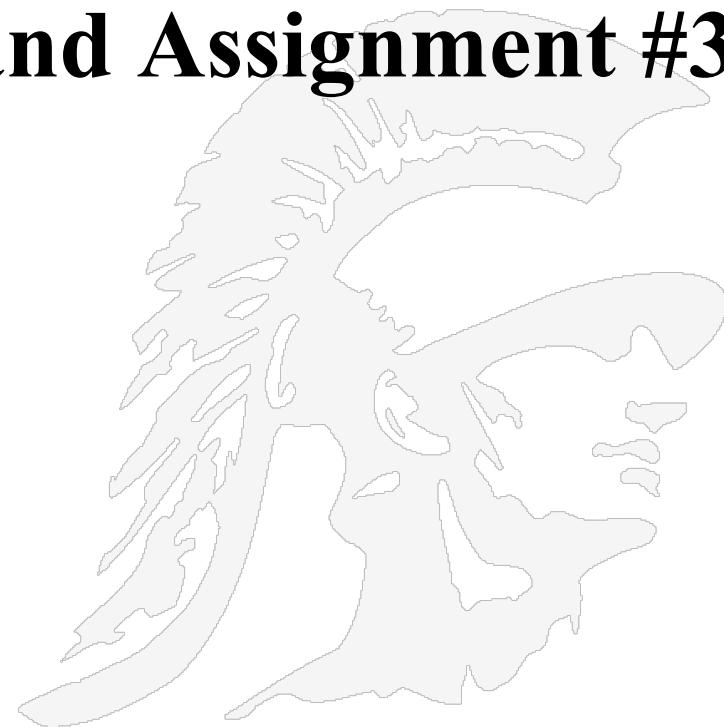


Cloud Computing and Assignment #3

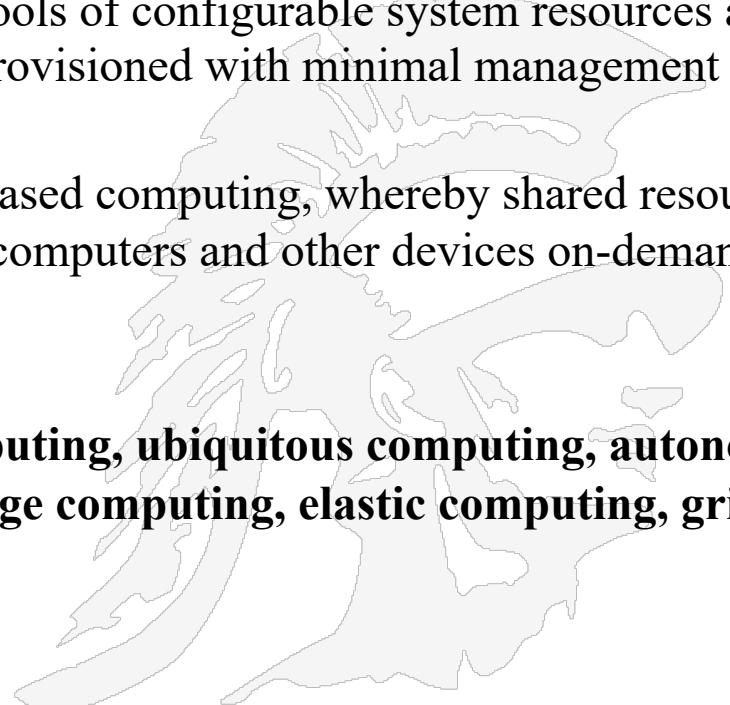


Computing is Rapidly Changing

- **There are many trends putting pressure on conventional computing centers, e.g.**
 - Explosive growth in applications: biomedical informatics, space exploration, business analytics, web 2.0 social networking
 - Extreme scale content *generation*: e-science and e-business data deluge
 - Extraordinary rate of digital content *consumption*: digital gluttony: Apple iPhone, iPad, Amazon Kindle
 - Exponential growth in compute capabilities: multi-core, storage, bandwidth, virtual machines (virtualization)
 - Very short cycle of obsolescence in technologies: Windows Vista → Windows 10 → Windows 11; Java versions; JavaScript, C → C#; Python, Swift, Go, R
 - Newer architectures: web services, persistence models, distributed file systems/repositories (Google, Hadoop), multi-core, wireless and mobile
- **It is far more difficult for a company to manage this complex situation with a traditional IT infrastructure:**

Enter the Cloud

- **Definition (Simple):** *Cloud computing* refers to the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer
- **Definition (Complicated):** *Cloud computing* is an information technology paradigm that enables ubiquitous access to shared pools of configurable system resources and higher-level services that can be rapidly provisioned with minimal management effort, often over the Internet.
- **Definition:** *Cloud computing* is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid.
- **Other names for cloud computing:**
 - **on-demand computing, utility computing, ubiquitous computing, autonomic computing, platform computing, edge computing, elastic computing, grid computing, ...**



Cloud Computing Terminology

• Infrastructure as a Service (IaaS)

- Virtual machine instances of various shapes and sizes
- Storage capabilities
- Networking support including configurable IP addresses

• Platform as a Service (PaaS)

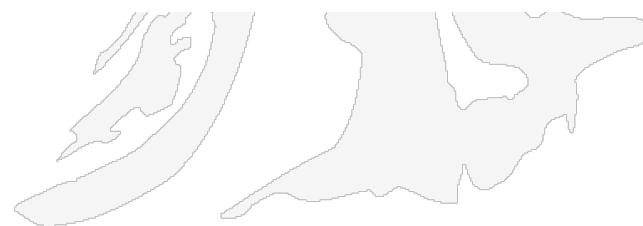
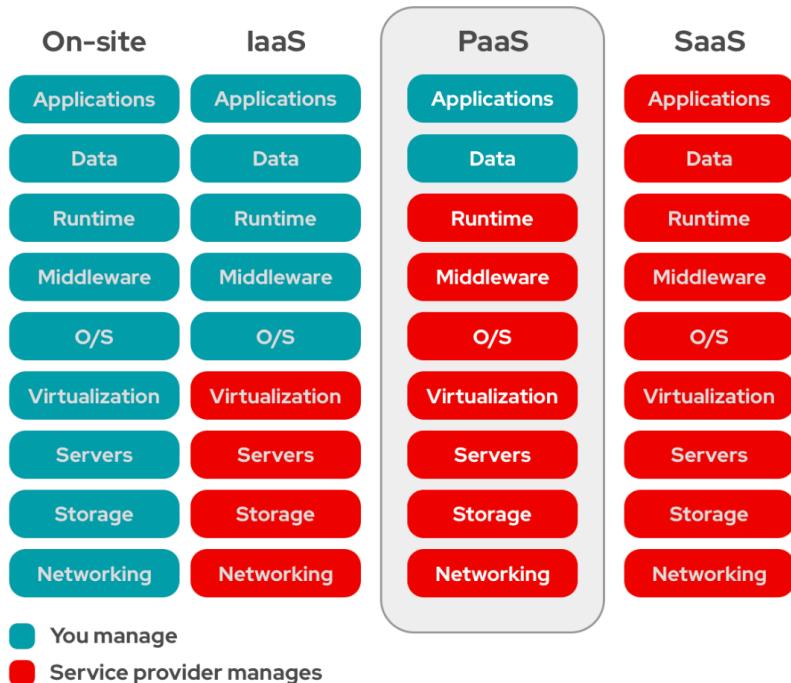
- Data management: text search, image analytics, deep learning
- APIs for building applications
- Business analytics

• Software as a Service (SaaS)

- Enterprise resource planning software
- Supply chain management software

• Data as a Service (DaaS)

- Aggregate and analyze consumer data



Cloud Service Models

Software as a Service (SaaS)

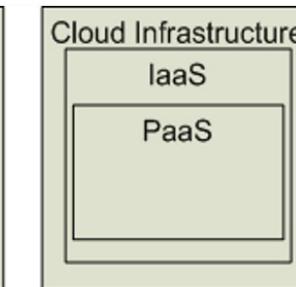
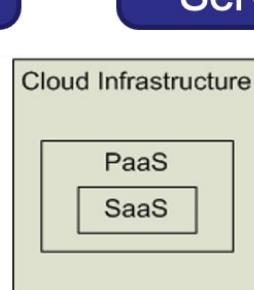
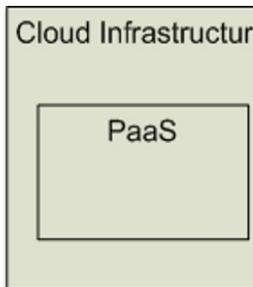
SalesForce CRM

LotusLive



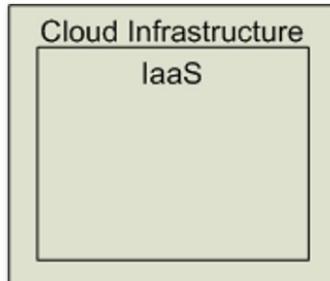
Platform as a Service (PaaS)

 Google App
 Windows Azure
The Future Made Familiar



Infrastructure as a Service (IaaS)

 amazon
webservices™
 rackspace
HOSTING



Software as a Service (SaaS)
Providers Applications

Platform as a Service (PaaS)
Deploy customer created Applications

Infrastructure as a Service (IaaS)

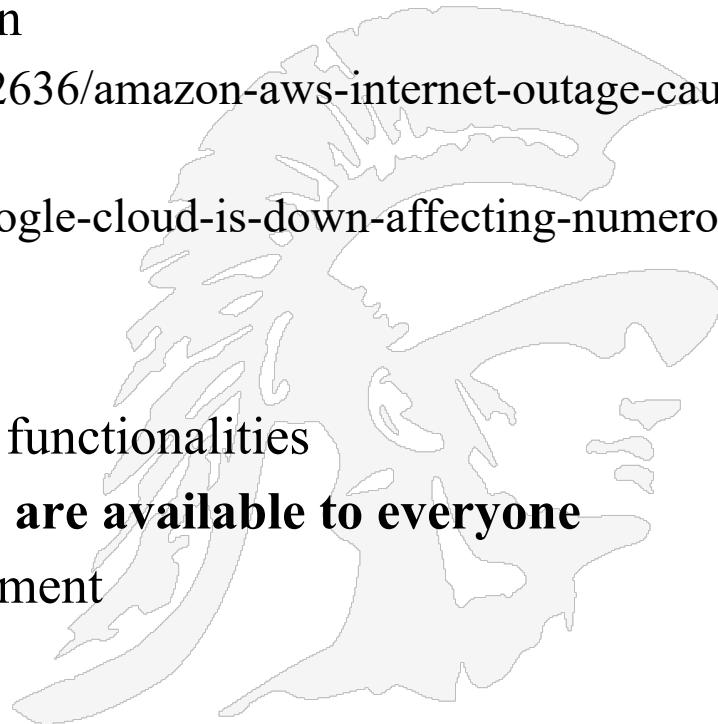
Rent Processing, storage, N/W capacity & computing resources

Cloud Computing Relies Upon the Internet, But . . .

- Cloud Computing takes place over the Internet,
- These platforms hide the complexity and details of the underlying infrastructure from users and applications by providing a “simple” graphical interface or API
- However, the balkanization of the internet is also occurring
 - Balkanization is a term for the process of fragmentation or division of a region or state into smaller regions or states that are often hostile or uncooperative with one another
 - Balkanization is an effort to control what information is seen by the population, avoiding points of view that contradict those in power
- Splinternet: a term used to describe efforts to control what people see on the internet (see <https://en.wikipedia.org/wiki/Splinternet>)
 - China’s great firewall – Google withdrawing their search engine due to censorship requirements
 - Russia’s Sovereign Internet Law allows for separation from the Internet
 - Iran, Syria, Tunisia, Saudi Arabia and others filter a wide range of internet content
 - South Korea filters news from North Korea

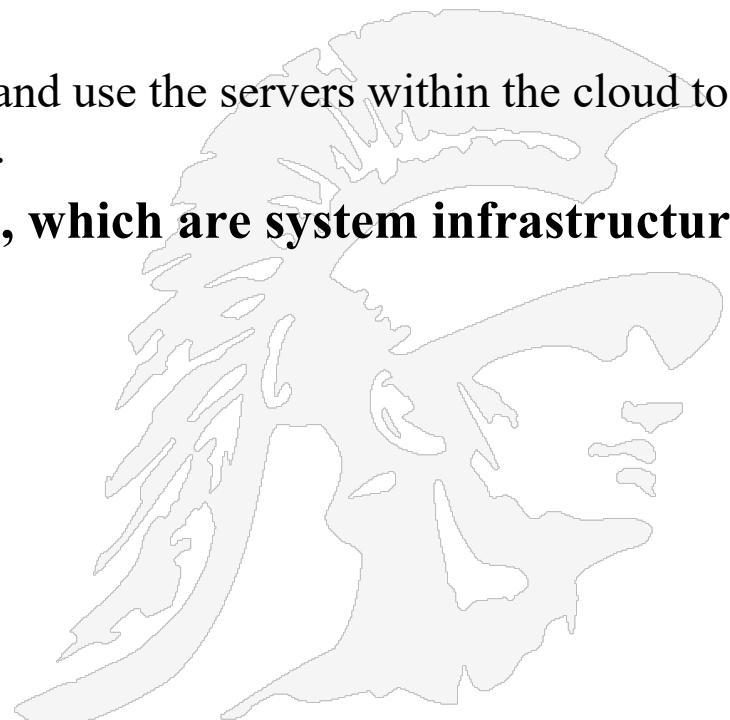
Cloud Computing Platform is Always on . . .

- The platform provides on-demand services, that are always on, anywhere, anytime and any place
 - Well almost, e.g. Amazon today (03/02/2017) blamed human error for the big AWS outage that took down a bunch of large internet sites for several hours on Tuesday afternoon
 - <http://www.recode.net/2017/3/2/14792636/amazon-aws-internet-outage-cause-human-error-incorrect-command>
 - <https://techcrunch.com/2019/06/02/google-cloud-is-down-affecting-numerous-applications-and-services/>
- Pay for use and as needed, *elastic*
 - scale up and down in capacity and functionalities
- The hardware and software services are available to everyone
 - general public, enterprises, government

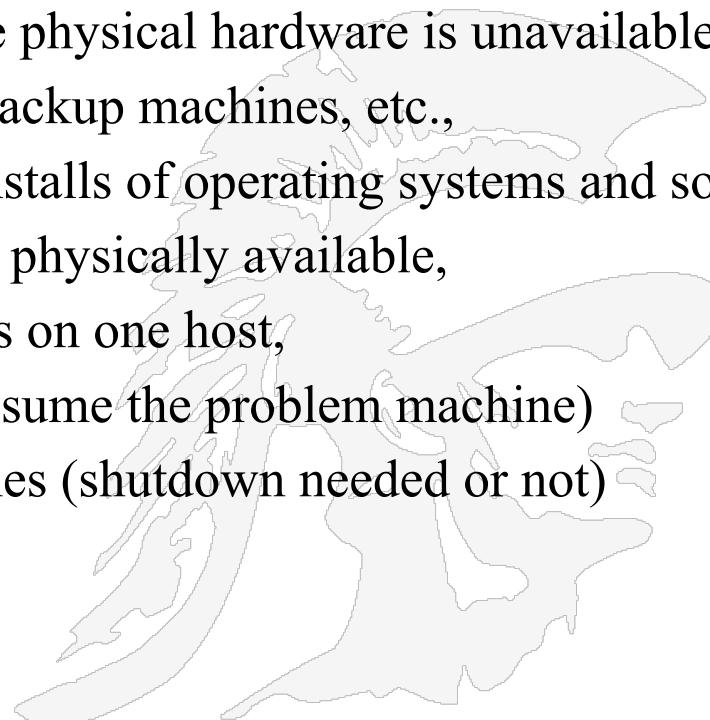


Purpose and Benefits

- By using the Cloud infrastructure on “pay as used and on demand”, companies save in capital and operational investment!
- Clients can:
 - Put their data on the platform instead of on their own desktop PCs and/or on their own servers.
 - Put their applications on the platform and use the servers within the cloud to do processing and data manipulations etc.
- Enables companies and applications, which are system infrastructure dependent, to be infrastructure-less
- Cost-Benefit Analysis
 - Will there be reduced IT costs
 - Will your applications be able to scale
 - Will you have access to automatic updates
 - Will your data be secure
 - There are other considerations



- **Virtualization:** the creation of a virtual -- rather than actual -- version of something, such as an operating system, a server, a storage device or network resources
- **Advantages of virtual machines:**
 1. Run operating systems where the physical hardware is unavailable,
 2. Easier to create new machines, backup machines, etc.,
 3. Software testing using “clean” installs of operating systems and software,
 4. Emulate more machines than are physically available,
 5. Timeshare lightly loaded systems on one host,
 6. Debug problems (suspend and resume the problem machine)
 7. Easy migration of virtual machines (shutdown needed or not)
 8. Run legacy systems!



Hypervisor

(variant of supervisor)

- A **hypervisor** or **virtual machine monitor (VMM)** is computer software, firmware or hardware that creates and runs virtual machines
- A computer on which a hypervisor runs one or more virtual machines is called a *host machine*, and each virtual machine is called a *guest machine*.
- The hypervisor manages the execution of the guest operating systems.
- Multiple instances of a variety of operating systems may share the virtualized hardware resources: for example, Linux, Windows and MacOS, can all run on a single physical machine.
- Some hypervisor vendors
 - VMware ESX Server, ESX301**
 - Microsoft Windows Hyper-V**
 - Oracle VM Virtual Box**
 - Xen Project**
 - developed by the Univ. of Cambridge and is now being developed by the Linux foundation with support from INTEL XEN3030

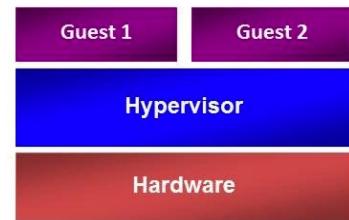


Hypervisor Design: Two approaches

Type 2 Hypervisor



Type 1 Hypervisor



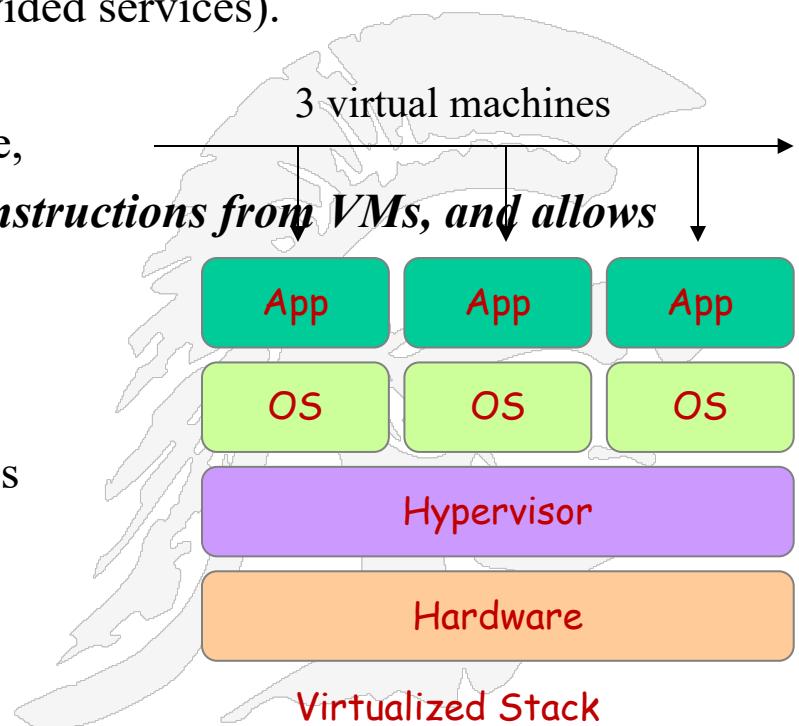
Examples:

Virtual PC & Virtual Server
 VMware Workstation
 KVM

Examples:
 Hyper-V
 Xen
 VMware ESX

Virtualization Makes Cloud Computing Possible

- **Virtual workspaces:**
 - An abstraction of an execution environment that can be made dynamically available to authorized clients by using well-defined protocols,
 - Resource quota (e.g. CPU, memory share),
 - Software configuration (e.g. O/S, provided services).
- **Virtual Machines (VMs):**
 - Abstraction of a physical host machine,
 - *Hypervisor intercepts and emulates instructions from VMs, and allows management of VMs,*
 - VMWare, Xen, etc.
- **Provide infrastructure API:**
 - Plug-ins to hardware/support structures



Virtualization Approaches

1. The **full virtualization** approach allows datacenters to run an unmodified guest operating system
 - **VMware** uses a combination of direct execution and binary translation techniques to achieve full virtualization of an x86 system
2. The **para-virtualization** approach modifies the guest operating system to eliminate the need for binary translation. Therefore it offers potential performance advantages for certain workloads but requires using specially modified operating system kernels
 - The **Xen open source project** was designed initially to support para-virtualized operating systems. While it is possible to modify open source operating systems, such as Linux and OpenBSD, it is not possible to modify “closed” source operating systems such as Microsoft Windows .
 - Microsoft Windows is the most widely deployed operating system in enterprise datacenters.
 - For such unmodified guest operating systems, a virtualization hypervisor must either adopt the full virtualization approach or rely on hardware virtualization in the processor architecture.

Hypervisor Features

Features

	Xen	KVM	VirtualBox	VMWare
Paravirtualization	Yes	No	No	No
Full Virtualization	Yes	Yes	Yes	Yes
Host CPU	X86, X86_64, IA64	X86, X86_64, IA64, PPC	X86, X86_64	X86, X86_64
Guest CPU	X86, X86_64, IA64	X86, X86_64, IA64, PPC	X86, X86_64	X86, X86_64
Host OS	Linux, Unix	Linux	Windows, Linux, Unix	Proprietary Unix
Guest OS	Linux, Windows, Unix	Linux, Windows, Unix	Linux, Windows, Unix	Linux, Windows, Unix
VT-x / AMD-v	Opt	Req	Opt	Opt
Supported Cores	128	16*	32	8
Supported Memory	4TB	4TB	16GB	64GB
3D Acceleration	Xen-GL	VMGL	Open-GL	Open-GL, DirectX
Licensing	GPL	GPL	GPL/Proprietary	Proprietary

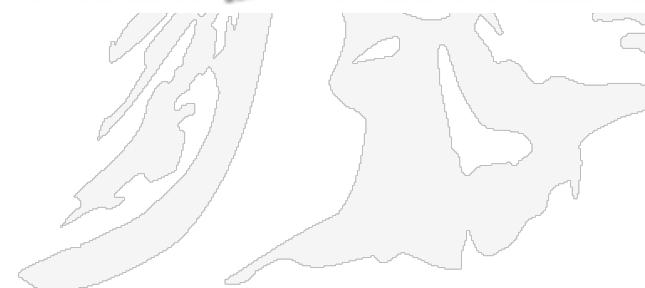
- Most “Cloud” deployments rely on virtualization.
 - Amazon EC2, GoGrid, Azure, Rackspace Cloud
 - ...
 - Nimbus, Eucalyptus, OpenNebula, OpenStack
 - ...
- There are a number of Virtualization tools or Hypervisors available today.
 - Xen, KVM, VMWare, Virtualbox, Hyper-V ...

Virtualization Tools

Current Hypervisors



5



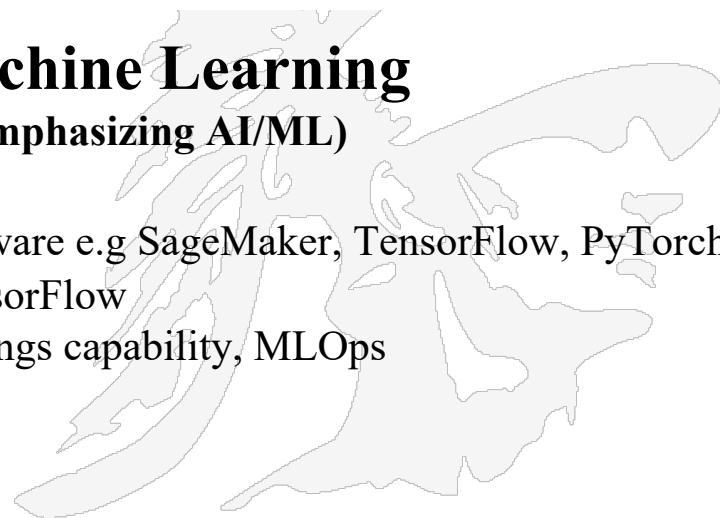
Leading Cloud Vendors

Cloud vendor	Annual revenue run rate
Microsoft commercial cloud	\$21.2 billion
Amazon Web Services	\$20.4 billion
IBM	\$10.3 billion
Oracle	\$6.08 billion
Google Cloud Platform/G Suite	\$4 billion
Alibaba	\$2.2 billion

Source: Company filings, earnings reports

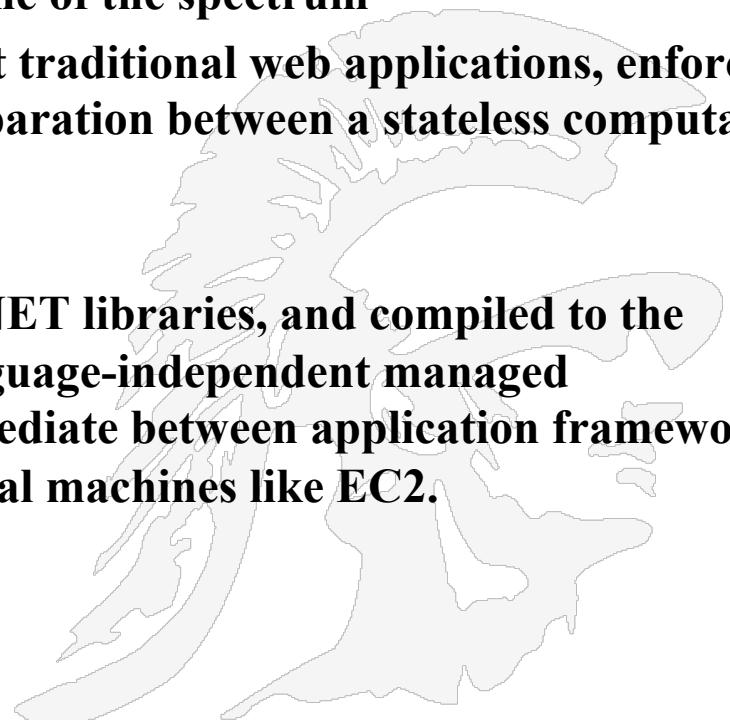
Role of AI and Machine Learning (All cloud vendors are emphasizing AI/ML)

- AWS offers multiple machine learning software e.g SageMaker, TensorFlow, PyTorch
- Google Cloud has BigQuery, AutoML, TensorFlow
- Microsoft emphasizes Azure Internet of Things capability, MLOps
- IBM offers its Watson platform
- Oracle offers automated cloud services



Comparing the Leading Cloud Vendors

- Amazon EC2 is at one end of the spectrum.
 - An EC2 instance looks much like physical hardware, and users can control nearly the entire software stack, from the kernel upwards
- Google AppEngine is at the other extreme of the spectrum
 - AppEngine is targeted exclusively at traditional web applications, enforcing an application structure of clean separation between a stateless computation tier and a stateful storage tier.
- Microsoft Azure
 - applications are written using the .NET libraries, and compiled to the Common Language Runtime, a language-independent managed environment. Thus, Azure is intermediate between application frameworks like AppEngine and hardware virtual machines like EC2.



Computation Model

Examples of Cloud Computing vendors and how each provides virtualized resources (computation, storage, networking) and ensures scalability and high availability of the resources.

Computation Model (VM)

Amazon EC2

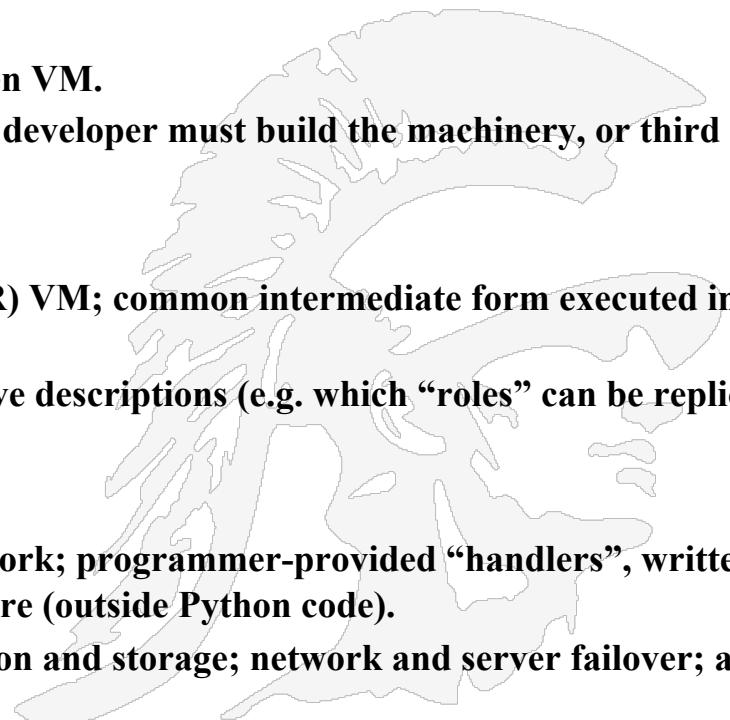
- x86 Instruction Set Architecture (ISA) via Xen VM.
- Computation elasticity allows scalability, but developer must build the machinery, or third party such as RightScale must provide it.

Microsoft's Azure

- Microsoft Common Language Runtime (CLR) VM; common intermediate form executed in managed environment.
- Machines are provisioned based on declarative descriptions (e.g. which “roles” can be replicated); automatic load balancing.

Google AppEngine

- Predefined application structure and framework; programmer-provided “handlers”, written in Python, all persistent state stored in MegaStore (outside Python code).
- Automatic scaling up and down of computation and storage; network and server failover; all consistent with 3-tier Web app structure.



Storage Model

Amazon EC2

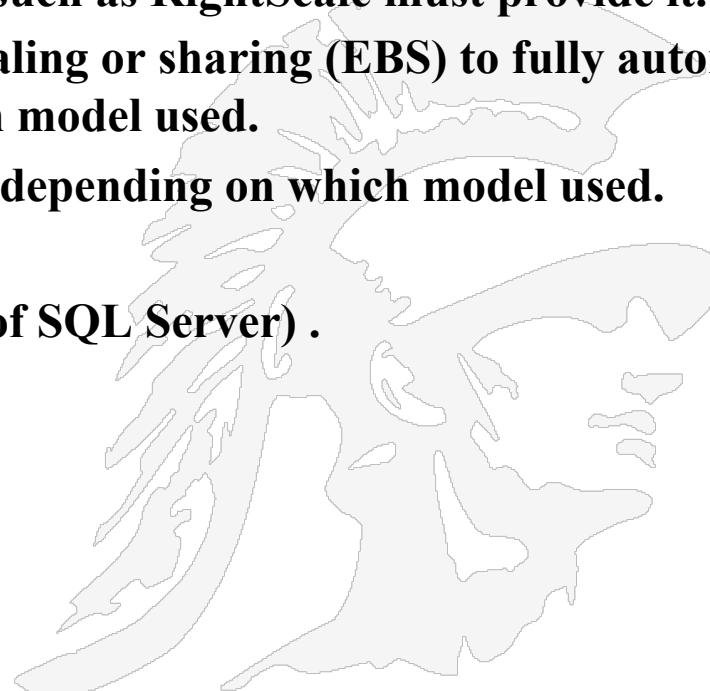
- Range of models from block store (EBS) to augmented key/blob store (SimpleDB). Computation elasticity allows scalability, but developer must build the machinery, or third party such as RightScale must provide it.
- Automatic scaling varies from no scaling or sharing (EBS) to fully automatic (SimpleDB, S3), depending on which model used.
- Consistency guarantees vary widely depending on which model used.

Microsoft's Azure

- SQL Data Services (restricted view of SQL Server).
- Azure storage service

Google AppEngine

- MegaStore/BigTable



Now Let's Shift to HW#3

Redeeming Google Cloud Credits

Students will sign up for Google's free trial, \$300 credit

Select Account Type Individual

use your @gmail.com address and finish by clicking “Start my free trial”

console.cloud.google.com

Google Cloud Platform

Try Cloud Platform for free

Country: United States

Acceptances: Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.

Yes No

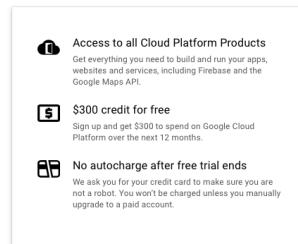
I agree that my use of any services and related APIs is subject to my compliance with the applicable Terms of Service. I have also read and agree to the Google Cloud Platform Free Trial Terms of Service.

Required to continue

Yes No

Agree and continue

Privacy policy



Google Cloud Platform

Try Cloud Platform for free

Customer info

Account type: Individual

Name and address:

Name: myfirstname mylastname

Address line 1: 100 main street

Address line 2:

City: los angeles

State: California ZIP code: 90089

Phone number:

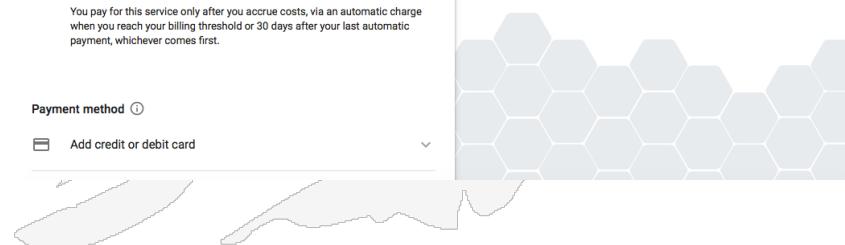
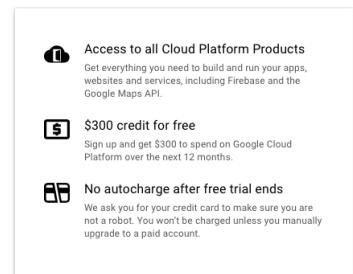
How you pay

Automatic payments

You pay for this service only after you accrue costs, via an automatic charge when you reach your billing threshold or 30 days after your last automatic payment, whichever comes first.

Payment method

Add credit or debit card

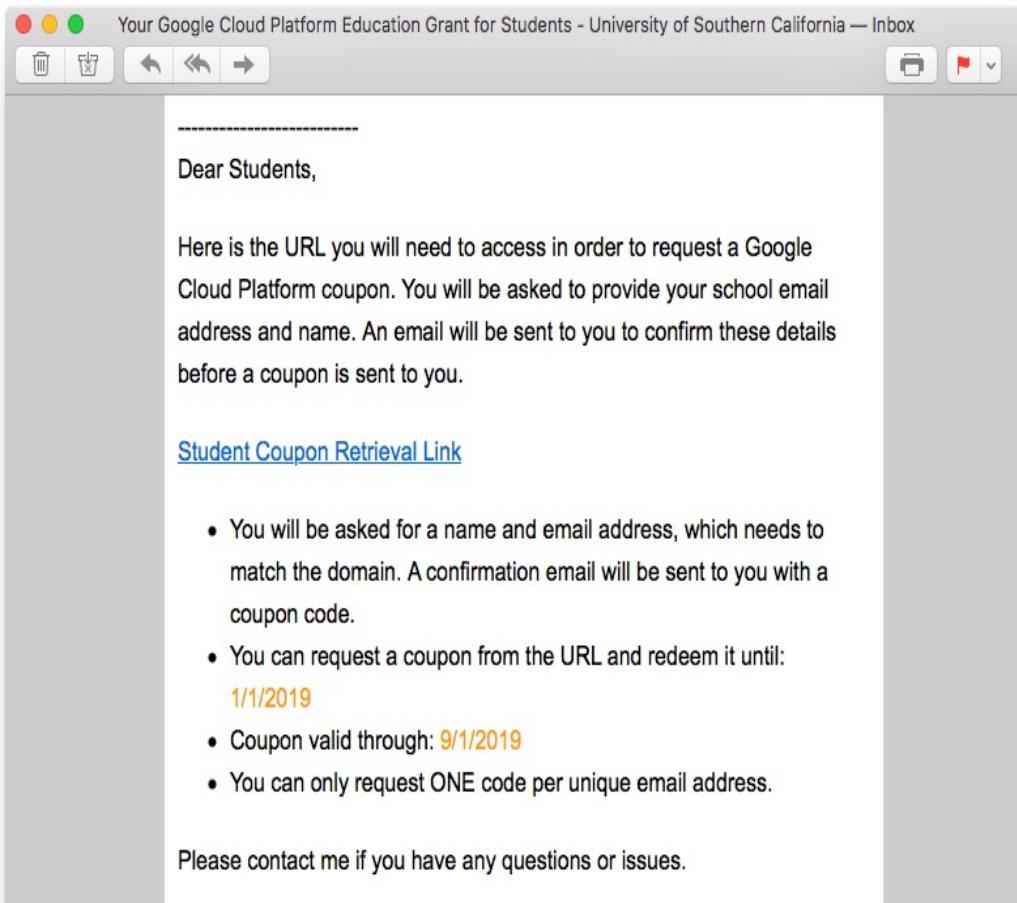


Google Cloud Platform Education Grants Credit

If you do not have a credit card, Google provides you with a coupon code via the Google Cloud Platform Education Grants program. If you do have a credit card, you can sign up for the Google Cloud Platform “Free Trial”.

On Piazza and by e-mail, you will receive a communication like the one displayed in the image.

Click on the Student Coupon Retrieval Link. New window will open shown in the next slide.

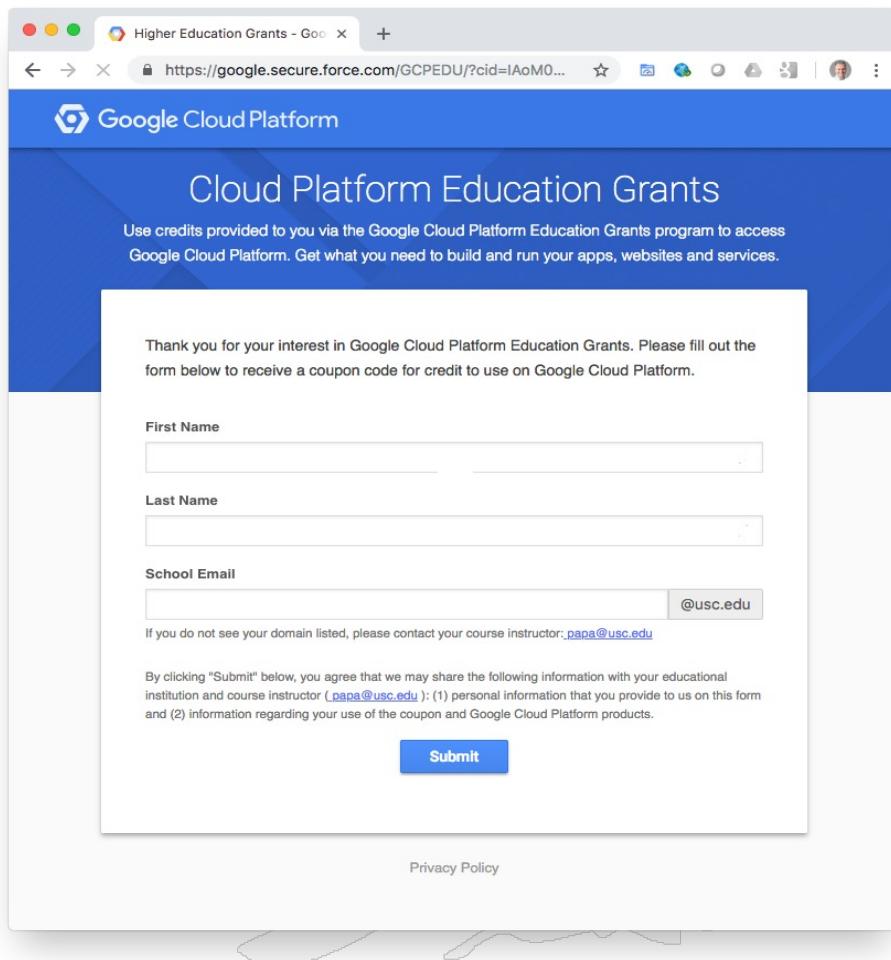


Google Cloud Platform Education Grants Credit

Enter your First Name, Last Name and your USC e-mail address. @usc.edu will be pre-filled. Click on Submit. If you entered a valid USC e-mail address, an email will be sent to that USC email address to verify that you own such address.

Once your USC email address is “verified”, you will receive a second email with a Google Cloud Platform Coupon Code in it.

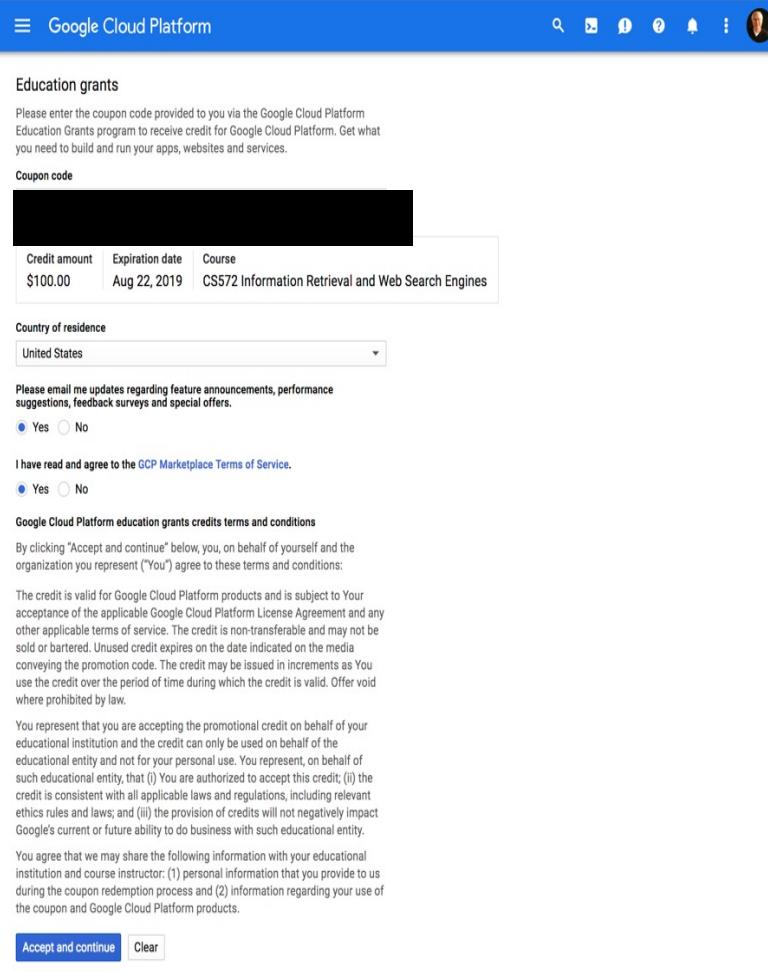
**More details in Homework #3,
Instructions for Setting Up Google
Cloud Account documents (Refer
1.1)**



The screenshot shows a web browser window titled "Higher Education Grants - Google". The URL is https://google.secure.force.com/GCPEDU/?cid=IAoM0... . The page is titled "Google Cloud Platform" and "Cloud Platform Education Grants". It explains that credits are provided via the Google Cloud Platform Education Grants program to access Google Cloud Platform for building and running apps, websites, and services. A message thanks the user for their interest and asks them to fill out a form to receive a coupon code for credit. The form fields are "First Name", "Last Name", and "School Email". The "School Email" field includes "@usc.edu". Below the form, a note says if the domain is not listed, contact the course instructor at papa@usc.edu. A terms and conditions section states that by clicking "Submit", the user agrees to share information with the educational institution and course instructor. A "Submit" button is at the bottom, and a "Privacy Policy" link is at the bottom right.

Google Cloud Platform Education Grants Credit

- Before clicking on the link labeled in the email, you should open your default browser, and login to a Gmail account. Every USC student has been provided with a Gmail account. Once logged into Gmail, you can click on link in the mail , or you can go to this page:
<https://console.cloud.google.com/education> to redeem your coupon. The web form below will be displayed.
- You need to paste your coupon into the field labeled Coupon code. Click on Accept and continue. You will now be taken to the Google Cloud Platform's Billing section, and the amount of your credit will be displayed

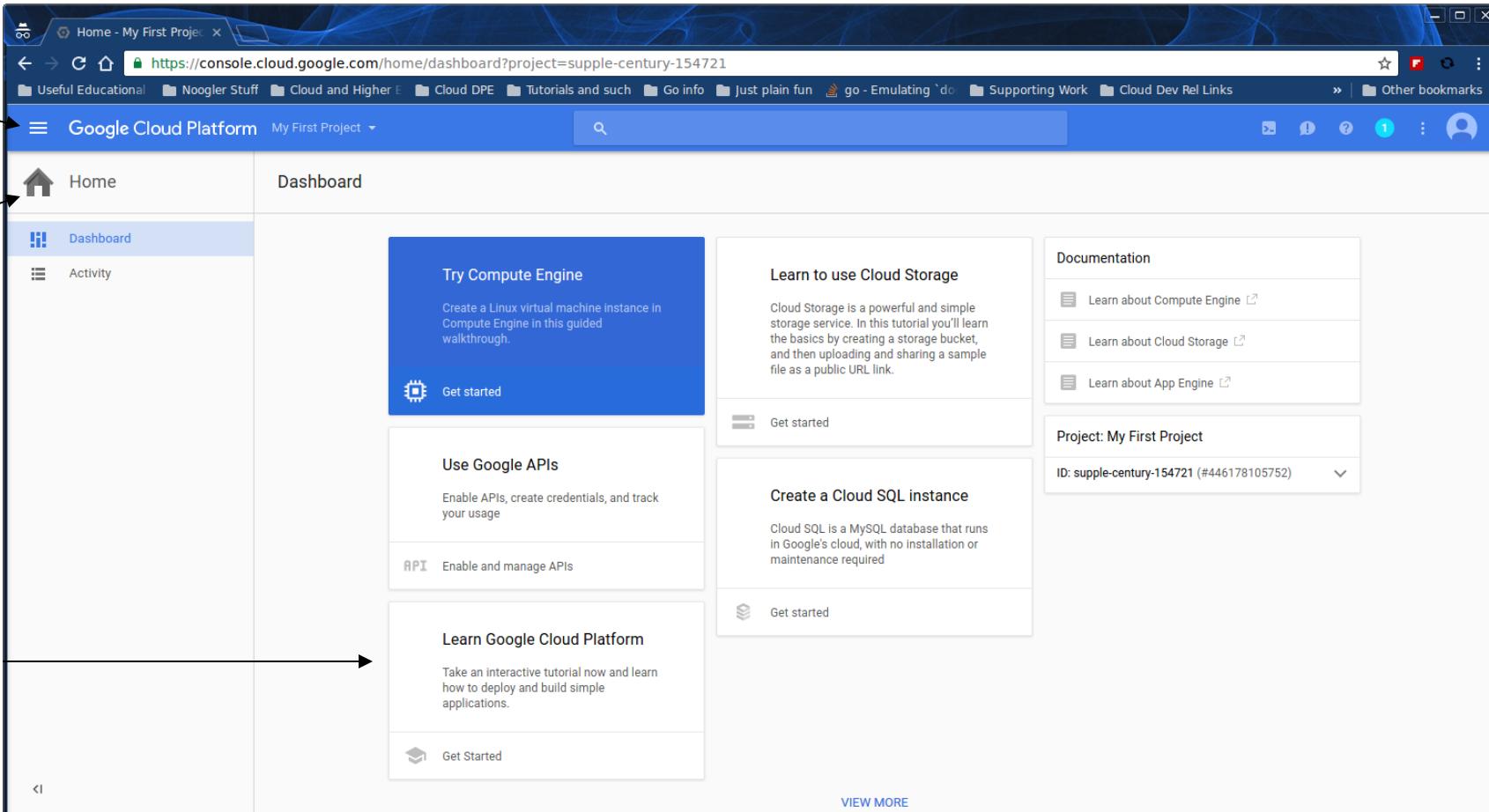


The screenshot shows a web form titled "Google Cloud Platform" with a blue header bar. The main title is "Education grants". Below it, a message says: "Please enter the coupon code provided to you via the Google Cloud Platform Education Grants program to receive credit for Google Cloud Platform. Get what you need to build and run your apps, websites and services." A "Coupon code" input field is shown with a blacked-out value. Below it, a table shows "Credit amount" (\$100.00), "Expiration date" (Aug 22, 2019), and "Course" (CS572 Information Retrieval and Web Search Engines). A "Country of residence" dropdown menu is set to "United States". There are two sections for "Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers." and "I have read and agree to the GCP Marketplace Terms of Service." Both sections have "Yes" radio buttons selected. At the bottom, there are sections for "Google Cloud Platform education grants credits terms and conditions" and "You represent that you are accepting the promotional credit on behalf of your educational institution and the credit can only be used on behalf of the educational entity and not for your personal use. You represent, on behalf of such educational entity, that (i) You are authorized to accept this credit; (ii) the credit is consistent with all applicable laws and regulations, including relevant ethics rules and laws; and (iii) the provision of credits will not negatively impact Google's current or future ability to do business with such educational entity." A note at the bottom states: "You agree that we may share the following information with your educational institution and course instructor: (1) personal information that you provide to us during the coupon redemption process and (2) information regarding your use of the coupon and Google Cloud Platform products." At the very bottom are "Accept and continue" and "Clear" buttons.

Google Cloud Home Page

<https://console.cloud.google.com>

There are two menus available from the console: main and context



The screenshot shows the Google Cloud Platform Home Page for a project named "My First Project". The URL is <https://console.cloud.google.com/home/dashboard?project=supple-century-154721>. The page features a main navigation bar at the top with links like "Useful Educational", "Noogler Stuff", "Cloud and Higher E...", "Cloud DPE", "Tutorials and such", "Go info", "Just plain fun", "go - Emulating 'do", "Supporting Work", "Cloud Dev Rel Links", and "Other bookmarks". Below the main menu, there are several cards:

- Try Compute Engine**: Create a Linux virtual machine instance in Compute Engine in this guided walkthrough. [Get started](#)
- Learn to use Cloud Storage**: Cloud Storage is a powerful and simple storage service. In this tutorial you'll learn the basics by creating a storage bucket, and then uploading and sharing a sample file as a public URL link. [Get started](#)
- Use Google APIs**: Enable APIs, create credentials, and track your usage. [API](#) [Enable and manage APIs](#)
- Create a Cloud SQL instance**: Cloud SQL is a MySQL database that runs in Google's cloud, with no installation or maintenance required. [Get started](#)
- Learn Google Cloud Platform**: Take an interactive tutorial now and learn how to deploy and build simple applications. [Get Started](#)

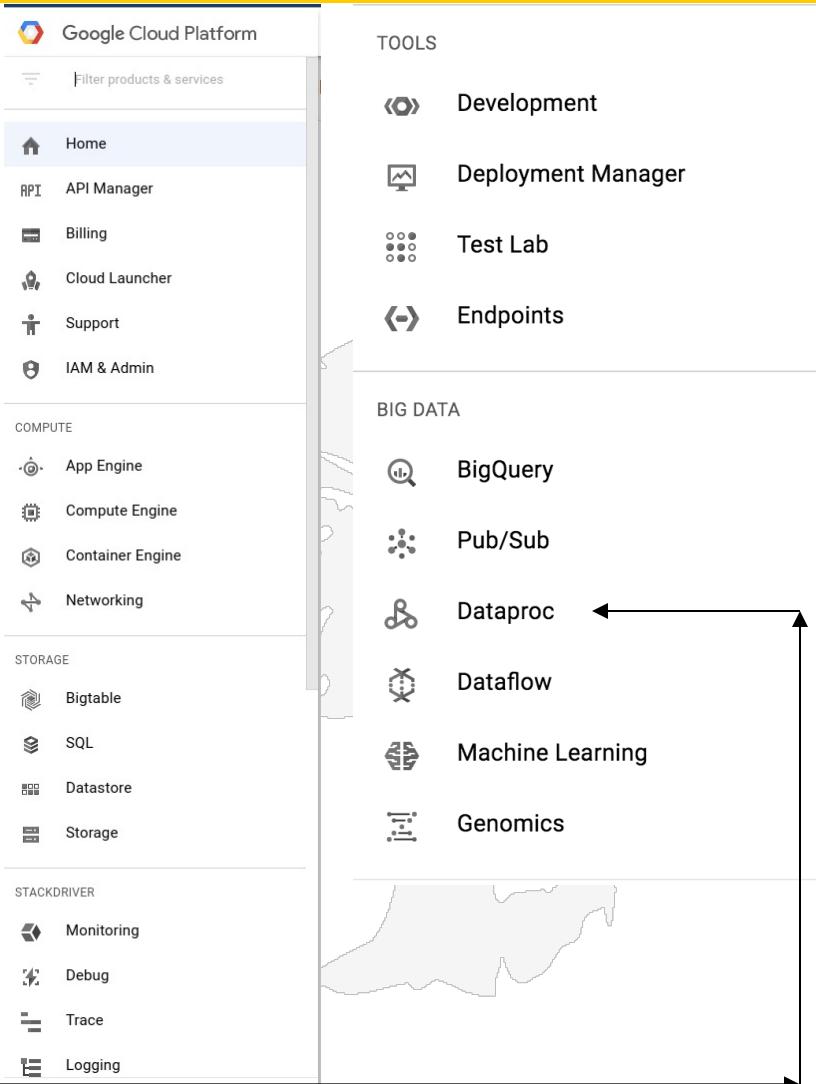
On the right side, there is a sidebar with sections for "Documentation" (links to Compute Engine, Cloud Storage, and App Engine), "Project: My First Project" (ID: supple-century-154721), and a "VIEW MORE" button.

Annotations on the left side of the screenshot:

- Main**: Points to the main navigation bar at the top.
- context**: Points to the context menu (the vertical list of items) on the left side of the dashboard.
- tutorials**: Points to the "Tutorials and such" link in the main navigation bar.

Google Cloud Main Menu

- From the hamburger menu in the top left corner, you can access a menu that brings you to the 5 major components of Google Cloud Platform (GCP):
 - Compute
 - Storage
 - Stackdriver (company to manage distributed apps running on the cloud)
 - Tools
 - Big Data
- For this exercise you will use DataProc within BIG DATA to set up a cluster of compute instances



Context Menu

The context menu changes based on the current major component
 Here are three examples

-  Datastore
-  Entities
-  Dashboard
-  Indexes
-  Admin

 Compute Engine

-  VM instances
-  Instance groups
-  Instance templates
-  Disks
-  Snapshots
-  Images
-  Metadata
-  Health checks
-  Zones
-  Operations
-  Quotas
-  Settings

 App Engine

-  Dashboard
-  Services
-  Versions
-  Instances
-  Task queues
-  Security scans
-  Quotas
-  Blobstore
-  Memcache
-  Search
-  Settings

Snapchat was built on top of App Engine

App Engine lets clients host their software at datacenters managed by Google

App Engine is a Platform as a Service (PaaS) while Compute Engine is an Infrastructure as a Service (IaaS)

For the App Engine you just write your code and it automatically executes

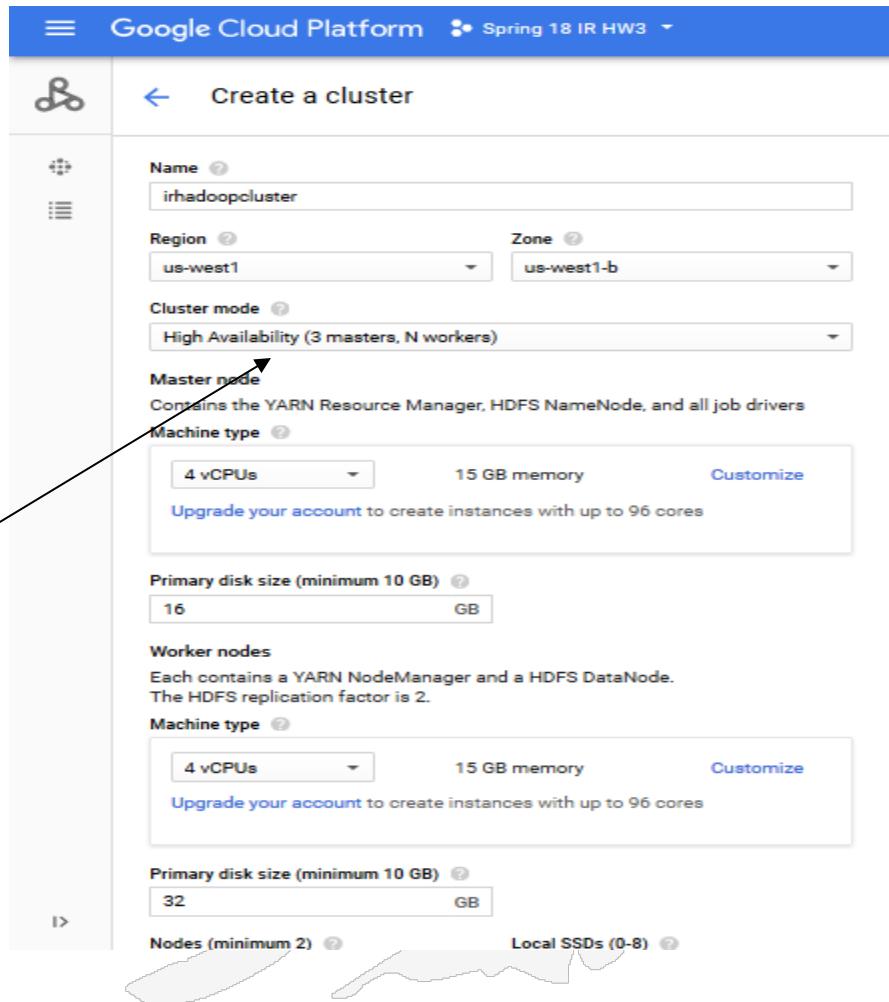
- Snap (Snapchat) recently signed a \$2 billion, five year contract with Google for its cloud services, which makes Snap Google's largest customer of its cloud platform
- Apple confirms it is using Google cloud for iCloud services

Create a Cluster

Using the **Google Cloud Platform** Console you can **create a cluster** by going to the **Cloud Platform** Console.

Select your project, and then click Continue to open the **Clusters** page.

Contains 1 master node and 3 worker nodes



The screenshot shows the 'Create a cluster' page in the Google Cloud Platform. The 'Name' field is set to 'irhadoopcluster'. The 'Region' is 'us-west1' and the 'Zone' is 'us-west1-b'. The 'Cluster mode' is set to 'High Availability (3 masters, N workers)'. An arrow points to the 'Master node' section, which describes it as containing the YARN Resource Manager, HDFS NameNode, and all job drivers. The 'Machine type' for the master node is set to 4 vCPUs and 15 GB memory. The 'Primary disk size' is 16 GB. The 'Worker nodes' section indicates each contains a YARN NodeManager and a HDFS DataNode, with an HDFS replication factor of 2. The 'Machine type' for worker nodes is also set to 4 vCPUs and 15 GB memory, with a primary disk size of 32 GB. The 'Nodes (minimum 2)' field is highlighted.

Successful Creation of a Cluster

Cloud Dataproc

Clusters

CREATE CLUSTER

REFRESH

DELETE

Name	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
cluster-1	asia-east1-b	3	dataproc-723aedf5-c6dd-4fc7-b841-a1d731e6abaf-asia	Feb 19, 2017, 9:46:01 AM	Running

This URL will let you ssh to your cluster



ssh into the Cluster

Dataproc ← cluster-572-grader + SUBMIT JOB ⌛ REFRESH 🗑 DELETE ⏷ VIEW LOGS

Clusters

For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See <https://cloud.google.com/compute/docs/disks/performance> for information on disk I/O performance.

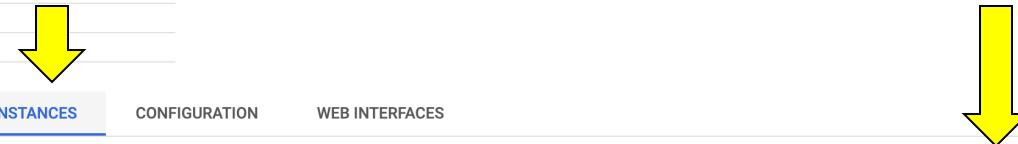
Name	cluster-572-grader
Cluster UUID	706f39f6-ae82-4e2a-9d9d-4a872902454b
Type	Dataproc Cluster
Status	Running

MONITORING JOBS **VM INSTANCES** CONFIGURATION WEB INTERFACES

Filter instances

	Name ↑	Role
●	cluster-572-grader-m	Master
○	cluster-572-grader-w-0	Worker
○	cluster-572-grader-w-1	Worker

Equivalent [REST](#)



Environment Variables Set

Create a home directory

Check env variables

JAVA_HOME

HADOOP_CLASSPATH

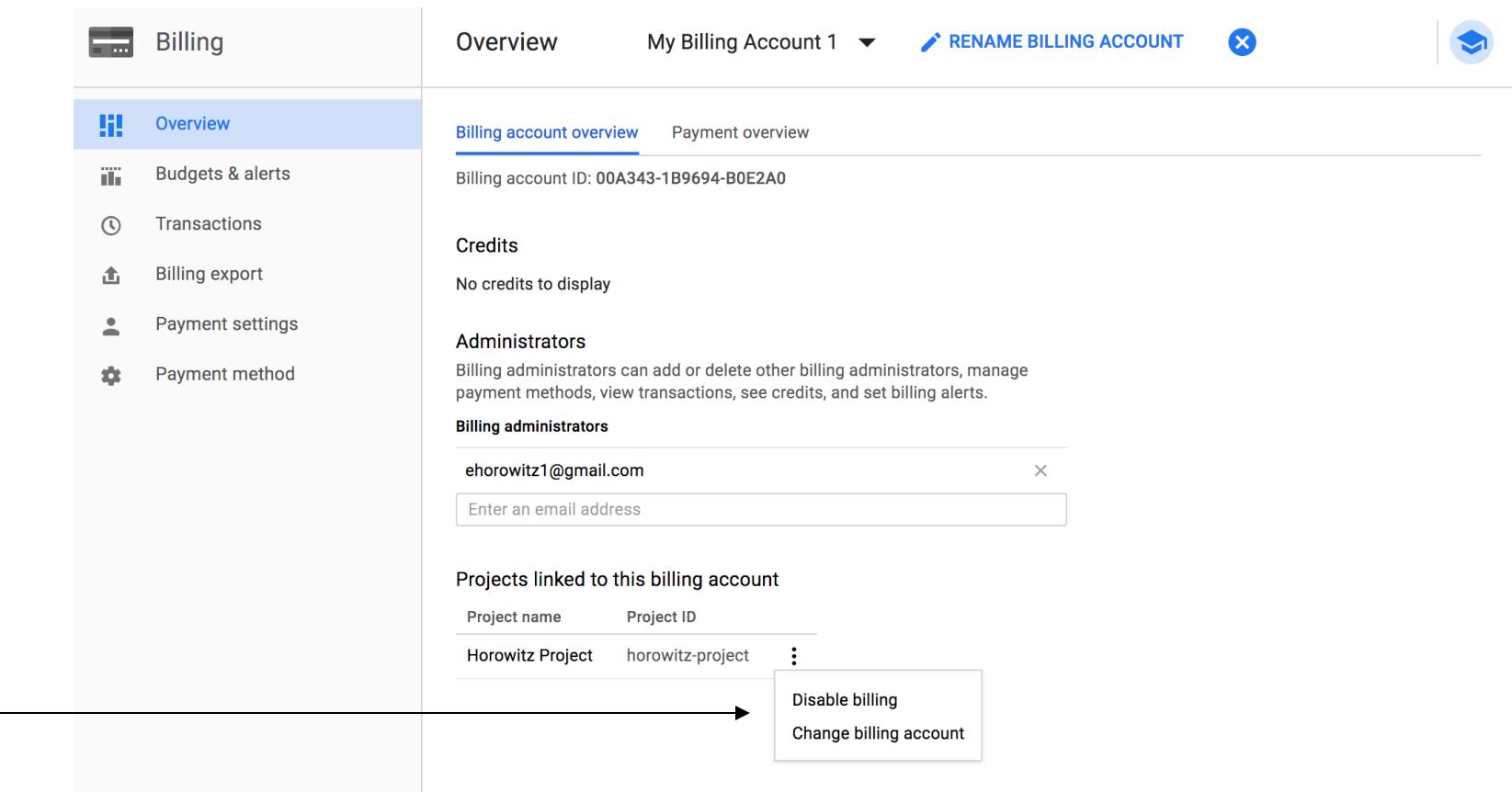
```

ehorowitz1@cluster-1-m: ~
Secure | https://ssh.cloud.google.com/projects/horowitz-project/zones/asia-east1-b/instances/cluster-1-m?authuser=0&hl=en_US&...
ehorowitz1@cluster-1-m:~$ env
TERM=xterm-256color
SHELL=/bin/bash
SSH_CLIENT=173.194.90.33 63226 22
SSH_TTY=/dev/pts/0
USER=ehorowitz1
LS_COLORS=rs=0:di=01;34:ln=01;36:mh=00:pi=40;33:so=01;35:do=01;35:bd=40;33:01:cd=40;33:01:or=40;31:01:su=37;41:sg=3
0;43:ca=30;41:tw=30;42:ow=34;42:st=37;44:ex=01;32:*.tar=01;31:*.tgz=01;31:*.arc=01;31:*.arj=01;31:*.taz=01;31:*.lha
=01;31:*.lz4=01;31:*.lzh=01;31:*.lzma=01;31:*.tlz=01;31:*.txz=01;31:*.tzo=01;31:*.tz=01;31:*.zip=01;31:*.z=01;31:*
.Z=01;31:*.dz=01;31:*.gz=01;31:*.lrz=01;31:*.lz=01;31:*.lzo=01;31:*.xz=01;31:*.bz2=01;31:*.bz=01;31:*.tbz=01;31:*.t
bz=01;31:*.tz=01;31:*.deb=01;31:*.rpm=01;31:*.jar=01;31:*.war=01;31:*.ear=01;31:*.sar=01;31:*.rar=01;31:*.alz=01;3
1:*.ace=01;31:*.zoo=01;31:*.cpio=01;31:*.7z=01;31:*.rz=01;31:*.cab=01;31:*.jpg=01;35:*.jpeg=01;35:*.gif=01;35:*.bmp
=01;35:*.pbm=01;35:*.pgm=01;35:*.ppm=01;35:*.tga=01;35:*.xbm=01;35:*.xpm=01;35:*.tif=01;35:*.tiff=01;35:*.png=01;35
:*.svg=01;35:*.svgz=01;35:*.mng=01;35:*.pcx=01;35:*.mov=01;35:*.mpg=01;35:*.mpeg=01;35:*.m2v=01;35:*.mkv=01;35:*.we
bm=01;35:*.ogm=01;35:*.mp4=01;35:*.m4v=01;35:*.mp4v=01;35:*.vob=01;35:*.qt=01;35:*.nuv=01;35:*.wmv=01;35:*.ASF=01;3
5:*.rm=01;35:*.rmvb=01;35:*.flc=01;35:*.avi=01;35:*.fli=01;35:*.flv=01;35:*.gl=01;35:*.dl=01;35:*.xcf=01;35:*.xwd=0
1;35:*.yuv=01;35:*.cgm=01;35:*.emf=01;35:*.axv=01;35:*.anx=01;35:*.ovg=01;35:*.ogg=01;35:*.aac=00;36:*.au=00;36:*.f
lac=00;36:*.m4a=00;36:*.mid=00;36:*.mka=00;36:*.mp3=00;36:*.mpc=00;36:*.ogg=00;36:*.ra=00;36:*.wav=00;
36:*.axa=00;36:*.oga=00;36:*.spx=00;36:*.xspf=00;36:
DATAPROC_MASTER_HA_COMPONENTS=hadoop-hdfs-journalnode hadoop-hdfs-zkfc zookeeper-server
SSH_AUTH_SOCK=/tmp/ssh-zGxHCr3rJ9/agent.3623
DATAPROC_MASTER_COMPONENTS=hadoop-hdfs-namenode hadoop-yarn-resourcemanager mysql-server
MAIL=/var/mail/ehorowitz1
PATH=/usr/local/bin:/usr/bin:/usr/local/games:/usr/games
PWD=/home/ehorowitz1
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
HADOOP_CLASSPATH=/lib/tools.jar
LANG=en_US.UTF-8
DATAPROC_COMMON_COMPONENTS=openjdk-8-jdk libjansi-java python-numpy libmysql-java hadoop-client hive pig spark-core
spark-python spark-r autofs nfs-common libhdfs0 libsnappy1 libatlas3-base libopenblas-base libapr1 vim git bash-completion
spark-yarn-shuffle spark-datanucleus spark-extras hadoop-lzo
DATAPROC_MASTER_STANDALONE_COMPONENTS=hadoop-hdfs-secondarynamenode
ALPN_JAR=/usr/local/share/google/alpn/alpn-boot-8.1.7.v20160121.jar
DATAPROC_WORKER_COMPONENTS=hadoop-hdfs-datanode hadoop-yarn-nodemanager
SHLVL=1
HOME=/home/ehorowitz1
BDUTIL_DIR=/usr/local/share/google/dataproc/bdutil-dataproc-20170214-094528-RC1
LOGNAME=ehorowitz1
SSH_CONNECTION=173.194.90.33 63226 10.140.0.4 22
DATAPROC_AGENT_JAR=/usr/local/share/google/dataproc/agent-20170214-094528-RC1.jar
DATAPROC_MASTER_EXCLUSIVE_COMPONENTS=hadoop-mapreduce-historyserver hive-metastore hive-server2 nfs-kernel-server spark-history-server _=/usr/bin/env

```

Disable Billing for Your Cluster

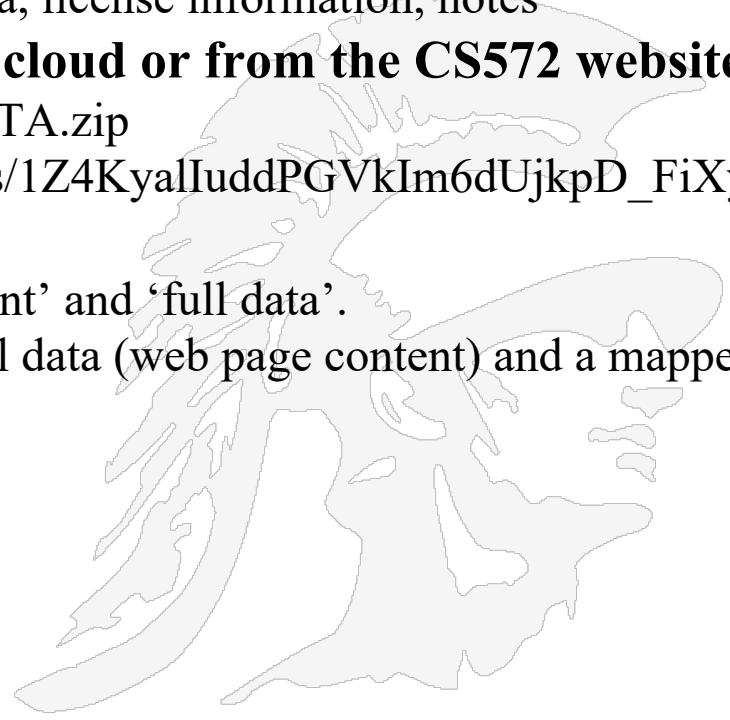
- Please **disable** the billing for the cluster when you are not using it.
- Leaving it running will cost extra credits.
- The cluster is billed based on how many hours it is running and not how much data it is processing



The screenshot shows the Google Cloud Billing Overview page. On the left, a sidebar menu includes 'Overview' (which is selected and highlighted in blue), 'Budgets & alerts', 'Transactions', 'Billing export', 'Payment settings', and 'Payment method'. The main content area has tabs for 'Billing account overview' (selected) and 'Payment overview'. It displays a 'Billing account ID: 00A343-1B9694-B0E2AO'. Under 'Credits', it says 'No credits to display'. Under 'Administrators', it says 'Billing administrators can add or delete other billing administrators, manage payment methods, view transactions, see credits, and set billing alerts.' A list of administrators includes 'ehorowitz1@gmail.com' with a remove button ('X') and a text input field 'Enter an email address'. Below this, a section titled 'Projects linked to this billing account' lists 'Horowitz Project' with 'Project ID: horowitz-project'. To the right of this list is a vertical ellipsis ('...'), followed by a dropdown menu with options 'Disable billing' and 'Change billing account'. A large red arrow points from the bottom left towards this dropdown menu.

Upload the Data Set

- **We'll be using a collection of 74 files of web pages whose HTML has been removed**
 - The data comes from <https://ebiquity.umbc.edu/resource/html/id/351>
 - The data has been cleaned of metadata, license information, notes
- **Retrieve the dataset either from the cloud or from the CS572 website**
 - <http://csci572.com/2022Fall/hw3/DATA.zip>
 - https://drive.google.com/drive/folders/1Z4KyalIuddPGVkIm6dUjkpD_FiXyNICq
- **Unzip the contents**
 - two folders inside named ‘development’ and ‘full-data’.
 - Each of the folders contains the actual data (web page content) and a mapper file to map the docID to the file name.



Uploading data to GCP

- You can upload data to GCP using the GUI, mentioned in the document **Hadoop Exercise to Create an Inverted Index**:

<http://csci572.com/2022Fall/hw3/HadoopExercise.pdf>

-OR-

- Install Google Cloud SDK Shell using this link:

https://cloud.google.com/storage/docs/gsutil_install

- Select your operating system from the options provided
- Follow the instructions mentioned
- Upload data using:

```
gsutil -m cp -r "\Users\Documents\fullData" "gs://bucket/foldername"
```
- *Instead of \Users\Documents\fullData provide the path on your computer to data.
- *Instead of gs://bucket/foldername provide the directory on your GCP storage
 - Eg: gs://dataproc-e2659b1-9ce7-4d18-93b1-83c013225-us-west1/Data
- An example command is:

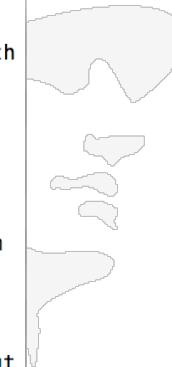
```
gsutil -m cp -r "C:\Users\IR\Docs\Data\devdata" "gs://dataproc-e2659b1-9ce7-4d18-93b1-83c013225-us-west1/Data"
```



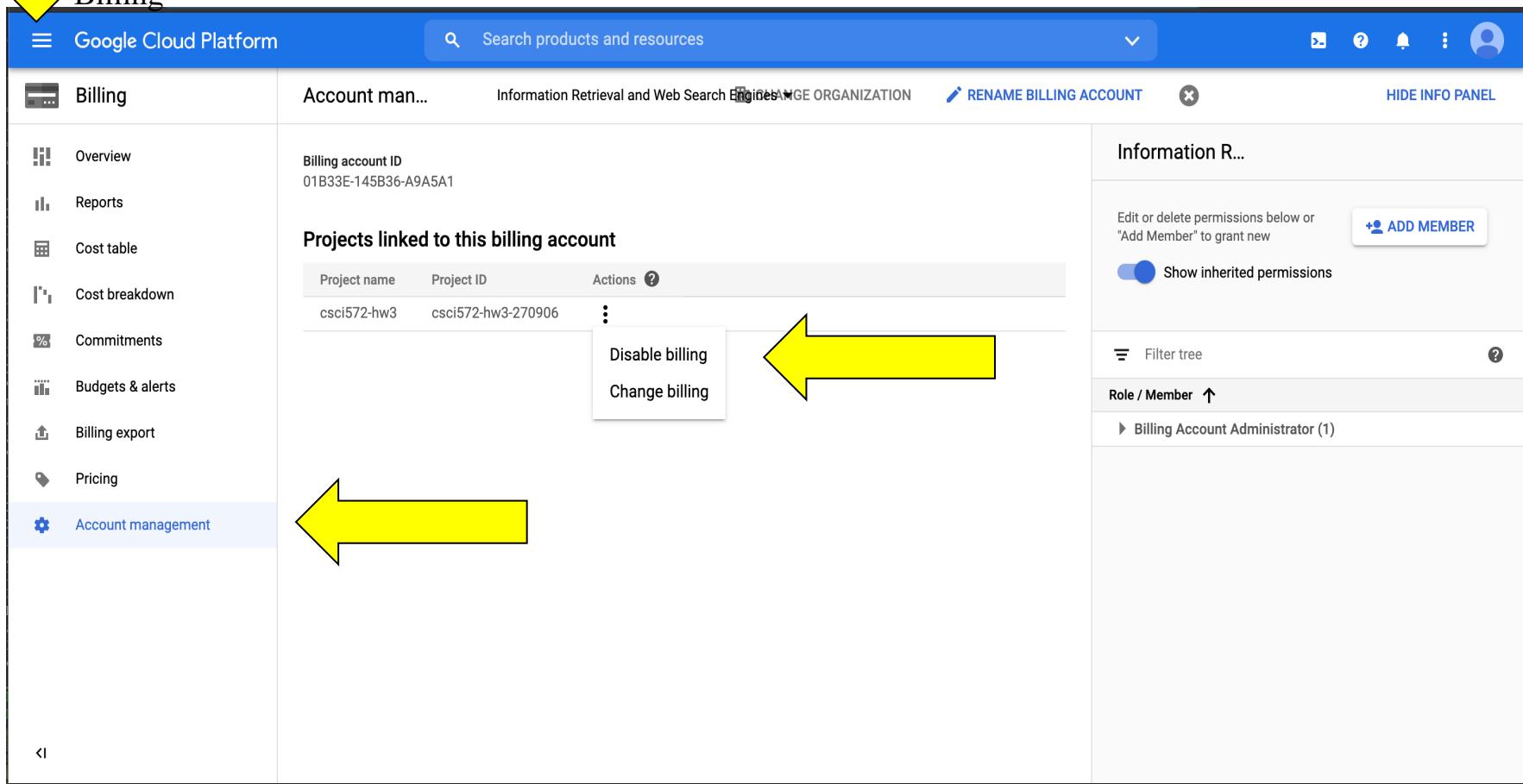
Development Data

	--	Folder	Today at 9:35 AM
▼ DATA			
▼ devdata			
5722018101.txt	43.6 MB	Plain Text	Today at 9:35 AM
5722018235.txt	41.8 MB	Plain Text	Today at 9:35 AM
5722018301.txt	36.4 MB	Plain Text	Today at 9:35 AM
5722018496.txt	12.8 MB	Plain Text	Today at 9:35 AM
5722018508.txt	52.4 MB	Plain Text	Today at 9:35 AM
▼ fulldata			
5722018435.txt	47.3 MB	Plain Text	Today at 9:35 AM
5722018436.txt	47.6 MB	Plain Text	Today at 9:35 AM
5722018437.txt	49.1 MB	Plain Text	Today at 9:35 AM
5722018438.txt	41.6 MB	Plain Text	Today at 9:35 AM
5722018439.txt	44.5 MB	Plain Text	Today at 9:35 AM
5722018440.txt	46.6 MB	Plain Text	Today at 9:35 AM

5722018101 "The DeLorme PN-20 represents a new breed of GPS devices.a fantastic device, and it leads the way in a new breed of GPS devices which can display aerial photography and satellite imagery. For people who have dreamed about having a Google Earth type product in a handheld device... this is it."November 25, 2005 marked Something Fishy's ten year anniversary on the web! We are one of the largest, oldest and most comprehensive web sites available on the Anorexia, Bulimia, Compulsive Overeating and Binge Eating Disorders, providing information and support to sufferers and their loved ones. Do you know your family members? You might not know them as well as you think. We gathered a listing of comments from members about what they are really feeling and what they wished their friends and family really knew about them. If You Really Knew Me...Our comprehensive eating disorders treatment finder at Something Fishy contains listings from over 1,800 therapists, dieticians, treatment centers and other professionals worldwide working to help those with Anorexia, Bulimia, Compulsive Overeating and Binge Eating Disorder recover. Fully searchable by category (type of treatment), country, state, area code, name, services, description or zipcode. Our Mission: We are dedicated to raising awareness about eating disorders... emphasizing always that eating disorders are NOT about food and weight; They are just the symptoms of something deeper going on, inside. Something Fishy is determined to remind each and every sufferer of anorexia, bulimia, compulsive overeating and binge eating disorder that they are not alone, and that complete recovery is possible. If you are the loved-one of someone that suffers with an eating disorder, use this website to educate yourself. The more you know, the more you are equipped to provide the support your loved-one needs. If you have an eating disorder, you can find help. You can recover. And you deserve to do both. Though our site should be friendly to most browsers it is best viewed on Internet Explorer 4 (or higher) or Netscape



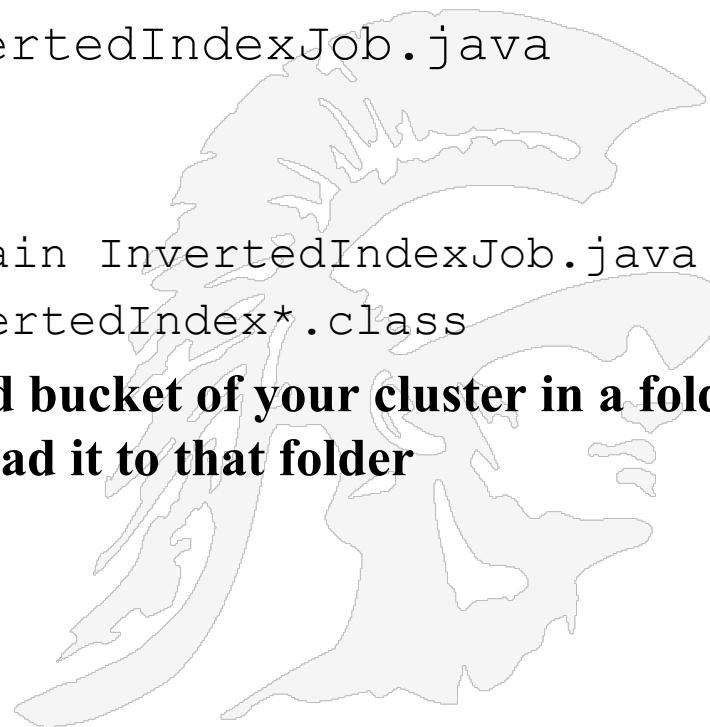
Click here
and
select
Billing



The screenshot shows the Google Cloud Platform Billing interface. On the left, a sidebar menu lists various billing-related options: Overview, Reports, Cost table, Cost breakdown, Commitments, Budgets & alerts, Billing export, Pricing, and Account management. The 'Account management' option is highlighted with a blue arrow pointing to it from the bottom left. In the main content area, the 'Billing' tab is selected. It displays the 'Information Retrieval and Web Search Engine' billing account, which has a Billing account ID of 01B33E-145B36-A9A5A1. A table titled 'Projects linked to this billing account' shows one project: 'csci572-hw3' with Project ID 'csci572-hw3-270906'. To the right of this table, a context menu is open with two options: 'Disable billing' and 'Change billing'. A large yellow arrow points from the text 'Click here and select Billing' towards this context menu. At the top right of the main content area, there are buttons for 'CHANGE ORGANIZATION', 'RENAME BILLING ACCOUNT', and 'HIDE INFO PANEL'. The 'Information R...' panel on the right contains sections for 'Edit or delete permissions below or "Add Member" to grant new' (with an 'ADD MEMBER' button), 'Show inherited permissions' (which is turned on), 'Filter tree', and a 'Role / Member' section showing 'Billing Account Administrator (1)'. The top navigation bar includes the 'Google Cloud Platform' logo, a search bar, and various user icons.

Inverted Index Implementation

- You need to write some code, in Java, that processes the data file of web pages and produces an inverted index of the words that occur
- Google Cloud requires the code to be packaged as a jar file, e.g.
- If your Java program is called InvertedIndexJob.java
 - first compile the code and then
 - run the jar program
- hadoop com.sun.tools.javac.Main InvertedIndexJob.java
- jar cf invertedindex.jar InvertedIndex*.class
- Place this jar file in the default cloud bucket of your cluster in a folder called JAR on your bucket and upload it to that folder



Mapper Class

The Google cluster requires that you write two routines to implement the Map/Reduce functionality

A lecture on Map/Reduce is coming later

Class is WordCountMapper;
Routine is map(key,value,context)

Program reads a line of text, and for each token (word) that it finds on the line it sends the pair (token, 1) to the collector/reducer

```
/*
This is the Mapper class. It extends the Hadoop's Mapper class.
This maps input key/value pairs to a set of intermediate(output) key/value pairs.
Here our input key is a LongWritable and input value is a Text.
And the output key is a Text and value is an IntWritable.

*/
class WordCountMapper extends Mapper<LongWritable, Text, Text, IntWritable>
{
    /*
    Hadoop supported data types. This is a Hadoop specific datatype that is used to handle
    numbers and Strings in a hadoop environment. IntWritable and Text are used instead of
    Java's Integer and String datatypes.
    Here 'one' is the number of occurrences of the 'word' and is set to the value 1 during the
    Map process.
    */
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException
    {
        //Reading input one line at a time and tokenizing.
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);

        //Iterating through all the words available in that line and forming the key value pair.
        while (tokenizer.hasMoreTokens())
        {
            word.set(tokenizer.nextToken());
            /*
            Sending to output collector(Context) which in-turn passes the output to Reducer.
            The output is as follows:
                'word1' 1
                'word1' 1
                'word2' 1
            */
            context.write(word, one);
        }
    }
}
```

Reducer Class

Class is WordCountReducer;
Program is
reduce(key, values, context)

For each key (word), the number
of occurrences are summed
together and written out

```
/*
This is the Reducer class. It extends the Hadoop's Reducer class.
This maps the intermediate key/value pairs we get from the mapper to a set
of output key/value pairs, where the key is the word and the value is the word's count.
Here our input key is a Text and input value is a IntWritable.
And the output key is a Text and value is an IntWritable.
*/
class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable>
{
    /*
    Reduce method collects the output of the Mapper and adds the 1's to get the word's count.
    */
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException
    {
        int sum = 0;
        /*
        Iterates through all the values available with a key and add them together and give the
        final result as the key and sum of its values
        */
        for (IntWritable value : values)
        {
            sum += value.get();
        }
        context.write(key, new IntWritable(sum));
    }
}
```



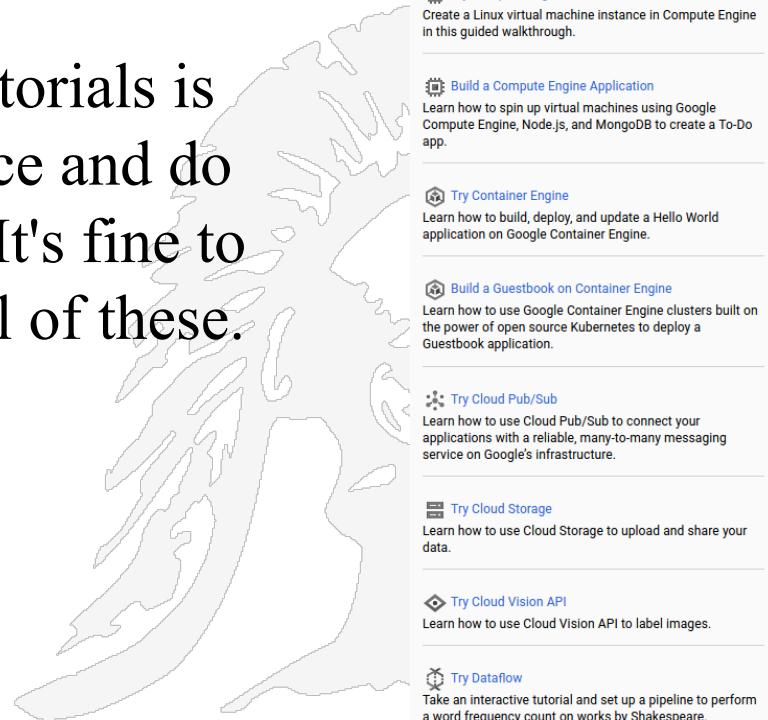
Main Class

Class WordCount;
Program: main
Create the Hadoop
job

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.*;
public class WordCount
{
    public static void main(String[] args)
        throws IOException, ClassNotFoundException, InterruptedException {
        if (args.length != 2) {
            System.err.println("Usage: Word Count <input path> <output path>");
            System.exit(-1);
        }
        //Creating a Hadoop job and assigning a job name for identification.
        Job job = new Job();
        job.setJarByClass(WordCount.class);
        job.setJobName("Word Count");
        //The HDFS input and output directories to be fetched from the Dataproc job submission console.
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        //Providing the mapper and reducer class names.
        job.setMapperClass(WordCountMapper.class);
        job.setReducerClass(WordCountReducer.class);
        //Setting the job object with the data types of output key(Text) and value(IntWritable).
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        job.waitForCompletion(true);
    }
}
```

Built-in Tutorials

- Click the tricolon at the top right of the console and select "*Try an interactive tutorial*" to be brought to this list of tutorials
- An advantage of these built-in tutorials is they'll step you through each piece and do necessary project management. It's fine to use the default first project for all of these.



Sample Document and Output

5722018411 A look at the most publicized aspects of the strike-- economics, stress, and management-- shows how these issues obscured and distorted the controllers' main concern of workplace control and helps explain why problems persist in the ATC workforce. It also demonstrates how management and labor's focus on economic issues since World War II has bankrupted labor's discourse and limited its ability to address concerns outside of a narrow range of concerns. These perceptions were in part responsible for the overwhelming public approval of Reagan's handling of the strike; 65% in a public opinion poll; mail, according to one representative, ran 1000 to 1 in favor of the administration. Most strikers denied that money was a critical component in their decision to strike. Yet Poli insisted that his demands, headed by a pay raise, reflected the desires of his constituency. Arthur Shostak, who conducted five surveys of PATCO members in 1979 and 1980 backs up Poli's assertion that salary was important to the strikers. It is tempting to concede then that workers did see the

Sample of Mapper Output

```
aspect 5722018411
distorted 5722018411
economics 5722018411
economics 5722018411
management 5722018411
publicized 5722018411
```

Sample of Reducer Output

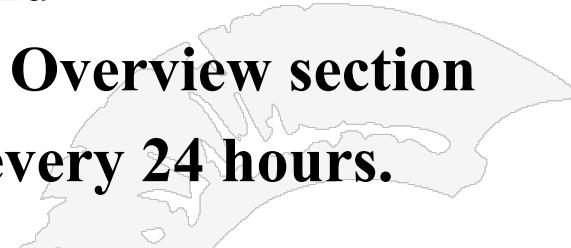
1	answer	5722018453:2	5722018483:1						
2	antecedence	5722018502:1	5722018435:1						
3	asterisks.	5722018417:1	5722018504:2	5722018447:1					
4	beautiful	5722018439:7	5722018417:2	5722018416:3	5722018438:5	5722018437:1	5722018415:1	5722018414:2	5722018435:3
5	bind	5722018419:6	5722018417:39	5722018416:1					
6	chunking	5722018507:1	5722018502:1						

aspect occurred 1 time in the document with docID 5722018411
economics occurred 2 times in the document with docID is 5722018411



Credits Spent

- To check how much you've been charged for your cluster,
 - navigate to the Billing section and
 - click on the project name in the Overview section
 - check this section at least once every 24 hours.



 Billing Overview CS572 Information Retrieval ▾  RENAME BILLING ACCOUNT

 Overview Billing account ID: 00AB19-9B02F2-9EA431

 Budgets & alerts

 Billing export

Credits

Promotion ID	Expires ▾	Promotion value	Amount remaining
CS572 Information Retrieval	Jan 9, 2018	\$150.00	\$92.16

Projects linked to this billing account

Project name	Project ID	⋮
My First Project	euphoric-drive-158517	⋮

 Billing Charges this month

[Go to billing](#)

Product	Resource	Usage	Amount
Google Compute	Standard Intel N1 4 VCPU running in JP	11,280 Minutes	\$49.86
Google Compute	Google Cloud Dataproc running on VM Image CORE	45,120 Minutes	\$7.52
Google Compute	Storage Pd Capacity Jp	8.95 GB-month	\$0.47
Google Compute	Storage Pd Capacity Jp	Credit applied	\$-0.47
Google Compute	Google Cloud Dataproc running on VM Image CORE	Credit applied	\$-7.52
Google Compute	Standard Intel N1 4 VCPU running in JP	Credit applied	\$-49.86

*Estimated charges before taxes, updated daily Total: \$0.00