

Assignment: Creating a cluster in Google cloud

Kamal Nandan

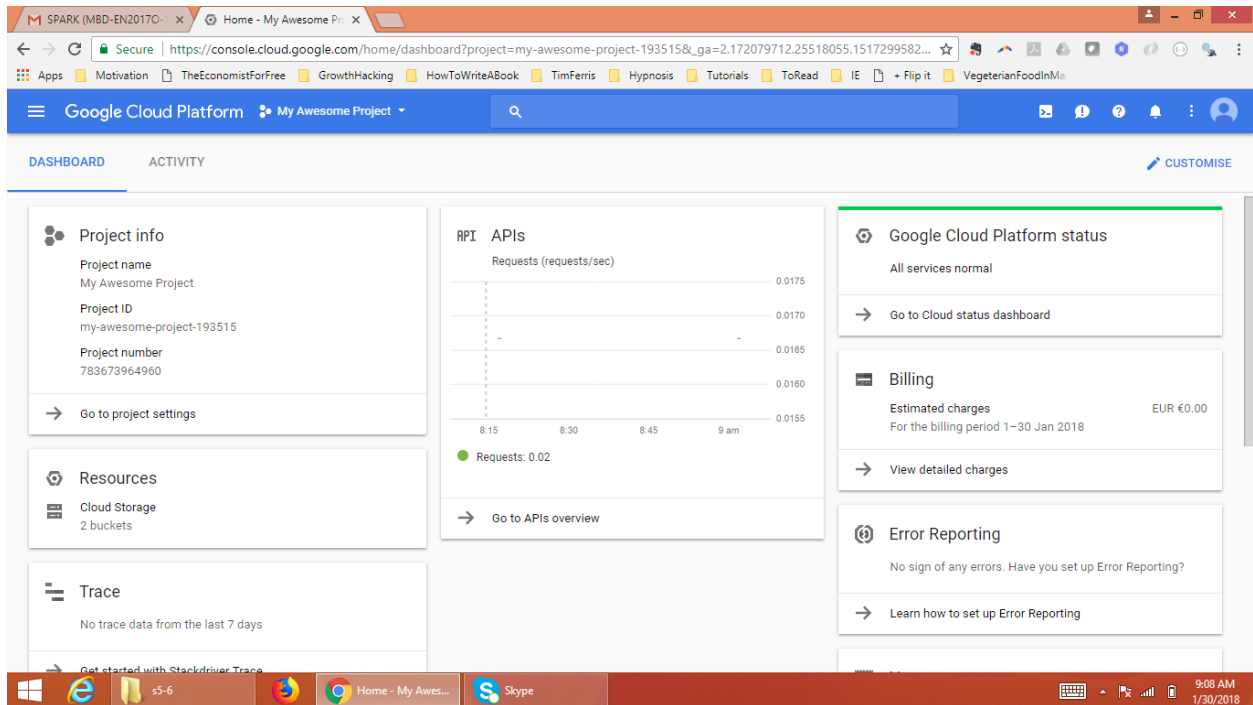
Kamal.nandan@student.ie.edu

(IE-MBD-O1)

(I had created the cluster already and deleted that, but since I hadn't recorded the screenshots, I am doing it again. Since I had already created the cluster, I wasn't required to enable the APIs again.)

So, here are the steps:

1. Create “My Awesome Project” and browse to the project.



2. Go to “APIs and Services” and click on “Enable APIs and services”

The screenshot displays the Google Cloud Platform console for a project named 'My Awesome Project'. The 'APIs & Services' section is active, showing the 'Dashboard' for 'Enabled APIs and services'. The dashboard includes a traffic graph for the last hour, error statistics, and a table of enabled APIs.

Enabled APIs and services
Some APIs and services are enabled automatically

Activity for the last hour

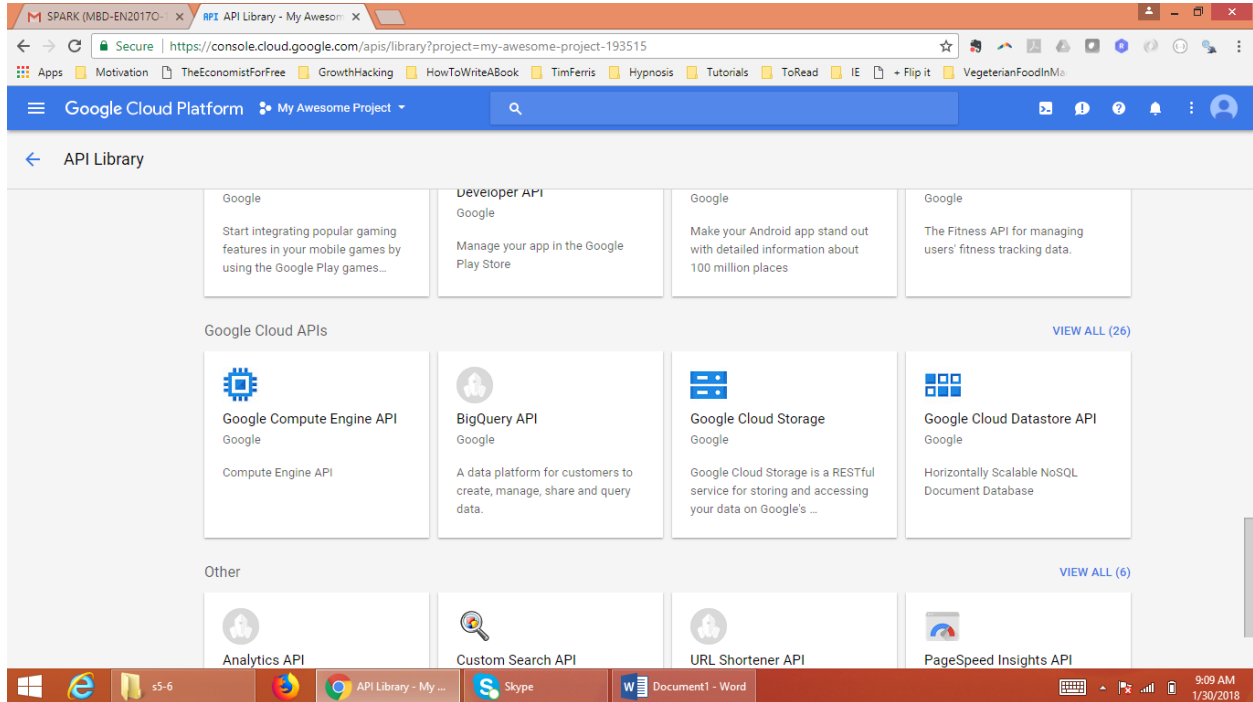
Traffic
Requests/sec

Errors
Percent of requests
There are no errors for this time period.

Median latency
Milliseconds
There is no latency data.

API	Requests	Errors	Error ratio	Latency, median	Latency, 98%	
Google Compute Engine API	3	0	0%	—	—	Disable
BigQuery API	—	—	—	—	—	Disable

3. Go to “Google Compute Engine API”



4. Enable “Google Compute Engine API”

The screenshot shows a web browser window with the Google Cloud Platform console. The address bar displays the URL: <https://console.cloud.google.com/apis/library/compute.googleapis.com/?id=a08439d8-80d6-43f1-af2e-6878251f018d&project=my-...>. The page header includes a notification about a credit balance and a trial period, along with buttons for 'DISMISS' and 'UPGRADE'. The main navigation bar shows 'Google Cloud Platform' and 'My Awesome Project'. The left sidebar indicates the current view is 'API Library'. The main content area displays the 'Google Compute Engine API' by Google. It includes a circular icon with a blue chip, a 'MANAGE' button, a 'TRY THIS API' button with an external link icon, and a green checkmark indicating 'API enabled'. Below this, the 'Type' is listed as 'APIs & services', 'Last updated' as '11/01/2018, 05:56', and 'Category' as 'Compute' and 'Networking'. The 'Overview' section states: 'Creates and runs virtual machines on Google Cloud Platform.' The 'About Google' section mentions Google's mission to organize the world's information and make it universally accessible and useful. The Windows taskbar at the bottom shows the Start button, taskbar icons for Edge, File Explorer, and several open applications including 'APIs & Services', 'Skype', and 'Word'. The system clock shows 9:09 AM on 1/30/2018.

You have €241.27 in credit and 363 days left of your free trial. [DISMISS](#) [UPGRADE](#)

Google Cloud Platform My Awesome Project

API Library

Google Compute Engine API

Google

Compute Engine API

[MANAGE](#) [TRY THIS API](#) ✓ API enabled

Type
APIs & services

Last updated
11/01/2018, 05:56

Category
Compute
Networking

Service name

Overview
Creates and runs virtual machines on Google Cloud Platform.

About Google
Google's mission is to organize the world's information and make it universally accessible and useful. Through products and platforms like Search, Maps, Gmail, Android, Google Play, Chrome and YouTube, Google plays a meaningful role in the daily lives of billions of people.

Windows taskbar: Edge, File Explorer, s5-6, APIs & Services - ..., Skype, Document1 - Word, 9:09 AM 1/30/2018

5. Go to “DataProc”

The screenshot shows the Google Cloud Platform console interface. At the top, a search bar contains the text "dataproc". Below the search bar, a dropdown menu displays two results: "Dataproc" and "RPI Google Cloud Dataproc API". The main dashboard area is visible in the background, showing various sections like Project info, Resources, Trace, RPI APIs, Google Cloud Platform status, Billing, and Error Reporting. The browser's address bar shows the URL "https://console.cloud.google.com/home/dashboard?project=my-awesome-project-193515". The Windows taskbar at the bottom indicates the system time is 9:10 AM on 1/30/2018.

6. Click on “Create cluster” button and you will reach the “Create a cluster” interface.

The screenshot shows the Google Cloud Platform console for a project named 'My Awesome Project'. The left sidebar contains a navigation menu with 'Cloud Dataproc' selected, showing sub-items for 'Clusters' and 'Jobs'. The main content area is titled 'Create a cluster' and contains the following configuration fields:

- Name:** gcelab
- Region:** us-central1
- Zone:** us-central1-c
- Cluster mode:** Standard (1 master, N workers)
- Master node:** Contains the YARN Resource Manager, HDFS NameNode and all job drivers
- Machine type:** 1 vCPU, 3.75 GB memory. A link 'Customise' is available. A note states: 'Upgrade your account to create instances with up to 96 cores'.
- Primary disk size (minimum 10 GB):** 500 GB
- Worker nodes:** Each contains a YARN NodeManager and a HDFS DataNode. The HDFS replication factor is 2.

The bottom of the image shows a Windows taskbar with icons for the Start menu, Edge browser, a folder named 's5-6', the 'Create Cluster' application, Skype, and a Word document titled 'Document1 - Word'. The system clock in the bottom right corner displays '9:11 AM 1/30/2018'.

7. Provide the configuration details. (To keep it cheaper I went for the minimal resources i.e. 1 Master Node and 2 worker nodes; for each node too, I chose the lowest possible configuration as seen in the following screenshot).

The screenshot shows the 'Create a cluster' page in the Google Cloud Platform console. The page is for a project named 'My Awesome Project'. The left sidebar shows 'Cloud Dataproc' with 'Clusters' and 'Jobs' options. The main content area is titled 'Create a cluster' and contains the following configuration details:

- Machine type:** 1 vCPU, 3.75 GB memory. A link to 'Customise' is available. A note says 'Upgrade your account to create instances with up to 96 cores'.
- Primary disk size (minimum 10 GB):** 10 GB.
- Nodes (minimum 2):** 2. **Local SSDs (0-8):** 0 x 375 GB.
- YARN cores:** 2. **YARN memory:** 6.00 GB.
- Pre-emptible workers, bucket, network, version, initialisation & access options:** This section is expanded, showing a checked checkbox for 'Pre-emptible workers'.
- Buttons:** 'Create' and 'Cancel' buttons are at the bottom.

The bottom of the screenshot shows a Windows taskbar with the following applications: 's5-6', 'Create Cluster - ...', 'Skype', and 'Document1 - Word'. The system clock shows '9:12 AM' on '1/30/2018'.

8. After providing the configuration details, click on Create button and we see that the process of cluster creation has started.

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'My Awesome Project', and a search bar. The left sidebar shows the 'Cloud Dataproc' menu with 'Clusters' and 'Jobs' options. The main content area is titled 'Clusters' and features a 'CREATE CLUSTER' button, a 'REFRESH' button, a 'DELETE' button, and a 'REGIONS' dropdown. Below this is a search bar with the placeholder text 'Search clusters, press Enter'. A table lists the clusters:

<input type="checkbox"/>	Name ^	Region	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
<input type="checkbox"/>	gcelab	us-central1	us-central1-c	2	dataproc-7a2e6a4c-e3db-4a9e-966f-e1d6f5f293c6-us-central1	30 Jan 2018, 09:12:21	Provisioning

The bottom of the screenshot shows the Windows taskbar with the Start button, taskbar icons for 's5-6', 'Dataproc Clusters...', 'Skype', and 'Document1 - Word', and a system tray showing the time as 9:12 AM on 1/30/2018.

9. Cluster created and its running now.

SPARK (MBD-EN20170)

Dataproc Clusters - My

Secure | https://console.cloud.google.com/dataproc/clusters?project=my-awesome-project-193515

Apps Motivation TheEconomistForFree GrowthHacking HowToWriteABook TimFerris Hypnosis Tutorials ToRead IE + Flip it VegetarianFoodInMa

You have €241.27 in credit and 363 days left of your free trial.

DISMISS UPGRADE

Google Cloud Platform My Awesome Project

Cloud Dataproc

Clusters

Jobs

Clusters

CREATE CLUSTER REFRESH DELETE REGIONS

SHOW INFO PANEL

Search clusters, press Ent

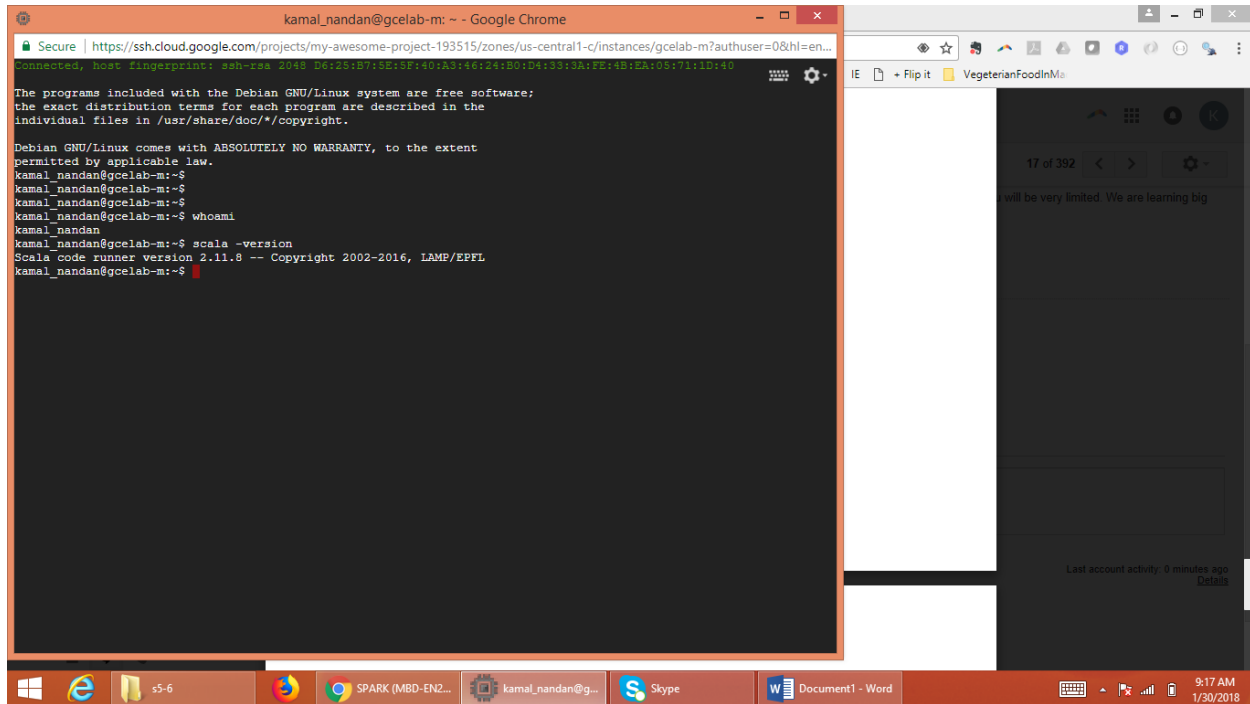
Name	Region	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
gcelab	us-central1	us-central1-c	2	dataproc-7a2e6a4c-e3db-4a9e-966f-e1d6f5f293c6-us-central1	30 Jan 2018, 09:12:21	Running

Windows Taskbar

Taskbar Icons: s5-6, Dataproc Clusters..., Skype, Document1 - Word

System Tray: 9:15 AM 1/30/2018

10. Go to “VM instances” Tab and there we will see the option to connect to our cluster through the SSH interface provided by dataproc itself.



11. Now we will like to test scaling up the cluster – I went from 2 worker nodes to 3 worker nodes

The screenshot shows the Google Cloud Platform console interface. The left sidebar displays the 'Cloud Dataproc' menu with 'Clusters' and 'Jobs' options. The main content area is titled 'Cluster details' for a cluster named 'gcelab'. The 'Configuration' tab is selected, showing the following details:

Property	Value
Region	us-central1
Zone	us-central1-c
Master node	Standard (1 master, N workers)
Machine type	n1-standard-1 (1 vCPU, 3.75 GB memory)
Primary disk size	10 GB
Worker nodes	2
Machine type	n1-standard-1 (1 vCPU, 3.75 GB memory)
Primary disk size	10 GB
Local SSDs	0
Preemptible worker nodes	0
Cloud Storage staging bucket	dataproc-7a2e6a4c-e3db-4a9e-966f-e1d6f5f293c6-us-central1
Subnetwork	default
Network tags	None
Internal IP only	No
Image version	1.2.20
Created	30 Jan 2018, 09:12:21

The bottom of the image shows a Windows taskbar with several open applications: 's5-6', 'gcelab - My Awes...', 'Skype', and 'Document1 - Word'. The system clock in the bottom right corner indicates the time is 9:18 AM on 1/30/2018.

12. Click on “Edit button” and provide the no. of nodes we want. Provide 3 and click on “Save” button. Cluster scaling-up process would start.

The screenshot shows the Google Cloud Platform console interface. The left sidebar displays the 'Cloud Dataproc' menu with 'Clusters' and 'Jobs' options. The main content area is titled 'Cluster details' for a cluster named 'gcelab'. The 'Configuration' tab is selected, showing various settings for the cluster. The 'Worker nodes' field is set to 3. The 'Machine type' is 'n1-standard-1 (1 vCPU, 3.75 GB memory)'. The 'Primary disk size' is 10 GB. The 'Local SSDs' are set to 0. The 'Preemptible worker nodes' field is set to 'Number of nodes'. The 'Cloud Storage staging bucket' is 'dataproc-7a2e6a4c-e3db-4a9e-966f-e1d6f5f293c6-us-central1'. The 'Subnetwork' is 'default'. The 'Network tags' are 'None'. The 'Internal IP only' option is 'No'. The 'Image version' is '1.2.20'. The 'Created' timestamp is '30 Jan 2018, 09:12:21'. The 'Labels' section shows two labels: 'goog-dataproc-cluster-name' with value 'gcelab' and 'goog-dataproc-cluster-uuid' with value '019a1b53-60a0-4133-8ee8-f6af68'.

Key	Value
goog-dataproc-cluster-name	gcelab
goog-dataproc-cluster-uuid	019a1b53-60a0-4133-8ee8-f6af68

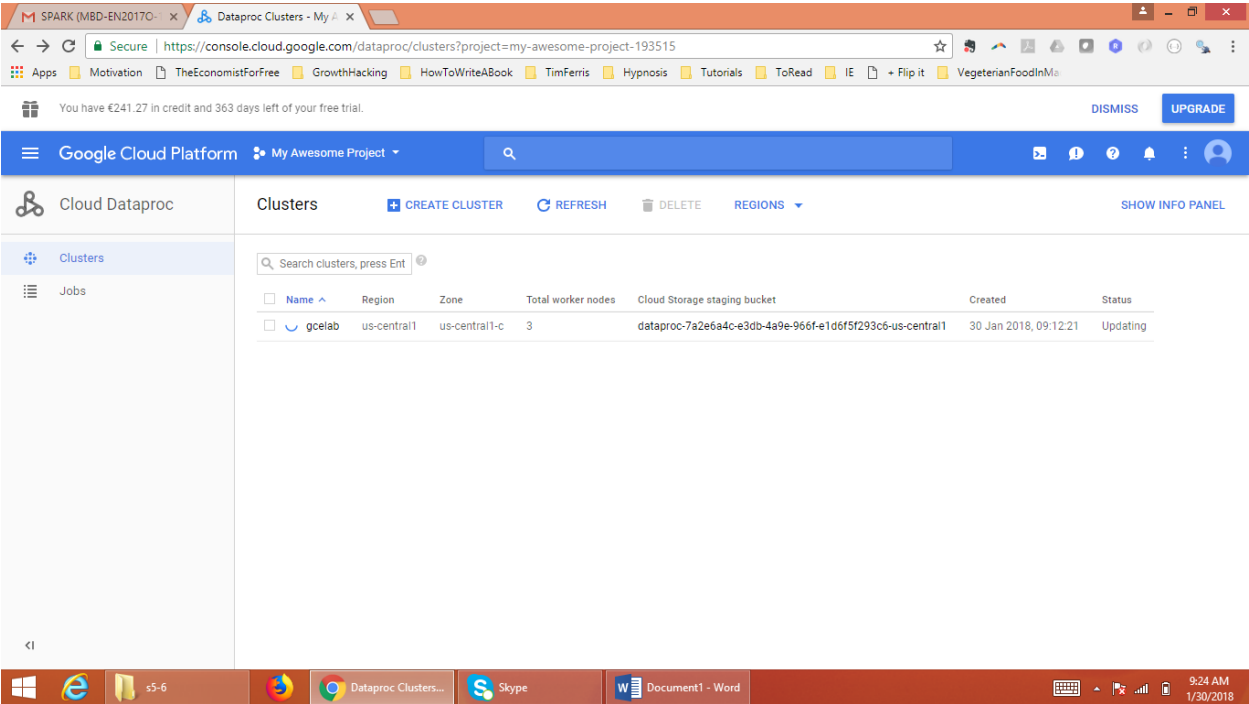
13. We have successfully scaled up the cluster from 2 to 3 nodes.
(It took me 50 seconds in scaling up. But, it is not consistent always – I tried a no. of times and it varied between 35 to 56 seconds)

The screenshot displays the Google Cloud Platform console for a project named 'My Awesome Project'. The 'Cloud Dataproc' section is active, showing the 'gcelab' cluster details. The 'VM Instances' tab is selected, displaying a table of instances:

Name	Role	SSH
gcelab-m	Master	SSH
gcelab-w-0	Worker	
gcelab-w-1	Worker	
gcelab-w-2	Worker	

Below the table, it indicates 'Equivalent REST'. The bottom status bar shows the time as 9:23 AM on 1/30/2018.

14. We followed the same steps as we did in scaling up, to scale down the cluster – we scaled it down from 3 to 2 nodes.



15. Successfully scaled it down to 2 worker nodes. (Scaling down took 1 min. and 22 seconds and this time was more or less consistent.)

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud logo, the project name 'My Awesome Project', and a search bar. Below the navigation bar, the left sidebar shows the 'Cloud Dataproc' section with a 'Clusters' link selected. The main content area displays the 'Clusters' page with a table of active clusters. A single cluster named 'gcelab' is listed with 2 worker nodes. The table columns are: Name, Region, Zone, Total worker nodes, Cloud Storage staging bucket, Created, and Status.

Name	Region	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
gcelab	us-central1	us-central1-c	2	dataproc-7a2e6a4c-e3db-4a9e-966f-e1d6f5f293c6-us-central1	30 Jan 2018, 09:12:21	Running

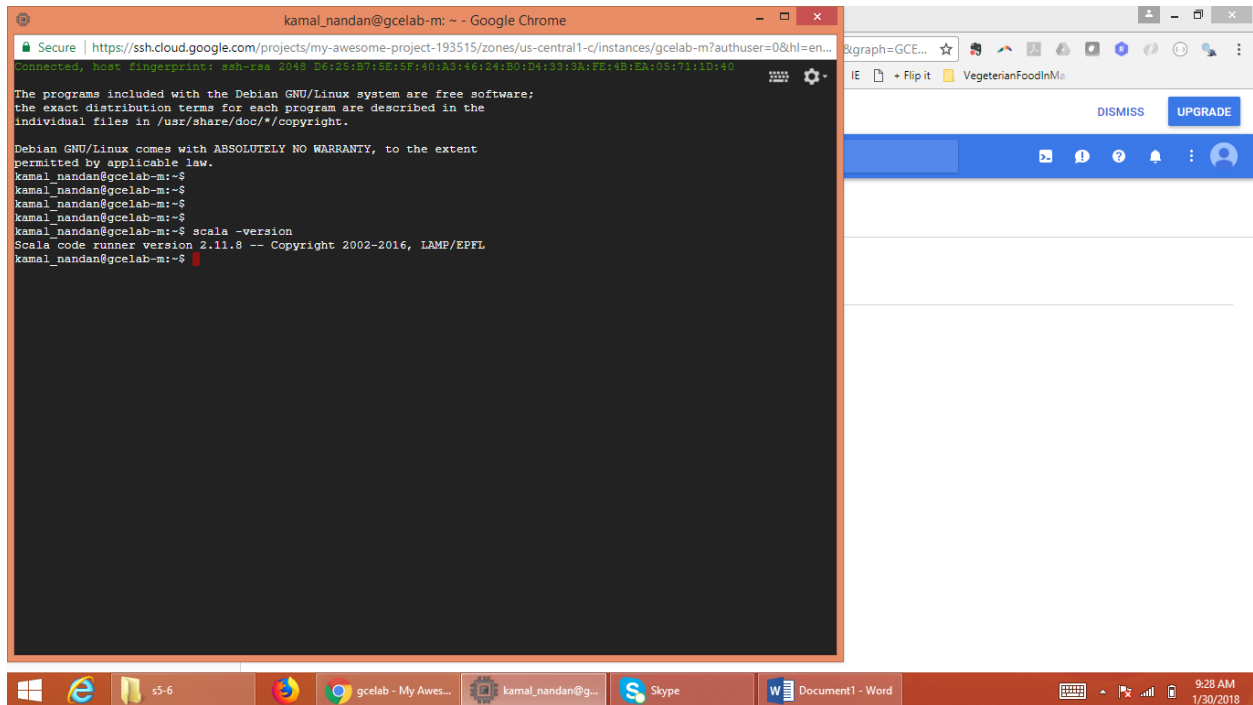
16. Go to VM instances tab and click on “SSH”, and we would have a browser based SSH client.

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud logo, the project name 'My Awesome Project', and a search bar. The left sidebar shows the 'Cloud Dataproc' section with 'Clusters' and 'Jobs' sub-panels. The main content area displays the 'Cluster details' for a cluster named 'gcelab'. The 'VM Instances' tab is selected, showing a table of instances:

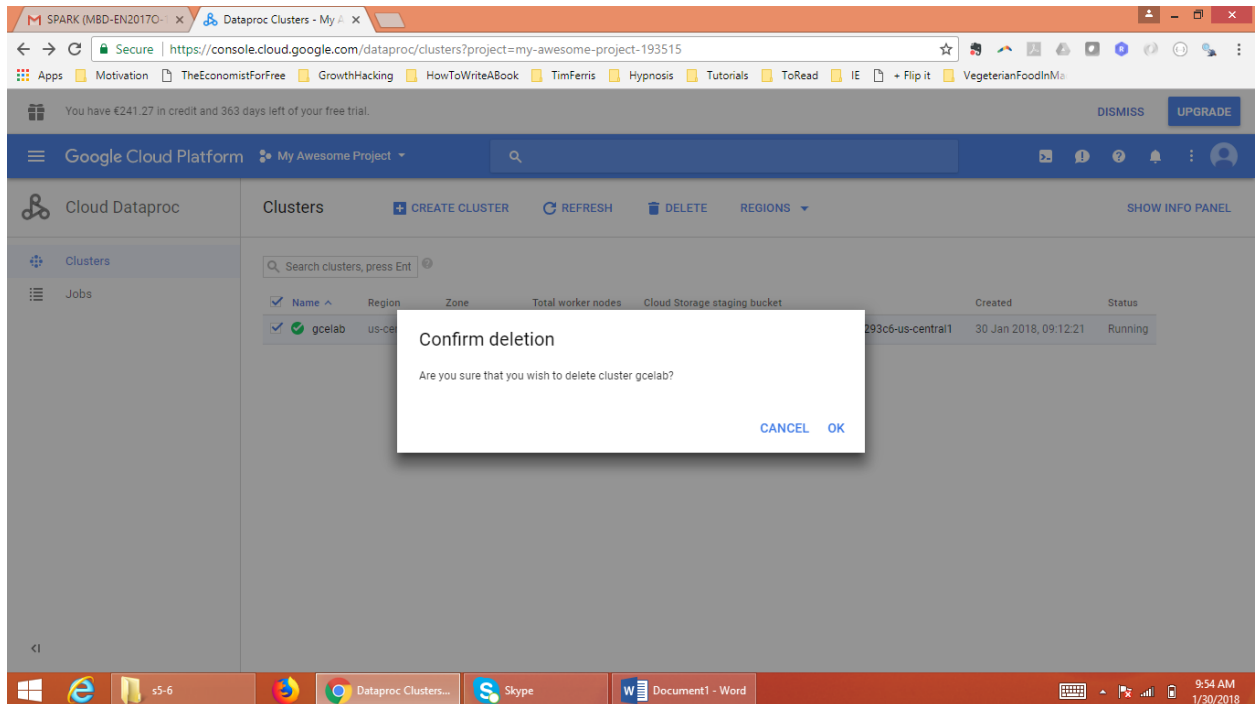
Name	Role	SSH
gcelab-m	Master	SSH
gcelab-w-0	Worker	
gcelab-w-1	Worker	

Below the table, there is a link for 'Equivalent REST'. The bottom of the screen shows a Windows taskbar with various application icons and a system tray indicating the time as 9:26 AM on 1/30/2018.

17. Check the scala version – its 2.11.8



18. Now delete the cluster – we don't need it for now.



19. As an extra step, I also tried to login through putty ssh client, which I find more convenient. For this, I had to generate public/private ssh keys and provide my public key to dataproc and then login through putty using my private key. I have not taken the screenshots of this process.