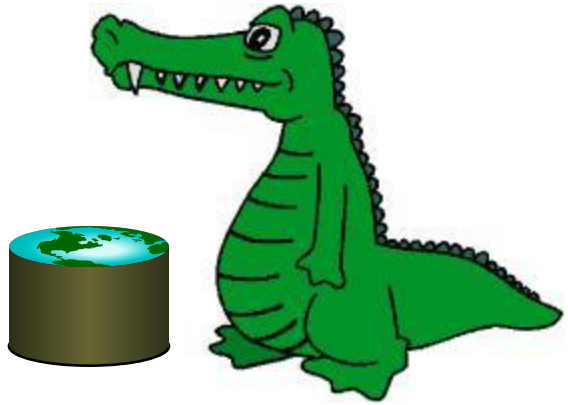
CAP4770/CAP5771 Fall 2015
Introduction to Data Science
Amazon Web Service Tutorial

University of Florida, CISE Department
Xiaofeng Zhou



Amazon Web Service Cloud Platform

- Compute: EC2, EMR/Hadoop (HDFS, MapReduce, Hive)
- Storage: S3

Amazon EC2 (Elastic Compute Cloud)

- A Web service that provides resizable compute capacity in the cloud.
- Designed to make Web-scale computing easier for developers.
- A simple Web service interface that provides high-degree of control of your computing resources

Amazon EC2 Benefits

- Reduces the time required to obtain and boot new server instances to minutes
- Quickly scales capacity, both up and down, as your computing requirements change
- Changes the economics of computing:
  - No start-up, monthly, or fixed costs
  - Pay only for capacity that you actually use
  - *a + bc* becomes just *bc*

Pricing Models

- **On-Demand Instances** – On-Demand Instances let you pay for compute capacity by the hour with no long-term commitments.

- **Reserved Instances** – Reserved Instances give you the option to make a low, one-time payment for each instance you want to reserve and in turn receive a significant discount on the hourly charge for that instance.

- **Spot Instances** – Spot Instances allow customers to bid on unused Amazon EC2 capacity and run those instances for as long as their bid exceeds the current Spot Price.

Instance vs. VM

- Instance = VM + hardware (instance type)
- AMI (Amazon Machine Image) = VM "image"
- VM "image"= OS + software
- Users specify the type of VM and hardware (i.e., instance type) when setting up an instance

Amazon S3 (Simple Storage Service) Basics

- Data stored as objects (files) in buckets
  - "key" to file is path
  - identified by <bucket> + <path>
  - No real directories, just path segments

- Great as persistent storage for data
  - Reliable – up to 99.999999999%
  - Scalable – up to petabytes of data
  - Fast – highly parallel requests

S3 Access

- Via your web browser
- Various command line tools
  - s3cmd
- Or via HTTP REST interface
  - Create (PUT/POST), Read (GET), Delete (DELETE)

S3 Limitations

- Can’t be modified (no random write or append)
- Max size of 5TB per object

Amazon EMR(Elastic Map-Reduce)

- A web service that allow cost-effective large data processing
- Hadoop (HDFS + Map-Reduce) over EC2 and S3
- EMR is mostly used for data intensive tasks
  - Examples: web indexing, data mining, log analysis, data warehousing, machine learning, financial analysis, scientific simulation, bioinformatics

Why Use Elastic MapReduce?

- Reduce hardware & IT personnel costs
  - Pay for what you actually use
  - Don't pay for people you don't need
  - Don't pay for capacity you don't need
- More agility, less wait time for hardware
  - Don't waste time buying/racking/configuring servers
  - Many server classes to choose from (micro to massive)
- Less time doing Hadoop deployment & version mgmt
  - Optimized Hadoop is pre-installed

Homework(Preparation for Lab 3)

- Setup AWS account(apply credit code)
- Watch and Follow:

  Getting Started (outdated since the AWS console interface changed, but still useful)

  Video on AWS Training on topics:

  – Introduction to Amazon Elastic Compute Cloud (EC2)
  – Introduction to Amazon Simple Storage Service (S3)
  – Introduction to Amazon Elastic MapReduce (EMR)

And follow their steps!

Learn More About AWS

- AWS: http://aws.amazon.com
- EC2 Resources:
  http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/
- Amazon EMR:
  http://aws.amazon.com/elasticmapreduce/
- Tutorial
  http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started.html
- Run your own job(wordcount)
  http://log.malchiodi.com/2014/11/12/executing-jar-encoded-mapreduce-jobs-in-aws-either-through-web-interface-or-cli/

Group up for Lab 3

Max 2-people group for lab 3.

Enter your group members info here