

빅데이터 분석 및 응용

L01.1: GCP & DFS Practice

Summer 2020

Kookmin University

Contents

- ❖ **Google Cloud Platform**
- ❖ **Create a Hadoop Cluster**
- ❖ **Setup your cluster**
- ❖ **Upload files to the master node**
- ❖ **Upload files to HDFS**

Google Cloud Platform

- A suite of cloud computing services
- **Cloud Dataproc:** Big data platform for running **Apache Hadoop** and **Apache Spark** jobs.
- Free 300\$ credit for first 12 months.
- <https://cloud.google.com>

Google Cloud Platform



Google을 선택해야 하는 이유 솔루션 제품 가격 책정 시작하기 >



한국어 콘솔

h

영업팀에 문의

무료로 시작하기

차세대 제품 개발 더 나은 소프트웨어로 더욱 빠르게

- ✓ Google의 핵심 인프라, 데이터 분석, 머신러닝을 사용하세요.
- ✓ 모든 기업에 적합한 보안과 완벽한 기능을 제공합니다.
- ✓ 오픈소스 및 업계 최고 수준의 가격 대비 성능을 제공하기 위해 최선을 다하고 있습니다.

Google Cloud Platform

Google Cloud Platform 무료로 사용해 보기

1/2단계



hamyung park
hamyung.park@gmail.com

[계정 전환](#)

국가

대한민국 ▼

서비스 약관

☒ [Google Cloud Platform 무료 평가판 서비스 약관](#)을 읽었으며 이에 동의합니다.

계속 진행하려면 체크박스를 선택하세요.

계속

모든 Cloud Platform 제품에 액세스

Firebase, Google Maps API 등을 포함해 앱, 웹사이트, 서비스를 구축하고 실행하는 데 필요한 모든 기능을 이용할 수 있습니다.

\$300의 무료 크레딧

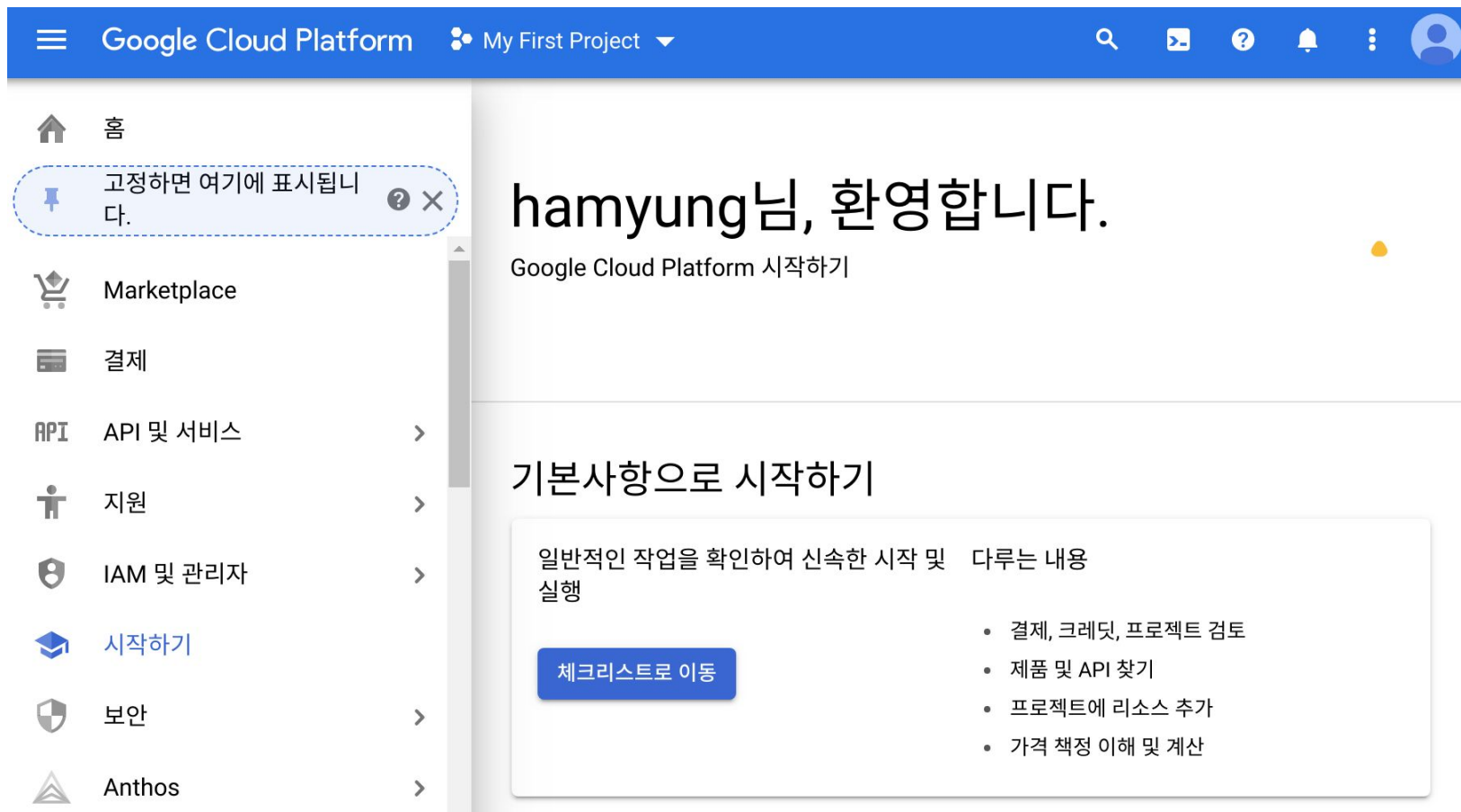
가입하여 Google Cloud Platform에서 12개월간 사용할 수 있는 \$300 크레딧을 받아 보세요.

무료 체험판 종료 후 자동 청구되지 않음

신용카드를 요청하는 이유는 자동 가입을 방지하기 위해서입니다. 유료 계정으로 직접 업그레이드하지 않는 한 요금이 청구되지 않습니다.

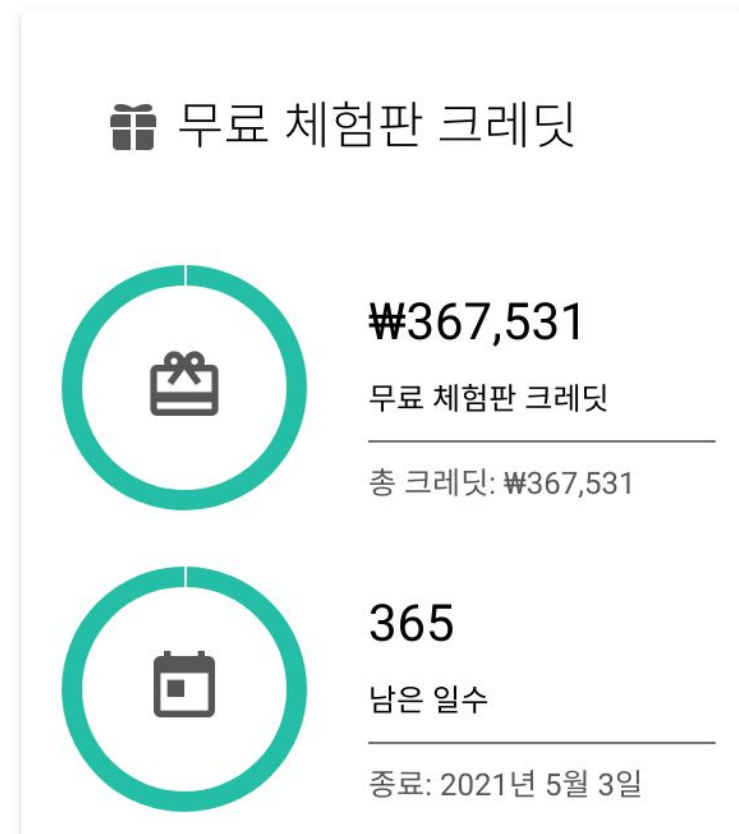
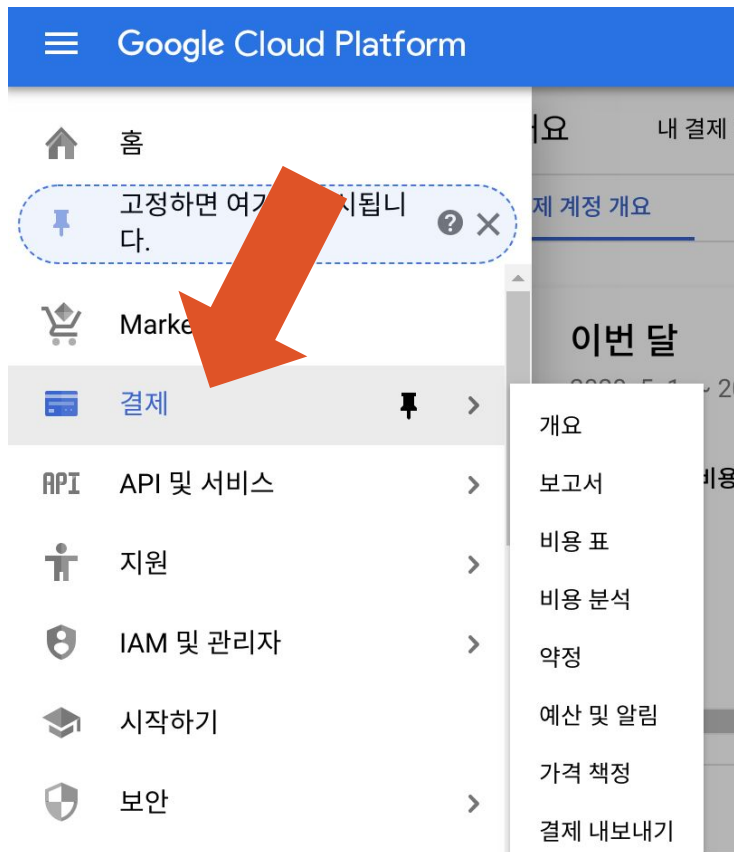
Google Cloud Platform

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



Google Cloud Platform

- Check the status of your credit.



Contents

- ❖ Google Cloud Platform
- ❖ **Create a Hadoop Cluster**
- ❖ Setup your cluster
- ❖ Upload files to the master node
- ❖ Upload files to HDFS

Creating a Hadoop cluster

The image illustrates the steps to create a Hadoop cluster on Google Cloud Platform:

- 1**: Click the menu icon (hamburger menu) in the top left corner of the Google Cloud Platform console.
- 2**: Click on **Dataproc** in the left-hand navigation menu.
- 3**: Click on **클러스터** (Clusters) in the sub-menu that appears next to Dataproc.
- 4**: Click the **클러스터 만들기** (Create Cluster) button in the right-hand panel.

The right-hand panel shows the **Dataproc** console with the **클러스터** (Clusters) tab selected. The main content area displays the **Cloud Dataproc** header and a description: "Google Cloud Dataproc 애널리틱스 데이터 저장" (Google Cloud Dataproc Analytics Data Storage). Below this, it states "현재 선택한 Cloud Data:" (Currently selected Cloud Data:). The **클러스터 만들기** (Create Cluster) button is visible at the bottom right.

Creating a Hadoop cluster

이름 ?

kmubigdata-cluster

리전 ?

asia-northeast1

영역 ?

asia-northeast1-a

클러스터 모드 ?

표준(마스터 1, 작업자 N)

마스터 노드

YARN Resource Manager, HDFS NameNode 및 모든 작업 드라이버를 포함합니다.

머신 구성

머신 계열

일반 용도

일반적인 작업 부하에 적합한 머신 유형이며 가격 및 유연성을 위해 최적화되었습니다.

시리즈

N1

Intel Skylake CPU 플랫폼 또는 이전 버전의 플랫폼에서 제공

머신 유형

n1-standard-2(vCPU 2개, 7.5GB 메모리)



vCPU

2

메모리

7.5GB

⌵ CPU 플랫폼 및 GPU

기본 디스크 크기(최소 15GB) ?

30

GB

기본 디스크 유형 ?

표준 영구 디스크

Creating a Hadoop cluster

작업자 노드

각각 YARN NodeManager 및 HDFS DataNode를 포함합니다.
HDFS 복제 인수는 2입니다.

머신 구성

머신 계열

일반 용도

일반적인 작업 부하에 적합한 머신 유형이며 가격 및 유연성을 위해 최적화되었습니다.

시리즈

N1

Intel Skylake CPU 플랫폼 또는 이전 버전의 플랫폼에서 제공

머신 유형

n1-standard-2(vCPU 2개, 7.5GB 메모리)



vCPU

2

메모리

7.5GB

⌵ CPU 플랫폼 및 GPU

기본 디스크 크기(최소 15GB) ?

30 GB

노드(최소 2개) ?

3

기본 디스크 유형 ?

표준 영구 디스크

로컬 SSD(0~8) ?

0

x 375 GB

YARN 코어 수 ?

6

YARN 메모리 ?

18GB

자동 확장 정책 ? (선택사항)



클러스터에 자동 확장을 사용 설정합니다.

현재 프로젝트에 이 리전의 자동 확장을 사용 설정할 수 있는 정책이 없습니다. [자동 확장 정책을 만드는 방법 알아보기](#)

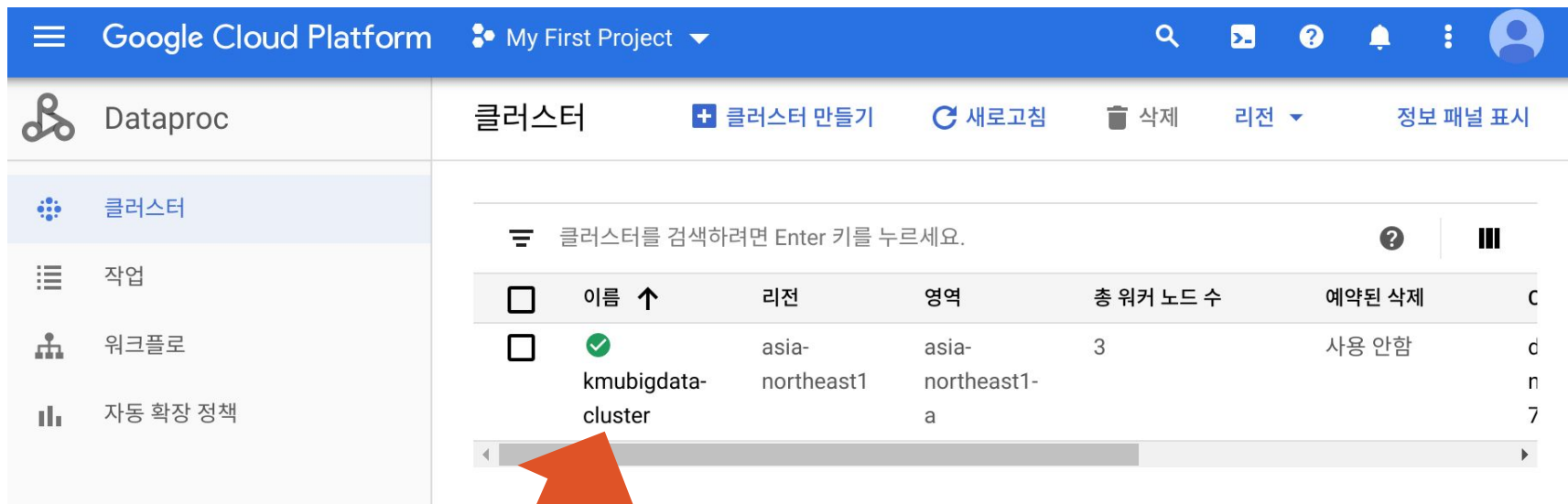
구성요소 게이트웨이



클러스터에서 기본 및 선택된 구성요소(선택사항)의 웹 인터페이스에 대한 액세스를 사용 설정합니다. [자세히 알아보기](#)

Creating a Hadoop cluster

- Yeah~! We just got a Hadoop cluster!





The screenshot shows the Google Cloud Platform interface for the 'My First Project'. The left sidebar contains the 'Dataproc' menu with options for '클러스터' (Clusters), '작업' (Jobs), '워크플로' (Workflows), and '자동 확장 정책' (Autoscaling Policies). The main panel displays the '클러스터' (Clusters) page with a table of existing clusters. A red arrow points to the cluster named 'kmubigdata-cluster'.


이름 ↑	리전	영역	총 워커 노드 수	예약된 삭제
kmubigdata-cluster	asia-northeast1	asia-northeast1-a	3	사용 안함


Click here to see Web UI of Hadoop


Creating a Hadoop cluster


 Dataproc


 클러스터


 작업


 워크플로


 자동 확장 정책


 클러스터 세부정보


 + 작업 제출

 새로고침

 삭제

 로그 보기

 **kmubigdata-cluster**

 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.

모니터링

작업

VM 인스턴스

구성

웹 인터페이스

SSH 터널

SSH 터널을 만들어 웹 인터페이스에 연결

구성요소 게이트웨이

클러스터에서 기본 및 선택된 구성요소(선택사항)의 웹 인터페이스에 대한 액세스를 제공합니다. [자세히 알아보기](#)

[YARN ResourceManager](#)

[HDFS NameNode](#)

[MapReduce Job History](#)

[YARN Application Timeline](#)

[Spark History Server](#)

[Tez](#)

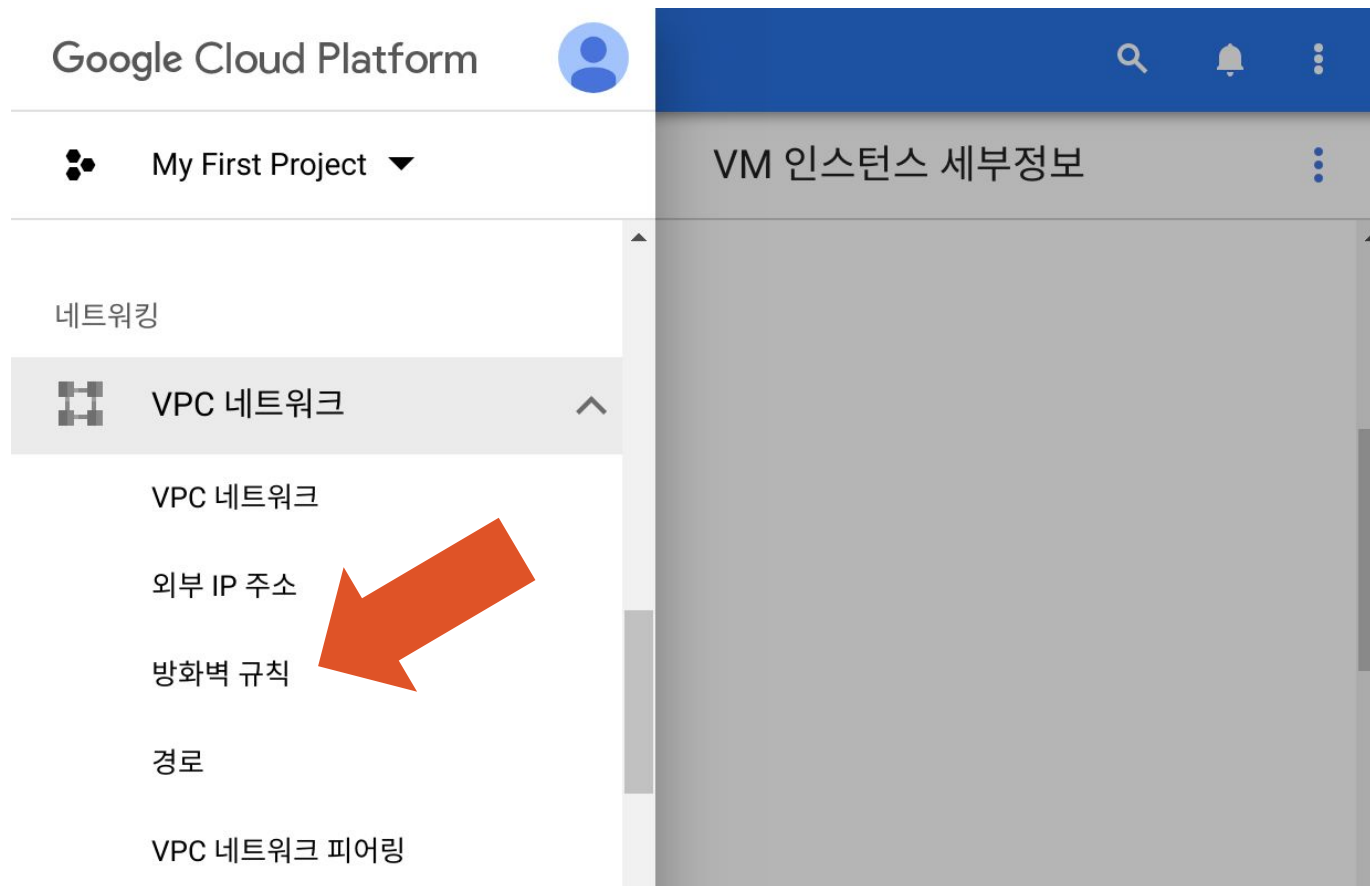
Click here
to see the Web UIs

Contents

- ❖ Google Cloud Platform
- ❖ Create a Hadoop Cluster
- ❖ **Setup your cluster**
- ❖ Upload files to the master node
- ❖ Upload files to HDFS

Setup your cluster

- Adding Firewall rules



Setup your cluster

- Adding Firewall rules

Google Cloud Platform

VPC 네트워크

VPC 네트워크

외부 IP 주소

방화벽 규칙

경로

VPC 네트워크 피어링

공유 VPC

서버리스 VPC 액세스

방화벽 규칙

+ 방화벽 규칙 만들기

새로고침

로그 사용 설정

방화벽 규칙은 인스턴스에 들어오는 또는 송신되는 트래픽을 제어합니다. 기본적으로 네트워크 외부에서 수신되는 트래픽은 차단됩니다. [자세히 알아보기](#)

참고: App Engine 서비스는 [여기](#)에서 관리합니다.

테이블 필터링

<input type="checkbox"/>	이름	유형	대상	필터	프로토콜/포트	작업
<input type="checkbox"/>	default-allow-icmp	수신	전체 적용	IP 범위: 0.0.0	icmp	허용
<input type="checkbox"/>	default-	수신	전체 적용	IP 범위: 10.12	tcp:0-65535	허용

Setup your cluster

• Adding Firewall rules

방화벽 규칙은 인스턴스로 수신 또는 발신되는 트래픽을 제어합니다. 기본적으로 네트워크 외부에서 수신되는 트래픽은 차단됩니다. [자세히 알아보기](#)

이름 *
default-allow-dataproc-access ?
소문자, 수자, 하이픈이 허용됩니다

일치 시 작업 ?

- ☒ 허용
☐ 거부

대상
네트워크의 모든 인스턴스 ?

소스 필터
IP 범위 ?

소스 IP 범위 *
1.209.63.170 ✕ 예: 0.0.0.0/0, 192.168.2.0/24 ?

프로토콜 및 포트 ?

- ☒ 모두 허용
☐ 지정된 프로토콜 및 포트

✓ 규칙 사용

만들기 취소

Get your current ip addr. here:

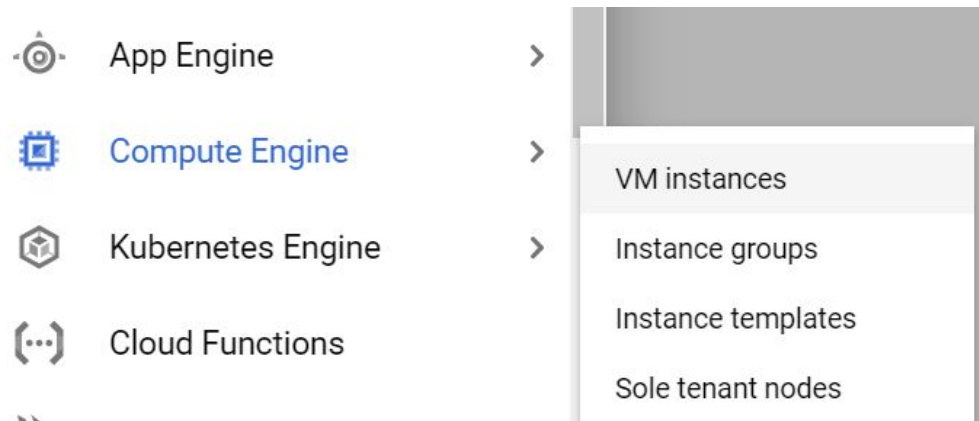
<https://www.whatismyip.com/>

Setup your cluster

- Setting up the master node to log in by password.
 - Step 1. modify a configuration file of the master node
 - Step 2. change your password

Setup your cluster

- Side-menu > Compute Engine > VM instances




Setup your cluster

- Select “SSH” button of the master node

VM instances [+ CREATE INSTANCE](#) [IMPORT VM](#) [REFRESH](#) [▶ START](#) [■ STOP](#) [RESET](#)

Filter VM instances

<input type="checkbox"/>	Name ^	Zone	Recommendation	In use by	Internal IP	External IP	Con
<input type="checkbox"/>	✓ kmubigdata-cluster-m	asia-northeast1-a			10.146.0.5 (nic0)	34.85.10.179	SSH ▾ ⋮
<input type="checkbox"/>	✓ kmubigdata-cluster-w-0	asia-northeast1-a			10.146.0.4 (nic0)	34.85.87.150	SSH ▾ ⋮
<input type="checkbox"/>	✓ kmubigdata-cluster-w-1	asia-northeast1-a			10.146.0.2 (nic0)	34.85.97.22	SSH ▾ ⋮
<input type="checkbox"/>	✓ kmubigdata-cluster-w-2	asia-northeast1-a			10.146.0.3 (nic0)	35.243.102.209	SSH ▾ ⋮



Setup your cluster

- Edit /etc/ssh/sshd_config file
 - Open the file using a text editor (e.g. nano, vi, ...)
 - PasswordAuthentication **no** ☐ **yes**
 - uncomment “Port 22” and add “Port 2222”
 - Save (Ctrl + X ☐ y ☐ Enter on nano, esc ☐ “:wq” on vi)

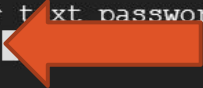
```
x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sat May  4 15:28:33 2019 from 106.249.180.251
hmpark@kmubigdata-cluster-m:~$ sudo nano /etc/ssh/sshd_config
```

```
Port 22
Port 2222
#AddressFamily any
#ListenAddress 0.0
#ListenAddress ::
```

```
# Change to yes if you don't trust ~/.ssh/known_hosts for
# HostbasedAuthentication
#IgnoreUserKnownHosts no
# Don't read the user's ~/.rhosts and ~/.shosts files
#IgnoreRhosts yes


# To disable tunneled clear text passwords, change to no here!
PasswordAuthentication yes
#PermitEmptyPasswords no
```



Setup your cluster

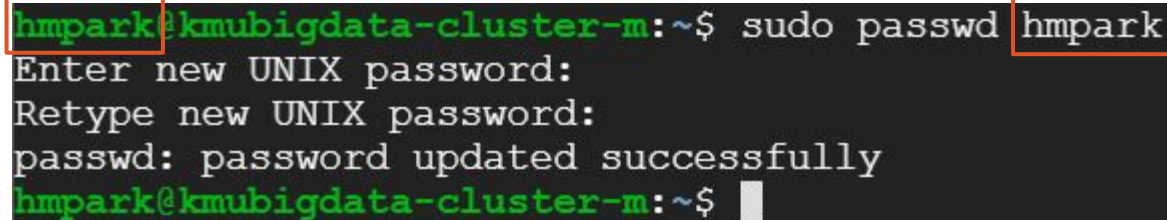
- Restart ssh server
 - `sudo service ssh restart`

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent  
permitted by applicable law.  
Last login: Sat May  4 15:28:33 2019 from 106.249.180.251  
hmpark@kmubigdata-cluster-m:~$ sudo nano /etc/ssh/sshd_config  
hmpark@kmubigdata-cluster-m:~$ sudo service ssh re  
reload    restart  
hmpark@kmubigdata-cluster-m:~$ sudo service ssh restart
```



Setup your cluster

- Change password
 - `sudo passwd <your-id>`
 - Enter new password



A terminal window showing the process of changing a password. The prompt is `hmpark@kmubigdata-cluster-m:~$`. The command `sudo passwd hmpark` is entered. The output shows the password being updated successfully. Annotations include a red box around the username `hmpark` in the prompt, a red box around the username `hmpark` in the command, a red arrow pointing from the first box to the second, and a large orange arrow pointing to the command.

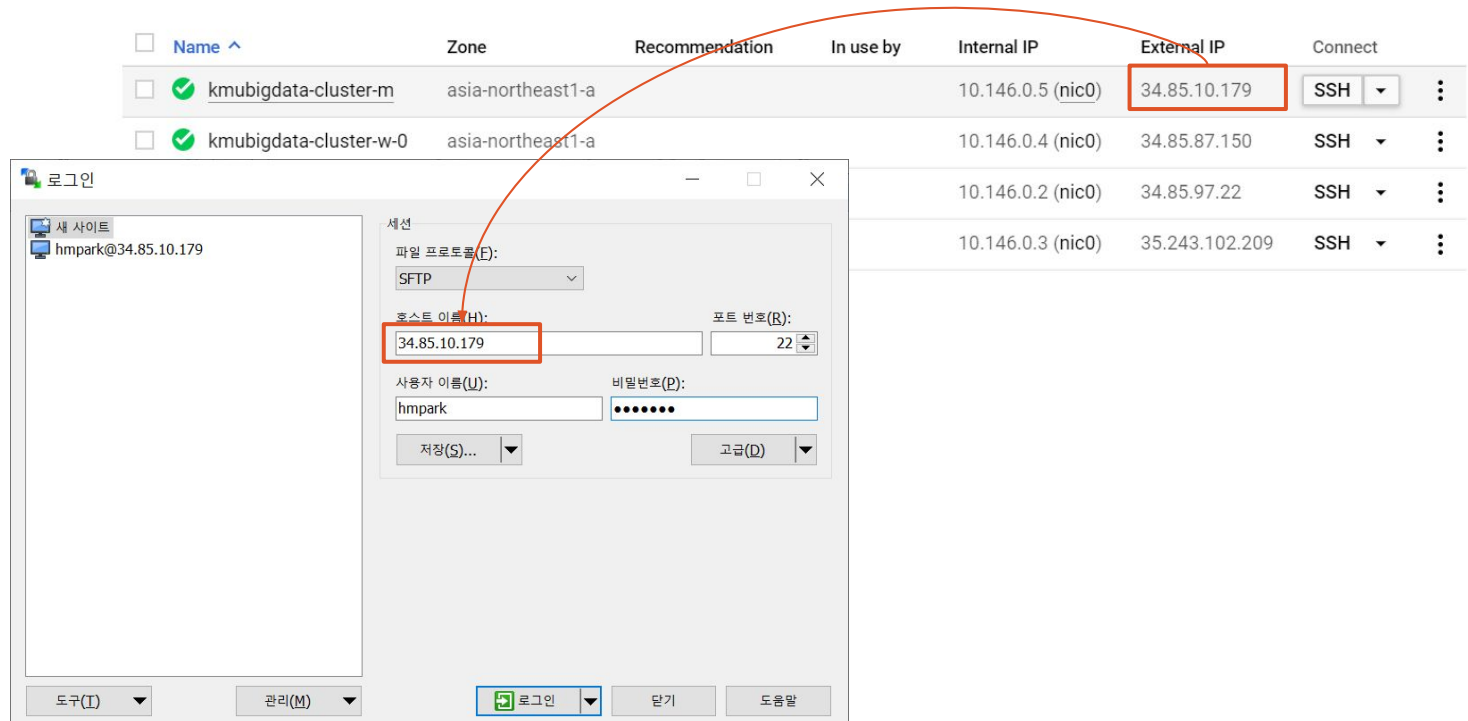
```
hmpark@kmubigdata-cluster-m:~$ sudo passwd hmpark
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
hmpark@kmubigdata-cluster-m:~$
```

Contents

- ❖ Google Cloud Platform
- ❖ Create a Hadoop Cluster
- ❖ Setup your cluster
- ❖ **Upload files to the master node**
- ❖ Upload files to HDFS

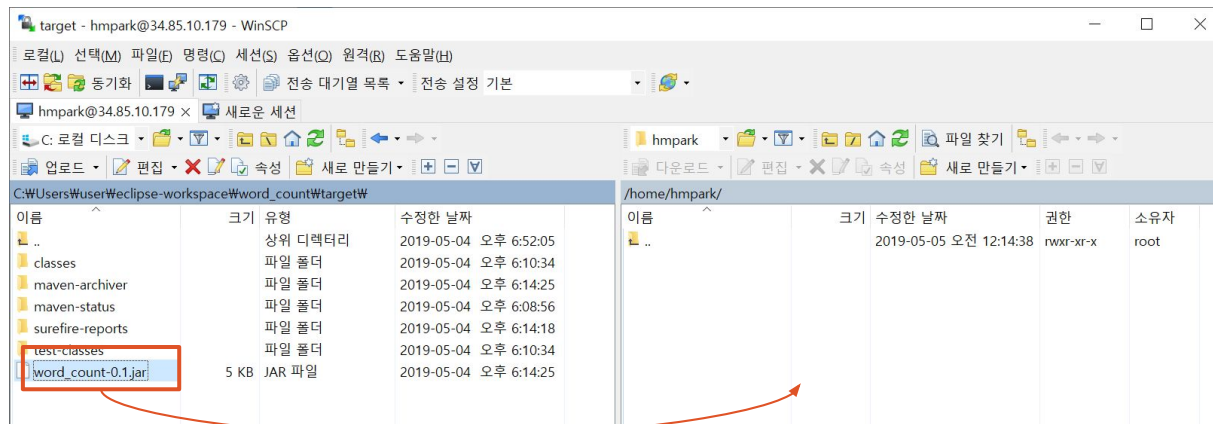
Upload files to the master

- On Windows 10
 - Download and install WinSCP from <https://winscp.net/eng/download.php>
 - Run WinSCP and set up the connection to the master

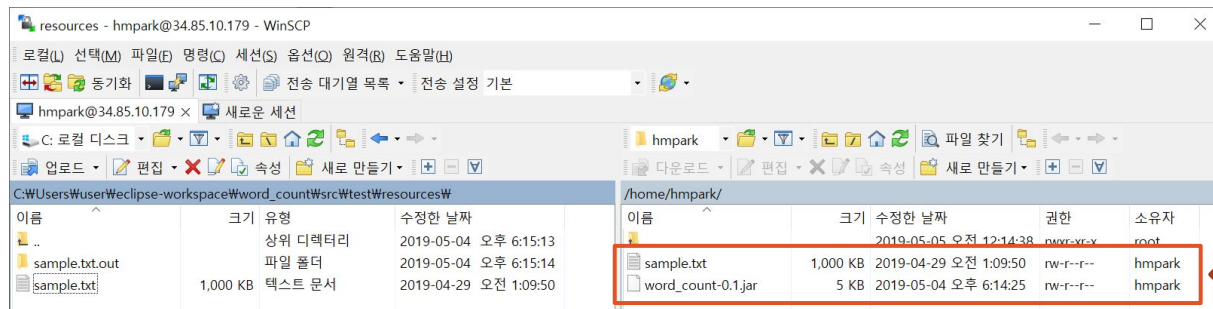


Upload files to the master

- On Windows 10
 - Move files to the master (a jar file, and a text file)



Drag-and-drop



Like this

Upload files to the master

- On Linux (Ubuntu)
 - Use “scp” command (secure copy)
 - scp [local file path] [remote ip]:[remote directory path]

<input type="checkbox"/> Name ^	Zone	Recommendation	In use by	Internal IP	External IP	Connect
<input type="checkbox"/> kmubigdata-cluster-m	asia-northeast1-a			10.146.0.5 (nic0)	34.85.10.179	SSH ▾ ⋮
<input type="checkbox"/> kmubigdata-cluster-w-0	asia-northeast1-a			10.146.0.4 (nic0)	34.85.87.150	SSH ▾ ⋮
<input type="checkbox"/> kmubigdata-cluster-w-1	asia-northeast1-a			10.146.0.2 (nic0)	34.85.97.22	SSH ▾ ⋮
<input type="checkbox"/> kmubigdata-cluster-w-2	asia-northeast1-a			10.146.0.3 (nic0)	35.243.102.209	SSH ▾ ⋮

```
hmpark@hmpark-acer: ~/eclipse-workspace/word_count
hmpark@hmpark-acer:~/eclipse-workspace/word_count$ scp target/word_counq-0.1.jar 34.85.10.179:~/
hmpark@34.85.10.179's password:
word_counq-0.1.jar                                100% 4550    48.4KB/s   00:00
hmpark@hmpark-acer:~/eclipse-workspace/word_count$ scp src/test/resources/sample.txt 34.85.10.179:~/
hmpark@34.85.10.179's password:
sample.txt                                         100% 996KB   1.0MB/s   00:00
hmpark@hmpark-acer:~/eclipse-workspace/word_count$
```

Contents

- ❖ Google Cloud Platform
- ❖ Create a Hadoop Cluster
- ❖ Setup your cluster
- ❖ Upload files to the master node
- ❖ **Upload files to HDFS**


Upload files to HDFS

- Connect the master via SSH

VM instances [CREATE INSTANCE](#) [IMPORT VM](#) [REFRESH](#) [START](#) [STOP](#) [RESET](#)

Filter VM instances

<input type="checkbox"/> Name ^	Zone	Recommendation	In use by	Internal IP	External IP	Connect
<input type="checkbox"/> kmubigdata-cluster-m	asia-northeast1-a			10.146.0.5 (nic0)	34.85.10.179	SSH
<input type="checkbox"/> kmubigdata-cluster-w-0	asia-northeast1-a			10.146.0.4 (nic0)	34.85.87.150	SSH
<input type="checkbox"/> kmubigdata-cluster-w-1	asia-northeast1-a			10.146.0.2 (nic0)	34.85.97.22	SSH
<input type="checkbox"/> kmubigdata-cluster-w-2	asia-northeast1-a			10.146.0.3 (nic0)	35.243.102.209	SSH



Upload files to HDFS

- Connect the master via SSH
- Use “ls” command to check the list of files in the current directory

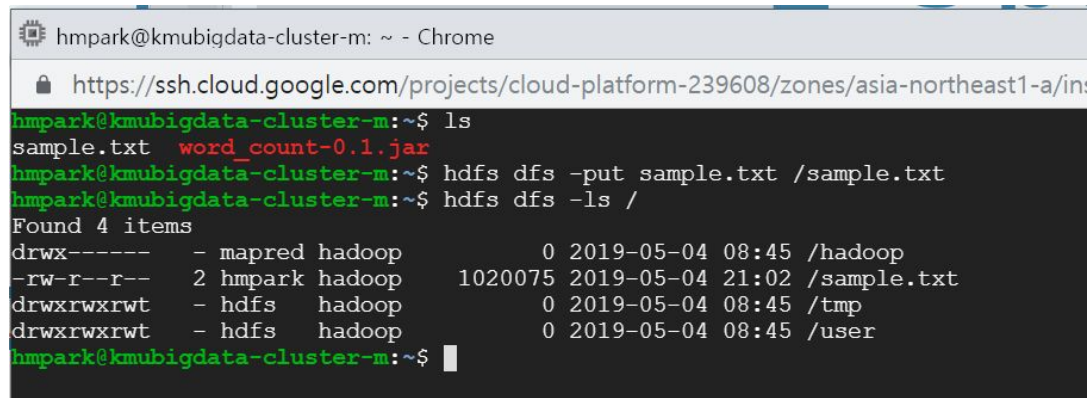
<input type="checkbox"/>	Name ^	Zone	Recommendation	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	✓ kmubigdata-cluster-m	asia-northeast1-a			10.146.0.5 (nic0)	34.85.10.179	SSH



```
hmpark@kmubigdata-cluster-m: ~ - Chrome
https://ssh.cloud.google.com/projects/cloud-platform-239608/z
hmpark@kmubigdata-cluster-m:~$ ls
sample.txt word_count-0.1.jar
hmpark@kmubigdata-cluster-m:~$
```

Upload files to HDFS

- Push the text file to HDFS
 - `hdfs dfs -put [local file path] [hdfs file path]`
 - To uploads the file at [local file path] to [hdfs file path] in HDFS
 - `hdfs dfs -ls [hdfs directory path]`
 - To check the list of files in [hdfs directory path]



```
hmpark@kmubigdata-cluster-m: ~ - Chrome
https://ssh.cloud.google.com/projects/cloud-platform-239608/zones/asia-northeast1-a/in:
hmpark@kmubigdata-cluster-m:~$ ls
sample.txt  word_count-0.1.jar
hmpark@kmubigdata-cluster-m:~$ hdfs dfs -put sample.txt /sample.txt
hmpark@kmubigdata-cluster-m:~$ hdfs dfs -ls /
Found 4 items
drwx----- - mapred  hadoop          0 2019-05-04 08:45 /hadoop
-rw-r--r--  2 hmpark  hadoop    1020075 2019-05-04 21:02 /sample.txt
drwxrwxrwt - hdfs    hadoop          0 2019-05-04 08:45 /tmp
drwxrwxrwt - hdfs    hadoop          0 2019-05-04 08:45 /user
hmpark@kmubigdata-cluster-m:~$
```

Questions?