

Date: 02/12/2025

Critères sur lesquels il faut se baser pour choisir nos VMs pour le cluster Spark:

1) Le nombre de vCPU

Le nombre de vCPU influence la capacité de calcul.

Plus il y a de vCPU, plus la machine peut exécuter de tâches en parallèle.

2) La quantité de RAM

Une RAM insuffisante peut provoquer des ralentissements ou des erreurs

Choix finale pour notre cluster

1. VM du Master **e2-standard-2** (2 vCPU , 8 Go de RAM)

2. VM des Workers: **e2-standard-2** et **e2-standard-4**

3. VM de l'Edge Node (point d'entrée du cluster, machine sur laquelle lancer les jobs.) **e2-medium** (série E2 shared-core) ou **e2-standard-2**

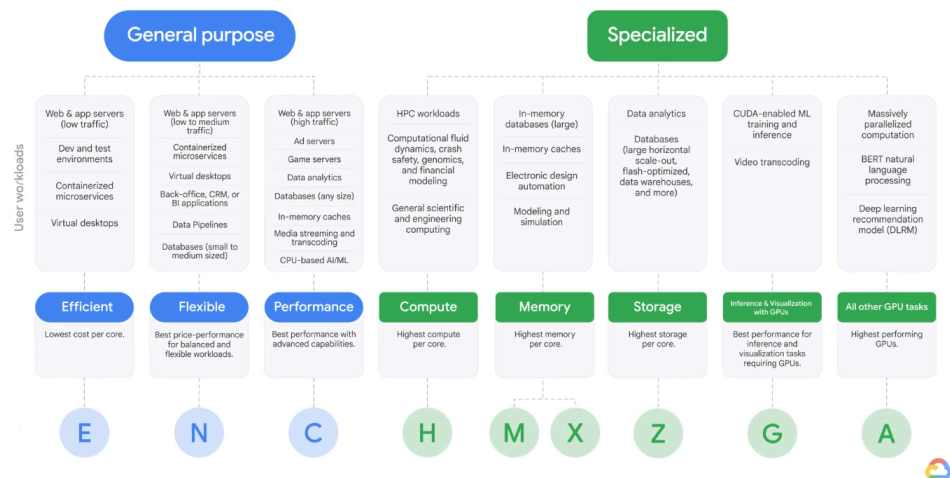
4. Stockage : bucket : 1 bucket Cloud Storage standard dans la même région que les VMs pour le stockage des données d'entrée et des résultats.

Images des VMs:

Soit Ubuntu 22.04 soit Debian12

Debian 12 est bien compatible avec spark , ansible.

Les différentes VM de GCP:



Sur Google Cloud, les machines virtuelles sont regroupées en deux grandes catégories : General Purpose et Specialized. Chaque catégorie correspond à des usages différents selon les besoins en performance, mémoire, stockage ou GPU.

1. General Purpose (usage général)

Ces machines offrent un équilibre entre coût, performance et flexibilité. Elles sont adaptées à la majorité des applications courantes.

E-Series (E) Efficient

- Conçues pour offrir le coût le plus bas par cœur CPU.
- Utilisées pour : serveurs web à faible trafic, environnements de développement et test, microservices simples, bureaux virtuels.
- Idéales lorsque les besoins en performance sont modérés et que le coût est prioritaire.

E2:

1. E2 Standard
2. E2 High-Memory
3. E2 High-CPU
4. E2 Shared-core (e2-micro / e2-small / e2-medium)

5. E2 Custom (tu choisis toi-même vCPU + RAM)

1) E2 Standard:

- **e2-standard-2** → 2 vCPU, 8 Go RAM
- **e2-standard-4** → 4 vCPU, 16 Go RAM
- **e2-standard-8** → 8 vCPU, 32 Go RAM

Les E2 Standard offrent un équilibre classique entre CPU et mémoire et conviennent à la majorité des usages.

2) Les E2 High-Memory: disposent de plus de mémoire que les Standard et sont utiles lorsque les applications ont besoin de beaucoup de RAM.

3) Les E2 High-CPU offrent davantage de puissance de calcul mais avec peu de mémoire, ce qui est intéressant pour des tâches très orientées calcul.

4) Les E2 Shared-core sont des machines très économiques avec un cœur processeur partagé, adaptées aux petits environnements de test ou aux services légers.

5) les E2 Custom permettent de choisir soi-même la quantité de CPU et de RAM pour répondre à des besoins particuliers, même si cela revient un peu plus cher.

N-Series (N) Flexible

- Machines équilibrées en termes de puissance CPU et mémoire.
- Utilisées pour : serveurs web avec trafic moyen, microservices conteneurisés, pipelines de données, bases de données de petite à moyenne taille, applications métier.
- Bon compromis performance / prix pour des environnements variés.

C-Series (C) Performance

- Machines optimisées pour les performances CPU.
- Utilisées pour : serveurs web à fort trafic, jeux vidéo en ligne, analytique de données, mise en cache en mémoire, streaming média, IA/ML basés sur CPU.
- Adaptées aux applications exigeantes en calcul.

2. Specialized (usage spécialisé)

Ces machines sont conçues pour des besoins spécifiques et intensifs : calcul scientifique, stockage rapide, mémoire très importante ou processing GPU.

H-Series (H) Compute

- Machines optimisées pour un niveau maximal de calcul CPU.
- Utilisées pour : simulation scientifique, recherche, modélisation, traitements HPC (High-Performance Computing).
- Très adaptées aux opérations mathématiques et scientifiques lourdes.

M-Series (M) Memory

- Machines avec une quantité de mémoire très élevée.
- Utilisées pour : bases de données en mémoire, caches volumineux, logiciels d'ingénierie électronique, modélisation et simulation nécessitant beaucoup de RAM.
- Choix adapté lorsque la mémoire est la contrainte principale.

Z-Series (Z) Storage

- Machines ayant un stockage local très rapide et une capacité élevée.
- Utilisées pour : data warehouses, bases de données avec haut débit d'écriture, analytique big data et stockage flash optimisé.
- Performantes pour les traitements intensifs en entrée/sortie (I/O).

G-Series (G) GPU pour visualisation et inférence

- Machines équipées de GPU pour la visualisation graphique ou l'inférence IA.
- Utilisées pour : traitement vidéo, rendu 3D, inférence de modèles IA existants.
- Destinées aux applications graphiques et IA en production.

A-Series (A) GPU haute performance

- Machines avec des GPU très puissants.
- Utilisées pour : entraînement de modèles de deep learning, NLP, calcul massivement parallèle, recommandations.
- Choix recommandé pour les projets d'IA intensifs et le machine learning avancé.