Instructions pour l'utilisation

Google Cloud Platform

TP3 INF8111

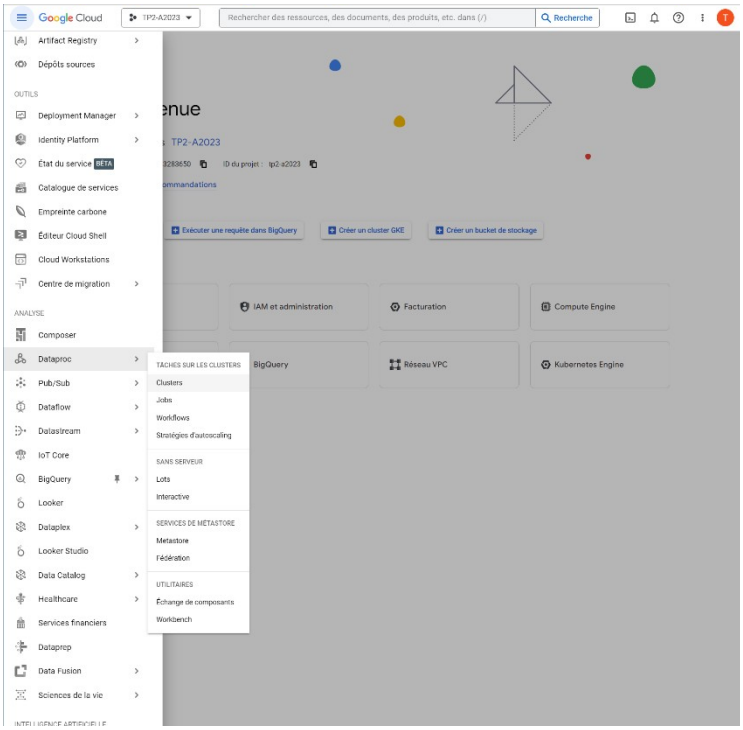
1. Obtention des crédits GCP

Avec l'URL fourni via Moodle, vous devez effectuer une demande pour obtenir des crédits GCP. Votre nom et votre adresse courriel (@polymt.ca) vous seront demandés. Un courriel de confirmation vous sera ensuite envoyé, avec un lien vers le coupon.

N'hésitez pas à me contacter si vous avez la moindres questions ou problèmes lors de la demande.

Une fois cette étape validée, vous pourrez voir le détail de ces crédits via le lien suivant : https://console.cloud.google.com/billing

2. Activation des APIs requises

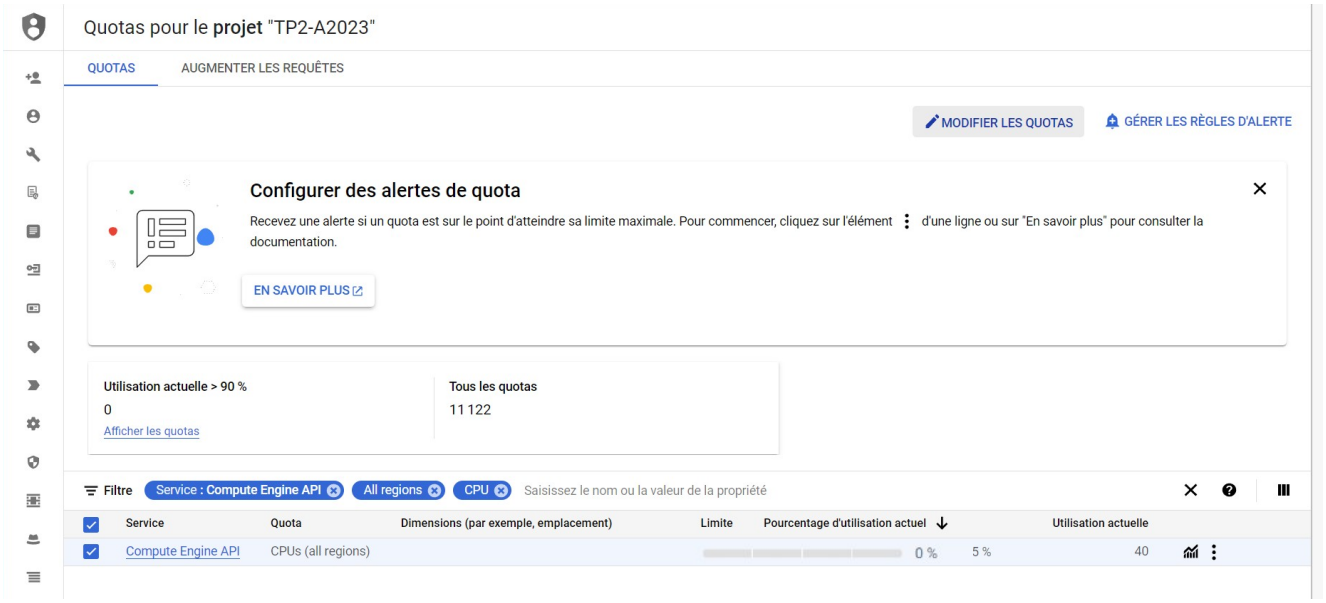


Pour exécuter l'algorithme, vous allez devoir utiliser le service Dataproc. Vous devez donc, préalablement avoir activé l'API. Pour cela, cherchez **Dataproc -> Clusters** (partie **Analyse**) depuis votre compte.

Ensuite, activer l'API.

3. Accroissement du nombre de CPUs disponibles par clusters

Par défaut, il y a 24 CPUs alloué par GCP à chaque étudiant. Néanmoins, pour pouvoir exécuter notre algorithme, nous avons besoin de plus de ressources. Pour cela, cherchez **IAM & administration -> Quotas**. Une fois sur la page, recherchez « **Compute Engine Api CPUs (all regions)** » comme le montre la capture d'écran si dessous.

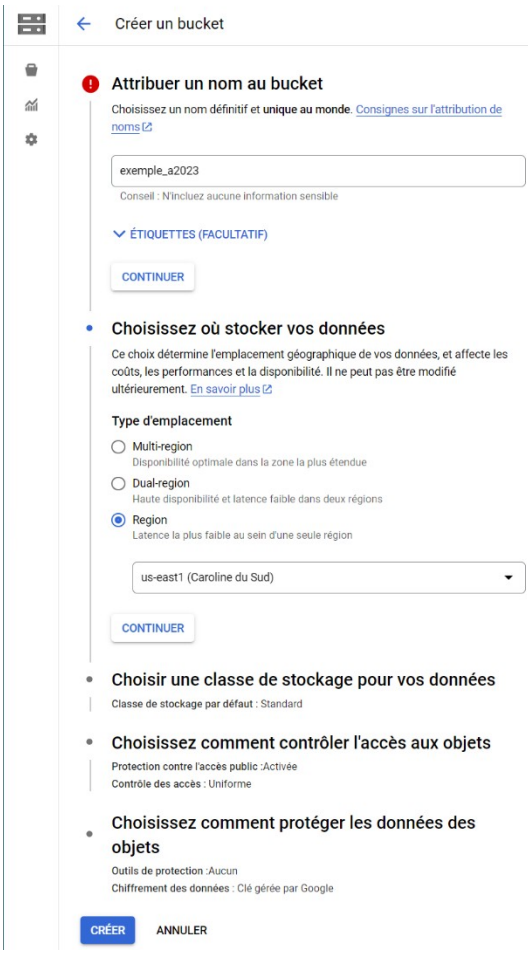


Cochez la ligne puis cliquez sur « Modifier les quotas ». Entrez, par exemple, 128 comme nouvelle limite. Comme la description, écrivez quelle que chose semblable à « *I am working on an academic project that demands a large cluster to run the application* ». Un courriel de confirmation vous sera alors envoyé. Cela peut prendre entre 30min et 48h. Réitérez cette manœuvre pour « **Compute Engine Api CPUs (us-east1)** ».

Notes :

- Il est possible que votre demande soit refusée avec un message du type « *Unfortunately, we are unable to grant you additional quota at this time. If this is a new project please wait 48h until you resubmit the request or until your Billing account has additional history* ». Dans ce cas, augmentez le nombre de CPUs progressivement : demandez 32 puis 64 et enfin 128.
- La configuration présente n'est que recommandée. Il est ainsi possible de demander plus de ressources.

4. Création d'un panier de stockage (i.e. « storage bucket »)



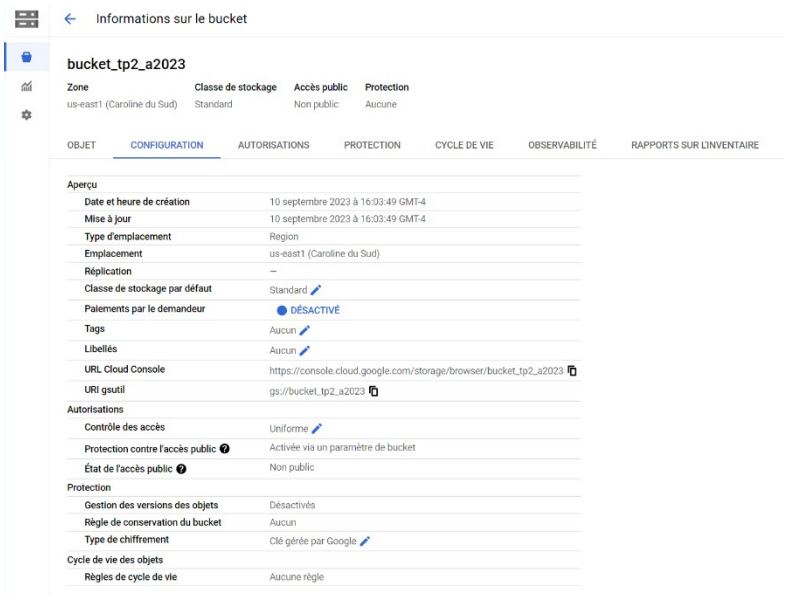
Il est nécessaire de charger les données sur lequel l'algorithme va être exécuté. Vous devez ainsi créer un *bucket storage*. Pour cela, rendez vous dans la section **Cloud Storage** -> **Buckets**.

Créer ensuite un nouveau *bucket*.

Puis, sélectionnez un nom. Dans la section « Choisissez où stocker vos données », cochez « Region » avec « us-east1 (Caroline du Sud) » comme valeur.

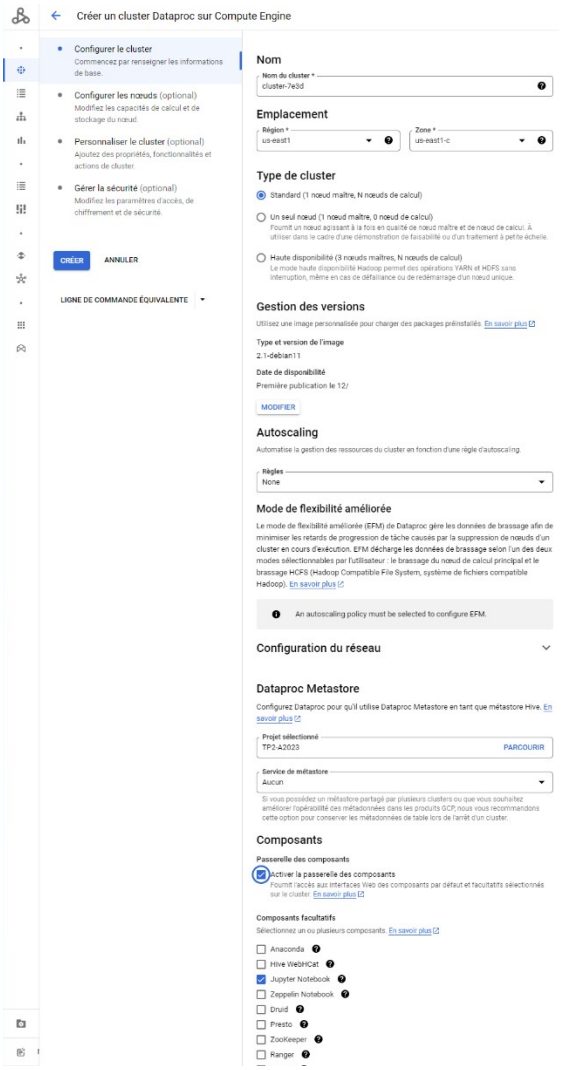
Un fois créé, vous serez redirigé vers une page d'où vous pouvez charger vos fichiers.

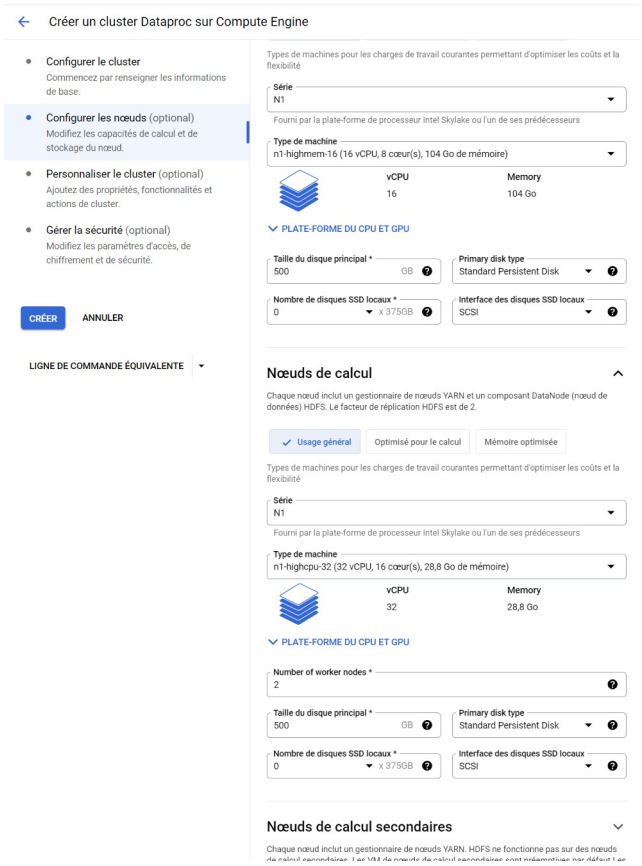
Si vous allez dans l'onglet **Configuration**, le **URI gsutil** vous donnera la racine de ce dépôt. Il faudra en prendre compter lors de l'accès aux données depuis le notebook.



5. Création d'un cluster de calcul

Une fois le bucket créé, vous allez pouvoir créer le cluster. Rendez vous dans la section **Dataproc -> Clusters**. Une configuration possible est la suivante :

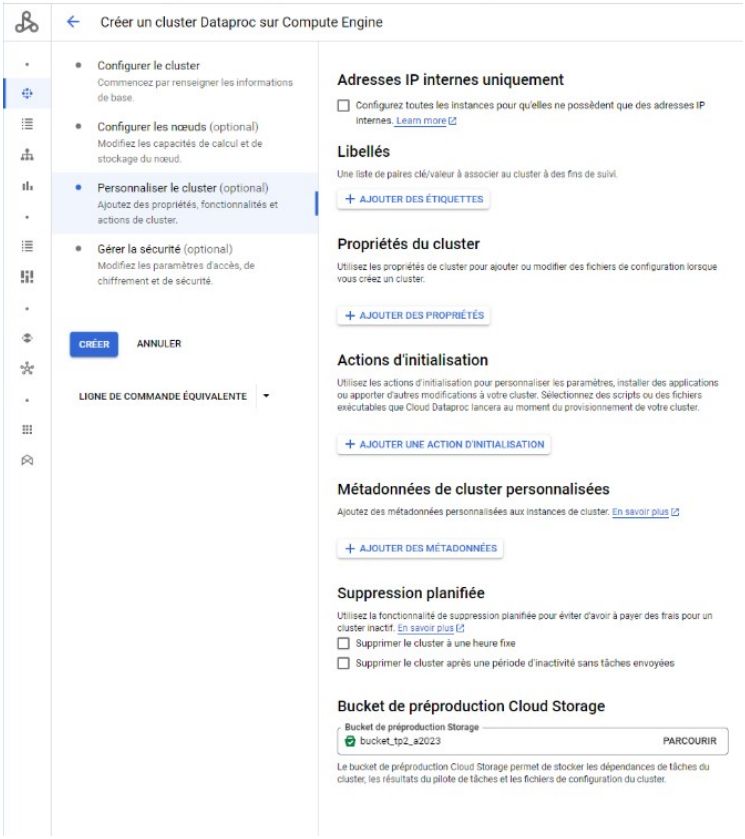




Notes :

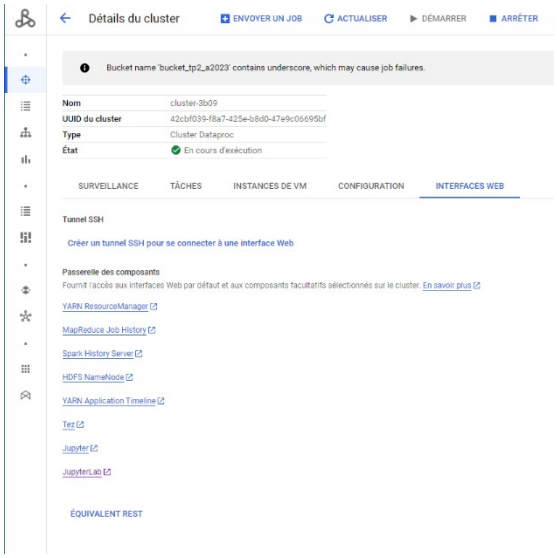
- Vous n'avez pas besoin de changer le nom du cluster
- Il est important de sélectionner « Activer la passerelle des composants » et « Jupyter Notebook »
- Le cluster crée possède 1 master pour 2 workers. Notre algorithme étant gourmand en mémoire, vous utiliserez des machines de type Highmen. **La capture d'écran droite présente la configuration recommandée. Libre à vous d'en tester d'autres.**

Enfin, vous devrez lier ce cluster avec le bucket créé précédemment. Cela s'effectue par le champ **Bucket de préproduction Cloud Storage**.



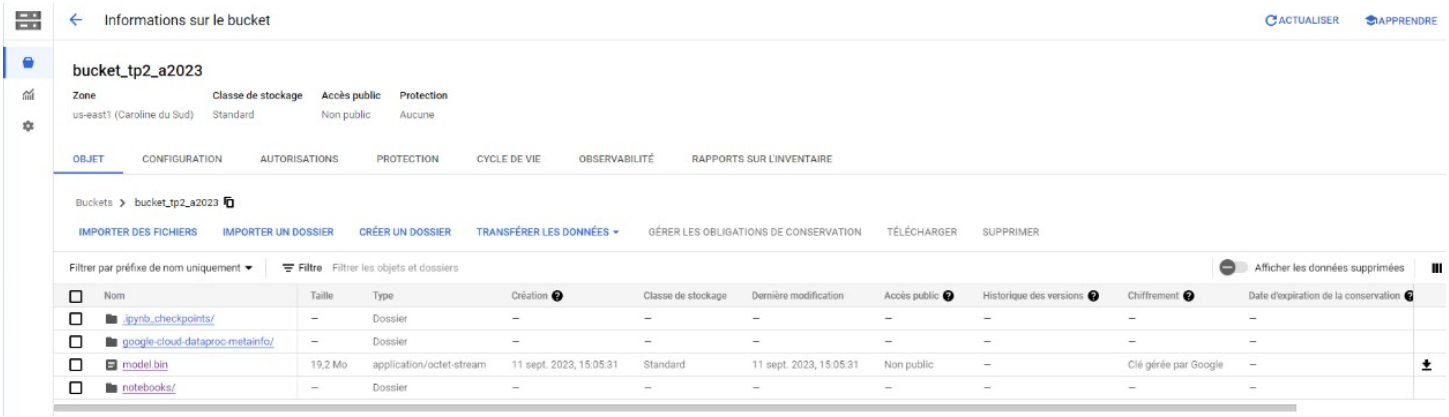
ATTENTION : Lorsque vous aurez fini la configuration du cluster et que vous cliquerez sur **Créer**, GCP commencera à prélever les 50$ de coupons. N'oubliez donc pas de supprimer ce cluster ou de l'arrêter une fois que vous avez fini l'exécution.

6. Utilisation du cluster

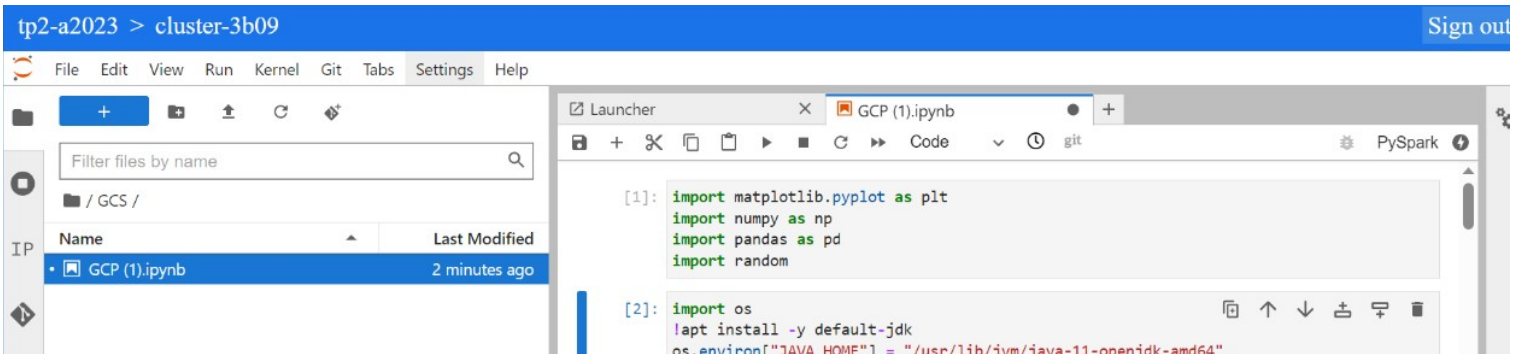


Une fois le cluster créé, cliquez dessus pour l'ouvrir. Rendez vous dans la partie **Interfaces Web** puis cliquez sur le lien **JupyterLab**.

Dans une nouvelle fenêtre, retournez dans le bucket précédemment créé, un dossier **notebook/jupyter** y a été ajouté. Importer y votre notebook.



Vous devriez ainsi voir apparaître sur l'autre page ouverte (JupyterLab) votre notebook. Ouvrez-le puis sélectionnez, en haut à droite, le noyau PySpark (à la place de Python 3).



Vous pouvez désormais exécuter votre code. Une fois que vous avez fini avec ce cluster, retournez dans **Dataproc -> Clusters** puis **Supprimer**.

ATTENTION : N'oubliez donc pas de supprimer ce cluster ou de l'arrêter une fois que vous avez fini l'exécution.