

Formation GCP

Ihab ABADI / UTOPIOS

SOMMAIRE

1. Introduction GCP.
2. Google compute Engine.
3. Google cloud Storage.
4. Google cloud Sql.
5. Google App Engine.
6. Google container registry.
7. Google Kubernetes Engine.
8. Infrastructure en Code.
9. DevOps avec gcp.

SOMMAIRE – Partie 1 – Introduction GCP

1. Présentation de Google Cloud Platform
2. Qu'est-ce que Google Cloud Platform ?
3. Éléments de Google Cloud Platform.
4. Services Google Cloud Platform.
5. Avantages de Google Cloud Platform.

Introduction GCP

- Google Cloud Platform (GCP), proposé par Google, est une suite de services de cloud computing qui s'exécute sur la même infrastructure que Google utilise en interne pour ses produits d'utilisateur final, tels que la recherche Google, Gmail, le stockage de fichiers et YouTube.
- Google est l'un des principaux développeurs de logiciels et de technologies au monde. Chaque année, Google propose différentes innovations et avancées dans le domaine technologique, ce qui est brillant et aide les gens du monde entier.
- Au cours des dernières années, Google Cloud Platform a connu une augmentation de son utilisation, de personnes adoptent le Cloud. Pour répondre à une forte demande, un certain nombre de services cloud Google ont été lancés pour les clients mondiaux.

C'est quoi GCP

- La plate-forme cloud de Google permet aux utilisateurs d'accéder facilement aux systèmes cloud et à d'autres services informatiques développés par Google.
- La plate-forme comprend une large gamme de services pouvant être utilisés dans différents secteurs du cloud computing, tels que le stockage et le développement d'applications.
- N'importe qui peut accéder à la plateforme cloud de Google et l'utiliser selon ses besoins
- La plate-forme Google Cloud, créée pour la première fois le 6 octobre 2011, détient une part de marché de 13%.
- Outre les différents outils de gestion disponibles sur Google Cloud Platform, la société a également inclus de nombreuses fonctionnalités et fonctionnalités cloud telles que le stockage dans le cloud, l'analyse de données, les options de développement et l'apprentissage automatique avancé.

Eléments de GCP

- La plate-forme cloud de Google est composée d'un ensemble différent d'éléments qui sont utiles aux utilisateurs de plusieurs manières.
- Google Compute Engine : Ce moteur de calcul a été introduit avec le service IaaS de Google qui fournit des VM similaires à amazon EC2.
- Google Cloud App Engine : Il s'agit d'une plate-forme très puissante et importante qui aide à développer des applications Web mobiles et différentes.
- Google Cloud Container Engine : cet élément particulier est utile car il permet à l'utilisateur d'exécuter les conteneurs Docker présents sur Google Cloud Platform.
- Google Cloud Storage : La capacité de stocker des données et des ressources importantes sur la plate-forme.

Eléments de GCP

- Service Google Big Query : Le service Google Big Query est un service d'analyse de données efficace qui permet aux utilisateurs d'analyser leur entreprise pour le Big Data.
- Google Cloud Dataflow : le flux de données cloud permet aux utilisateurs de gérer des pipelines de traitement de données parallèles cohérents.

Eléments de GCP

- Google Cloud Test Lab : Ce service fourni par Google permet aux utilisateurs de tester leurs applications à l'aide d'appareils physiques et virtuels présents dans le cloud.
- Google Cloud Endpoints : cette fonctionnalité particulière aide les utilisateurs à développer et à maintenir une interface de programme d'application sécurisée s'exécutant sur Google Cloud Platform.
- Google Cloud Machine Learning Engine : comme son nom l'indique, cet élément présent dans Google Cloud aide les utilisateurs à développer des modèles et des structures qui permettent aux utilisateurs de se concentrer sur les capacités et le cadre d'apprentissage automatique.

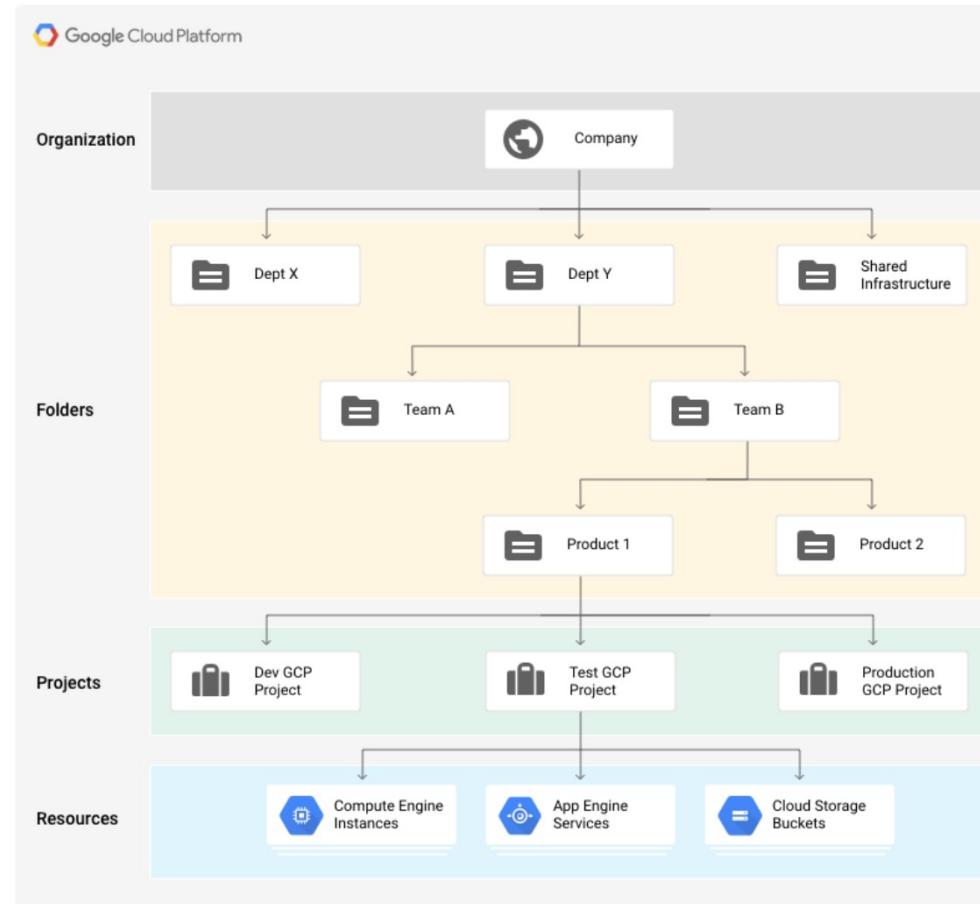
Eléments de GCP - Services

Category	Services
Compute	1. Compute Engine 2. app Engine 3. Cloud Functions
Storage	1. Cloud Storage 2. Cloud Memory store 3. Cloud Fire store 4. Cloud Storage for Firebase 5. Cloud File store
Databases	1. Cloud SQL 2. Cloud Big Table 3. Cloud Data store
Migration	1. Data Transfer 2. Transfer appliance 3. Cloud Storage Transfer Service 4. Big Query Data Transfer Service
Networking	1. virtual Private Cloud (VPC) 2. Cloud Load Balancing 3. Cloud Interconnect 4. Cloud DNS 5. Network Service Tiers

Eléments GCP - Ressources

- Il existe quatre types de ressources pouvant être gérées via Resource Manager :
- La ressource de l'organisation. Il s'agit du nœud racine dans la hiérarchie des ressources. Il représente une organisation, par exemple une entreprise.
- La ressource projets. Par exemple, pour séparer les projets des environnements de production et de développement. Ils sont nécessaires pour créer des ressources.
- La ressource de dossier. Ils fournissent un niveau supplémentaire d'isolement du projet. Par exemple, créer un dossier pour chaque service d'une entreprise.
- Ressources. Machines virtuelles, instances de base de données, équilibriseurs de charge, etc.

Eléments GCP - Ressources



Avantages de GCP

- Google Cloud Platform est un excellent moyen pour ceux qui souhaitent accéder aux meilleures fonctionnalités et services cloud possibles.
- Vous pouvez travailler de n'importe où
- Avec l'aide des serveurs cloud de Google, vous pouvez accéder à vos données et informations où que vous soyez. Il vous suffit de vous connecter à votre compte et de commencer à travailler.
- Meilleur prix et offres
- C'est un très gros avantage pour les utilisateurs car ils ne sont tenus de payer l'argent que pour le temps qu'ils ont utilisé la plateforme. Si vous utilisez le service pendant une longue période, ils offrent même des réductions supplémentaires. En comparant la plate-forme Google Cloud avec d'autres concurrents, il a été découvert que Google est comparativement beaucoup moins cher.
- Méthodes de sécurité
- Google est dans l'industrie depuis près de 15 ans. Par conséquent, le niveau de sécurité qu'ils offrent est exemplaire. Leurs serveurs, plate-forme cloud et autres réseaux sont cryptés et sécurisés avec des mesures de sécurité de pointe qui aident les clients à protéger leurs données et autres composants importants. Google fait venir des experts en sécurité du monde entier pour assurer la sécurité de leur système. Leurs systèmes de sécurité sont en ligne 24h/24 et 7j/7.

SDK GCP

- Le SDK Cloud est un ensemble d'outils permettant d'interagir avec Google Cloud Platform.
- Il comprend des outils de ligne de commande bq, kubectl, gcloud et gsutil qui peuvent interagir avec divers services GCP à l'aide de la CLI ou de scripts d'automatisation. Par exemple, il peut :
 - Créer/gérer un bucket Google Cloud Storage (GCS).
 - Créer/gérer une instance Google Compute Engine (GCE).
 - Créer/gérer Google Datalab.
 - Créez un ensemble de données BigQuery.
 - Envoyez une tâche à BigQuery.
 - Créer/gérer des règles de pare-feu.

SDK GCP – Installation

- Pour commencer à travailler avec Cloud SDK, vous devez installer des outils en fonction de la plate-forme de votre système d'exploitation.
- Python version 2.7.x ou moins
- Pour vérifier que l'installation a été correctement effectuée, on lance, dans un terminal, la commande gcloud.
- La commande gcloud init permet de se connecter à notre compte gcp.

SDK GCP – Installation

- gcloud nous permet d'accéder à des composants additionnels.
- L'installation d'un composant se fait à l'aide de la commande

```
gcloud components install kubectl
```

SDK GCP - CLI

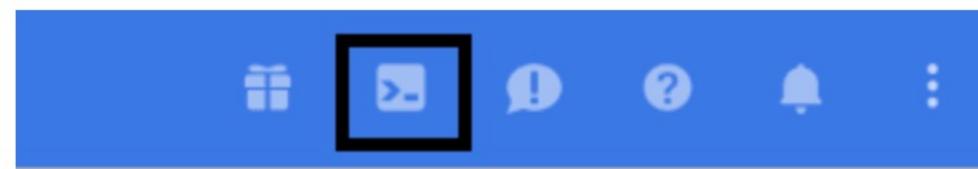
- Le SDK Cloud est fourni avec un ensemble d'outils de commande
- gcloud : fonctionne avec Google Compute Engine (équivalent d'AWS EC2)
- bq : fonctionne avec Cloud Bigquery (peut être comparé à Amazon Redshift)
- gsutil : fonctionne avec Cloud Storage (équivalent d'AWS S3/AWS EBS)
- ...

SDK GCP – API

- Google Cloud nous permet d'utiliser GCP à l'aide d'API REST.
- Google Cloud est fourni avec 7 bibliothèques d'API client.
- Les API Client permettent de créer différents service GCP.
- Exemple d'utilisation avec JAVA par exemple.

SDK GCP - Console

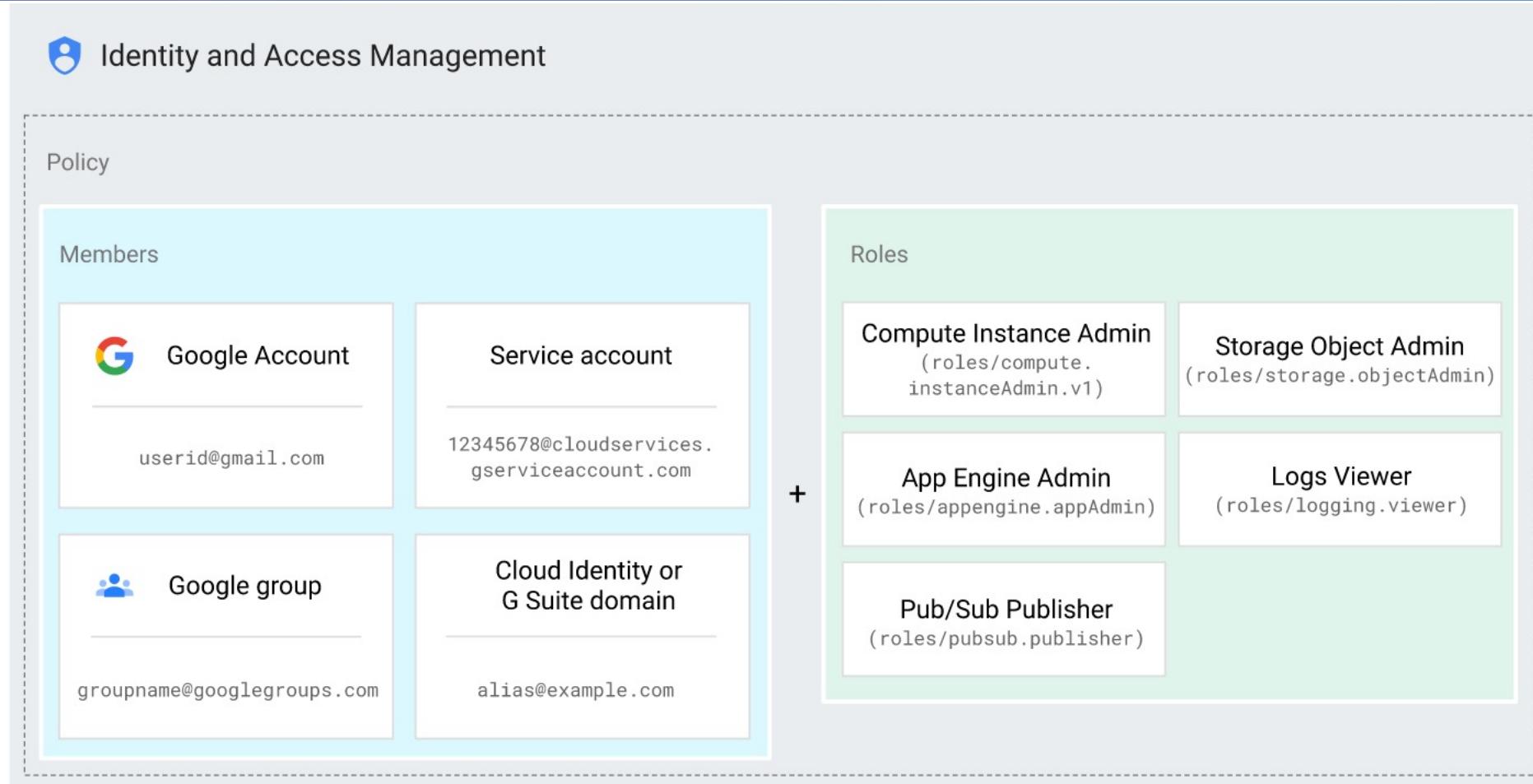
- Si nous ne souhaitons pas d'installer d'outils sur notre machine, nous pouvons utiliser directement la console.



Cloud IAM

- En termes simples, Cloud IAM contrôle qui peut faire quoi sur quelle ressource.
- Une ressource peut être une machine virtuelle, une instance de base de données, un utilisateur, etc.
- Il est important de noter que les autorisations ne sont pas directement attribuées aux utilisateurs. Au lieu de cela, ils sont regroupés dans des rôles, qui sont attribués aux membres.
- Une stratégie est une collection d'une ou plusieurs liaisons d'un ensemble de membres à un rôle.

Cloud IAM



Cloud IAM - Identités

- Dans un projet GCP, les identités sont représentées par des comptes Google, créés en dehors de GCP et définis par une adresse e-mail (pas nécessairement @gmail.com). Il existe différents types:
 - Comptes Google*. Représenter les gens : ingénieurs, administrateurs, etc.
 - Comptes de services. Utilisé pour identifier les utilisateurs non humains : applications, services, machines virtuelles et autres. Le processus d'authentification est défini par des clés de compte, qui peuvent être gérées par Google ou par les utilisateurs (uniquement pour les comptes de service créés par les utilisateurs).
- Les groupes Google sont une collection de comptes Google et de services.
 - G Suite Domain* est le type de compte que vous pouvez utiliser pour identifier les organisations. Si votre organisation utilise déjà Active Directory, elle peut être synchronisée avec Cloud IAM à l'aide de Cloud Identity.
 - tous les utilisateurs authentifiés. Pour représenter tout utilisateur authentifié dans GCP.
 - tous les utilisateurs. Pour représenter n'importe qui, authentifié ou non.

Cloud IAM - Rôles

Un rôle est un ensemble d'autorisations. Il existe trois types de rôles :

Primitif. Rôles GCP d'origine qui s'appliquent à l'ensemble du projet.
Il existe trois rôles concentriques : **Viewer**, **Editor**, and **Owner**.

Prédéfini. Fournit l'accès à des services spécifiques, par exemple,
storage.admin.

Custom. vous permet de créer vos propres rôles, en combinant les
autorisations spécifiques dont vous avez besoin.

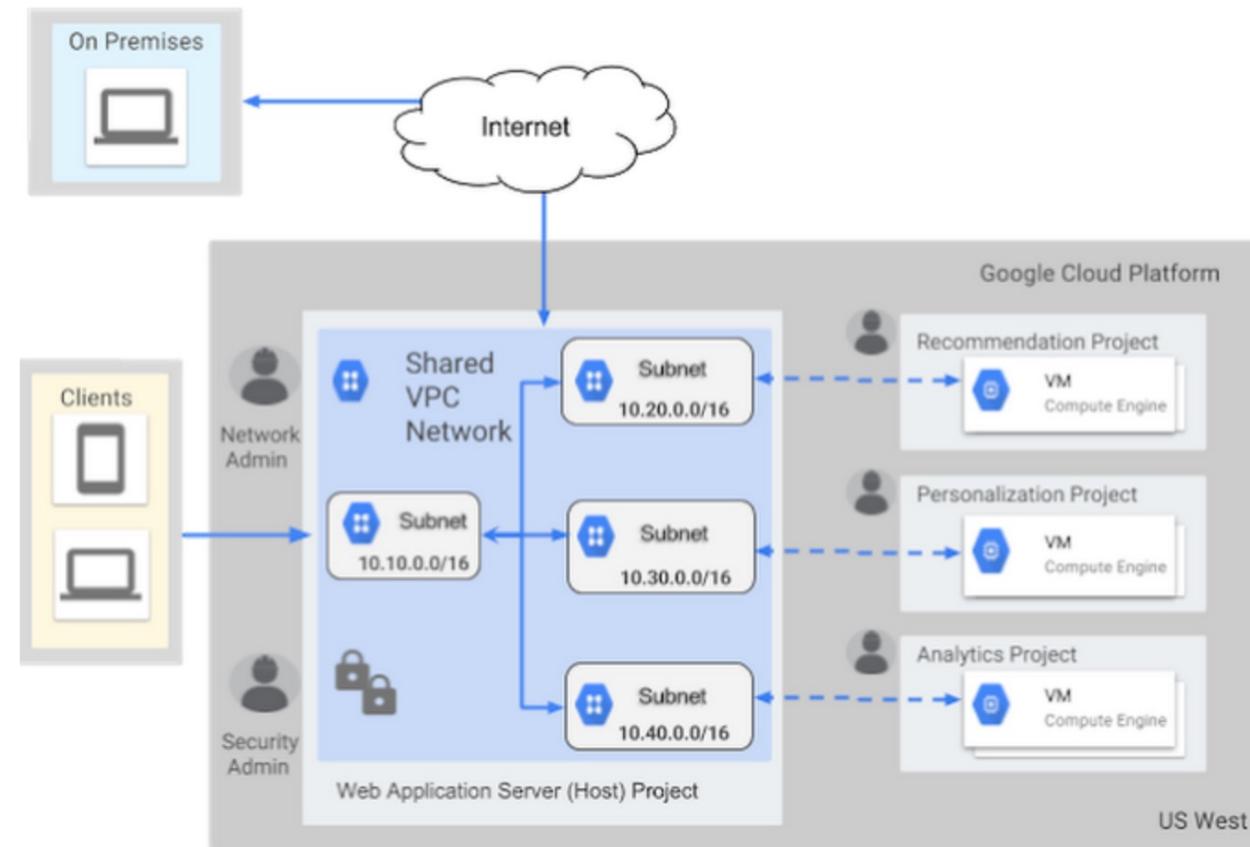
Réseau sur GCP - VPC

- Vous pouvez utiliser la même infrastructure réseau que celle utilisée par Google pour exécuter ses services : YouTube, Recherche, Maps, Gmail, Drive, etc.
- L'infrastructure de Google est divisée en :
 - Régions : zones géographiques indépendantes, distantes d'au moins 160 kilomètres, où Google héberge des centres de données. Il se compose de 3 zones ou plus. Par exemple, us-central1.
 - Zones : plusieurs centres de données individuels dans une région. Par exemple, us-central1-a.
 - Points de présence périphériques : points de connexion entre le réseau de Google et le reste d'Internet.
- L'infrastructure GCP est conçue de manière à ce que tout le trafic entre les régions passe par un réseau privé mondial, ce qui améliore la sécurité et les performances.

Réseau sur GCP - VPC

- En plus de cette infrastructure, vous pouvez créer des réseaux pour vos ressources, les VPCs:
 - Sous-réseaux. Partitions logiques d'un réseau définies à l'aide de la notation CIDR. Ils appartiennent à une seule région mais peuvent s'étendre sur plusieurs zones. Si vous disposez de plusieurs sous-réseaux (y compris vos réseaux sur site s'ils sont connectés à GCP), assurez-vous que les plages CIDR ne se chevauchent pas.
 - Adresses IP. Peut être interne (pour une communication privée au sein de GCP) ou externe (pour communiquer avec le reste d'Internet). Pour les adresses IP externes, vous pouvez utiliser une IP éphémère ou payer pour une IP statique..
 - Règles de pare-feu, pour autoriser ou refuser le trafic vers vos machines virtuelles, à la fois entrant (ingress) et sortant (egress). Par défaut, tout le trafic entrant est refusé et tout le trafic sortant est autorisé. Les règles de pare-feu sont définies au niveau du VPC, mais elles s'appliquent à des instances individuelles ou à des groupes d'instances à l'aide de balises réseau ou de plages d'adresses IP.

Réseau sur GCP - VPC



Compute Engine (GCE)

- Le moteur de calcul vous permet de faire tourner des machines virtuelles dans GCP.
- GCP fournit différentes familles de machines avec des quantités prédéfinies de RAM et de processeurs :
 - Usage général. Offre le meilleur rapport qualité-prix pour une variété de charges de travail.
 - Mémoire optimisée. Idéal pour les charges de travail gourmandes en mémoire. Ils offrent plus de mémoire par cœur que les autres types de machines.
 - Optimisé pour le calcul. Ils offrent les performances les plus élevées par cœur et sont optimisés pour les charges de travail intensives en calcul
 - Noyau partagé. Ces types de machines partagent un cœur physique en temps partagé. Cela peut être une méthode rentable pour exécuter de petites applications

Compute Engine (GCE) - Disques

- Les disques persistants fournissent un stockage de blocs durable et fiable. Ils ne sont pas locaux à la machine. Au contraire, ils sont connectés en réseau, ce qui a ses avantages et ses inconvénients :
- Les disques peuvent être redimensionnés, attachés ou détachés d'une VM même si l'instance est en cours d'utilisation.
- Ils ont une grande fiabilité.
- Les disques peuvent survivre à l'instance après sa suppression.
- Si vous avez besoin de plus d'espace, connectez simplement plus de disques.
- Des disques plus grands offriront de meilleures performances.
- Étant attachés au réseau, ils sont moins performants que les options locales. Des disques persistants SSD sont également disponibles pour les charges de travail plus exigeantes.

Compute Engine (GCE) – SSD Local

- Les SSD locaux sont attachés à une VM à laquelle ils fournissent un stockage éphémère hautes performances.
- Ces données seront perdues si la VM est supprimée.
- Les SSD locaux ne peuvent être attachés à une machine qu'au moment de sa création, mais vous pouvez attacher à la fois des SSD locaux et des disques persistants à la même machine.

Compute Engine (GCE) – Sauvegarde

- Les snapshots sont des sauvegardes de vos disques. Pour réduire l'espace, ils sont créés progressivement :
- La sauvegarde 1 contient tout le contenu de votre disque
- La sauvegarde 2 ne contient que les données qui ont changé depuis la sauvegarde 1
- La sauvegarde 3 ne contient que les données qui ont changé depuis la sauvegarde 2, et ainsi de suite

Compute Engine (GCE) – Image

- Les images font référence aux images du système d'exploitation nécessaires pour créer des disques de démarrage pour vos instances. Il existe deux types d'images :
- Images publiques. Ils sont fournis et gérés par Google, des communautés open source et des fournisseurs tiers. Prêt à être utilisé dès que vous créez votre projet. Accessible à tous
- Images personnalisées. Images que vous avez créées.
- Ils sont liés au projet dans lequel vous les avez créés mais vous pouvez les partager avec d'autres projets.
- Vous pouvez créer des images à partir de disques persistants et d'autres images, à la fois du même projet ou partagées à partir d'un autre projet.
- Les images associées peuvent être regroupées en familles d'images pour simplifier la gestion des différentes versions d'images.
- Pour les images basées sur Linux, vous pouvez également les partager en les exportant vers Cloud Storage sous forme de fichier tar.gz.

Compute Engine (GCE) – Groupe d'instance

- Les groupes d'instances vous permettent de traiter un groupe d'instances comme une seule unité et ils se présentent sous deux formes :
- Groupe d'instances non géré. Formé par un groupe hétérogène d'instances nécessitant des paramètres de configuration individuels.
- Groupe d'instances géré (MIG). C'est l'option préférée lorsque cela est possible. Toutes les machines se ressemblent, ce qui permet de les configurer facilement, de les créer dans plusieurs zones (haute disponibilité), de les remplacer si elles deviennent défectueuses (réparation automatique), d'équilibrer le trafic entre elles et de créer de nouvelles instances si le trafic augmente (échelle horizontale).

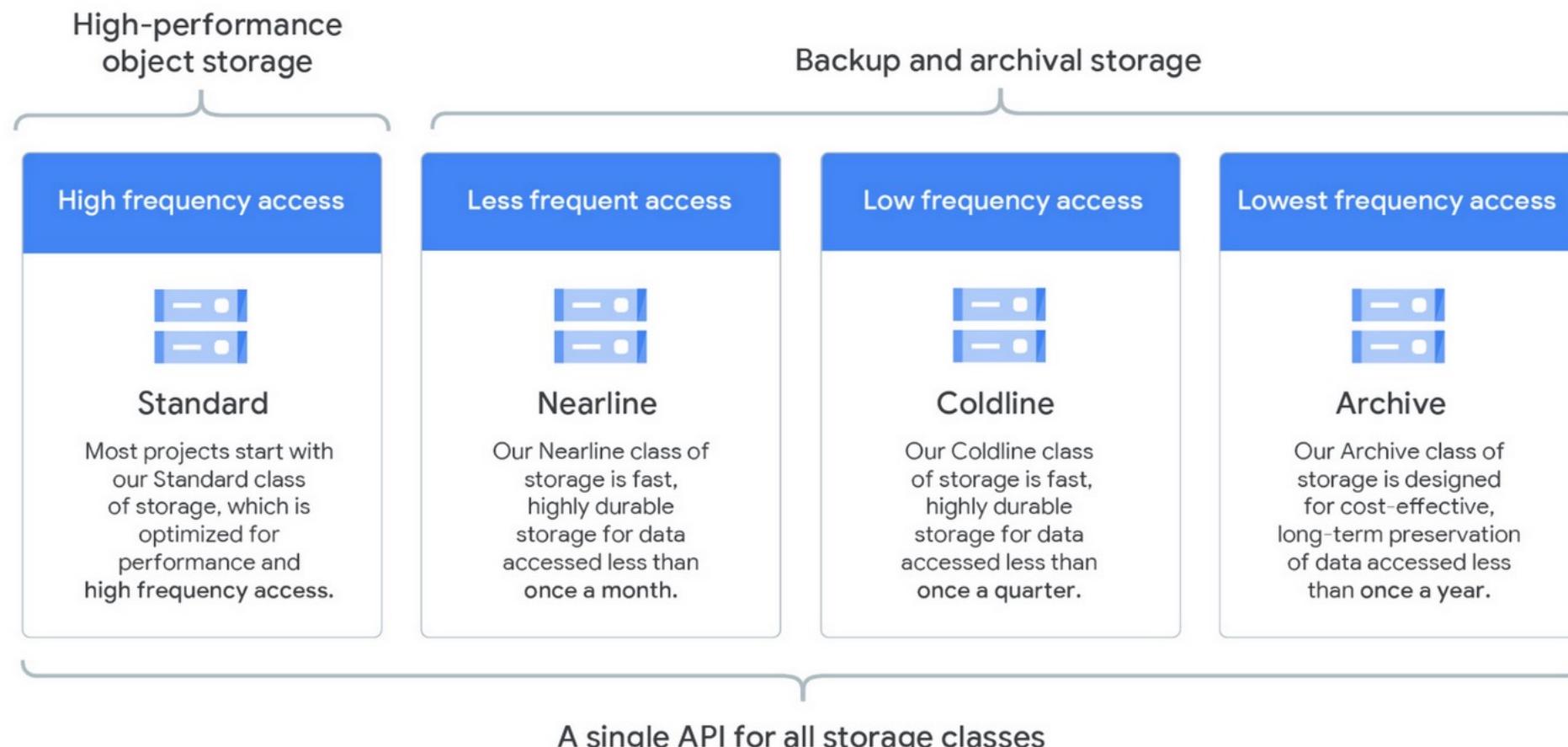
TP – Compute Engine

- Créer un groupe d'instance de deux vms sur deux régions différentes.
- Installer un serveur web sur les deux machines.
- Créer un load Balancer pour accéder aux machines.

Google Cloud Storage (GCS)

- GCS est le service de stockage de Google pour les données non structurées : images, vidéos, fichiers, scripts, sauvegardes de bases de données, etc.
- Les objets sont placés dans des compartiments, dont ils héritent des autorisations et des classes de stockage.
- Les classes de stockage fournissent différents SLA pour le stockage de vos données afin de minimiser les coûts de votre cas d'utilisation.
- La classe de stockage d'un bucket peut être modifiée (sous certaines restrictions), mais cela n'affectera que les nouveaux objets ajoutés au bucket.
- En plus de la console de Google, vous pouvez interagir avec GCS à partir de votre ligne de commande, à l'aide de gsutil.
- GCS permet d'utiliser un service de transfert de stockage (STS) pour importer des données à partir d'autre service.
 - Un compartiment AWS S3
 - Une ressource accessible via HTTP(S)
 - Un autre bucket Google Cloud Storage

Google Cloud Storage (GCS)



Google Cloud Storage (GCS) - LifeCycle

- Vous pouvez définir des règles qui déterminent ce qu'il adviendra d'un objet (sera-t-il archivé ou supprimé) lorsqu'une certaine condition sera remplie.
- Par exemple, vous pouvez définir une politique pour changer automatiquement la classe de stockage d'un objet de Standard à Nearline après 30 jours et pour le supprimer après 180 jours.

```
{  
  "lifecycle":{  
    "rule": [  
      {  
        "action":{  
          "type":"Delete"  
        },  
        "condition":{  
          "age":30,  
          "isLive":true  
        }  
      },  
      {  
        "action":{  
          "type":"Delete"  
        },  
        "condition":{  
          "numNewerVersions":2  
        }  
      },  
      {  
        "action":{  
          "type":"Delete"  
        },  
        "condition":{  
          "age":180,  
          "isLive":false  
        }  
      }  
    ]  
  }  
}
```

Google Cloud Storage (GCS) – Démo

- Utilisation de la console.
- Utilisation de gsutil.
- Utilisation de L'api.

Google Cloud Storage (GCS) – Exercice

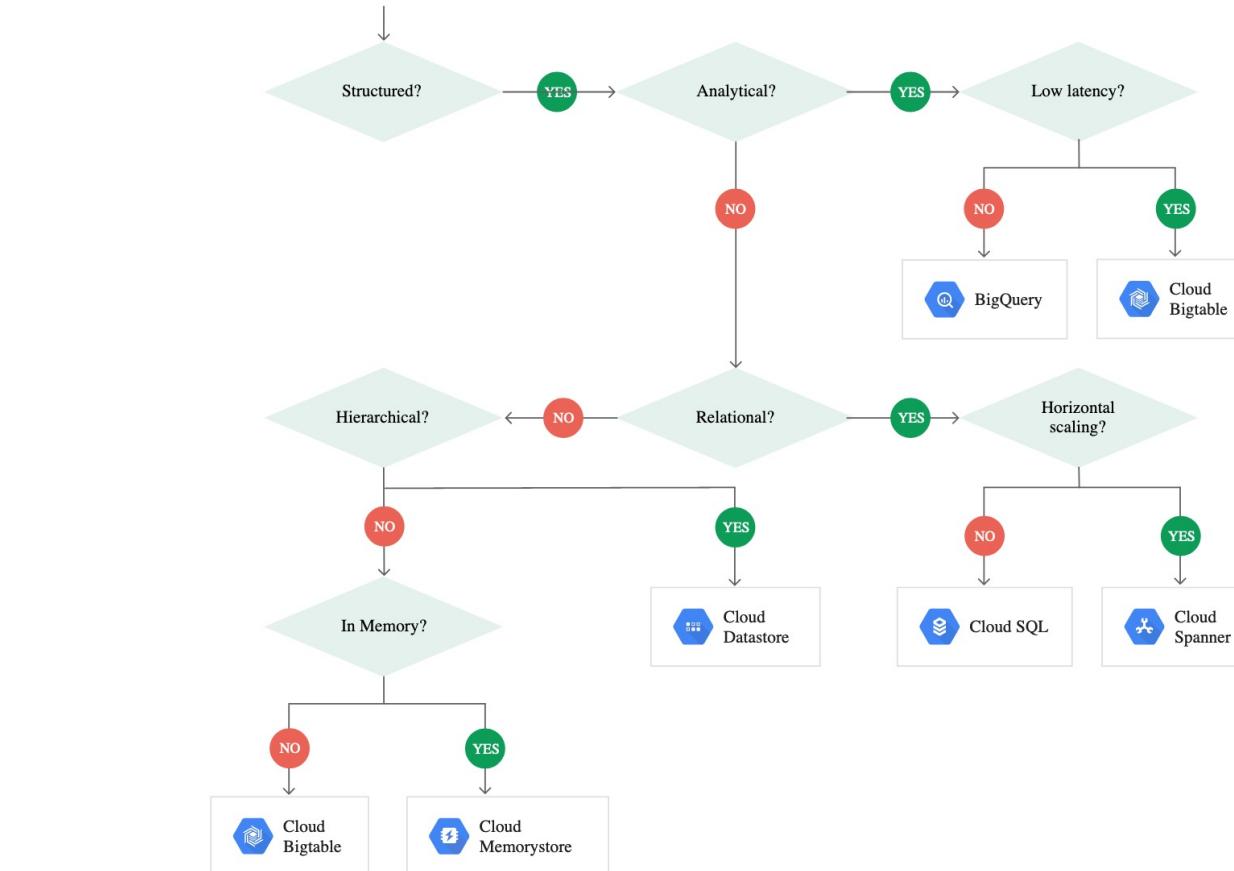
- Créer un bucket (classe standard) pour y stocker l'ensemble des logs d'un site web.
- Transférer les logs à partir de votre machine vers le bucket.
- Configurer les règles de cycle de vie pour :
 - Passer les objets du bucket en classe nearline après 30 jours.
 - Passer les objets du bucket en classe coldline après 90 jours.
 - Passer les objets du bucket en classe archive après 365 jours.

Gestion de base de données GCP

- GCP offre une multitude de services pour la gestion de données.
- Des services pour la gestion de base de données relationnelles.
 - Cloud SQL.
 - Cloud Spanner.
- Des services pour la gestion de base de données NoSQL.
 - DataStore.
 - BigTable.
 - MemoryStore.

Gestion de base de données GCP

Choix de la base de données.



CloudSQL

- Cloud SQL permet d'accéder à une instance de base de données MySQL ou PostgreSQL gérée dans GCP.
- Chaque instance est limitée à une seule région et a une capacité maximale de 30 To.
- Google se chargera de l'installation, des sauvegardes, de la mise à l'échelle, de la surveillance, du basculement et de la lecture des répliques. Pour des raisons de disponibilité.
- Les données peuvent être facilement importées (d'abord en téléchargeant les données sur Google Cloud Storage, puis sur l'instance).

TP Cloud SQL

- Importer un dump d'une base de données à partir de cloud storage.
- Importer d'une façon continue les données d'une base de données relationnelles mysql vers une instance cloud SQL.

Cloud Spanner

- Cloud Spanner est disponible dans le monde entier et peut très bien évoluer (horizontalement).
- Ces deux fonctionnalités le rendent capable de prendre en charge des cas d'utilisation différents de Cloud SQL et plus coûteux également. Cloud Spanner n'est pas une option pour les migrations lift-and-shift.

Cloud Spanner

- Cloud Spanner offre la totalité des fonctionnalité attendue d'une base de données relationnelles.
- Cloud Spanner offre une disponibilité de 99,999% pour les instances multirégionales (plus coûteuse qu'une instance sur une seul région).
- Cloud Spanner permet une segmentation automatique des données selon la charge des requêtes et la taille des données.
- D'autres fonctionnalités...

Cloud Spanner

- Cloud Spanner possède des différences dans l'utilisation de certains concepts de base de données relationnelles tel que :
 - Les procédures et triggers stockés.
 - Les séquences.
 - Le contrôle des accès.
 - Les contraintes de validations.
 - Les types de données.

Cloud Spanner – type de données

- Cloud Spanner offre les types de données suivants:
- INT64, BOOL, INT64, FLOAT64, NUMERIC, STRING, BYTES, DATE, TIMESTAMP, ARRAY<STRING>

Cloud Spanner – Entrelacement des tables

- Cloud Spanner propose une fonctionnalité dans laquelle vous pouvez définir deux tables comme ayant une relation parent/Enfant.
- Cette fonctionnalité entrelace les lignes de données enfants à côté de leur ligne parente dans le stockage.
- L'entrelacement signifie que les tables sont pré jointes, ce qui permet d'améliorer l'efficacité de la récupération des données lorsque le parent et les enfants sont recherchés ensemble.

TP Cloud Spanner

- Importer un dump d'une base de données Mysql vers cloud Spanner.

BigTable

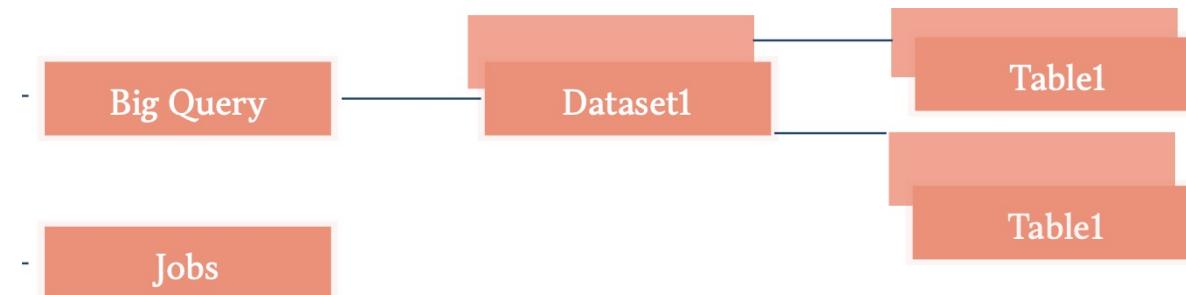
- Bigtable est une base de données NoSQL
- Analyse financière
- Données IOT
- Données de marketing
- Bigtable nécessite la création et la configuration de vos nœuds.
- Vous pouvez ajouter ou supprimer des nœuds à votre cluster sans aucun temps d'arrêt. Le moyen le plus simple d'interagir avec Bigtable est le CLI.
- Les performances de Bigtable dépendent de la conception de votre schéma de base de données.
- Vous ne pouvez définir qu'une seule clé par ligne et devez conserver toutes les informations associées à une entité dans la même ligne.

DataStore

- Datastore est une base de données de documents hautement évolutive, idéale pour les applications Web et mobiles : catalogues de produits, inventaire en temps réel.
- Par défaut, Datastore dispose d'un index intégré qui améliore les performances des requêtes simples. Vous pouvez créer vos propres index, appelés index composites, définis au format YAML.
- Si vous avez besoin d'un débit extrême (grand nombre de lectures/écritures par seconde), utilisez plutôt Bigtable.

BigQuery

- BigQuery est un service Web de Google utilisé pour gérer ou analyser des pétaoctets de données (Data Warehouse).
- BigQuery est entièrement géré.
- Il fait partie de Google Cloud Platform.
- BigQuery utilise SQL.



BigQuery - Stockage

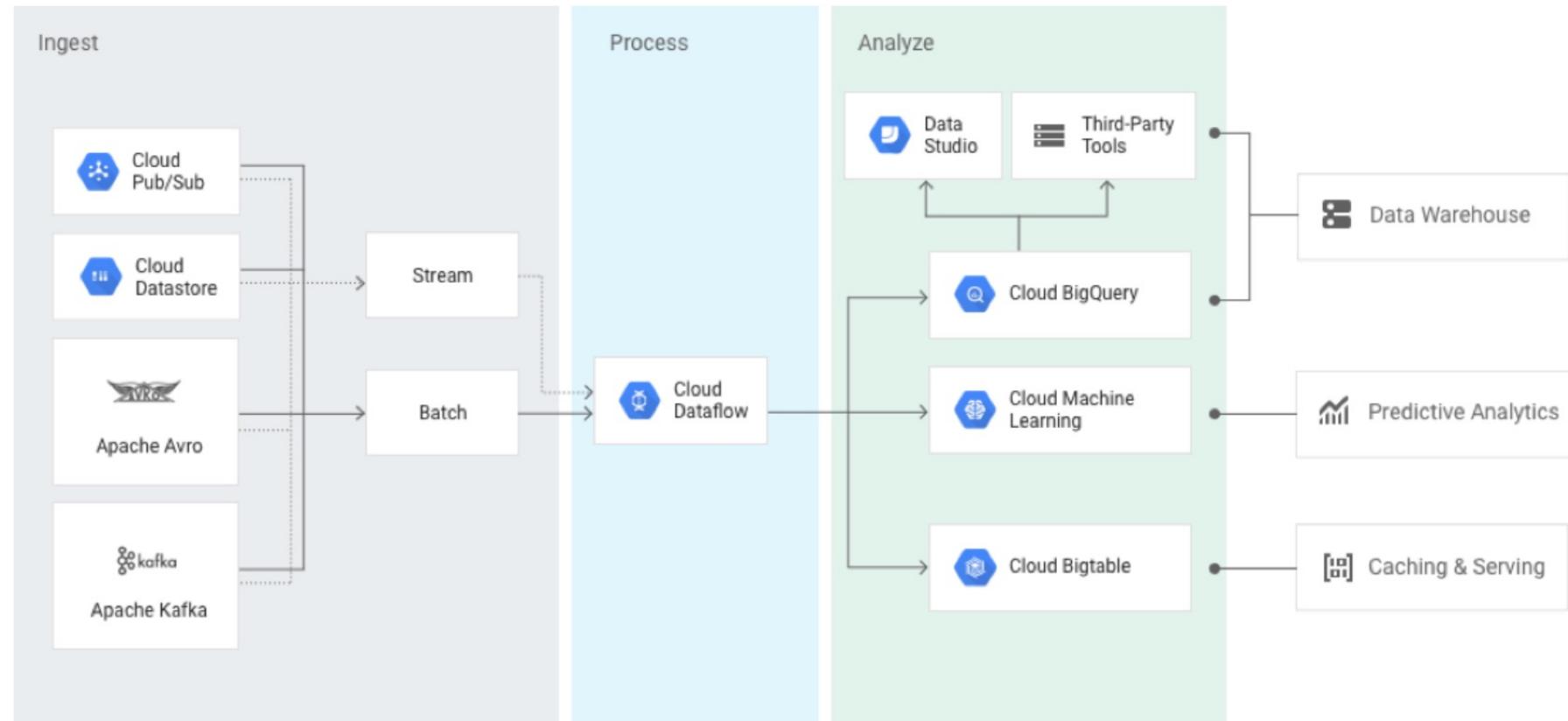
- BigQuery stocke les données sous format colonnes.
- BigQuery présente les données sous format de tables, lignes et colonnes.
- BigQuery est compatible avec les transactions des base de données.



BigQuery – Ingestion de données

- BigQuery permet d'ingérer les données à partir:
 - Google analytics.
 - Google cloud storage (csv, json...)
 - Google cloud dataFlow.
 - ...

BigQuery – Ingestion de données



Google App Engine

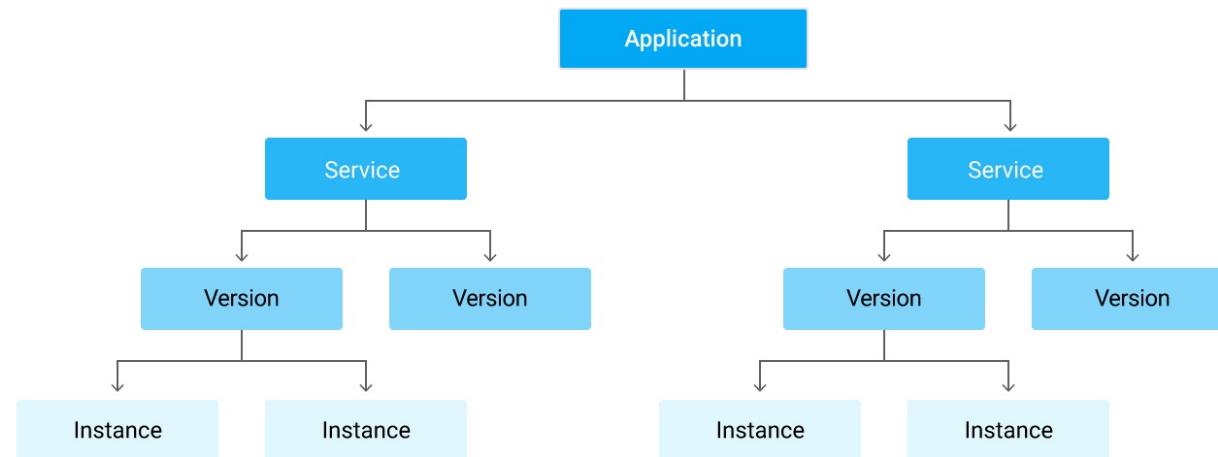
- Google App Engine est un service qui permet d'exécuter des applications en tant que service dans GCP.
- Google App Engine permet d'exécuter des applications en plusieurs langage en serverless.
- Google App Engine gère la scalabilité et le loadBalancing.

Google App Engine

- Google App Engine fournit :
 - Un SDK.
 - Des langages runtimes.
 - Administration par console.

Google App Engine

- Google App Engine est composé pour chaque projet de :
 - Services.
 - Versions.
 - Instances.



Google App Engine

- Google utilise un fichier de description pour la description de notre application sous format yaml.

```
runtime: nodejs16 # or another supported version

instance_class: F2

env_variables:
  BUCKET_NAME: "example-gcs-bucket"

handlers:
- url: /stylesheets
  static_dir: stylesheets

- url: /.*
  secure: always
  redirect_http_response_code: 301
  script: auto
```

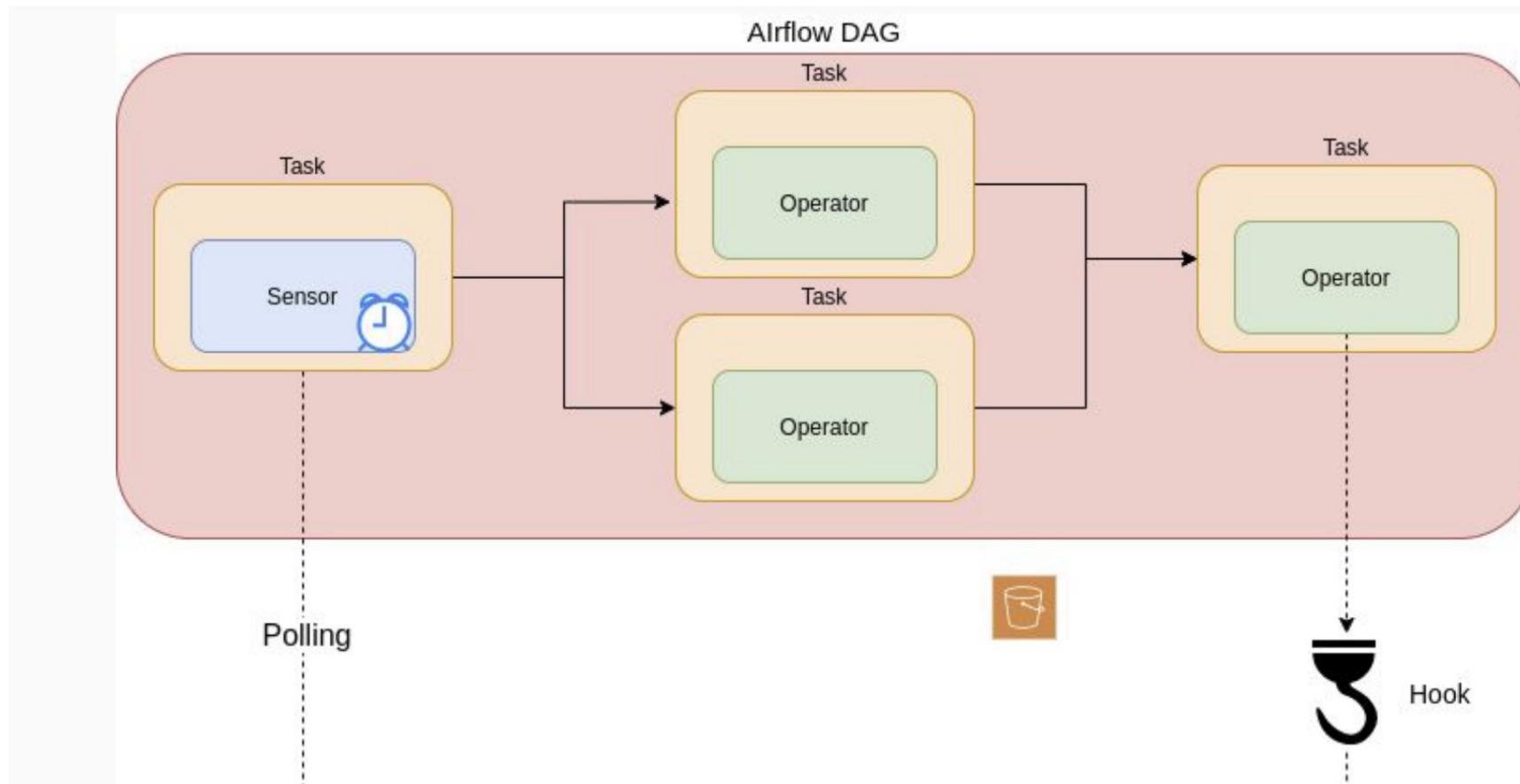
Cloud Composer

- Cloud Composer est service managé de GCP pour exécuter des workflow Apache AirFlow
- Cloud Composer permet de :
 - Créer des workflows à l'aide d'une API Python.
 - Planifier pour une exécution automatique ou de les démarrer manuellement.
 - Surveiller l'exécution des tâches en temps réel via une interface graphique.

Cloud Composer – Apache AirFlow

- Airflow est une plate-forme d'orchestration pour planifier et surveiller des WorkFlows par code.
- Les WorkFlows sont définis en tant que code Python
 - Plus flexible
 - Le workflow en tant que code est plus testable
 - Réutilisation
- Interface utilisateur web riche en fonctionnalités pour visualiser l'état des flux de travail, surveiller la progression, résoudre les problèmes, déclencher et redéclencher les workflows et les tâches qu'ils contiennent

AirFlow - Fonctionnement

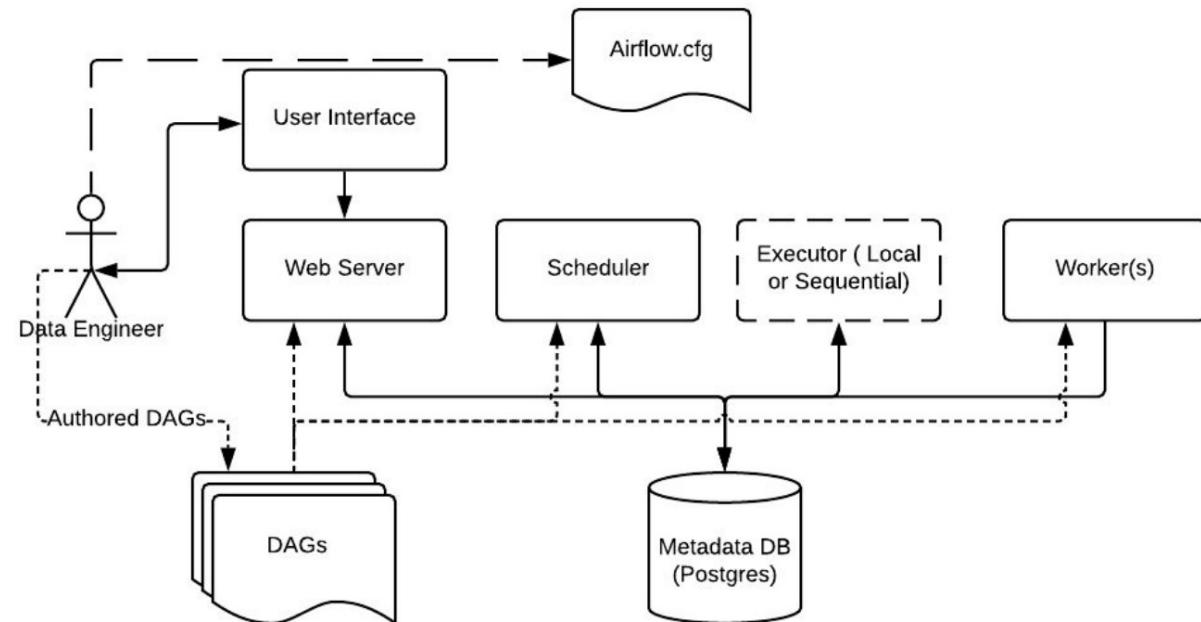


AirFlow - Fonctionnement

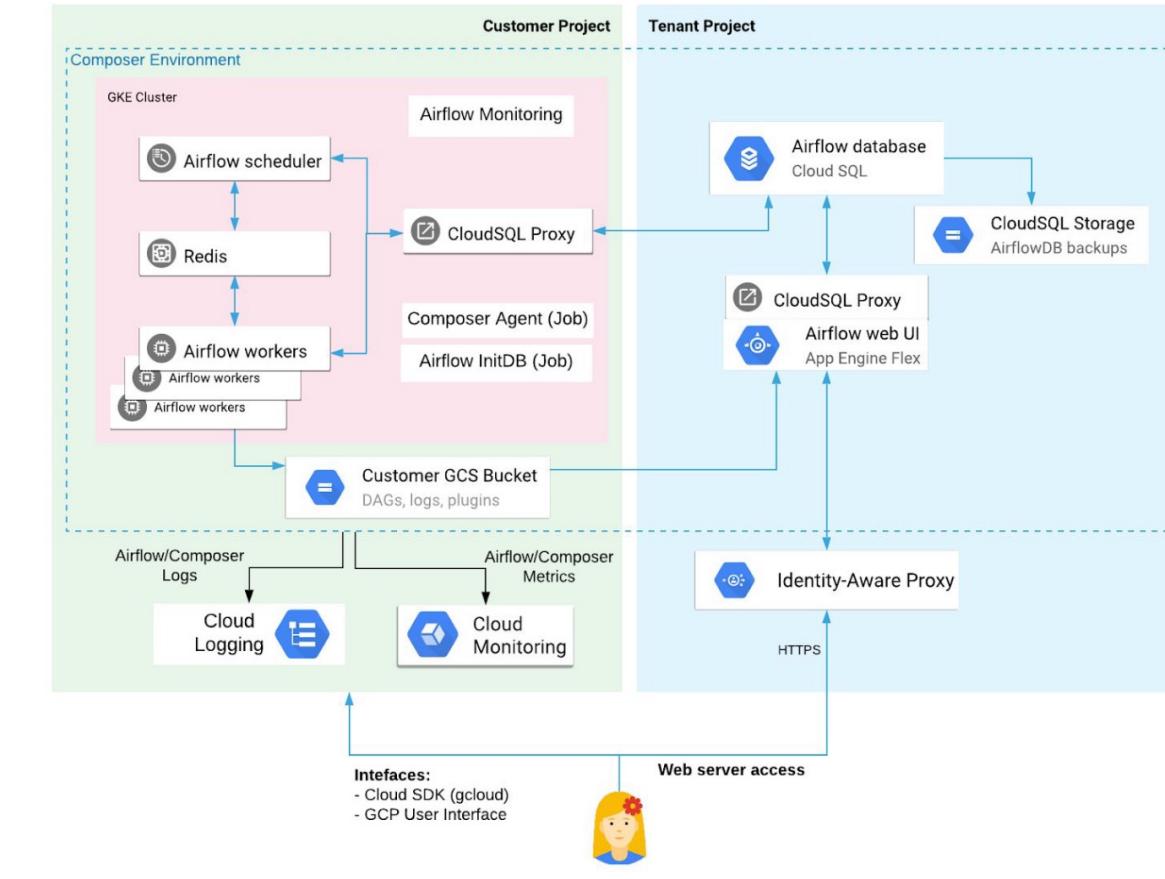
- DAG : le graphe acyclique un graphe orienté qui ne possède pas de circuit - workflows
- Opérateur : ils définissent ce qui doit être exécuté. Exemple : commande Bash, lecture d'un fichier, appel d'une API, charger des données dans une table, etc.
- Tâche : instance d'un opérateur, il s'agit d'un nœud dans un DAG/Workflow
- Sensor : un opérateur spécial qui s'exécute de manière répétée jusqu'à ce que la condition prédéfinie soit remplie.
 - Exemple : un capteur de fichier peut attendre que le fichier atterrisse, puis continuer le flux de travail
- Hook : une interface vers une plate-forme ou un système externe.
- Exécution DAG : lorsqu'un DAG est déclenché, il est appelé exécution DAG. Il représente l'instance du flux de travail

AirFlow Architecture

- Web UI/webserver
- Scheduler
- Worker
- Metadata database
- Executors
- SequentialExecutor
- LocalExecutor
- CeleryExecutor



Cloud Composer



TP – Cloud Composer

- On souhaite importer un fichier csv « meteo.csv », qui contient les informations suivantes : identifiant station météo, type mesure, valeur, dans une table bigquery.
- Après l'importation on souhaite récupérer les températures minimales mesurées et les enregistrer dans une nouvelles tables.
- On souhaite exécuter ces opérations d'une façon régulière (tous les jours).
- Pour réaliser notre workflow, on utilisera cloud composer avec les opérateurs :
 - GoogleCloudStorageToBigQueryOperator
 - BigQueryOperator