

Instructions pour l'utilisation Google Cloud Platform

TP3 INF8111

1. Obtention des crédits GCP

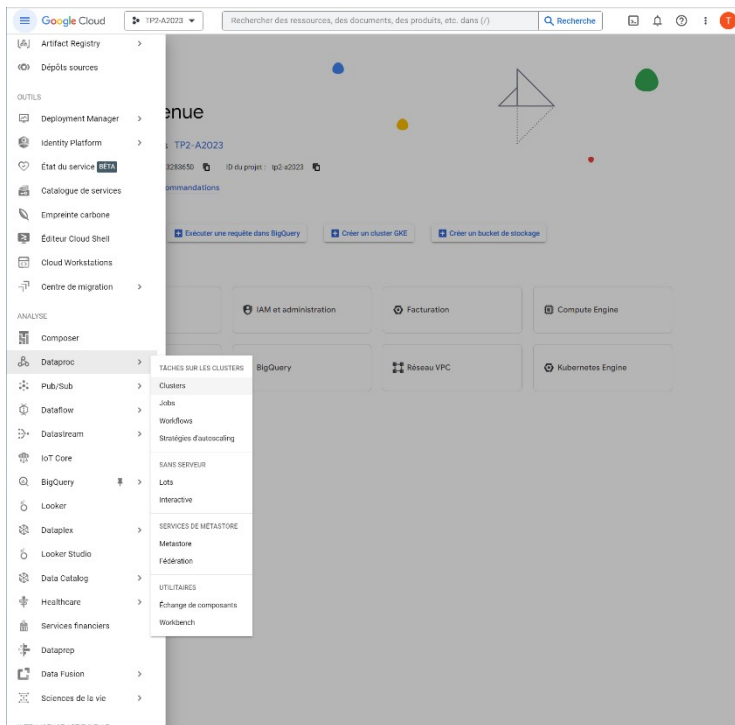
Avec l'URL fourni via Moodle, vous devez effectuer une demande pour obtenir des crédits GCP. Votre nom et votre adresse courriel (@polymt.ca) vous seront demandés. Un courriel de confirmation vous sera ensuite envoyé, avec un lien vers le coupon.

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.

The screenshot shows the 'Quotas' page for project 'TP2-A2023'. At the top, there are tabs for 'QUOTAS' and 'AUGMENTER LES REQUÊTES'. Below the tabs, there are buttons for 'MODIFIER LES QUOTAS' and 'GÉRER LES RÈGLES D'ALERTE'. A modal window titled 'Configurer des alertes de quota' is open, showing instructions on how to set up alerts. Below the modal, there is a summary section showing 'Utilisation actuelle > 90 %' with a value of 0 and 'Tous les quotas' with a value of 11 122. At the bottom, there is a table with filters for 'Service : Compute Engine API', 'All regions', and 'CPU'. The table has columns for 'Service', 'Quota', 'Dimensions (par exemple, emplacement)', 'Limite', 'Pourcentage d'utilisation actuel', and 'Utilisation actuelle'. The table shows one entry for 'Compute Engine API' with a quota of 'CPUs (all regions)' and a limit of 40.

Service	Quota	Dimensions (par exemple, emplacement)	Limite	Pourcentage d'utilisation actuel	Utilisation actuelle
Compute Engine API	CPUs (all regions)		40	0 %	5 %

Cochez la ligne puis cliquez sur « Modifier les quotas ». Entrez, par exemple, 128 comme nouvelle limite. Comme la description, écrivez quelque 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. Répétez 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 »)

← Créer un bucket

1

Attribuer un nom au bucket

Choisissez un nom définitif et unique au monde. [Consignes sur l'attribution de noms](#)

Conseil : N'incluez aucune information sensible

ÉTIQUETTES (FACULTATIF)

CONTINUER

• Choisissez où stocker vos données

Ce choix détermine l'emplacement géographique de vos données, et affecte les coûts, les performances et la disponibilité. Il ne peut pas être modifié ultérieurement. [En savoir plus](#)

Type d'emplacement

☐ Multi-region
Disponibilité optimale dans la zone la plus étendue

☐ Dual-region
Haute disponibilité et latence faible dans deux régions

☒ Region
Latence la plus faible au sein d'une seule région

CONTINUER

• Choisir une classe de stockage pour vos données

Classe de stockage par défaut : Standard

• Choisissez comment contrôler l'accès aux objets

Protection contre l'accès public : Activée

Contrôle des accès : Uniforme

• Choisissez comment protéger les données des objets

Outils de protection : Aucun

Chiffrement des données : Clé gérée par Google

CRÉER ANNULER

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 compte lors de l'accès aux données depuis le notebook.

← Informations sur le bucket

bucket_tp2_a2023

Zone

Classe de stockage

Accès public

Protection

us-east1 (Caroline du Sud)

Standard

Non public

Aucune

OBJET

CONFIGURATION

AUTORISATIONS

PROTECTION

CYCLE DE VIE

OBSERVABILITÉ

RAPPORTS SUR L'INVENTAIRE

Aperçu

Date et heure de création

10 septembre 2023 à 16:03:49 GMT-4

Mise à jour

10 septembre 2023 à 16:03:49 GMT-4

Type d'emplacement

Region

Emplacement

us-east1 (Caroline du Sud)

Réplication

—

Classe de stockage par défaut

Standard

Pailements par le demandeur

DÉSACTIVÉ

Tags

Aucun

Libellés

Aucun

URL Cloud Console

https://console.cloud.google.com/storage/browser/bucket_tp2_a2023

URI gsutil

gs://bucket_tp2_a2023

Autorisations

Contrôle des accès

Uniforme

Protection contre l'accès public

Activée via un paramètre de bucket

État de l'accès public

Non public

Protection

Gestion des versions des objets

Désactivés

Règle de conservation du bucket

Aucun

Type de chiffrement

Clé gérée par Google

Cycle de vie des objets

Règles de cycle de vie

Aucune règle

5. Création d'un cluster de calcul

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

Créer un cluster Dataprocc sur Compute Engine

Configurer le cluster

Configurer les nœuds (optional)

Personnaliser le cluster (optional)

Gérer la sécurité (optional)

CRÉER

ANNULER

LIGNE DE COMMANDE ÉQUIVALENTE

Nom

cluster-test

Emplacement

Région *us-east1

Zone *us-east1-c

Type de cluster

Standard (1 nœud maître, N nœuds de calcul)

Un seul nœud (1 nœud maître, 0 nœuds de calcul)

Haute disponibilité (2 nœuds maîtres, N nœuds de calcul)

Gestion des versions

Type et version de l'image

Date de disponibilité

Autoscaling

Mode de flexibilité améliorée

Configuration du réseau

Dataprocc Metastore

Composants

Créer un cluster Dataprocc sur Compute Engine

Configurer le cluster

Configurer les nœuds (optional)

Personnaliser le cluster (optional)

Gérer la sécurité (optional)

CRÉER

ANNULER

LIGNE DE COMMANDE ÉQUIVALENTE

Types de machines pour les charges de travail courantes permettant d'optimiser les coûts et la flexibilité

SérieN1

Type de machine

n1-highmem-16 (16 vCPU, 8 cœur(s), 104 Go de mémoire)

PLATE-FORME DU CPU ET GPU

Taille du disque principal *500 GB

Primary disk typeStandard Persistent Disk

Nombre de disques SSD locaux *0

x 375GB

Interface des disques SSD locauxSCSI

Nœuds de calcul

Usage généralOptimisé pour le calculMémoire optimisée

Types de machines pour les charges de travail courantes permettant d'optimiser les coûts et la flexibilité

SérieN1

Type de machine

n1-highcpu-32 (32 vCPU, 16 cœur(s), 28,8 Go de mémoire)

PLATE-FORME DU CPU ET GPU

Number of worker nodes *2

Taille du disque principal *500 GB

Primary disk typeStandard Persistent Disk

Nombre de disques SSD locaux *0

x 375GB

Interface des disques SSD locauxSCSI

Nœuds de calcul secondaires

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**.

← Créer un cluster Dataproc sur Compute Engine

- Configurer le cluster
Commencez par renseigner les informations de base.
- Configurer les nœuds (optional)
Modifiez les capacités de calcul et de stockage du nœud.
- **Personnaliser le cluster (optional)**
Ajoutez des propriétés, fonctionnalités et actions de cluster.
- Gérer la sécurité (optional)
Modifiez les paramètres d'accès, de chiffrement et de sécurité.

CRÉER **ANNULER**

LIGNE DE COMMANDE ÉQUIVALENTE ▾

Adresses IP internes uniquement

☐ Configurez toutes les instances pour qu'elles ne possèdent que des adresses IP internes. [Learn more](#)

Libellés

Une liste de paires clé/valeur à associer au cluster à des fins de suivi.

+ AJOUTER DES ÉTIQUETTES

Propriétés du cluster

Utilisez les propriétés de cluster pour ajouter ou modifier des fichiers de configuration lorsque vous créez un cluster.

+ AJOUTER DES PROPRIÉTÉS

Actions d'initialisation

Utilisez les actions d'initialisation pour personnaliser les paramètres, installer des applications ou apporter d'autres modifications à votre cluster. Sélectionnez des scripts ou des fichiers exécutables que Cloud Dataproc lancera au moment du provisionnement de votre cluster.

+ AJOUTER UNE ACTION D'INITIALISATION

Métadonnées de cluster personnalisées

Ajoutez des métadonnées personnalisées aux instances de cluster. [En savoir plus](#)

+ AJOUTER DES MÉTADONNÉES

Suppression planifiée

Utilisez la fonctionnalité de suppression planifiée pour éviter d'avoir à payer des frais pour un cluster inactif. [En savoir plus](#)

☐ Supprimer le cluster à une heure fixe

☐ Supprimer le cluster après une période d'inactivité sans tâches envoyées

Bucket de préproduction Cloud Storage

Bucket de préproduction Storage

bucket_tp2_#2023 **PARCOURIR**

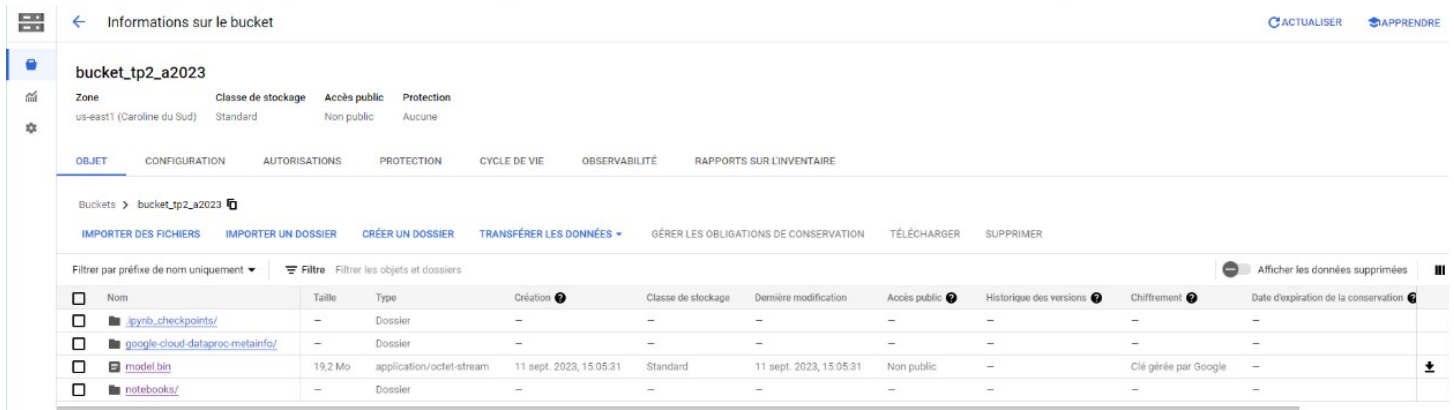
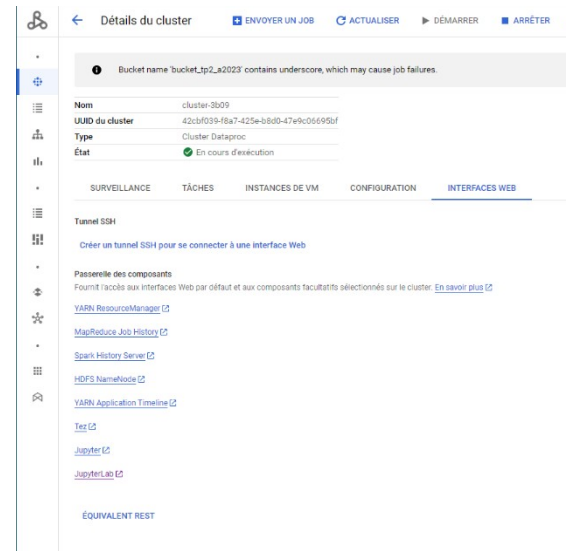
Le bucket de préproduction Cloud Storage permet de stocker les dépendances de tâches du cluster, les résultats du pilote de tâches et les fichiers de configuration du cluster.

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.

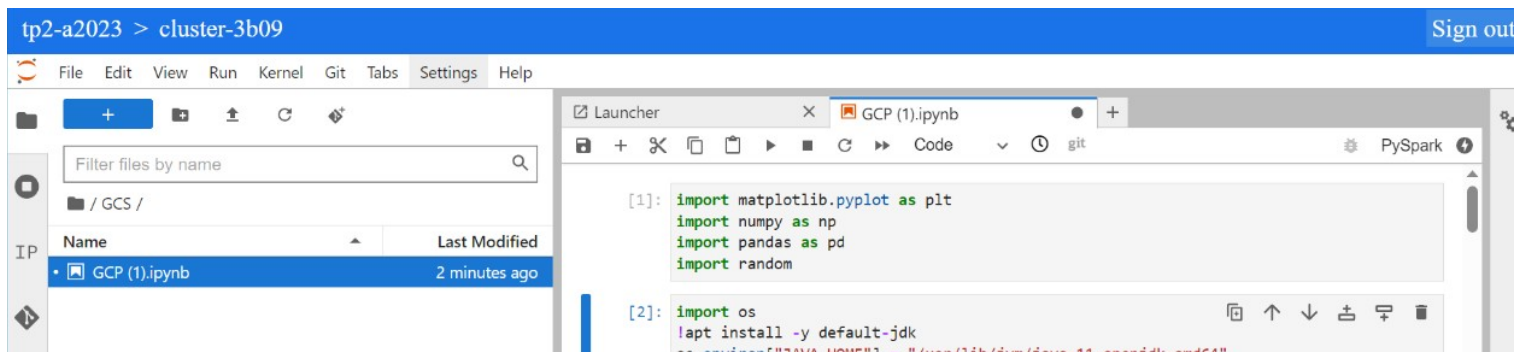
6. Utilisation du cluster

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.