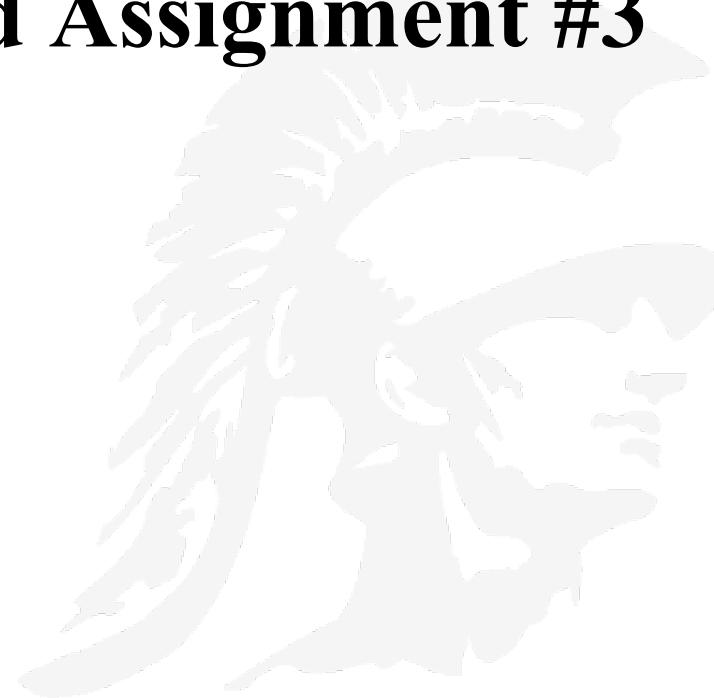


Google Cloud and Assignment #3

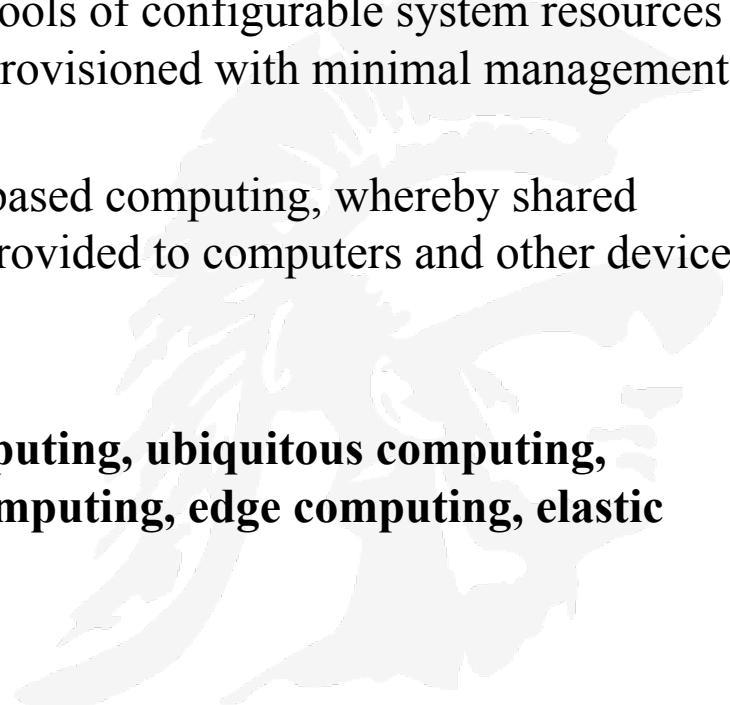


Computing is Rapidly Changing

- 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 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**

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>
- **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.**

Virtualization Makes Cloud Computing Possible

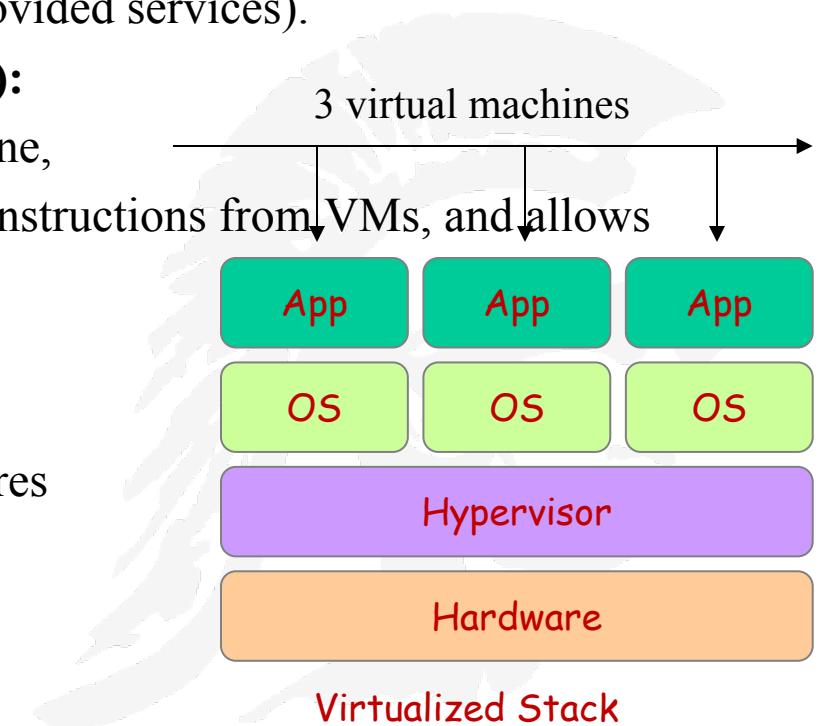
- **Virtualization:** the creation of a virtual -- rather than actual -- version of something, such as an operating system, a server, a storage device or network resources
- **Advantages of virtual machines:**
 1. Run operating systems where the physical hardware is unavailable,
 2. Easier to create 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

- 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.
- Hypervisor Vendors
 - **VMware ESX Server**, ESX301
 - **Xen Project** is a hypervisor providing services that allow multiple computer operating systems to execute on the same hardware concurrently.
 - It was developed by the Univ. of Cambridge and is now being developed by the Linux foundation with support from INTEL XEN3030

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).
- **Implement on 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

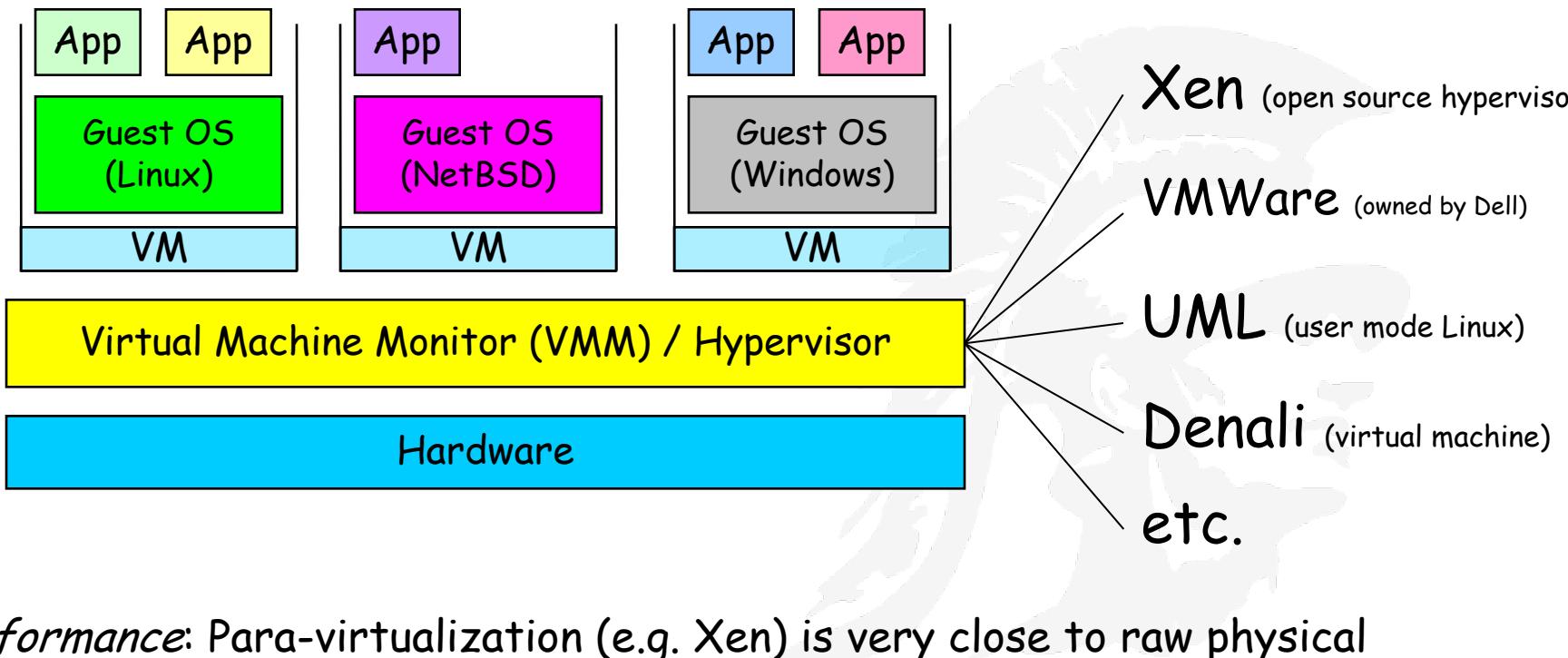


Virtualization Approaches

- 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
- 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.

Virtual Machines

- VM technology allows multiple virtual machines to run on a single physical machine.



Performance: Para-virtualization (e.g. Xen) is very close to raw physical performance!

Leading Cloud Vendors

Forbes / Tech / #InTheCloud



WAVEFRONT
by VMWARE

How to Monitor Containers

Discover Best Practices and New Tools for Monitoring Containers at Scale.



NOV 7, 2017 @ 09:06 AM 60,928

2 Free Issues of Forbes

The Top 5 Cloud-Computing Vendors: #1 Microsoft, #2 Amazon, #3 IBM, #4 Salesforce, #5 SAP



Bob Evans, CONTRIBUTOR
[FULL BIO ▾](#)

Opinions expressed by Forbes Contributors are their own.

CLOUD WARS

Top 10 Rankings — Nov. 7, 2017

1. Microsoft — Nadella on \$20.4B run rate w/[end-to-end customer-centric cloud](#)
2. Amazon — AWS needs software! [10 software companies Amazon might look at](#)
3. IBM — Rometty strikes gold helping customers convert legacy IT to private cloud
4. Salesforce — Benioff must extend SFDC impact from SaaS deeply into PaaS
5. SAP — McDermott accelerating major product-line overhaul to HANA and cloud
6. Oracle — Ellison on cybercrime: '[Make no mistake, this is a war—and we're losing](#)'
7. Google — Tons of potential but still unclear if/how it wants to play in enterprise
8. ServiceNow — Jumps ahead of Workday: revenue up 40%, new products boom
9. Workday — Q2 revenue surges 41% as [Bhusri](#) jumps into PaaS marketplace
10. VMware — revenue & stock jump on deals w/[AMZN MSFT IBM GOOG](#) for hybrid

Bob Evans

As the Cloud Wars heat up, IBM jumps to #3, Salesforce.com falls from #2 (tie) to #4[+]



Metric or Log Analytics?
A Quick Guide to Cloud Applications Monitoring



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



Now Lets Shift to HW#3

Redeeming Google Cloud Credits

Google Cloud Home Page

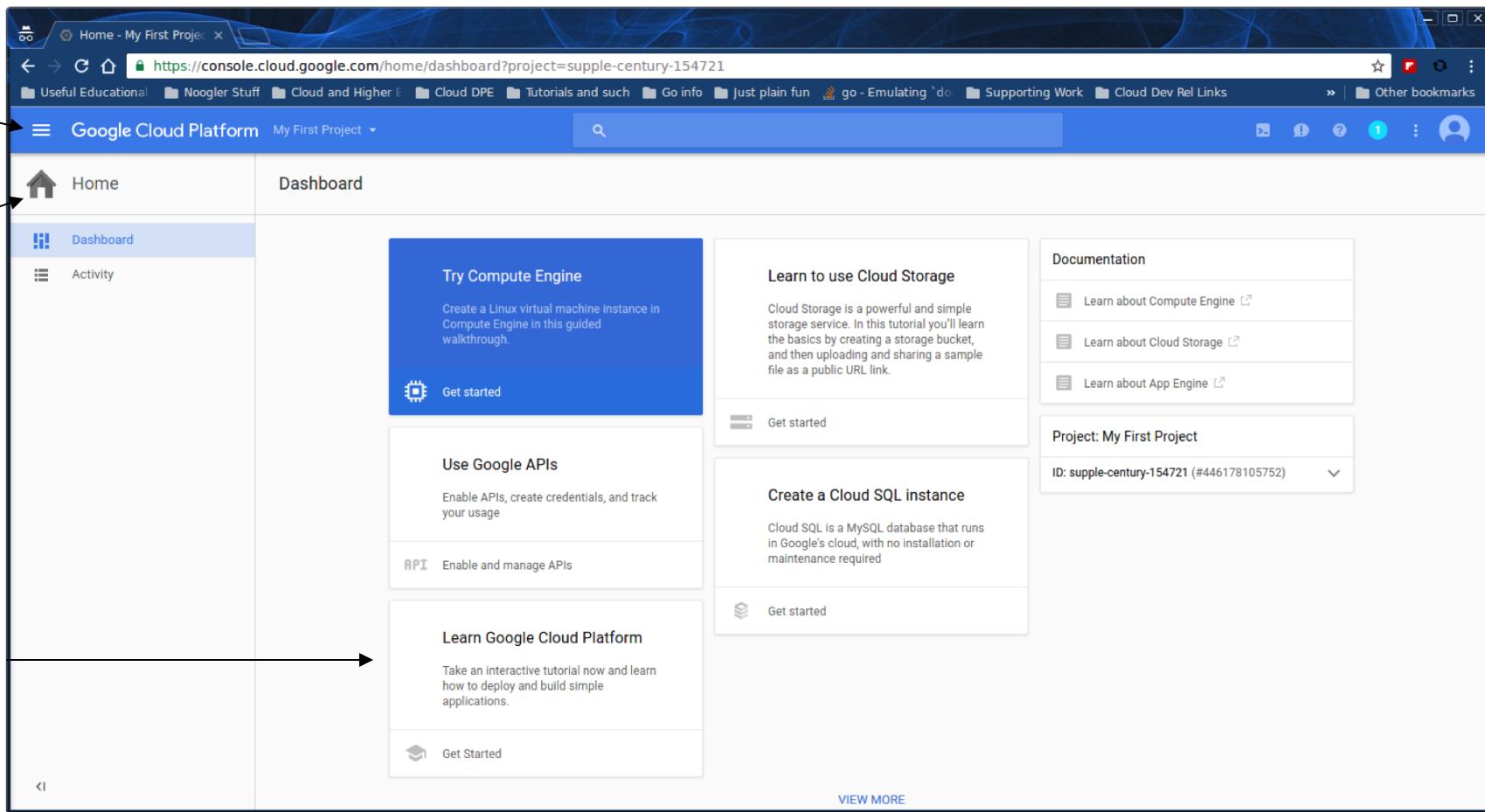
<https://console.cloud.google.com>

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

Main

context

tutorials

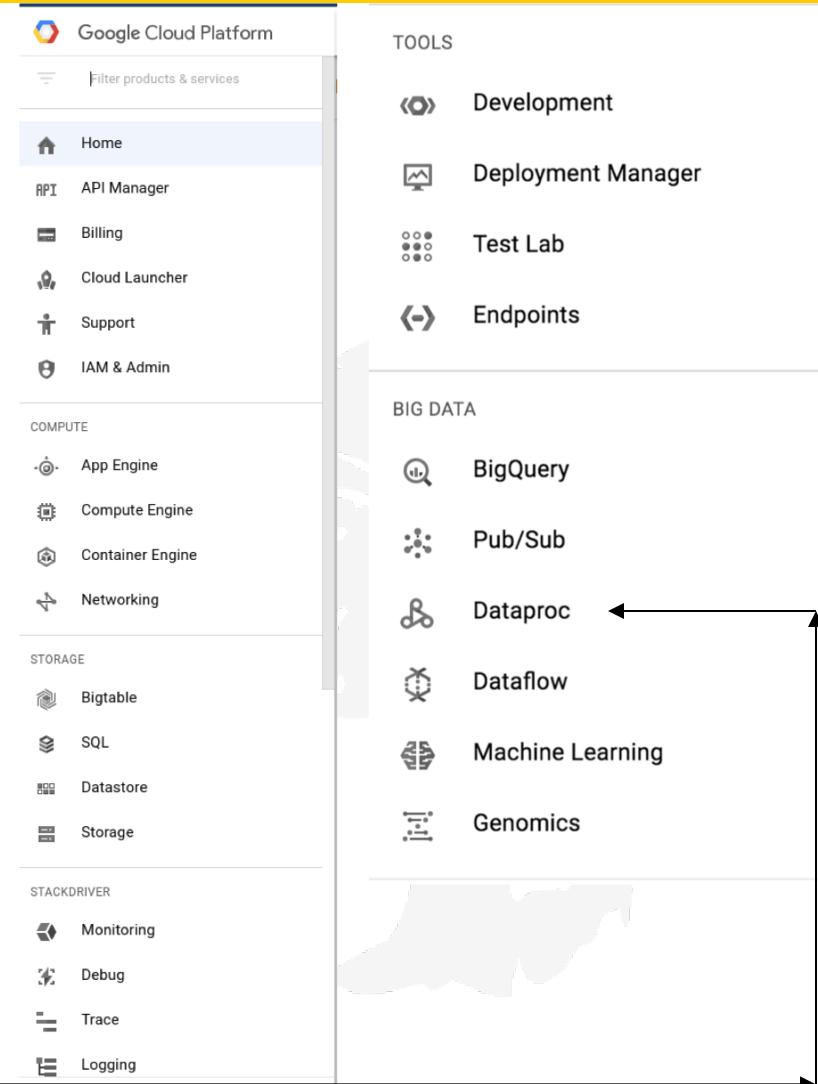


The screenshot shows the Google Cloud Platform Home Page for the project "My First Project".

- 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 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:** Information about the current project: "My First Project" with ID "supple-century-154721 (#446178105752)".

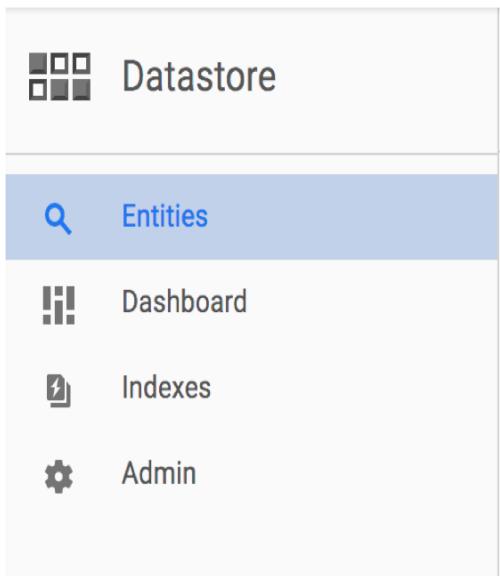
Google Cloud Main Menu

- 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



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

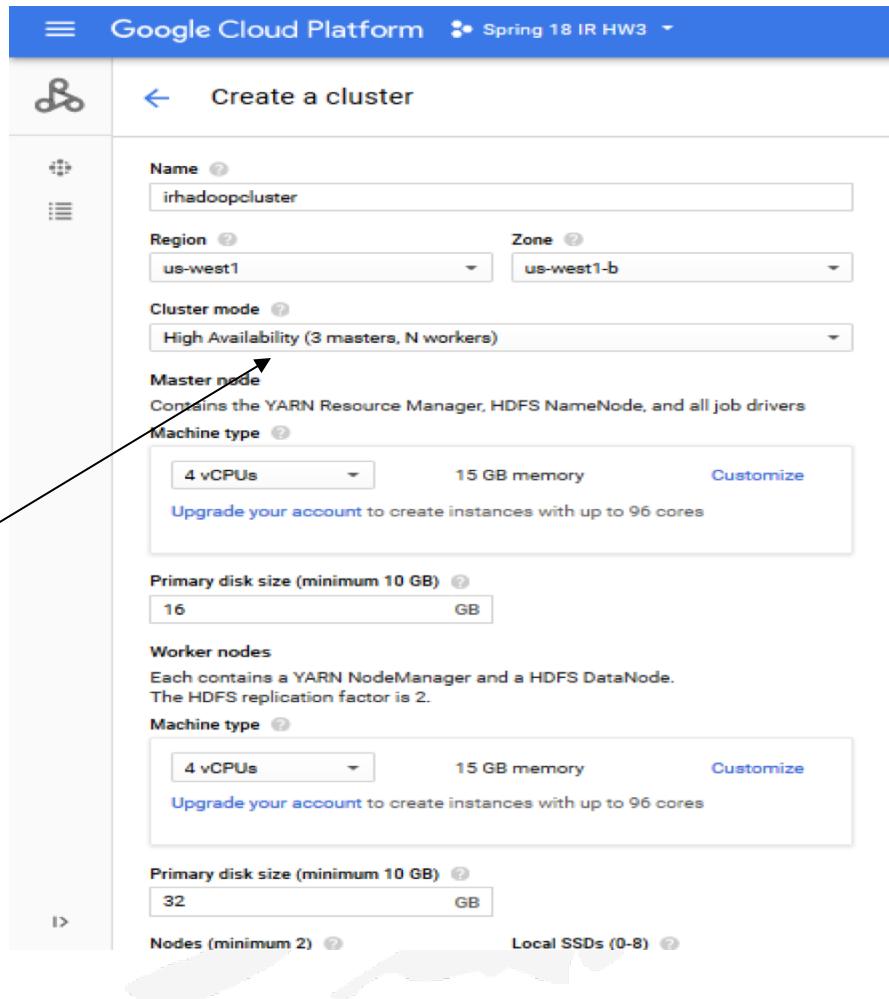
- Snap (Snapchat) recently signed a \$2 billion, five year contract with Google for its cloud services, which makes Snap Google's largest customer of its cloud platform
- Apple confirms it is using Google cloud for iCloud services

Create a Cluster

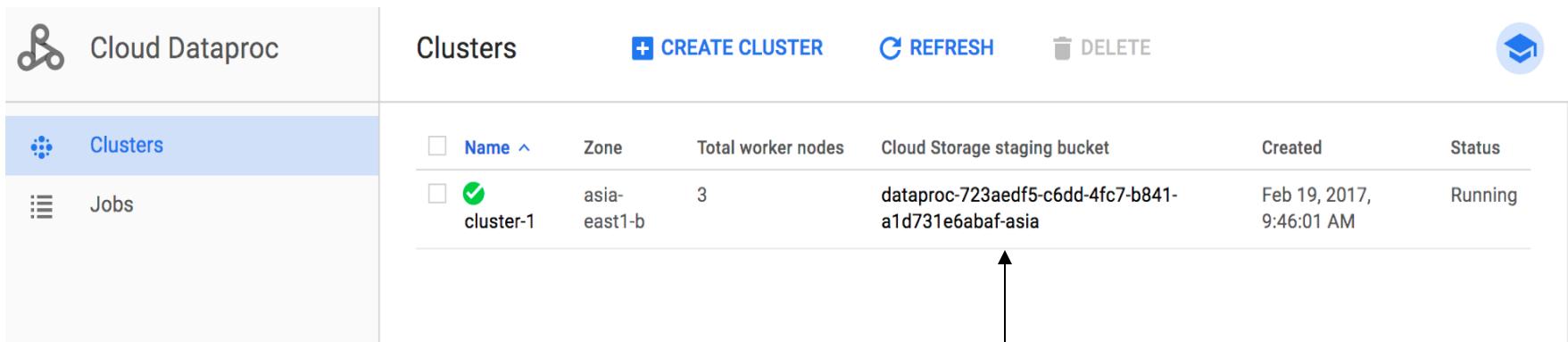
Using the **Google Cloud Platform** Console you can **create a cluster** by going to the **Cloud Platform** Console.

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

Contains 1 master node and 3 worker nodes

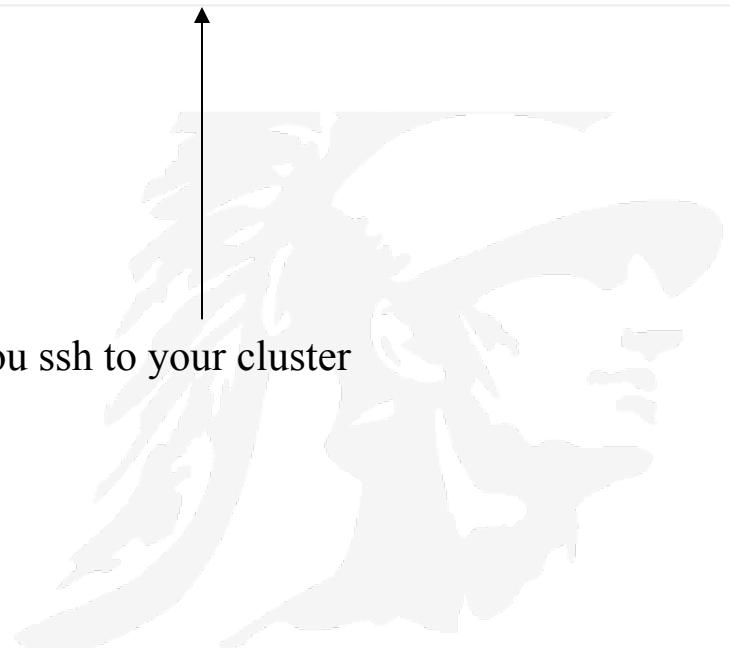


Successful Creation of a Cluster



The screenshot shows the Google Cloud Dataproc interface. On the left, there's a sidebar with 'Cloud Dataproc' at the top, followed by 'Clusters' (which is selected and highlighted in blue) 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
<input type="checkbox"/>	<input checked="" type="checkbox"/> 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

Environment Variables Set

Create a home directory

Check env variables

JAVA_HOME

HADOOP_CLASSPATH

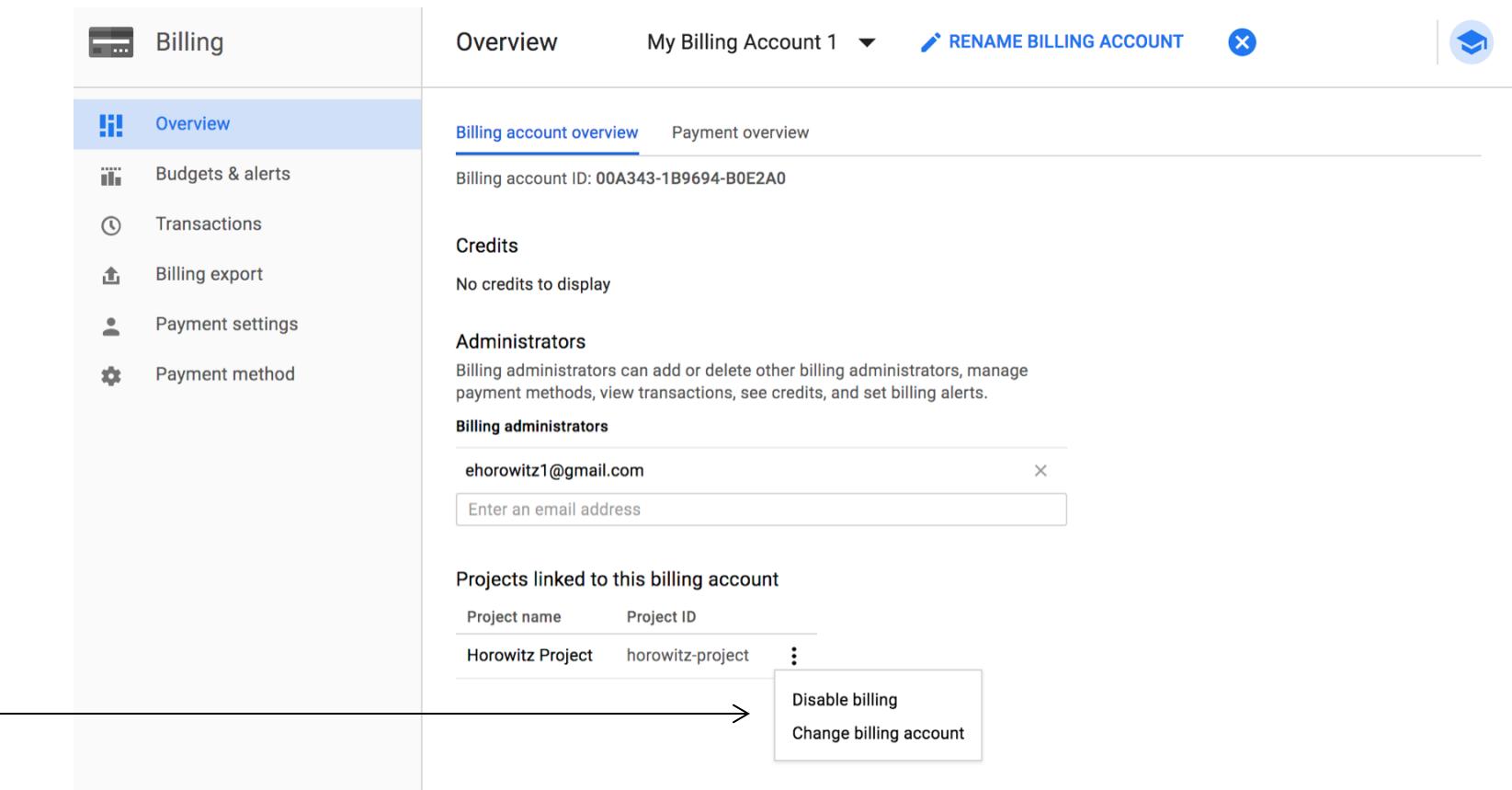
```

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:*.ogx=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:/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-datapro-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

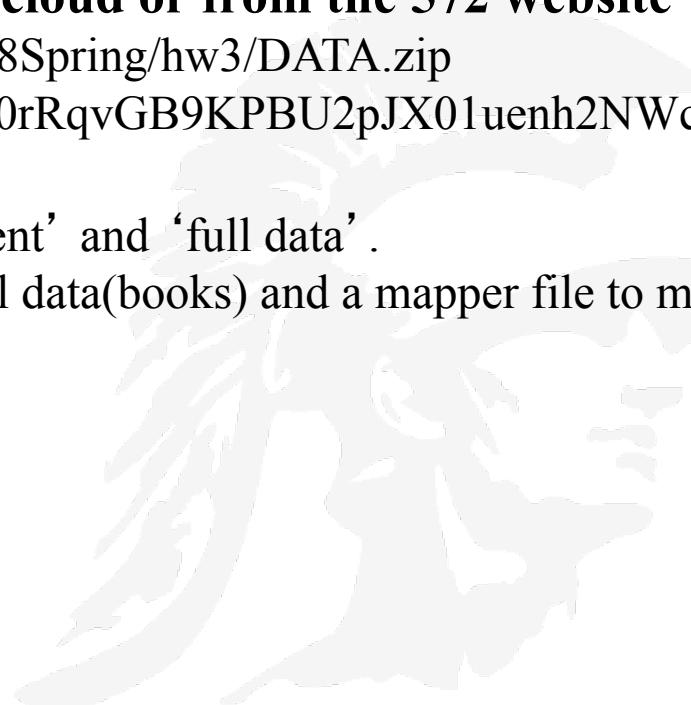
- Please **disable** the billing for the cluster when you are not using it.
- Leaving it running will cost extra credits.
- The cluster is billed based on how many hours it is running and not how much data it is processing



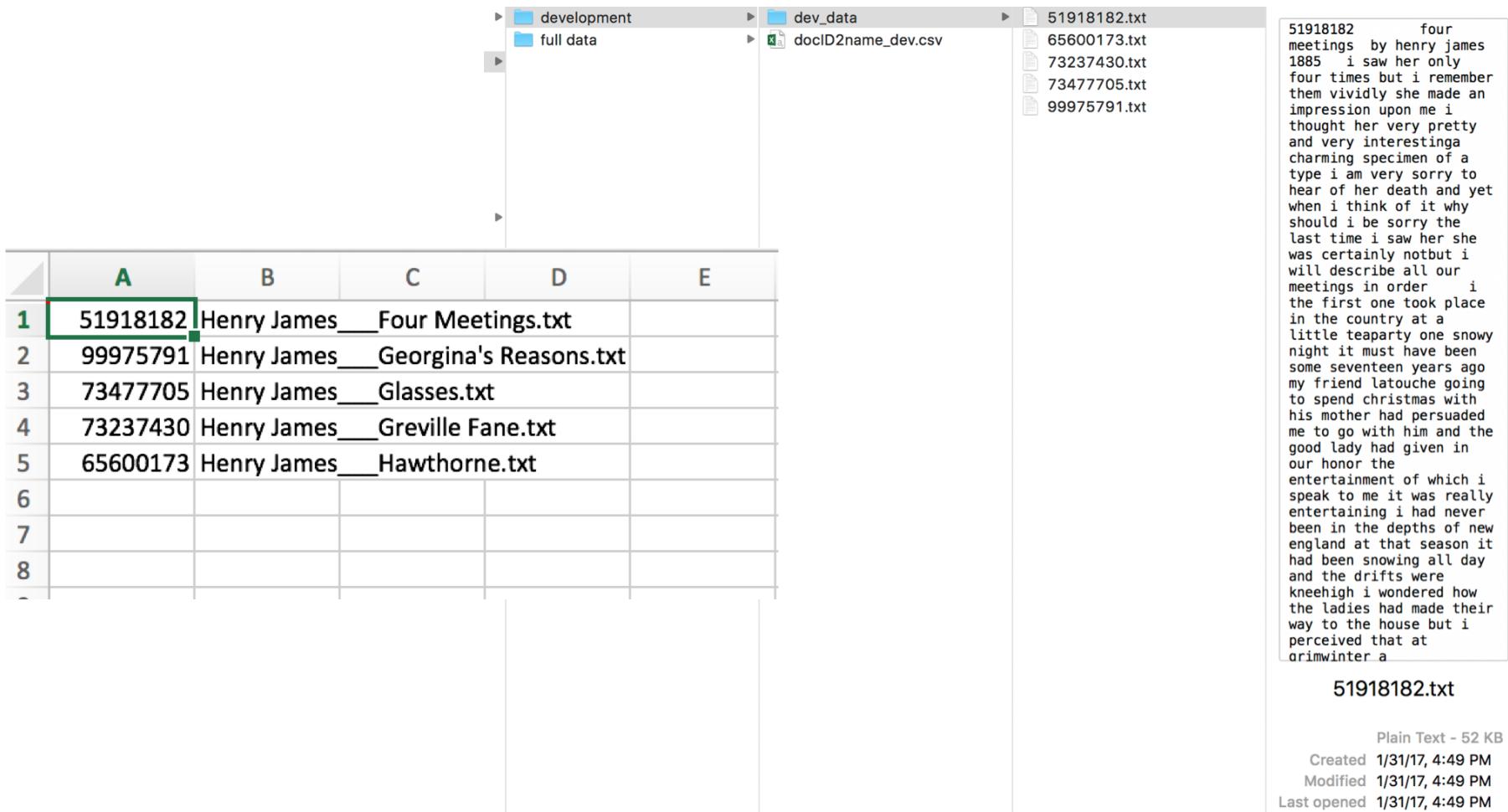
The screenshot shows the Google Cloud Billing Overview page. On the left, a sidebar menu includes 'Overview' (which is selected and highlighted in blue), 'Budgets & alerts', 'Transactions', 'Billing export', 'Payment settings', and 'Payment method'. The main content area displays the 'Overview' tab for 'My Billing Account 1'. It shows a 'Billing account overview' card with the ID '00A343-1B9694-B0E2AO' and a 'Payment overview' card. 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 admins is shown, with 'ehorowitz1@gmail.com' currently listed. A text input field allows adding more email addresses. At the bottom, there's a section for 'Projects linked to this billing account' with one entry: 'Horowitz Project' (Project ID: horowitz-project). To the right of this entry is a vertical ellipsis menu with two options: 'Disable billing' and 'Change billing account'. A large red arrow points from the bottom left towards this ellipsis menu.

Upload the Data Set

- **We'll be using a collection of 3,036 English books written by 142 authors**
 - The data comes from web.eecs.umich.edu/~lahiri/gutenberg_dataset.html
 - The data has been cleaned of metadata, license information, notes
- **Retrieve the dataset either from the cloud or from the 572 website**
 - <http://www-scf.usc.edu/~csci572/2018Spring/hw3/DATA.zip>
 - <https://drive.google.com/open?id=0B0rRqvGB9KPBU2pJX01uenh2NWc>
- **Unzip the contents**
 - two folders inside named ‘development’ and ‘full data’ .
 - Each of the folders contains the actual data(books) and a mapper file to map the docID to the file name.



Development Data



The image shows a file system and an Excel spreadsheet. The file system on the right contains a folder 'development' with subfolders 'full data' and 'dev_data'. Inside 'dev_data' is a CSV file 'docID2name_dev.csv'. A file '51918182.txt' is selected, showing its contents. The Excel spreadsheet on the left lists five entries, each with a file ID and a corresponding Henry James work.

A	B	C	D	E
1	51918182	Henry James	Four Meetings.txt	
2	99975791	Henry James	Georgina's Reasons.txt	
3	73477705	Henry James	Glasses.txt	
4	73237430	Henry James	Greville Fane.txt	
5	65600173	Henry James	Hawthorne.txt	
6				
7				
8				

File System View:

- development
 - full data
- dev_data
 - docID2name_dev.csv
- 51918182.txt
 - 65600173.txt
 - 73237430.txt
 - 73477705.txt
 - 99975791.txt

File Content (51918182.txt):

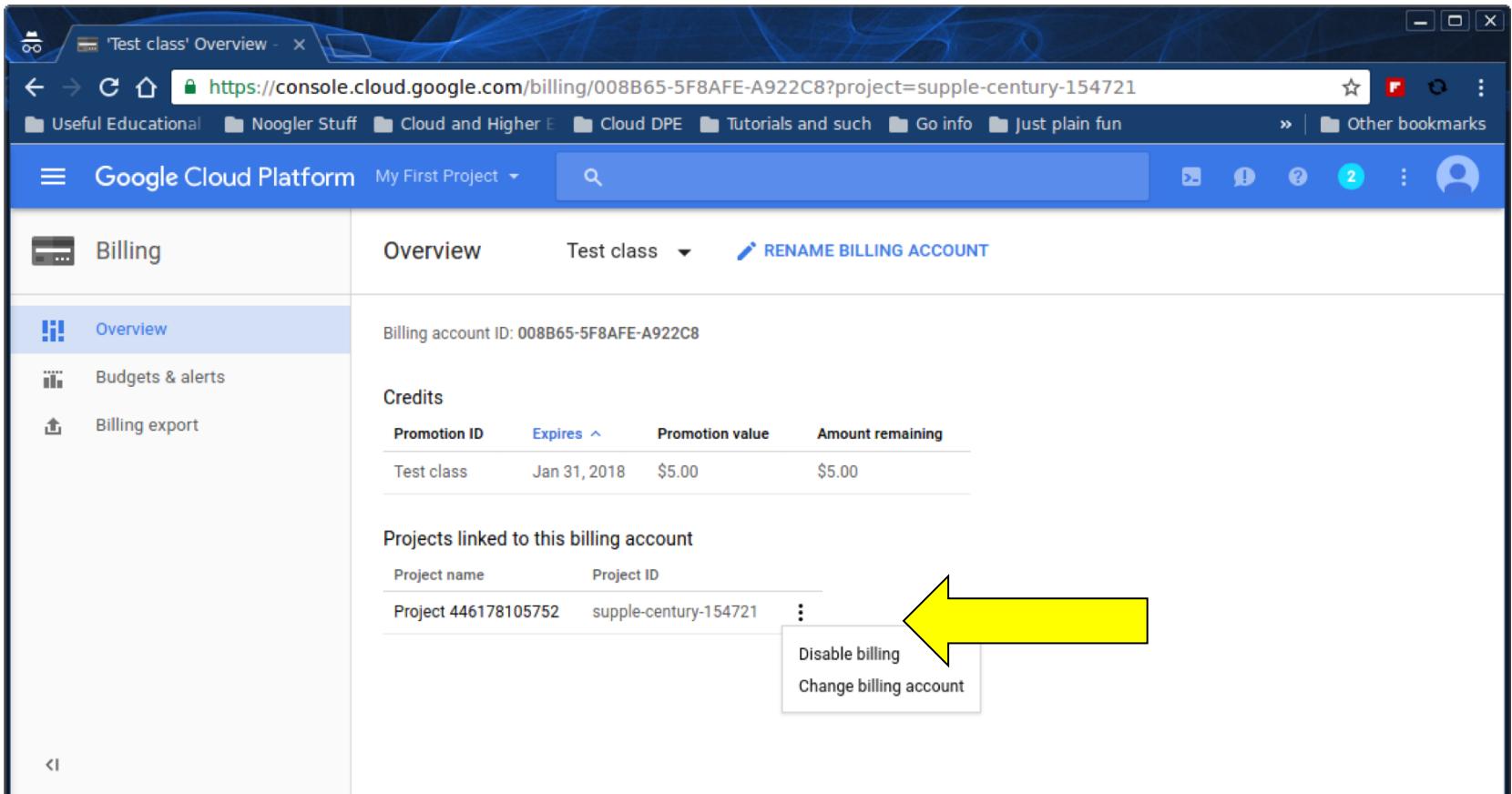
```
51918182        four
meetings by henry james
1885 i saw her only
four times but i remember
them vividly she made an
impression upon me i
thought her very pretty
and very interestinga
charming specimen of a
type i am very sorry to
hear of her death and yet
when i think of it why
should i be sorry the
last time i saw her she
was certainly notbut i
will describe all our
meetings in order i
the first one took place
in the country at a
little teaparty one snowy
night it must have been
some seventeen years ago
my friend latouche going
to spend christmas with
his mother had persuaded
me to go with him and the
good lady had given in
our honor the
entertainment of which i
speak to me it was really
entertaining i had never
been in the depths of new
england at that season it
had been snowing all day
and the drifts were
kneehigh i wondered how
the ladies had made their
way to the house but i
perceived that at
arimwinter a
```

File Details (51918182.txt):

- Plain Text - 52 KB
- Created 1/31/17, 4:49 PM
- Modified 1/31/17, 4:49 PM
- Last opened 1/31/17, 4:49 PM

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

Click here
and
select
Billing



The screenshot shows the Google Cloud Platform Billing Overview page for a project named 'Test class'. The left sidebar has 'Overview' selected. The main area displays the Billing account ID: 008B65-5F8AFE-A922C8. Below it, a 'Credits' section shows a promotion for 'Test class' expiring on Jan 31, 2018, with a value of \$5.00 remaining. A table lists 'Projects linked to this billing account' with one entry: 'Project 446178105752' and 'supple-century-154721'. A yellow arrow points to a context menu for this project row, which includes options 'Disable billing' and 'Change billing account'.

Promotion ID	Expires	Promotion value	Amount remaining
Test class	Jan 31, 2018	\$5.00	\$5.00

Project name	Project ID
Project 446178105752	supple-century-154721

Inverted Index Implementation

- You need to write some code, in Java, that processes the data file of books and produces an inverted index of the words that occur in the books
- 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

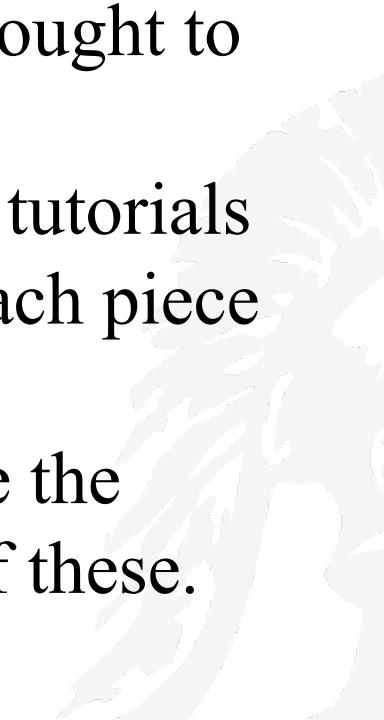
Class WordCount;
Program: main
Create the Hadoop job

```
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);
    }
}
```

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.

Built-in Tutorials



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, a section titled "Start a Tutorial" says "Learn Google Cloud products and services with interactive walkthroughs." A dropdown menu is open under "Try App Engine", showing "Learn how to create and deploy a Hello World app." Other visible sections include:

- Try Compute Engine**: Create a Linux virtual machine instance in Compute Engine in this guided walkthrough.
- Build a Compute Engine Application**: Learn how to spin up virtual machines using Google Compute Engine, Node.js, and MongoDB to create a To-Do app.
- Try Container Engine**: Learn how to build, deploy, and update a Hello World application on Google Container Engine.
- Build a Guestbook on Container Engine**: Learn how to use Google Container Engine clusters built on the power of open source Kubernetes to deploy a Guestbook application.
- Try Cloud Pub/Sub**: Learn how to use Cloud Pub/Sub to connect your applications with a reliable, many-to-many messaging service on Google's infrastructure.
- Try Cloud Storage**: Learn how to use Cloud Storage to upload and share your data.
- Try Cloud Vision API**: Learn how to use Cloud Vision API to label images.
- Try Dataflow**: Take an interactive tutorial and set up a pipeline to perform a word frequency count on works by Shakespeare.

Sample Document and Output

```
1 |51918182 four meetings by henry james 1885 i saw her only four times but i remember them  
vividly she made an impression upon me i thought her very pretty and very interestinga charming  
specimen of a type i am very sorry to hear of her death and yet when i think of it why should i  
be sorry the last time i saw her she was certainly notbut i will describe all our meetings in  
order i the first one took place in the country at a little teaparty one snowy night it must  
have been some seventeen years ago my friend latouche going to soend christmas with his mother
```

Sample of Mapper Output

```
james 51918182  
people 51918182  
people 51918182  
of 51918182  
of 51918182
```

Sample of Reducer Output

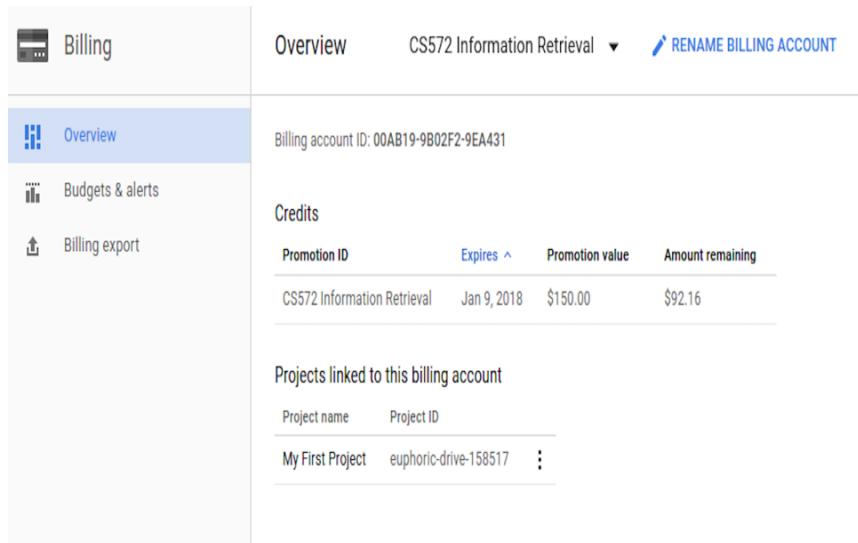
```
1 ably 9931985:1  
2 abnegate 85886314:1 80811098:1  
3 abney 31694096:3 15109590:1 38583612:1 98115965:98  
4 abnormal 47943267:1 94435826:1 80942074:1  
5 abroad 73713297:1 11200532:1
```

James occurs 1 time in the document whose ID is 51918182
People occurs 2 times in the document whose ID is 51918182

Ably appears once in document
9931985

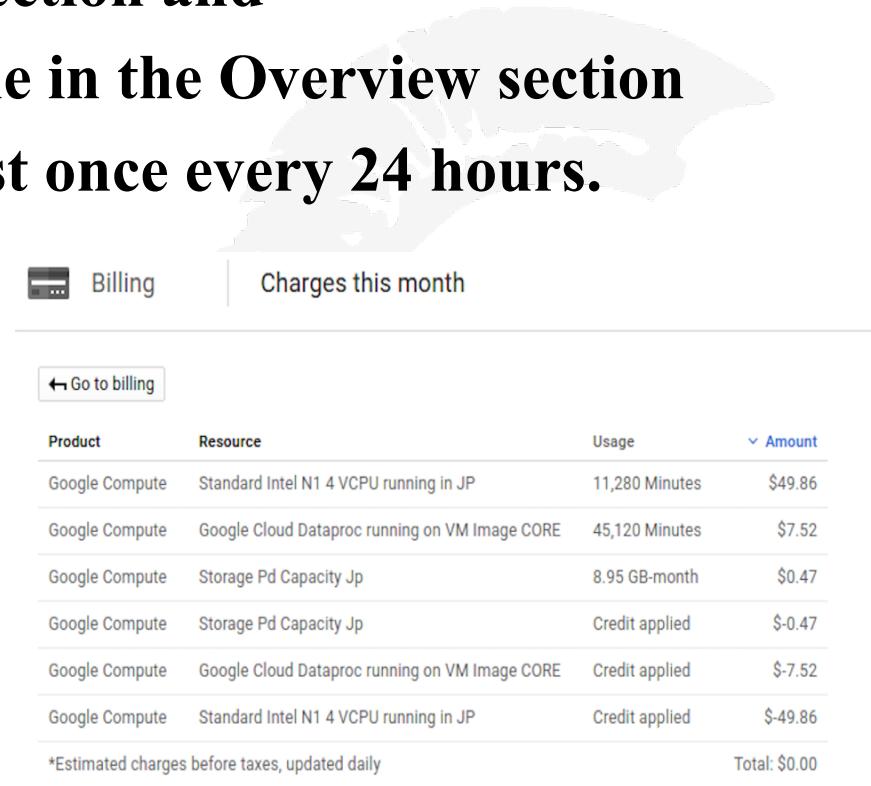
Credits Spent

- 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 links for Billing, Overview (which is selected and highlighted in blue), Budgets & alerts, and Billing export. The main content area has tabs for Overview (selected) and CS572 Information Retrieval (the active project). Below these are sections for Credits (showing a promotion ID, expiration date, value, and amount remaining), Projects linked to this billing account (listing 'My First Project' with its ID), and a summary table showing charges for various Google Compute resources.

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



The screenshot shows a report titled 'Charges this month' from Google Cloud Billing. It includes a 'Go to billing' button and a table of charges. The table has columns for Product, Resource, Usage, and Amount. The usage column shows estimated values like '11,280 Minutes' and '45,120 Minutes'. The amount column shows monetary values like '\$49.86' and '\$7.52'. A note at the bottom states '*Estimated charges before taxes, updated daily' and shows a total of '\$0.00'.

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