Intro to Amazon Web Services

Department of Biostatistics and Bioinformatics

Steve Pittard wsp@emory.edu

April 11, 2017

Motivations - The Five “V”s of Data

Big Data can be described by the following characteristics:

- **Volume** - The quantity of generated and stored data
- **Variety** - The type and nature of the data
- **Velocity** - The speed at which the data is generated and processed
- **Variability** - The inconsistency with the data
- **Veracity** - The quality of the data (or lack thereof)

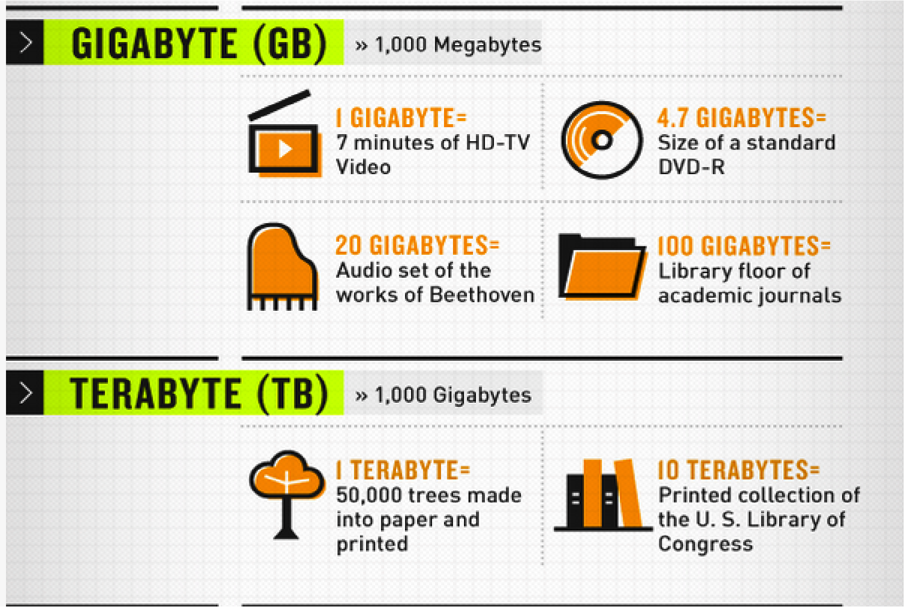
`https://en.wikipedia.org/wiki/Big_data`

Motivations - Data Has Gravity

Data has gravity. Once it becomes unmanageable locally you have to find some place to put it

- But then it is too large to move around comfortably
- Transfers over the network are slow
- Your local IT sends you nasty messages about using too much space
- Even if you have some space is the data being backed up in case of disaster ?
- Even if you have some space are there adequate computational resources available ?
- Can the network between the storage and compute resources work well under high loads ?

Motivations



GIGABYTE (GB) » 1,000 Megabytes

1 GIGABYTE= 7 minutes of HD-TV Video

4.7 GIGABYTES= Size of a standard DVD-R

20 GIGABYTES= Audio set of the works of Beethoven

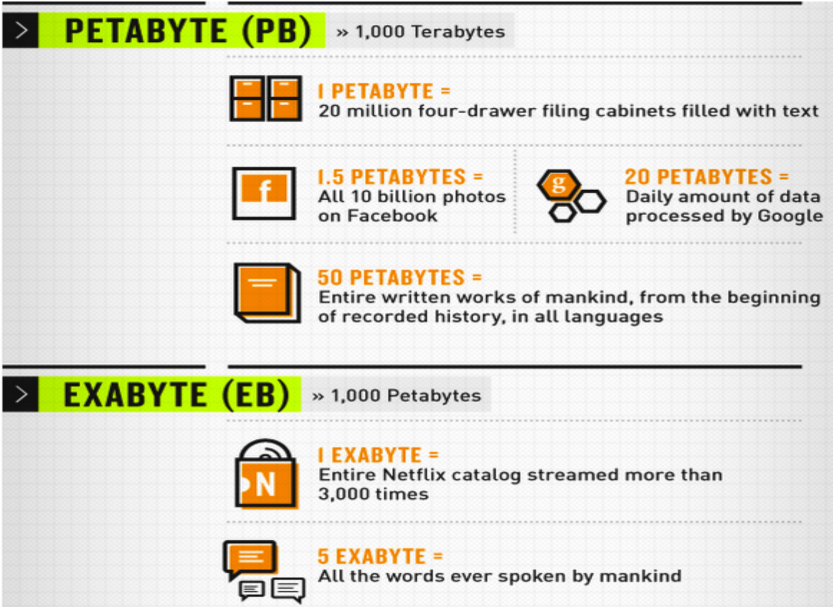
100 GIGABYTES= Library floor of academic journals

TERABYTE (TB) » 1,000 Gigabytes

1 TERABYTE= 50,000 trees made into paper and printed

10 TERABYTES= Printed collection of the U. S. Library of Congress

Motivations



PETABYTE (PB)

» 1,000 Terabytes

1 PETABYTE =
20 million four-drawer filing cabinets filled with text

1.5 PETABYTES =
All 10 billion photos on Facebook

20 PETABYTES =
Daily amount of data processed by Google

50 PETABYTES =
Entire written works of mankind, from the beginning of recorded history, in all languages

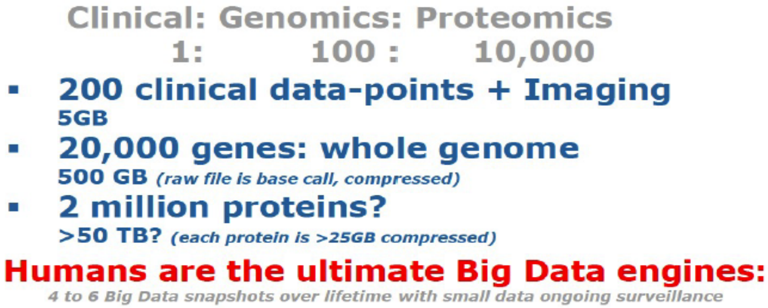
EXABYTE (EB)

» 1,000 Petabytes

1 EXABYTE =
Entire Netflix catalog streamed more than 3,000 times

5 EXABYTE =
All the words ever spoken by mankind

Motivations



Clinical: Genomics: Proteomics
1: 100 : 10,000

- **200 clinical data-points + Imaging**
  5GB
- **20,000 genes: whole genome**
  500 GB ***(raw file is base call, compressed)***
- **2 million proteins?**
  >50 TB? ***(each protein is >25GB compressed)***

Humans are the ultimate Big Data engines:

4 to 6 Big Data snapshots over lifetime with small data ongoing surveillance

Motivations - You Have One of These

Motivations - But This is your Kitchen

http://huntgatherlove.com/content/my-teeny-tiny-crib-kitchen-and-standing-desk-hacks

Motivations - One Possible Solution

https://www.homestratosphere.com/luxury-kitchen-designs-1/

Motivations

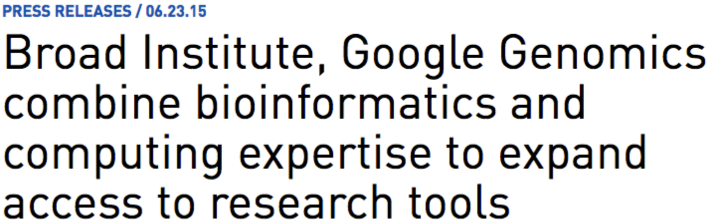
It would be nice to be able to rent a large kitchen space when you need it.

- Preferably with no contract or committment
- Pay only for what you use (you pay for food of course)
- Do not have to talk to anyone to arrange use
- Have a variety of kitchen sizes from which to select
- All equipment is in working order
- But you can customize the environment to suit your specific needs
- You can take a “snapshot” of your environemnt as a reference for future work
- You can prepay if you want but at a discount
- You can bid on price to possibly obtain a cheaper rate

Why Use the Cloud ?

- Your Data is too large for anything you have locally
- Computation takes too long on anything you have locally
- You need more RAM/Memory than anything you have locally
- You need to create a very large database
- You want your computation environment to be easily reproducible
- You wish to implement a method you found in a Research Paper that requires Map Reduce, Spark, or some other distributed computing framework

Why Use the Cloud ?



PRESS RELEASES / 06.23.15

Broad Institute, Google Genomics combine bioinformatics and computing expertise to expand access to research tools

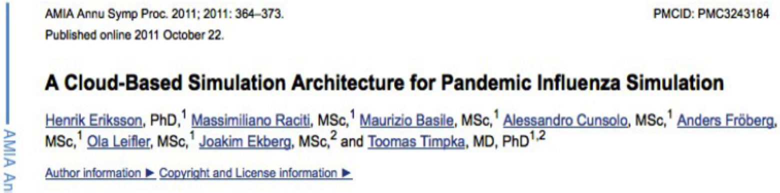
Cloud Computing

Solves the “horizontal computing” problem

Cloud Computing

- A remote computer someplace else ? Yes
- But ! You select what size of computer you want, when you want it, for as long as you want it, and you pay for only what you use
- The same is true for Storage and Databases
- Storage and Compute appear to be “infinite”
- You don’t have to talk to someone to set any of this up
- You create resources from a console or via an API
- You can create ”images” that others can use so they can easily collaborate with you and or reproduce your research
- Access from anywhere with Internet

Early Work on the Cloud



AMIA Annu Symp Proc. 2011; 2011: 364–373.
Published online 2011 October 22.

PMCID: PMC3243184

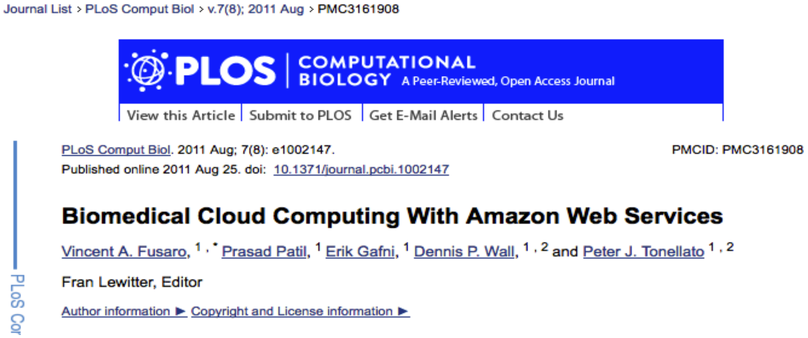
A Cloud-Based Simulation Architecture for Pandemic Influenza Simulation

Henrik Eriksson, PhD,[1] Massimiliano Raciti, MSc,[1] Maurizio Basile, MSc,[1] Alessandro Cunsolo, MSc,[1] Anders Fröberg, MSc,[1] Ola Leifler, MSc,[1] Joakim Ekberg, MSc,[2] and Toomas Timpka, MD, PhD[1,2]

Author information ▶ Copyright and License information ▶

AMIA An

Early Work on the Cloud



Journal List > PLoS Comput Biol > v.7(8); 2011 Aug > PMC3161908

PLOS | COMPUTATIONAL BIOLOGY A Peer-Reviewed, Open Access Journal
View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us
PLoS Cor

PLoS Comput Biol. 2011 Aug; 7(8): e1002147. PMCID: PMC3161908

Published online 2011 Aug 25. doi: 10.1371/journal.pcbi.1002147

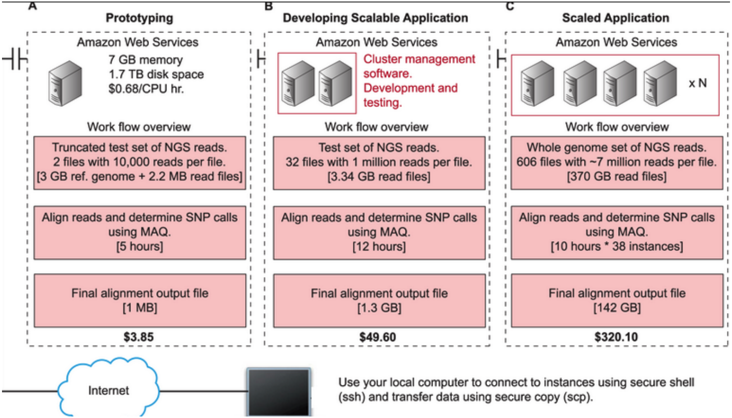
Biomedical Cloud Computing With Amazon Web Services

Vincent A. Fusaro,[1,*] Prasad Patil,[1] Erik Gafni,[1] Dennis P. Wall,[1,2] and Peter J. Tonellato[1,2]

Fran Lewitter, Editor

Author information ► Copyright and License information ►

Early Work on the Cloud



A
Prototyping

Amazon Web Services

7 GB memory
1.7 TB disk space
$0.68/CPU hr.

Work flow overview

Truncated test set of NGS reads.
2 files with 10,000 reads per file.
[3 GB ref. genome + 2.2 MB read files]

Align reads and determine SNP calls
using MAQ.
[5 hours]

Final alignment output file
[1 MB]

$3.85

B
Developing Scalable Application

Amazon Web Services

Cluster management
software.
Development and
testing.

Work flow overview

Test set of NGS reads.
32 files with 1 million reads per file.
[3.34 GB read files]

Align reads and determine SNP calls
using MAQ.
[12 hours]

Final alignment output file
[1.3 GB]

$49.60

C
Scaled Application

Amazon Web Services

x N

Work flow overview

Whole genome set of NGS reads.
606 files with ~7 million reads per file.
[370 GB read files]

Align reads and determine SNP calls
using MAQ.
[10 hours * 38 instances]

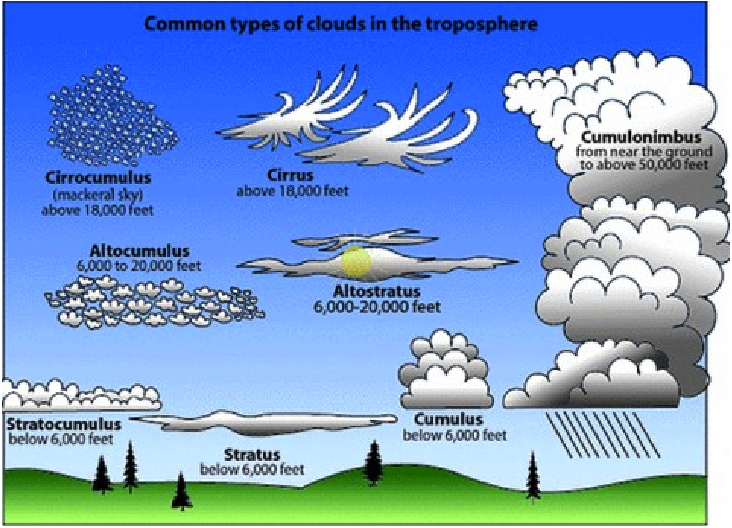
Final alignment output file
[142 GB]

$320.10

Internet

Use your local computer to connect to instances using secure shell (ssh) and transfer data using secure copy (scp).

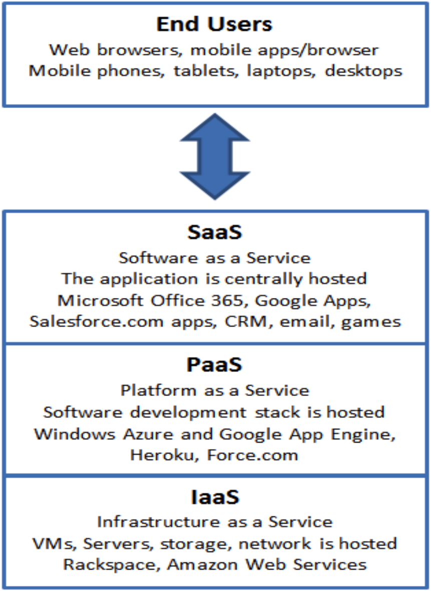
Know Your Clouds !



Cloud Computing ?

- **SaaS** - Software as a Service - An application and everything it takes to support it (e.g. MS Office 365)
  - Vendor provides everything
  - You login usually with a web browser or mobile phone client
- **PaaS** - Platform as a Service - Everything Supporting the Application except the Application and data
  - Vendor provides almost everything except data and the application
  - Web Hosting - You create content and Apps but vendor provides everything else
- **IaaS** - Infrastructure as a Service - The hardware, network, compute, and storage upon which you create servers
  - Vendor provides network, hardware, and virtualization services
  - You create servers and all that goes with it

Types of Service



End Users

Web browsers, mobile apps/browser
Mobile phones, tablets, laptops, desktops

SaaS

Software as a Service
The application is centrally hosted
Microsoft Office 365, Google Apps,
Salesforce.com apps, CRM, email, games

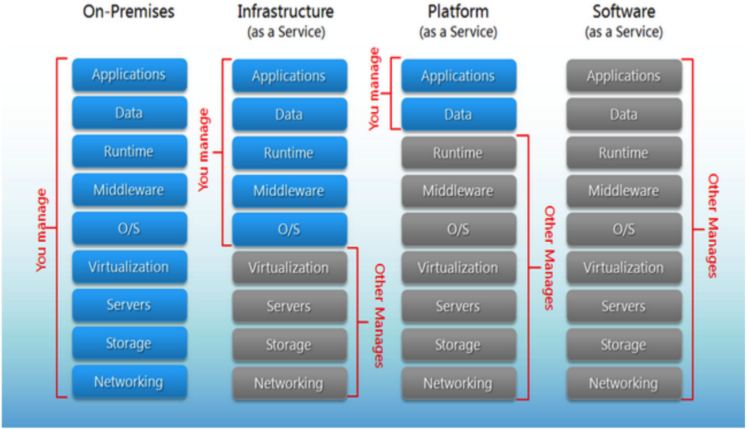
PaaS

Platform as a Service
Software development stack is hosted
Windows Azure and Google App Engine,
Heroku, Force.com

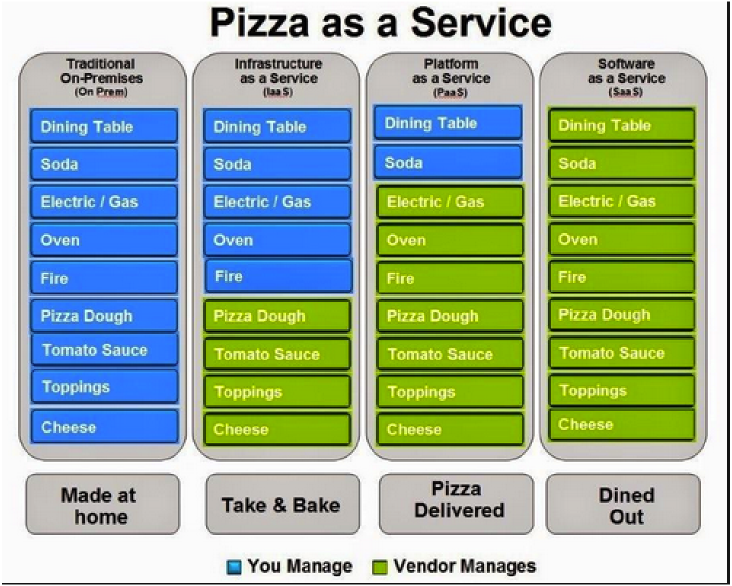
IaaS

Infrastructure as a Service
VMs, Servers, storage, network is hosted
Rackspace, Amazon Web Services

Types of Service



Types of Services



Pizza as a Service

Traditional On-Premises (On Prem)	Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
🟦 Dining Table	🟦 Dining Table	🟦 Dining Table	🟩 Dining Table
🟦 Soda	🟦 Soda	🟦 Soda	🟩 Soda
🟦 Electric / Gas	🟦 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas
🟦 Oven	🟦 Oven	🟩 Oven	🟩 Oven
🟦 Fire	🟦 Fire	🟩 Fire	🟩 Fire
🟦 Pizza Dough	🟩 Pizza Dough	🟩 Pizza Dough	🟩 Pizza Dough
🟦 Tomato Sauce	🟩 Tomato Sauce	🟩 Tomato Sauce	🟩 Tomato Sauce
🟦 Toppings	🟩 Toppings	🟩 Toppings	🟩 Toppings
🟦 Cheese	🟩 Cheese	🟩 Cheese	🟩 Cheese
Made at home	Take & Bake	Pizza Delivered	Dined Out

🟦 You Manage 🟩 Vendor Manages

Cloud Computing ?

- Amazon has been at it longer than any of them
- AWS has a high level of maturity and reliability
- Google is moving in fast on Genomic Computing
- Microsoft uses, surprise, Microsoft Products so if that's your thing then maybe go there
- All services from any of these providers are virtual servers though some offer "bare metal" access as an option
- It's okay if you don't know what this means just understand that in general you will be sharing a "real server" with someone else albeit virtually

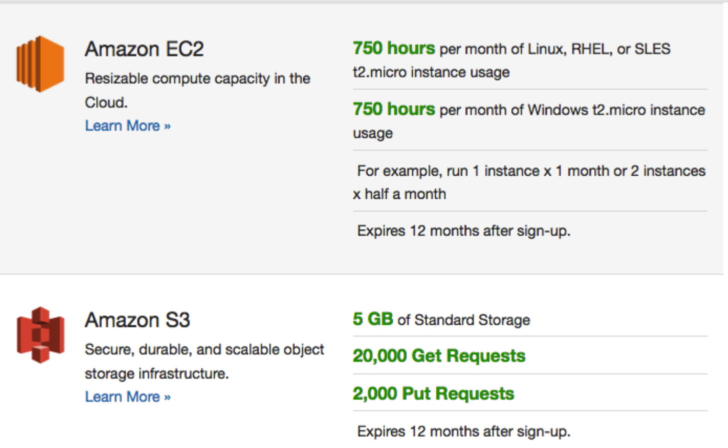
Cloud Computing - How to Use

Most Data Science people will use IaaS or SaaS (e.g. Galaxy Cloudman)

- Use the S3 storage to "park" data sets for later use
- Use the EC2 Service to boot up Linux servers or pre-packaged AMIs
- Create computers with as much RAM and disk as you want
- Analyze data and the put the EC2 "instances" to "sleep" to avoid running costs
- Make an AMI (Amazon Machine Instance) that others can use
- When finished with a project you can terminate the instances and delete data (if you wish)

Sign Up: See http://aws.amazon.com/free

Amaon provides “training wheels” so you can test things out at no or low cost



Amazon EC2

Resizable compute capacity in the Cloud.

Learn More »

750 hours per month of Linux, RHEL, or SLES t2.micro instance usage

750 hours per month of Windows t2.micro instance usage

For example, run 1 instance x 1 month or 2 instances x half a month

Expires 12 months after sign-up.

Amazon S3

Secure, durable, and scalable object storage infrastructure.

Learn More »

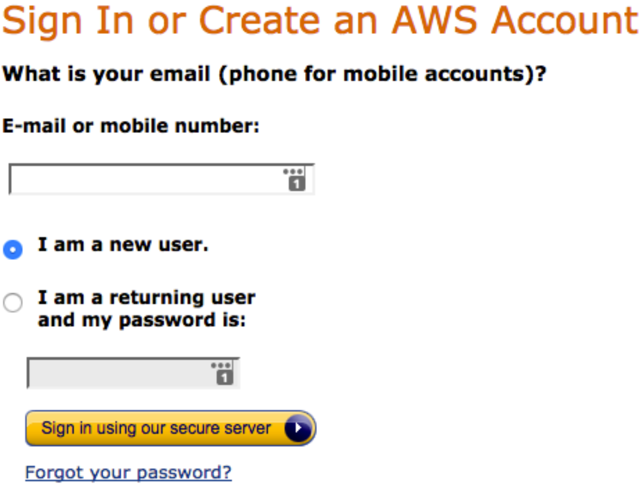
5 GB of Standard Storage

20,000 Get Requests

2,000 Put Requests

Expires 12 months after sign-up.

Sign Up: Go to http://aws.amazon.com



Sign In or Create an AWS Account

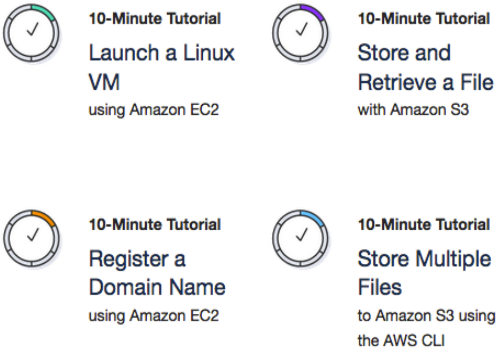
What is your email (phone for mobile accounts)?

E-mail or mobile number:

● **I am a new user.**

○ **I am a returning user and my password is:**

Tutorials Go to http://aws.amazon.com/start-now



10-Minute Tutorial

Launch a Linux VM

using Amazon EC2

10-Minute Tutorial

Store and Retrieve a File

with Amazon S3

10-Minute Tutorial

Register a Domain Name

using Amazon EC2

10-Minute Tutorial

Store Multiple Files

to Amazon S3 using the AWS CLI

The Dashboard

The Dashboard is the launchpad for all of Amazon's services

- It takes getting used to
- In reality you really only use perhaps 2-3 services at first
- S3 is for general storage and is up 99.9 percent of the time
- EC2 is for computing. This is where you generally want to be
- Other cool services are the Machine Learning Service

The Dashboard and APIs

It is possible to create a large variety of Virtual Servers. See `https://aws.amazon.com/ec2/instance-types/` for a full descriprion.

These “instances” can be created from the Dashboard or from an API (Application Programming Interface) using high level programming languages

- MS Windows
- UNIX / Linux
- High Performance Computational Clusters
- Map/Reduce Hadoop Clusters
- Machine Learning Clusters

As a Remote Computer

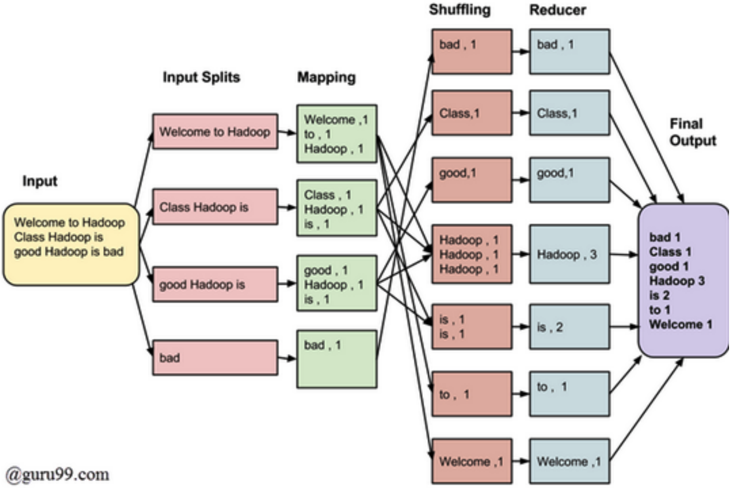
- You log in to a server somewhere that has software installed
- You upload your data, analyze it, and when done download it
- The server is put to “sleep” until you need it again
- Upon completion of project create an AMI (Amazon Machine Instance) as a reference
- Terminate the server

Parallel Processing

Example:

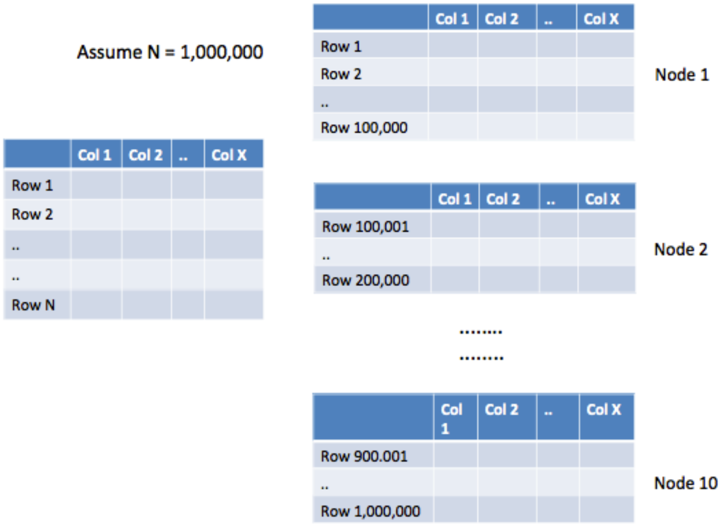
- We have a body of text in some language
- We want to count the number of times that each word appears in the text
- Really hard for a person to do except for really small books
- Divide the text into 100 chunks and assign to 100 people
- Have each person figure out the words in their chunk and the number of times they appear
- Evreryone reports back their totals

Map Reduce Simplified



@guru99.com

Distributed Data Frames



Assume N = 1,000,000

	Col 1	Col 2	..	Col X
Row 1				
Row 2				
..				
..				
Row N				

Node 1

	Col 1	Col 2	..	Col X
Row 1				
Row 2				
..				
Row 100,000				

Node 2

	Col 1	Col 2	..	Col X
Row 100,001				
..				
Row 200,000				

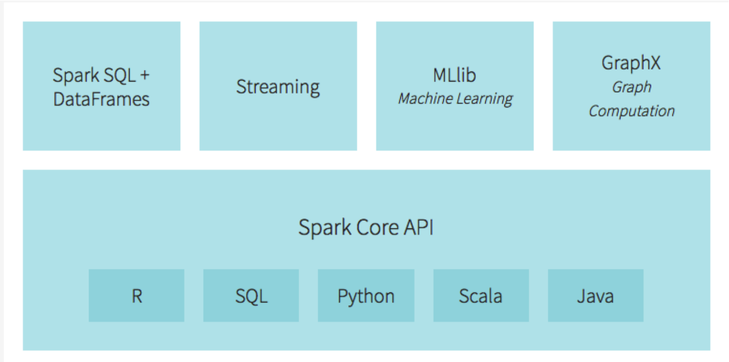
....

....

Node 10

	Col 1	Col 2	..	Col X
Row 900.001				
..				
Row 1,000,000				

Apache Spark



Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

SQL

Python

Scala

Java

Apache Spark

The key idea with Spark Version 1 the **R**esilient **D**istributed **D**ata Set (RDD)

- Support in-memory processing computation
- Data sharing in memory is 10 to 100 times faster than network and disk
- Faster than Map/Reduce that relies on storage
- Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster
- In Apache 2.0 there is explicit support for Data Frames (close to what R thinks a dataframe is)
- Data frames provide a domain specific language API to manipulate your distributed data

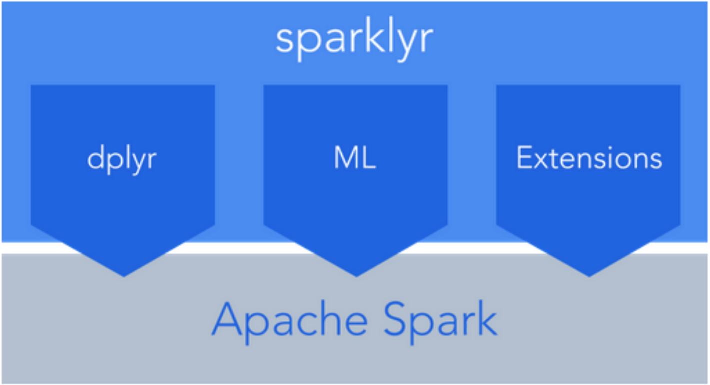
Apache Spark

But Wait ! There's More ! Spark provides access to a companion Machine Learning Library that is "baked in" to Spark.

- Provides most major ML capabilities
- Transformation Tools
- Can Access the Spark ML from RStudio !!
- Can Use the familiar R syntax to work with the Data Frame

Apache Spark - sparkly

sparkly is a package that provides connectivity to Apache Spark clusters directly from **RStudio**. Best of all you can use the **dplyr** package to work with the Spark Data Frames



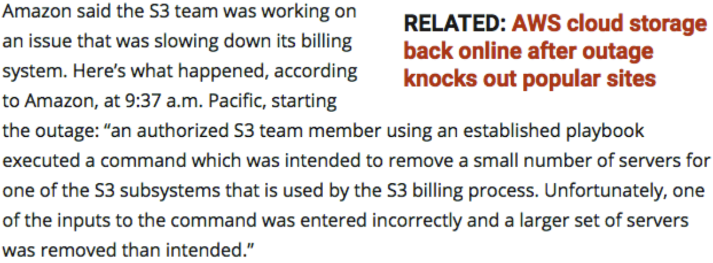
Concerns

Concerns

The meter is always running - must keep track of costs

Concerns

Outage of March 2017 - Outages Happen but not often



Amazon said the S3 team was working on an issue that was slowing down its billing system. Here's what happened, according to Amazon, at 9:37 a.m. Pacific, starting the outage: "an authorized S3 team member using an established playbook executed a command which was intended to remove a small number of servers for one of the S3 subsystems that is used by the S3 billing process. Unfortunately, one of the inputs to the command was entered incorrectly and a larger set of servers was removed than intended."

RELATED: AWS cloud storage back online after outage knocks out popular sites

Concerns

If you want to spin up your own instances from scratch you will need help unless you know something about system administration:

- Bioinformatics workloads almost always require UNIX operating system
- You will need to know about UNIX from a command line point of view
- It's good if you know Ubuntu Server which is very friendly for Bioinformatics
- You need to know how to provision storage and link it to EC2
- BUT you can take advantages of pre-existing Instances that have been created for you