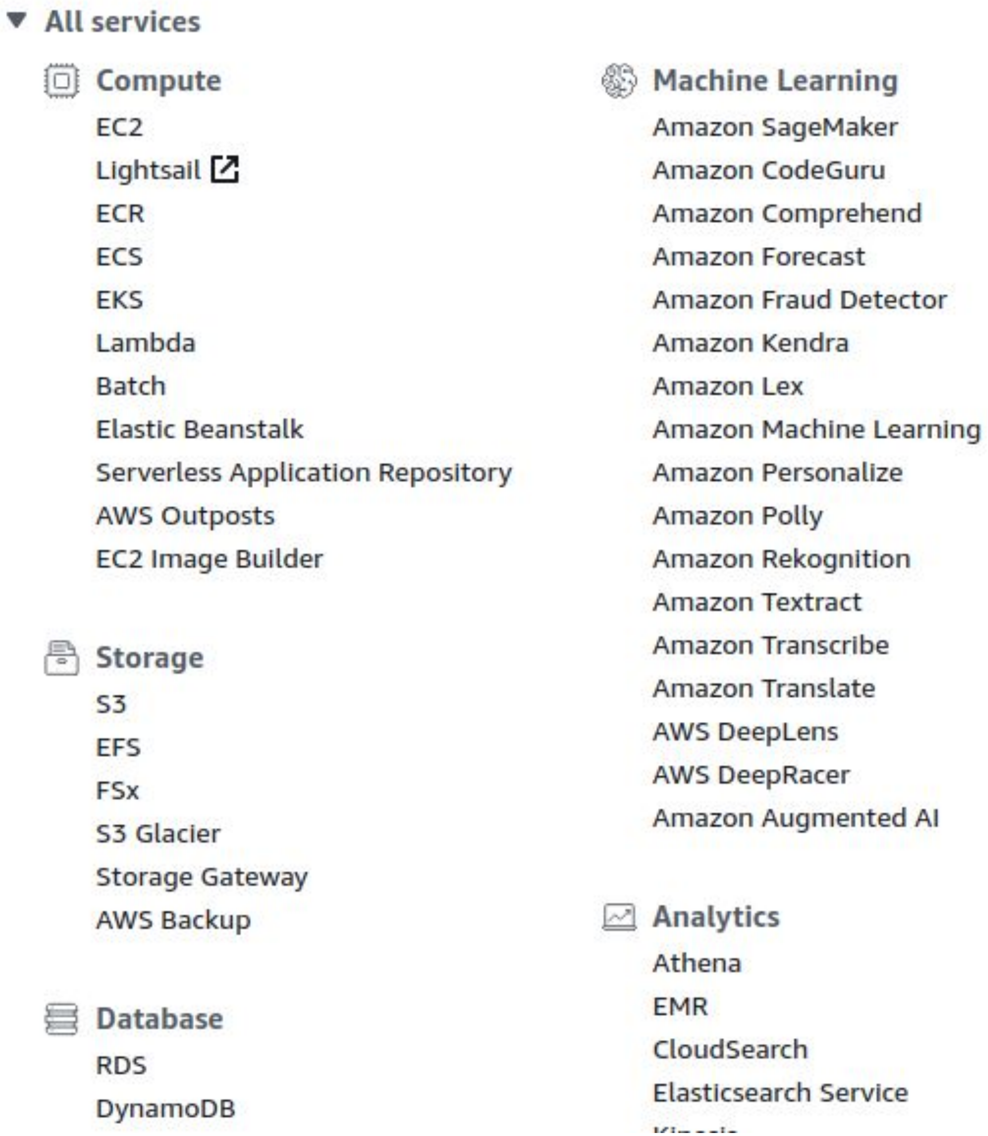


Google Cloud Platform	Amazon Web Services[9]	Microsoft Azure[10]	Oracle Cloud[11]
Google Compute Engine	Amazon EC2	Azure Virtual Machines	Oracle Cloud Infra OCI
Google App Engine	AWS Elastic Beanstalk	Azure App Services	Oracle Application Container
Google Kubernetes Engine	Amazon Elastic Container Service for Kubernetes	Azure Kubernetes Service	Oracle Kubernetes Service
Google Cloud Bigtable	Amazon DynamoDB	Azure Cosmos DB	Oracle NoSQL Database
Google BigQuery	Amazon Redshift	Microsoft Azure DataWarehouse	Oracle Autonomous DataWarehouse
Google Cloud Functions	AWS Lambda	Azure Functions	Oracle Cloud Fn
Google Cloud Datastore	Amazon DynamoDB	Cosmos DB	Oracle NoSQL Database
Google Cloud Storage	Amazon S3	Azure Blob Storage	Oracle Cloud Storage OCI

AWS Offerings

AWS offers a ton of different types of services.

In these slides, we'll look at a few of the common ones that you may encounter in the data science field.



▼ All services

Compute

EC2
Lightsail
ECR
ECS
EKS
Lambda
Batch
Elastic Beanstalk
Serverless Application Repository
AWS Outposts
EC2 Image Builder

Storage

S3
EFS
FSx
S3 Glacier
Storage Gateway
AWS Backup

Database

RDS
DynamoDB

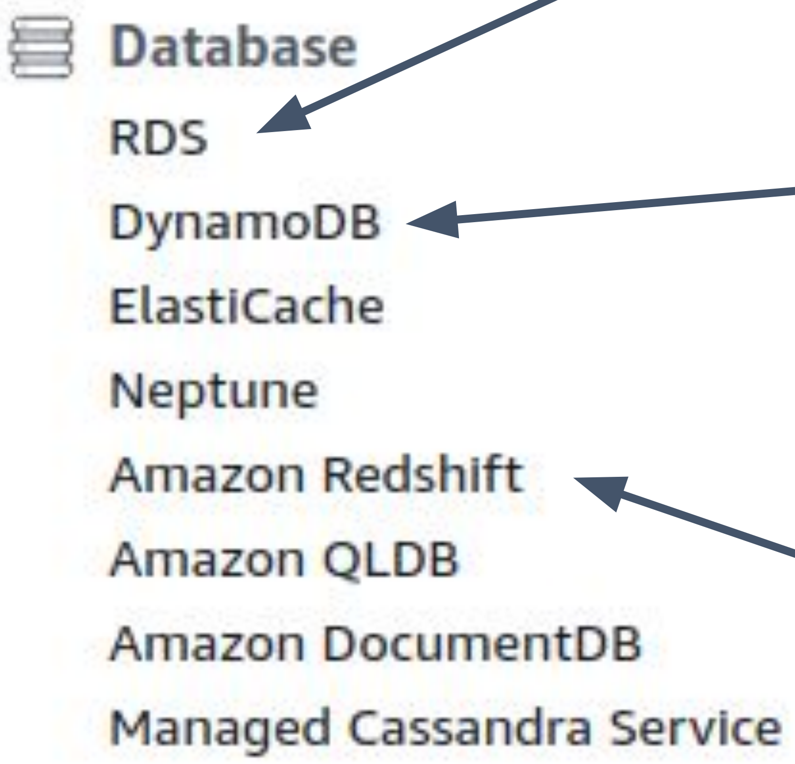
Machine Learning

Amazon SageMaker
Amazon CodeGuru
Amazon Comprehend
Amazon Forecast
Amazon Fraud Detector
Amazon Kendra
Amazon Lex
Amazon Machine Learning
Amazon Personalize
Amazon Polly
Amazon Rekognition
Amazon Textract
Amazon Transcribe
Amazon Translate
AWS DeepLens
AWS DeepRacer
Amazon Augmented AI

Analytics

Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis

AWS Data Stores



Database

RDS

DynamoDB

ElastiCache

Neptune

Amazon Redshift

Amazon QLDB

Amazon DocumentDB

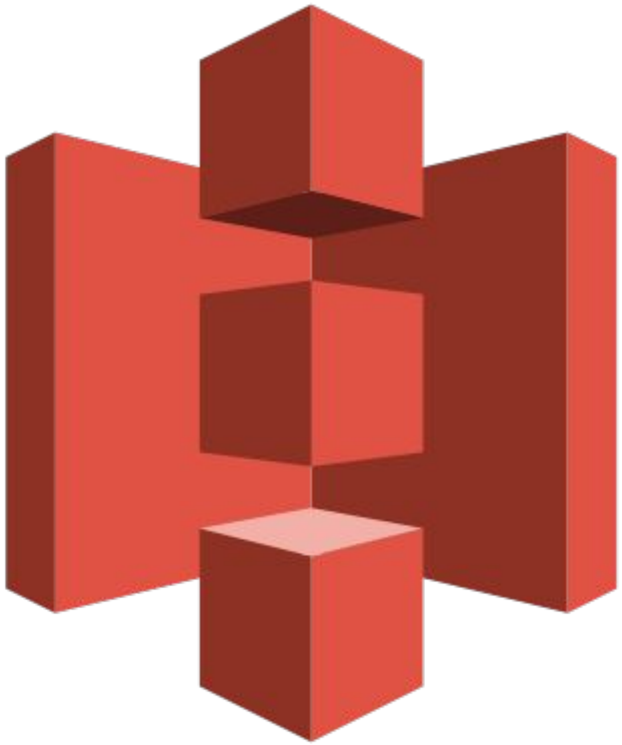
Managed Cassandra Service

RDS = Relational Database Services (SQL Databases)

NoSQL database. Key-value and document storage

Large-scale columnar data warehouse. Similar to SQL, but more efficient for analytics queries

AWS S3



AWS S3 (**Simple Storage Service**) is a scalable storage service that offers low latency and high availability.

Can be used to store any type of file, so is good for holding unstructured or semi-structured data.

Objects are organized into *buckets*.

AWS S3

There are a number of ways to interact with S3:

- In the browser via AWS Management Console (we'll see this on Saturday)
- The AWS command line interface CLI
- Python's *boto3* library

We'll look at option 2 first.

First, you need to download and configure the CLI: https://docs.aws.amazon.com/cli/latest/userguide/install-cliv2.html

AWS S3

Once you’ve run the installer, check that it is installed by running

```
$ aws --version
```

Next, you need to configure your access keys. Do this by running

```
$ aws configure
```

Input the ID and Secret Key shared with you. Set the default region name as us-east-1 and the output format as json.

AWS S3

There are numerous open datasets available on AWS: https://registry.opendata.aws/

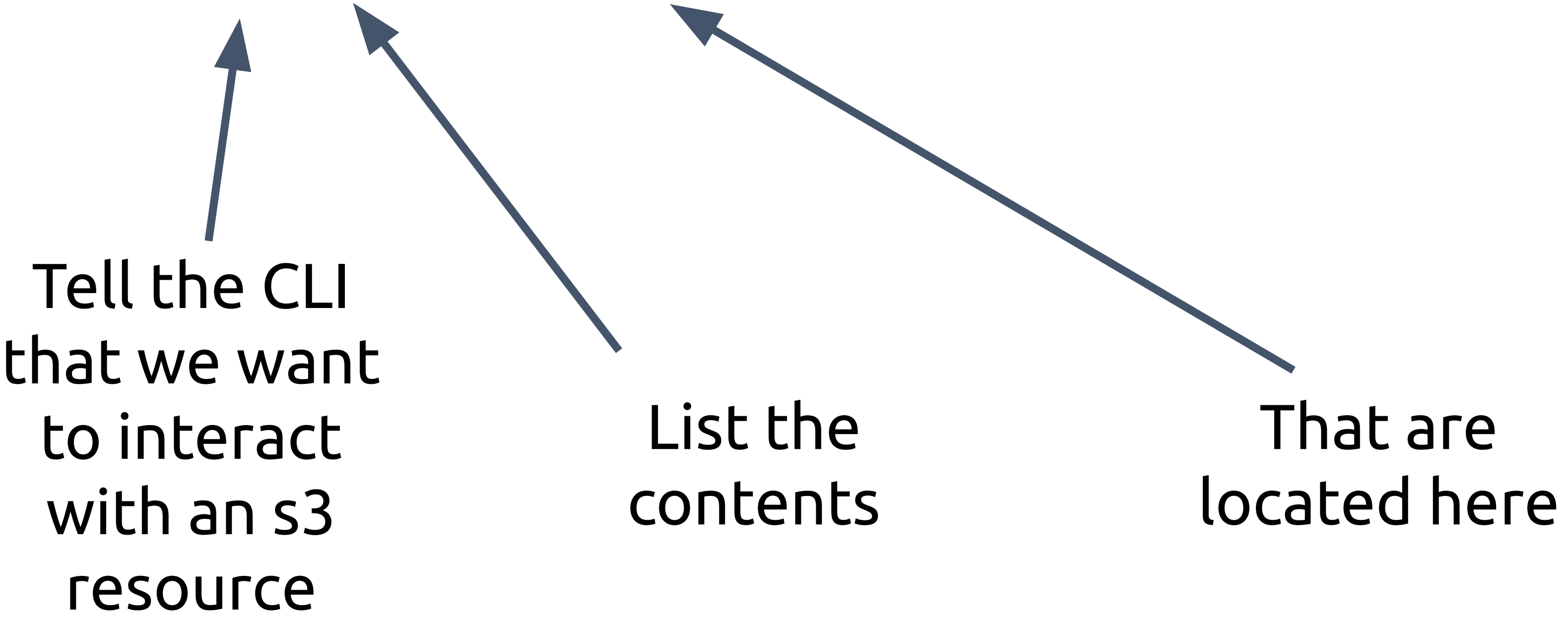
Let’s look at the New York City Taxi and Limousine Commission (TLC) Trip Record Data (https://registry.opendata.aws/nyc-tlc-trip-records-pds/)

AWS S3

This data is contained in a bucket named nyc-tlc

To see what is in this bucket, use

```
$ aws s3 ls s3://nyc-tlc
```



Tell the CLI that we want to interact with an s3 resource

List the contents

That are located here

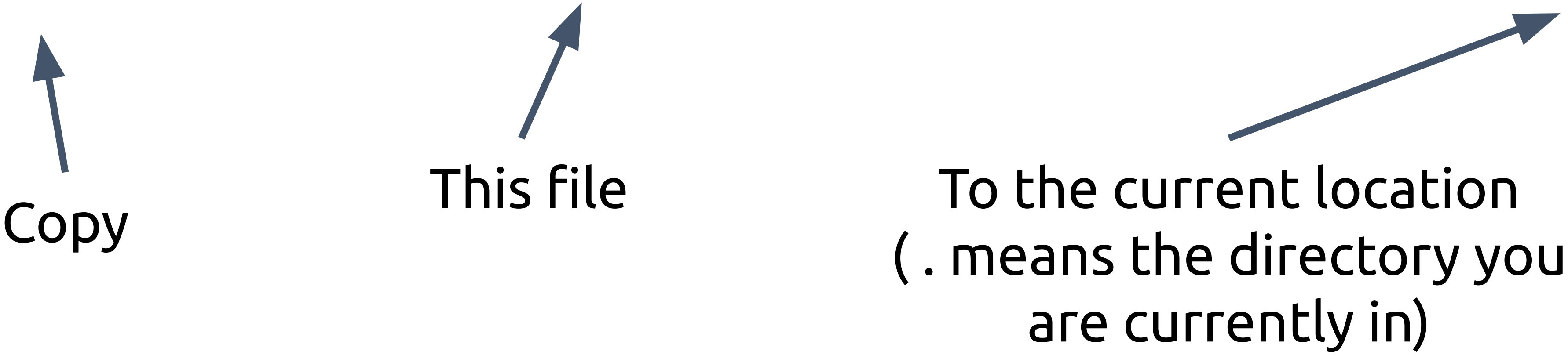
AWS S3

Let’s see what kind of trip data is available.

```
$ aws s3 ls s3://nyc-tlc/trip\ data/
```

Let’s grab the file for December, 2019:

```
$ aws s3 cp s3://nyc-tlc/trip\ data/yellow_tripdata_2019-12.csv .
```

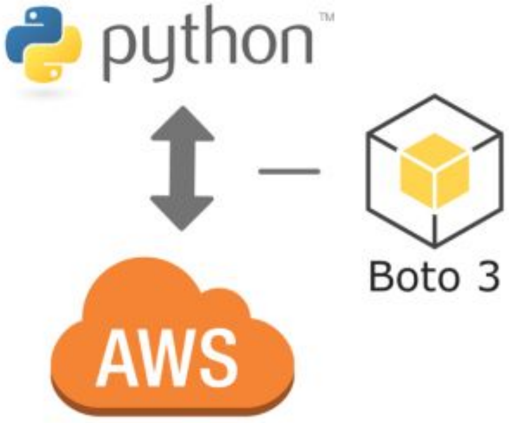


Copy

This file

To the current location
(. means the directory you
are currently in)

AWS S3



You can also use the *boto3* library in Python to interact with AWS. It works for a large number of services, but we’ll look specifically at s3 (https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/index.html).

You can install boto3 by running

```
$ pip install boto3
```

Note: For *boto3* to work, you will need to have installed the aws cli and run *aws configure*.

AWS S3

Download this file from the bucket

```
>>> import boto3
>>> s3 = boto3.resource('s3')
>>> bucket = s3.Bucket('nyc-tlc')
>>> bucket.download_file('trip data/yellow_tripdata_2019-12.csv',
'yellow_tripdata_2019-12.csv')
```

Save it at this filepath (on your local machine)

AWS

In the next couple of classes, we will be exploring some of the other tools offered on AWS.

In order to prepare for this, you will need to install ssh.

On mac, it should come preinstalled, so to check that it is, type

```
$ ssh
```

Windows users should install PuTTY:

https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html