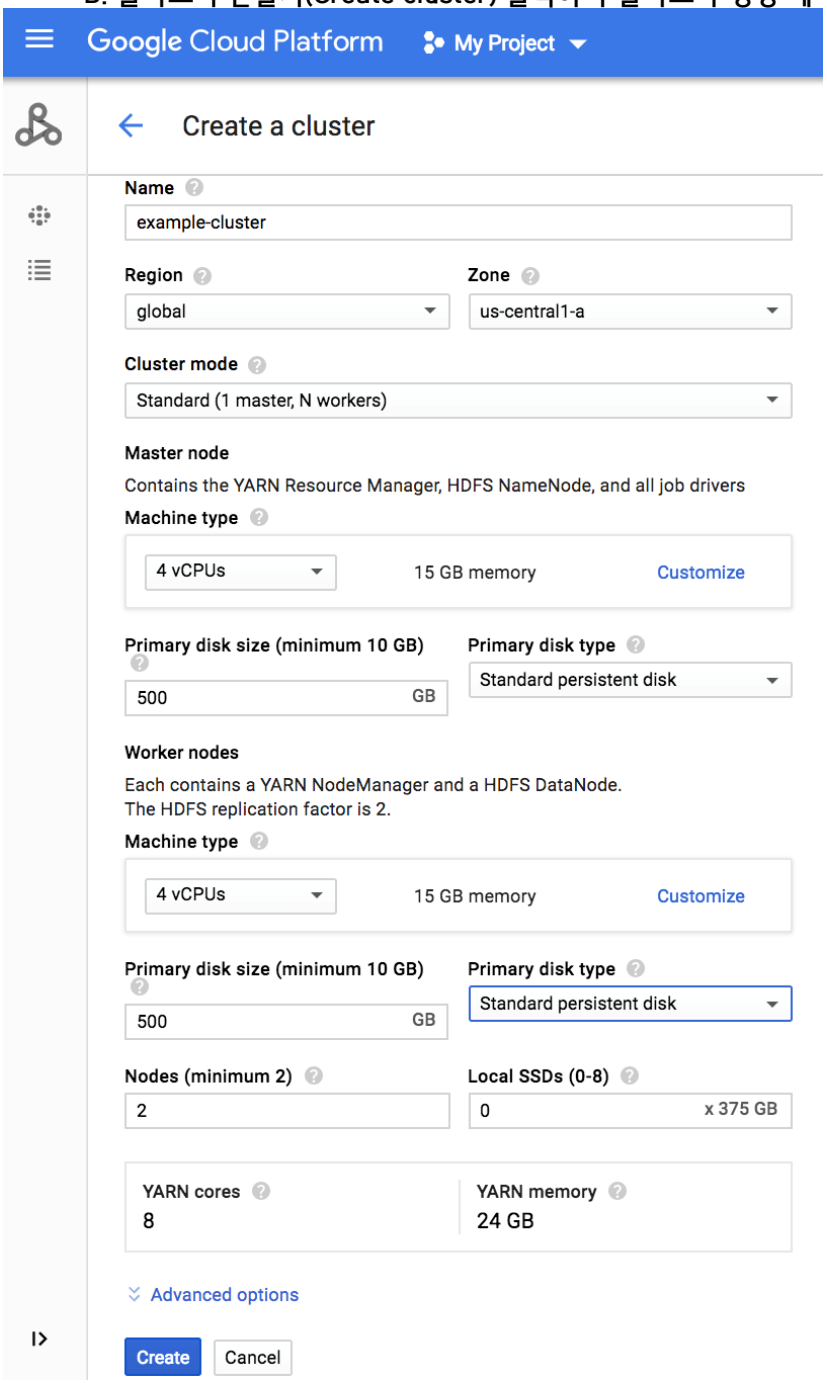
[Datalab에서 Dataproc 사용하기]

노트북: Ideas
만든 날짜: 2018-11-20 오전 10:35 **수정한 날짜:** 2018-11-20 오후 10:33
작성자: supremed14@gmail.com

[Datalab에서 Dataproc 사용하기]

A. Dataproc 콘솔 페이지로 이동
B. 클러스터 만들기(Create cluster) 클릭하여 클러스터 생성 페이지로 이동



C. 클러스터 세부 옵션 설정

1. 이름(Name) 칸에 사용할 클러스터 이름 입력(예: example-cluster)
2. 클러스터를 생성하고 싶은 지역(Region)과 영역(Zone) 설정(예: 지역 - us-east1, 영역 - us-east1-b)
3. 클러스터 모드(Cluster mode)는 표준(Standard: 1 master, N workers)으로 설정
4. 희망하는 사양에 따라 마스터 노드(Master node) 및 작업자 노드(Worker node) 구성

* 마스터/작업자 노드의 사양이 높아질수록, 작업자 노드의 수가 많아질수록 클러스터의 성능이 좋아지나, 사용 시간 대비 과금도 많이 된다고 생각하면 됨.

(예시 1)
- 마스터 노드: 4 vCPUs(26GB memory, n1-highmem-4)
- 작업자 노드: 4 vCPUs(26GB memory, n1-highmem-4), 노드 수(Nodes; minimum 2) - 8-10개 사이로 설정
- 고급 옵션(Advanced options) 버튼 눌러 확장 탭 펼쳐보기
- 초기화 작업 필드(Initialization actions)에 아래의 코드 입력

```
gs://dataproc-initialization-actions/datalab/datalab.sh
```

(예시 2 - 클러스터의 성능을 높이는 경우)
- 마스터 노드: 8 vCPUs(52GB memory, n1-highmem-8)
- 작업자 노드: 8 vCPUs(52GB memory, n1-highmem-8), 노드 수(Nodes; minimum 2) - 8-10개 사이로 설정
- 고급 옵션(Advanced options) 버튼 눌러 확장 탭 펼쳐보기
- 초기화 작업 필드(Initialization actions)에 아래의 코드 입력

```
gs://dataproc-initialization-actions/datalab/datalab.sh
```

* 마스터 노드와 작업자 노드의 사양은 이질적으로 구성 가능함. (i.e. 마스터 노드는 8 vCPUs로, 작업자 노드는 4 vCPUs로 구성 가능)

5. 파란색 만들기(Create) 버튼을 눌러 클러스터 생성
6. 5-10분 정도 기다리면 Dataproc 콘솔 화면 및 Compute Engine 화면에서 클러스터가 정상적으로 생성되었는지 여부 확인 가능함.

D. 브라우저에서 Cloud Datalab 노트북 실행시키기

1. SSH 터널 생성

- Cloud Shell 열기
- Shell 창에 아래 명령어를 실행

[작성 규칙]
gcloud compute ssh 클러스터 이름-m \ (마스터 노드 표시: 생성한 클러스터 이름-m)
--project=프로젝트 ID --zone 영역(zone) -- \ (프로젝트 ID 표시: 열려 있는 클라우드 셸에서 확인 가능), (영역(zone) 표시: 클러스터 페이지에서 확인 가능)
-4 -N -L 8080:클러스터 이름-m:8080 (마스터 노드 다시 한번 입력: 생성한 클러스터 이름-m)

[예시]

```
  gcloud compute ssh example-cluster2-m \
--project=strange-descent-219905 --zone us-east1-b -- \
-4 -N -L 8080:example-cluster2-m:8080
```

2. 생성 후, Cloud Shell의 웹 미리보기 버튼(Preview on port 8080)을 클릭하여 Datalab 페이지 열기