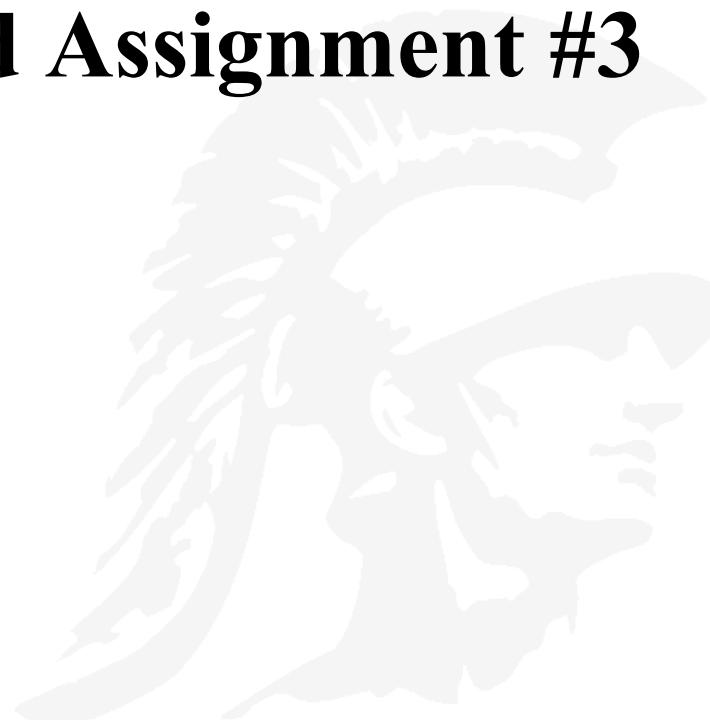


Google Cloud and Assignment #3



Computing is Rapidly Changing

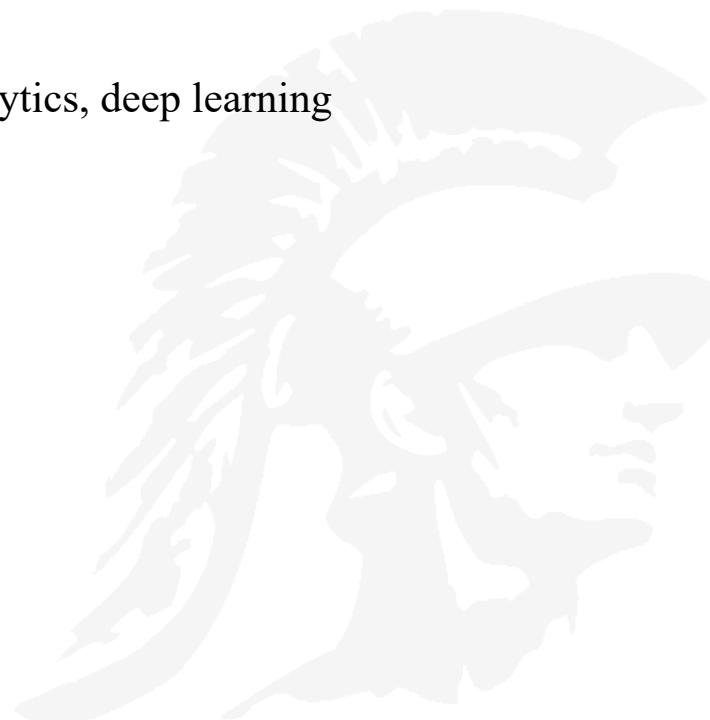
- **There are many trends putting pressure on conventional computing centers, e.g.**
 - Explosive growth in applications: biomedical informatics, space exploration, business analytics, web 2.0 social networking
 - Extreme scale content *generation*: e-science and e-business data deluge
 - Extraordinary 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; Java versions; C → C#; Python
 - 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 Buzz-Words

- **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 Computing

- **Cloud Computing takes place over the Internet,**
 - a collection/group of integrated and networked hardware, software and Internet infrastructure (called a *platform*).
 - Using the Internet for communication and transport provides hardware, software and networking services to clients
- **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, Splinternet (see <https://en.wikipedia.org/wiki/Splinternet>)**
 - China’s great firewall
 - Iran, Saudi Arabia and others filter internet content
- ***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

Cloud Computing Platform

- **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.
 - They can put their applications on the cloud 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**
- **Controlling costs**
 - If your organization only needs one or two virtual servers, and spikes rarely occur, then cloud computing is cheaper than conventional renting of equipment, but
 - As your virtual machines grow and data size increases, costs will swamp in-house computing centers

Virtualization Makes Cloud Computing Possible

- **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)

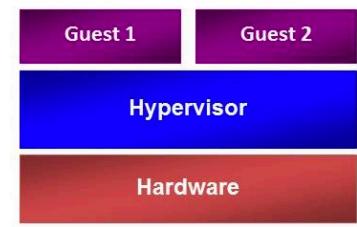
- A **hypervisor** or **virtual machine monitor (VMM)** is computer software, firmware or hardware that creates and runs virtual machines
- A computer on which a hypervisor runs one or more virtual machines is called a *host machine*, and each virtual machine is called a *guest machine*.
- The hypervisor 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:

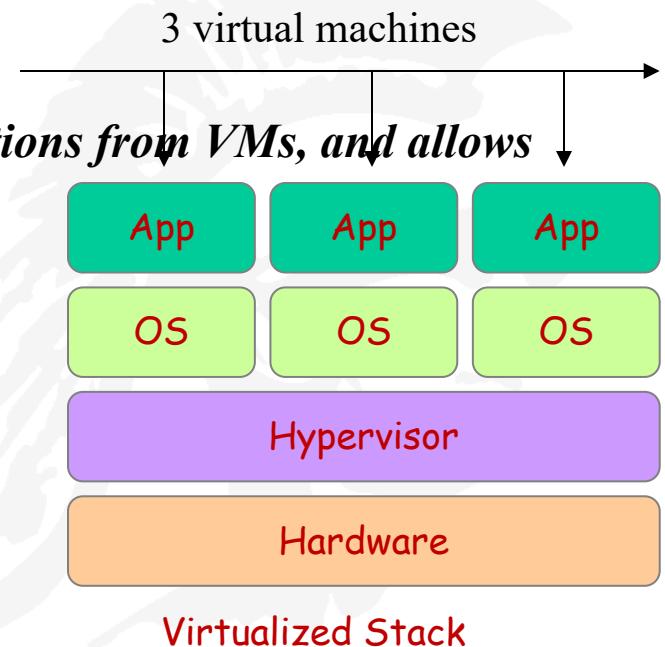
Virtual PC & Virtual Server
 VMware Workstation
 KVM

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.

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 pushing 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

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

Access to all Cloud Platform Products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Sign up and get \$300 to spend on Google Cloud Platform over the next 12 months.

No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

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

Access to all Cloud Platform Products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Sign up and get \$300 to spend on Google Cloud Platform over the next 12 months.

No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

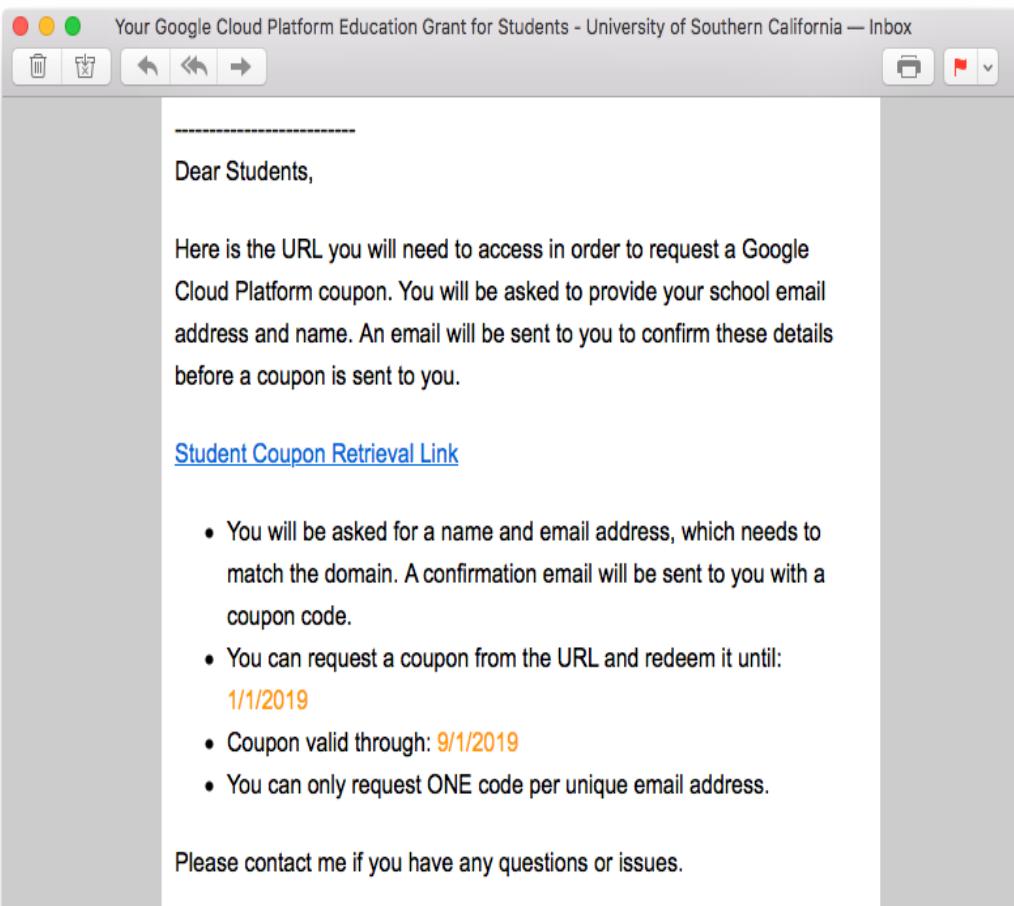


Google Cloud Platform Education Grants Credit

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.

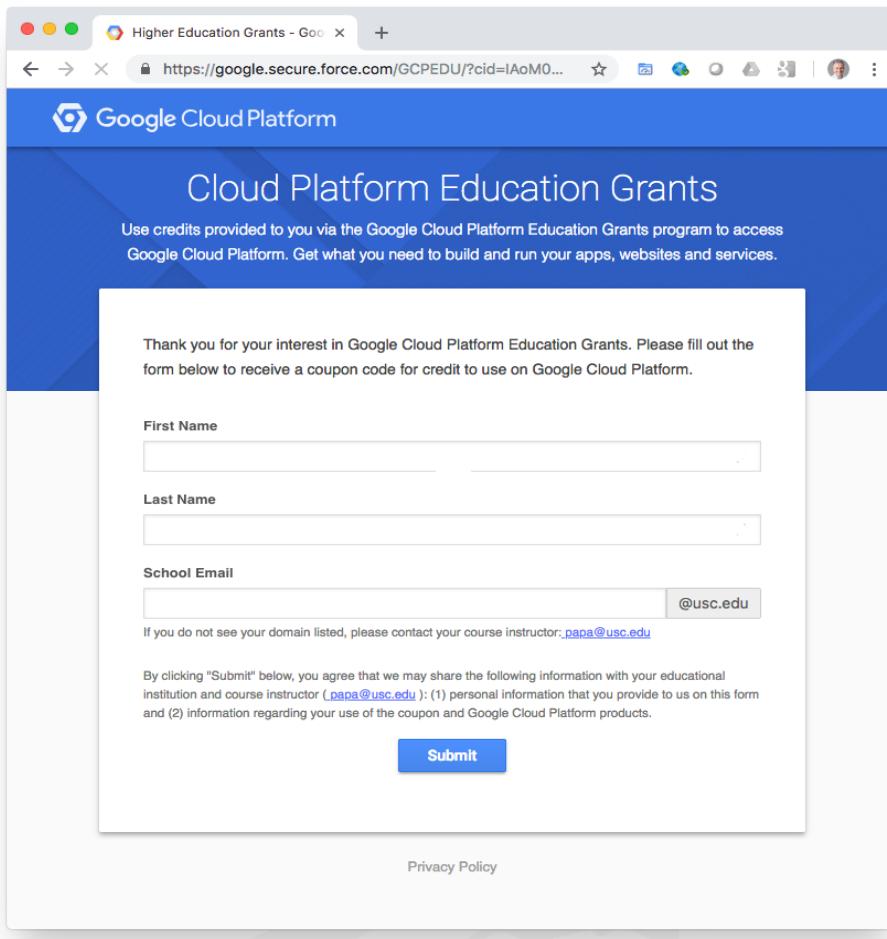


Google Cloud Platform Education Grants Credit

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.

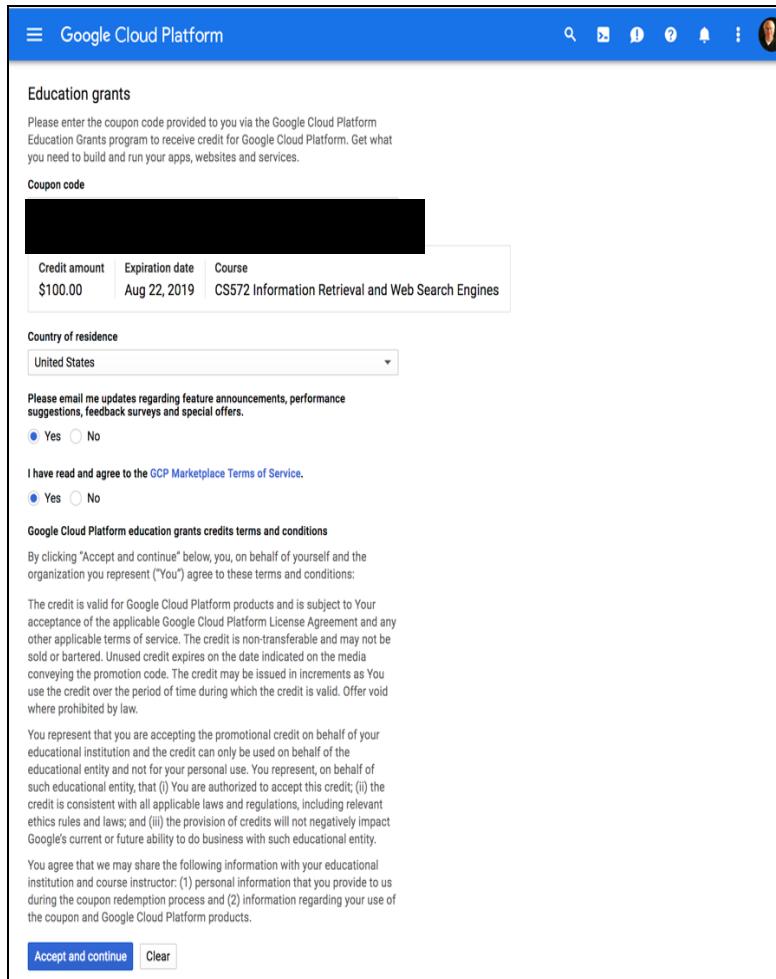
**More details in Homework #3,
Instructions for Setting Up Google
Cloud Account documents (Refer
1.1)**



The screenshot shows a web browser window with the URL <https://google.secure.force.com/GCPEDU/?cid=IAoM0...>. The page title is "Higher Education Grants - Google". The main heading is "Cloud Platform Education Grants". Below it, a sub-headline reads: "Use credits provided to you via the Google Cloud Platform Education Grants program to access Google Cloud Platform. Get what you need to build and run your apps, websites and services." A message box contains the text: "Thank you for your interest in Google Cloud Platform Education Grants. Please fill out the form below to receive a coupon code for credit to use on Google Cloud Platform." There are three input fields: "First Name", "Last Name", and "School Email". The "School Email" field includes a ".edu" suffix. Below the fields is a note: "If you do not see your domain listed, please contact your course instructor: papa@usc.edu". At the bottom, a note states: "By clicking "Submit" below, you agree that we may share the following information with your educational institution and course instructor (papa@usc.edu): (1) personal information that you provide to us on this form and (2) information regarding your use of the coupon and Google Cloud Platform products." A blue "Submit" button is at the bottom right, and a "Privacy Policy" link is at the bottom center.

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" field is filled 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 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 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

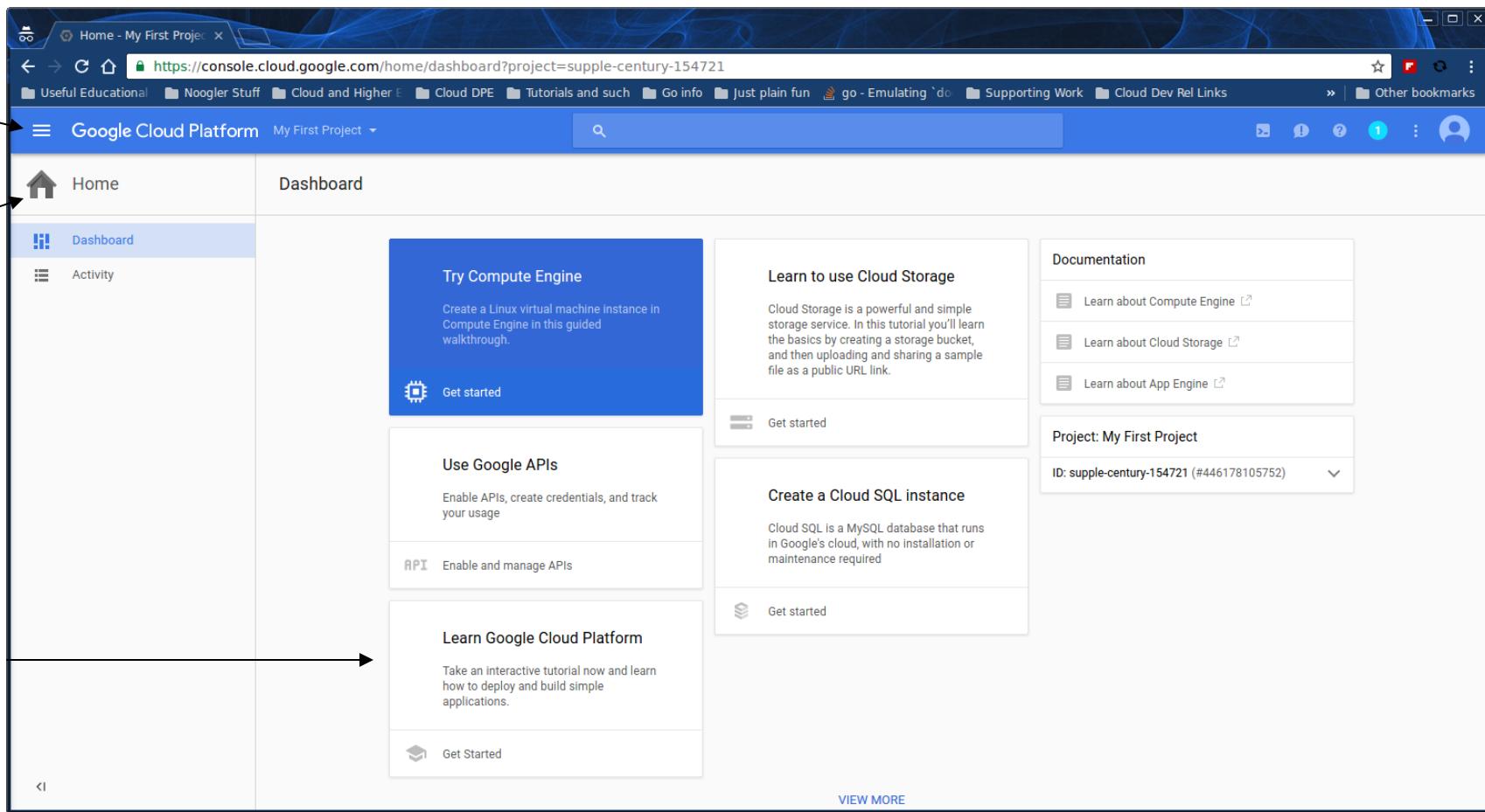
<https://console.cloud.google.com>

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

Main

context

tutorials

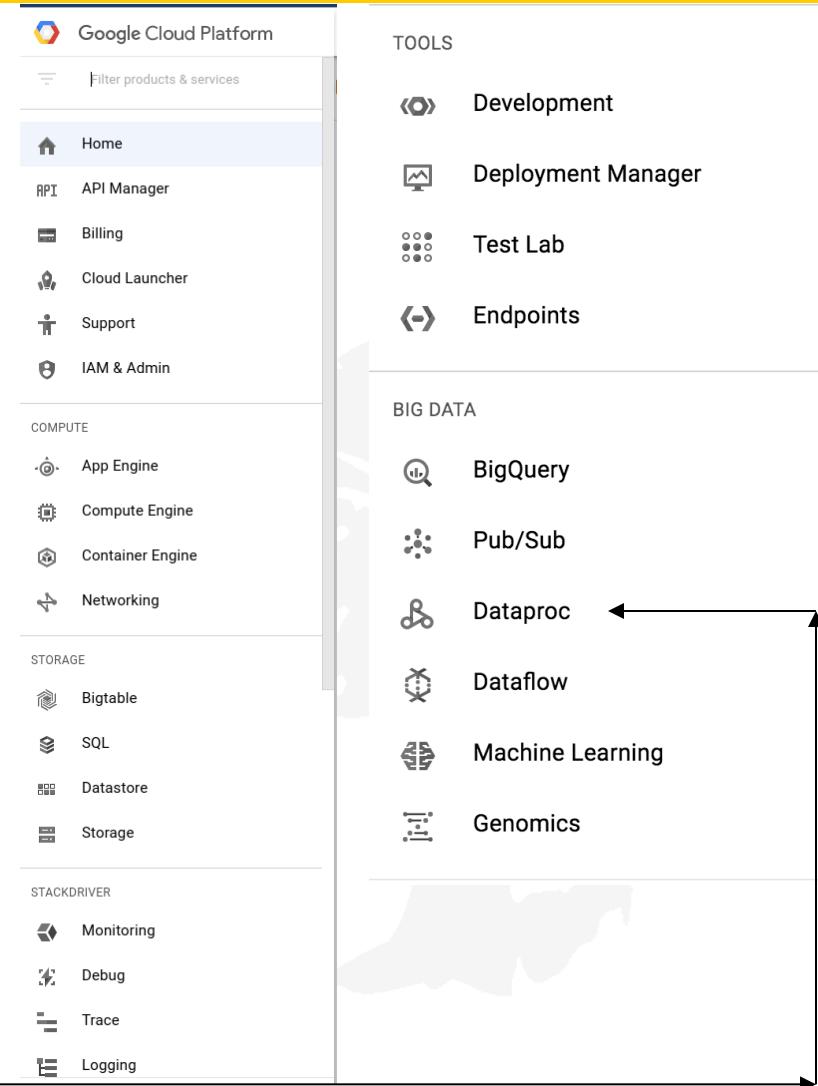


The screenshot shows the Google Cloud Platform Home Page for a project named "My First Project".

- Main menu:** Located at the top left, it includes links like "Home - My First Projec...", "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".
- Context menu:** Located on the left side of the dashboard, it has three items: "Home", "Dashboard" (which is currently selected), and "Activity".
- Tutorials:** Located at the bottom left, it has a large arrow pointing right towards the "Learn Google Cloud Platform" section.
- Content:** The main area contains several cards:
 - Try Compute Engine:** Create a Linux virtual machine instance in Compute Engine in this guided walkthrough. Includes a "Get started" button.
 - Use Google APIs:** Enable APIs, create credentials, and track your usage. Includes an "API" link and a "Enable and manage APIs" button.
 - Learn Google Cloud Platform:** Take an interactive tutorial now and learn how to deploy and build simple applications. Includes a "Get Started" button.
 - 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. Includes a "Get started" button.
 - Create a Cloud SQL instance:** Cloud SQL is a MySQL database that runs in Google's cloud, with no installation or maintenance required. Includes a "Get started" button.
- Documentation:** A sidebar on the right with links to "Learn about Compute Engine", "Learn about Cloud Storage", and "Learn about App Engine".
- Project:** A sidebar on the right showing "Project: My First Project" and "ID: supple-century-154721 (#446178105752)".

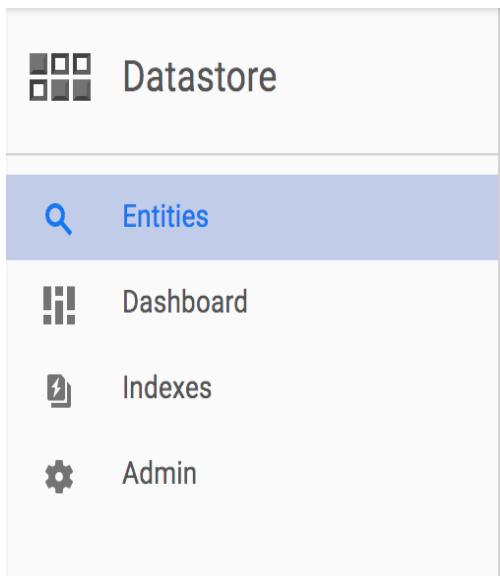
Google Cloud Main Menu

- 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

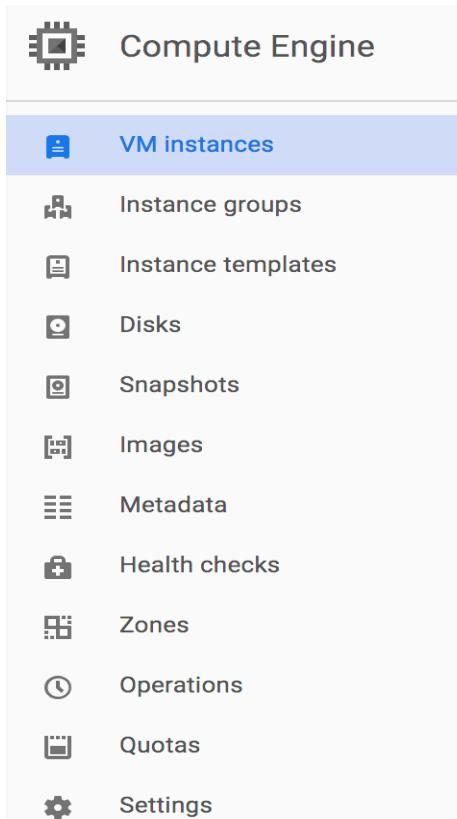


Context Menu

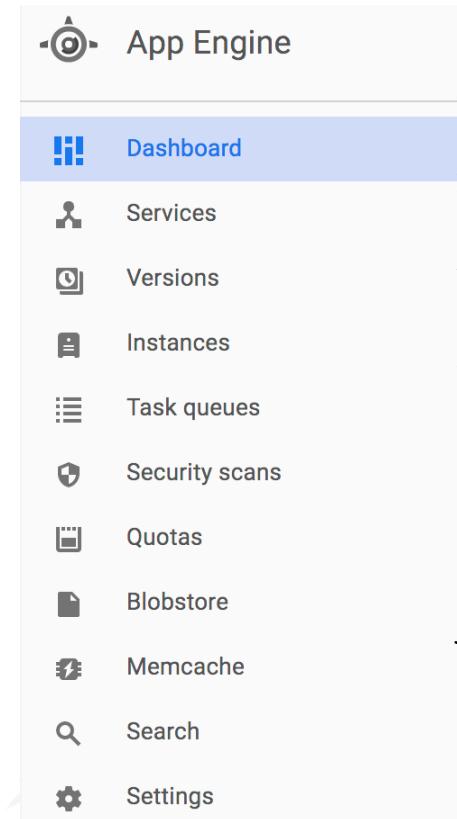
The context menu changes based on the current major component
Here are three examples



A screenshot of a web-based application interface for a 'Datastore'. On the left, there's a sidebar with icons and labels: 'Datastore' (grid icon), 'Entities' (magnifying glass icon, highlighted in blue), 'Dashboard' (bar chart icon), 'Indexes' (key icon), and 'Admin' (gear icon). The main area shows a list of entities with columns for 'Name', 'Type', and 'Actions'.



A screenshot of the 'Compute Engine' interface. The top navigation bar has a 'Compute Engine' icon and the title 'Compute Engine'. Below it, a sub-menu for 'VM instances' is open, showing options like 'Instance groups', 'Instance templates', 'Disks', 'Snapshots', 'Images', 'Metadata', 'Health checks', 'Zones', 'Operations', 'Quotas', and 'Settings'. Each option has a small icon next to it.



A screenshot of the 'App Engine' interface. The top navigation bar has an 'App Engine' icon and the title 'App Engine'. Below it, a sub-menu for 'Dashboard' is open, showing options like 'Services', 'Versions', 'Instances', 'Task queues', 'Security scans', 'Quotas', 'Blobstore', 'Memcache', 'Search', and 'Settings'. Each option has a small icon next to it.

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

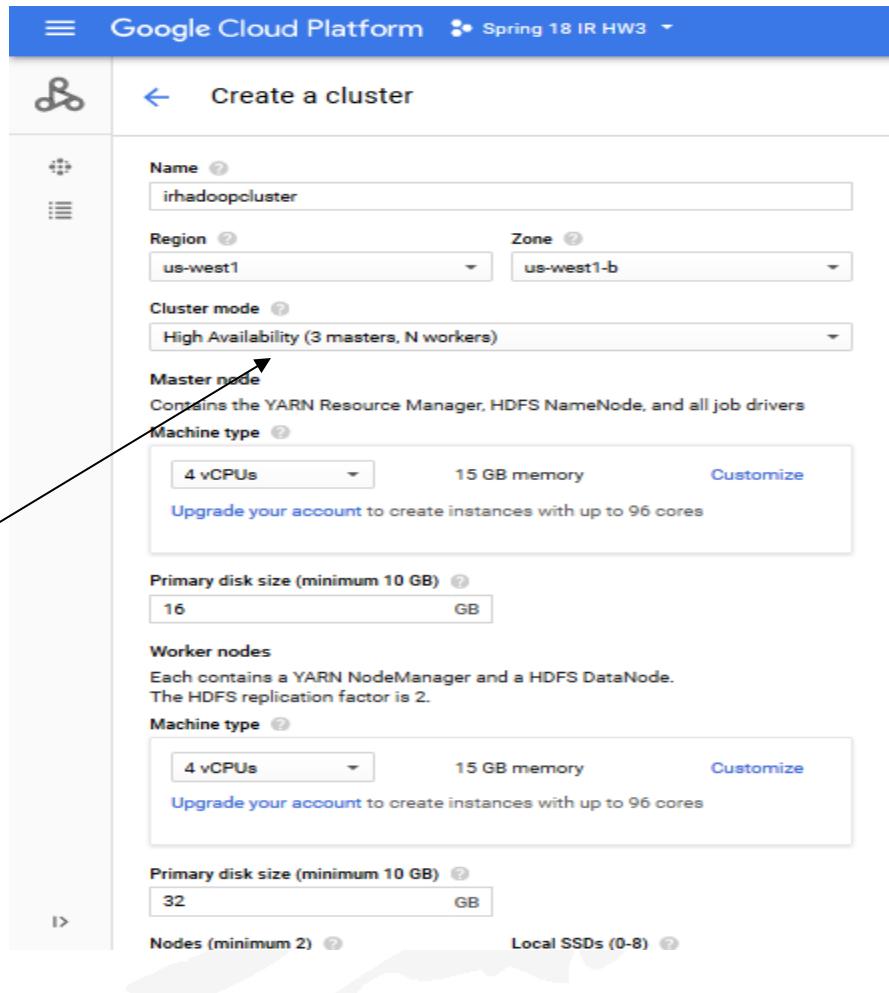
- Snap (Snapchat) recently signed a \$2 billion, five year contract with Google for its cloud services, which makes Snap Google's largest customer of its cloud platform
- Apple confirms it is using Google cloud for iCloud services

Create a Cluster

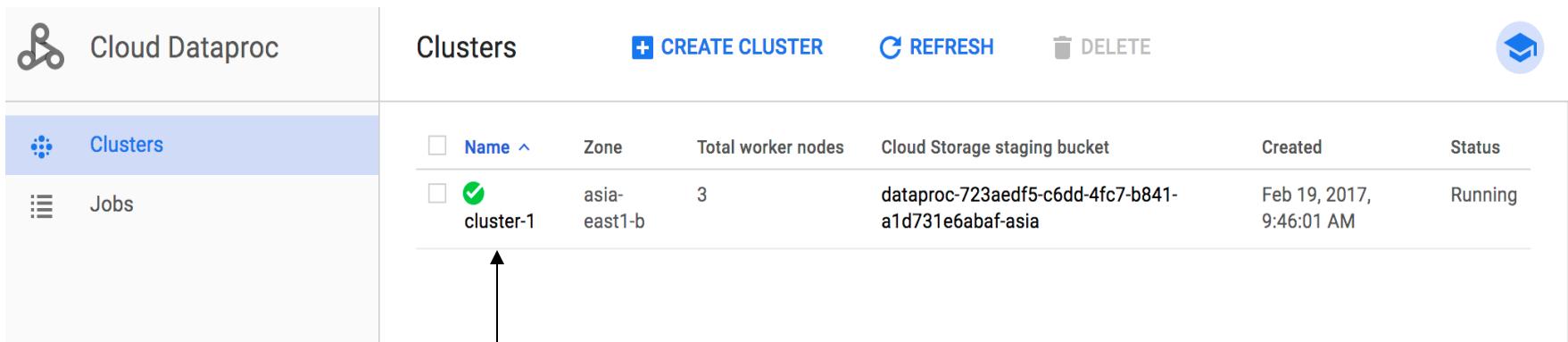
Using the **Google Cloud Platform** Console you can **create a cluster** by going to the **Cloud Platform** Console.

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

Contains 1 master node and 3 worker nodes



Successful Creation of a Cluster



The screenshot shows the Google Cloud Platform Cloud Dataproc interface. On the left, there's a sidebar with icons for Cloud Dataproc, Clusters (selected), and Jobs. The main area is titled "Clusters" and contains a table with the following data:

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

A blue arrow points from the text "This URL will let you SSH to your cluster" up towards the cluster row in the table.

This URL will let you SSH to your cluster

SSH into the Cluster

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

Clusters Jobs Workflows Autoscaling policies Component exchange Notebooks

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.

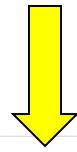
Name	cluster-572-grader
Cluster UUID	706f39f6-ae82-4e2a-9d9d-4a872902454b
Type	Dataproc Cluster
Status	Running

MONITORING JOBS **VM INSTANCES** CONFIGURATION WEB INTERFACES

Filter instances

Name	Role	Actions
cluster-572-grader-m	Master	SSH
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

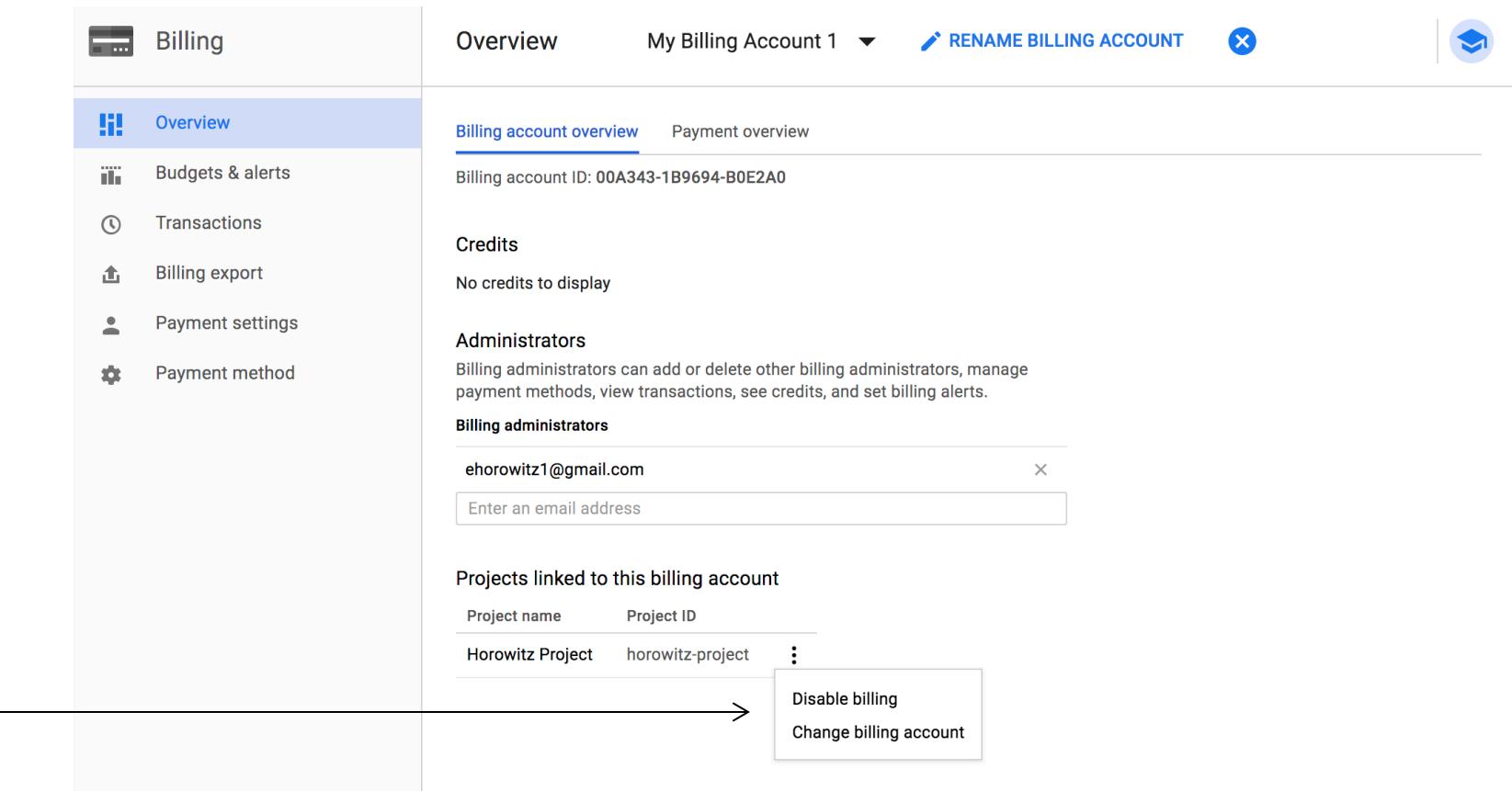
```

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:*.midi=00;36:*.spk=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:/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 libsnappy 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 s
park-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 is titled 'Overview' and shows 'My Billing Account 1'. It includes sections for 'Billing account overview' (selected), 'Payment overview', and 'Billing account ID: 00A343-1B9694-B0E2AO'. Below these are sections for 'Credits' (No credits to display) and 'Administrators'. 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 email addresses for billing administrators is shown, with 'ehorowitz1@gmail.com' listed and an 'x' icon to its right. An input field below says 'Enter an email address'. At the bottom, there's a section for 'Projects linked to this billing account' with a table showing 'Horowitz Project' and 'horowitz-project'. To the right of the table is a vertical ellipsis '...', followed by a box containing two options: 'Disable billing' and 'Change billing account'. A large red arrow points from the bottom left towards this 'Disable billing' option.

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/2020Fall/hw3/DATA.zip>
 - https://drive.google.com/drive/folders/1Z4KyalIuddPGVkJm6dUjkpD_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/2020Fall/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"
```

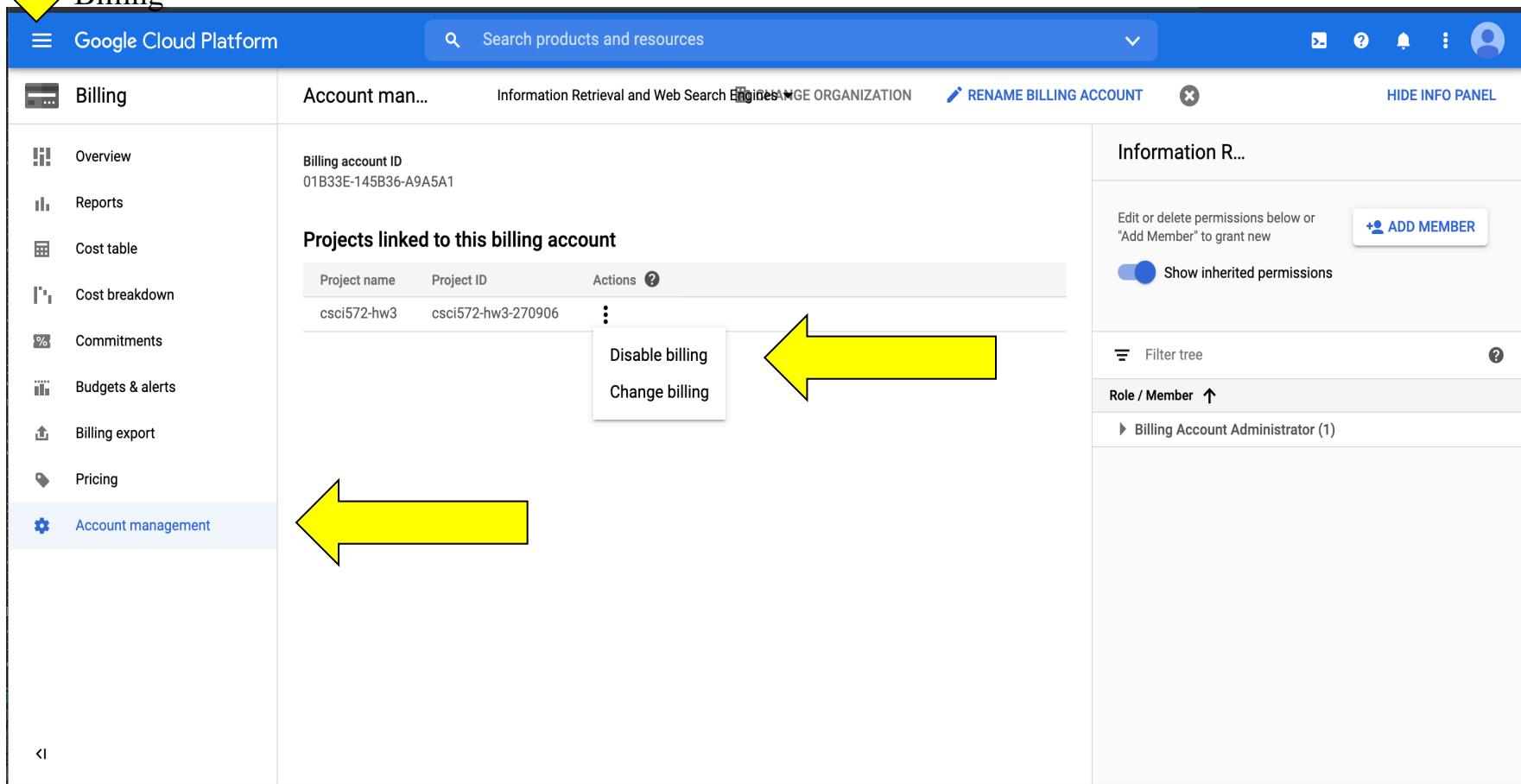
Development Data

	--	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

REMEMBER TO disable the billing for
the cluster when you are not using it



The screenshot shows the Google Cloud Platform Billing interface. On the left, there's a sidebar with various options: Overview, Reports, Cost table, Cost breakdown, Commitments, Budgets & alerts, Billing export, Pricing, and Account management (which is currently selected). The main area displays a billing account with ID 01B33E-145B36-A9A5A1. It lists a single project: csci572-hw3 (Project ID: csci572-hw3-270906). A context menu is open over this project entry, showing 'Disable billing' and 'Change billing' options. A large yellow arrow points from the top-left towards this menu. Another yellow arrow points from the bottom-left towards the 'Disable billing' option in the menu. To the right of the main content, there's a sidebar for managing permissions, showing 'Information R...' and a section for adding members.

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

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

Mapper Class

```
/*
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));
    }
}
```

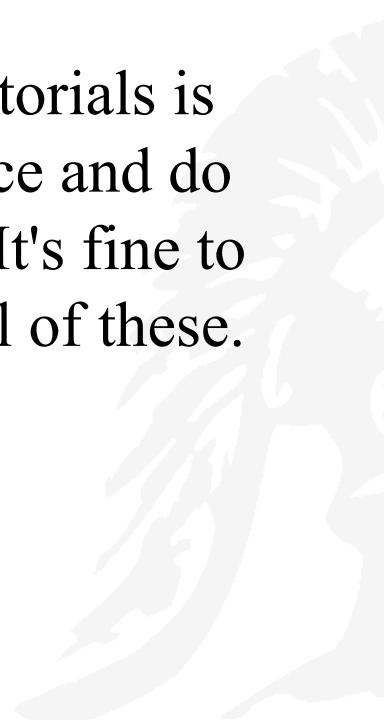
Main Class

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.



The screenshot shows a list of built-in Google Cloud tutorials. At the top, there's a blue header bar with icons for back, forward, search, and user profile. Below it, the title "Start a Tutorial" is displayed, followed by a brief description: "Learn Google Cloud products and services with interactive walkthroughs." A horizontal line separates this from the first tutorial entry. The first entry is titled "-> Try App Engine" with a dropdown arrow icon. Its description is "Learn how to create and deploy a Hello World app." Another horizontal line follows. The second entry is titled "Try Compute Engine" with a server icon. Its description is "Create a Linux virtual machine instance in Compute Engine in this guided walkthrough." A third entry is titled "Build a Compute Engine Application" with a server icon. Its description is "Learn how to spin up virtual machines using Google Compute Engine, Node.js, and MongoDB to create a To-Do app." A fourth entry is titled "Try Container Engine" with a server icon. Its description is "Learn how to build, deploy, and update a Hello World application on Google Container Engine." A fifth entry is titled "Build a Guestbook on Container Engine" with a server icon. Its description is "Learn how to use Google Container Engine clusters built on the power of open source Kubernetes to deploy a Guestbook application." A sixth entry is titled "Try Cloud Pub/Sub" with a gear icon. Its description is "Learn how to use Cloud Pub/Sub to connect your applications with a reliable, many-to-many messaging service on Google's infrastructure." A seventh entry is titled "Try Cloud Storage" with a folder icon. Its description is "Learn how to use Cloud Storage to upload and share your data." A eighth entry is titled "Try Cloud Vision API" with a camera icon. Its description is "Learn how to use Cloud Vision API to label images." A ninth entry is titled "Try Dataflow" with a gear icon. Its description is "Take an interactive tutorial and set up a pipeline to perform a word frequency count on works by Shakespeare."

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

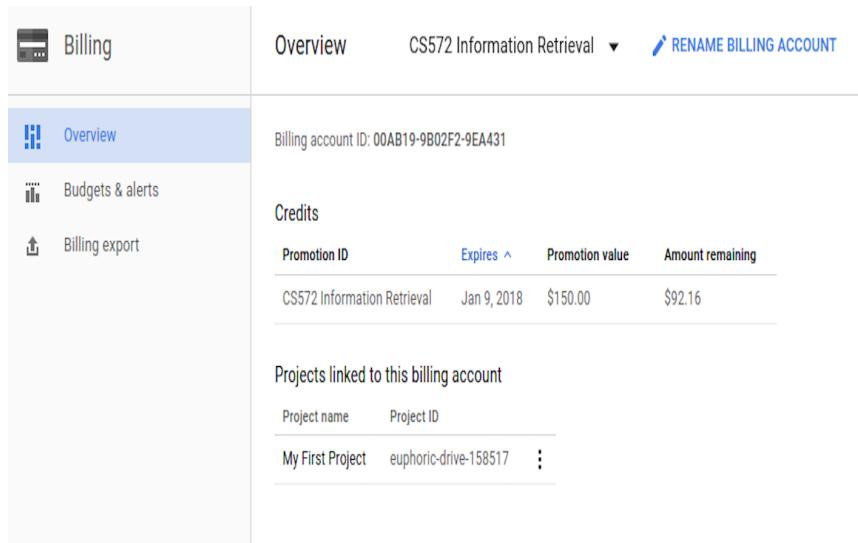
1	answer	5722018453:2	5722018483:1						
2	antecedence	5722018502:1	5722018435:1						
3	asterisks.	5722018417:1	5722018504:2	5722018447:1					
4	beautiful	5722018439:7	5722018417:2	5722018416:3	5722018438:5	5722018437:1	5722018415:1	5722018414:2	5722018435:3
5	bind	5722018419:6	5722018417:39	5722018416:1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

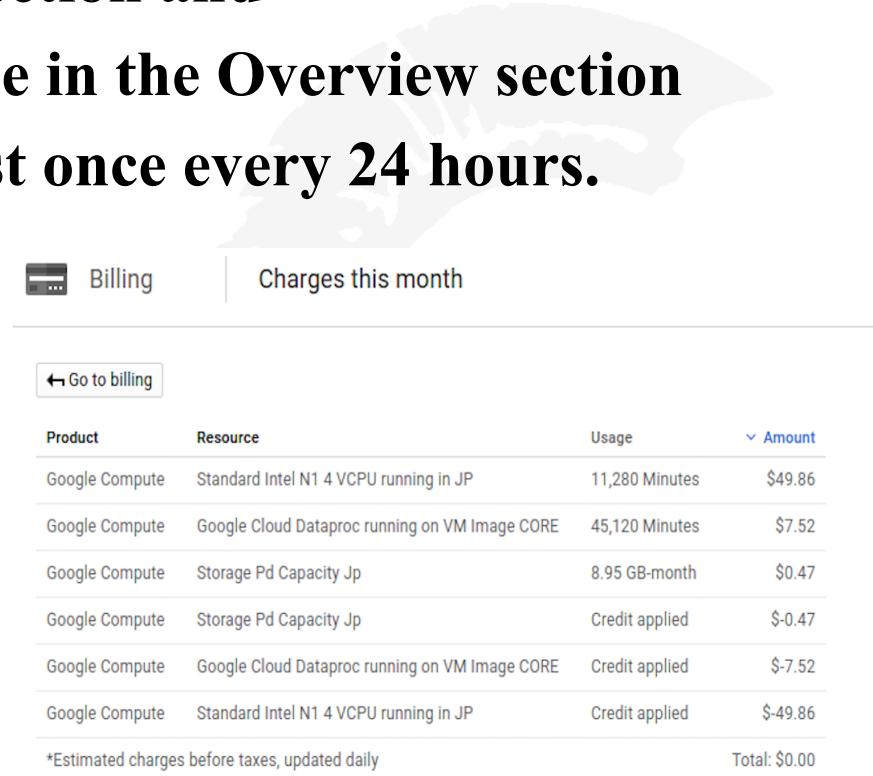
answer appears twice

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.



The screenshot shows the Google Cloud Billing Overview page. On the left, there's a sidebar with navigation links: Billing (selected), Overview (highlighted in blue), Budgets & alerts, and Billing export. The main content area has tabs for Overview (selected) and CS572 Information Retrieval. Under Overview, it shows the Billing account ID: 00AB19-9B02F2-9EA431. Below that is a section for Credits, showing a promotion for CS572 Information Retrieval that expires on Jan 9, 2018, with a value of \$150.00 and remaining amount of \$92.16. At the bottom, it lists "Projects linked to this billing account" with one entry: My First Project (Project ID: euphoric-drive-158517).



The screenshot shows a report titled "Charges this month" from the Google Cloud Billing section. It includes a "Go to billing" button and a table of charges. The table has columns for Product, Resource, Usage, and Amount. The data is as follows:

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