Big Data - Budgétisation

Ce document sert de guide pour estimer les coûts de l'utilisation des services Google Cloud Storage, Google BigQuery et Google Dataflow dans un contexte Big Data.

1. Google Cloud Storage est utilisé pour le stockage et la récupération de vos données.
2. Google BigQuery est un entrepôt de données qui permet d'effectuer des analyses sur de grandes quantités de données.
3. Google Dataflow est utilisé pour le traitement en temps réel et par lots de vos données.

Le document détaille les coûts associés à chacun de ces services, y compris le stockage des données, le réseau, l'analyse des données et le traitement. Il fournit également une estimation du coût total basée sur ces facteurs. Cela aidera votre organisation à planifier et à budgétiser ses dépenses en matière de services de stockage et d'analyse de données. Il est important de noter que les coûts réels peuvent varier en fonction de l'utilisation réelle et des fluctuations des prix de Google.

Avant de de faire une estimation voici les flux de données à prendre en compte :

- Poids des données structuré au format json par entité (Train/Tramway/Metro, Véhicule, Bus) ⇒ 10Mo/h
- Le nombre de caméra (donc de flux vidéo) par entité donc :
  - Nombre de caméra par train/tramway/metro ⇒ 1
  - Nombre de train/tramway/metro /jour ⇒ ˜950
  - Nombre de caméra par taxi ⇒ 1
  - Nombre de taxi ⇒ ˜58 000

    - Nombre de caméra par bus ⇒ 1
    - Nombre de bus ⇒ ˜9000
- Poids du flux vidéo par entité (Train/Tramway/Metro, Véhicule, Bus)

  ⇒ 500Mo/h en 720p pour toutes les entités

Ce qui reviens à 67 950 entités à envoyé un flux vidéo de 500Mo/h et un flux de données de 10 Mo/h.

> 💡 Ce qui nous fais
>
> - **33 975 000 Mo/h** juste pour les flux vidéos.
> - **679 500 Mo/h** pour les flux de données.
> - Un total de : **34 654 500 Mo/h** soit **34 654,5 Go/h** soit **34,6545 To/h**
> - Le chiffre qui nous intéresses est surtout la quantité de données par mois qui serais au maximum de : **25 782 948 Go/mois**

Il faut prendre en compte que ces chiffres sont une estimation du nombre d'entité qui serait susceptible d'envoyé des données en même temps, soit le maximum de données possible. Dans les faits les données seront bien moins volumineuse par exemple en dehors des heures de pointe notamment la nuit.

I. Google Cloud Storage

Description : Google Cloud Storage est utilisé pour stocker et récupérer n'importe quel volume de données à tout moment.

Plusieurs classe de stockage sont possible et on des prix différents :

- Classe Standard : 0,023 € / Go
- Classe Nearline : 0,013 € / Go
- Classe Coldline : 0,006 € / Go
- Classe Archive : 0,0025 € / Go

Le format de stockage standard dans Google Cloud Storage est souvent le plus intéressant pour les organisations pour plusieurs raisons :

1. **Disponibilité :** Le format standard offre la plus haute disponibilité parmi les différentes classes de stockage, ce qui signifie que vos données sont accessibles chaque fois que vous en avez besoin.
2. **Performance :** La classe standard offre les temps de latence les plus bas et les débits les plus élevés, ce qui est crucial pour les applications et les services qui nécessitent un accès rapide aux données.
3. **Flexibilité :** Avec le stockage standard, il n'y a pas de restrictions sur la fréquence d'accès aux données ni de frais pour l'accès aux données tôt. C'est utile si vos données d'application sont consultées fréquemment.
4. **Redondance géographique :** Le stockage standard est stocké de manière redondante dans plusieurs régions, offrant une durabilité accrue en cas de panne dans une région.

Cependant, le stockage standard est généralement plus coûteux que les autres classes de stockage comme Nearline, Coldline et Archive. Ces classes de stockage sont plus appropriées pour les données qui sont moins fréquemment accédées ou qui peuvent être stockées pendant longtemps sans être consultées. Mais dans notre cas le stockage dans Google Cloud Storage à pour vocation à faire de l’ingestion de données avant d’être transféré dans Google Big Query.

Coût prévu pour 25 782 948 Go/mois : **£530,319.20 soit €613,376.43**

II. Google BigQuery

Description : Google BigQuery est un entrepôt de données sans serveur qui permet d'exécuter des analyses sur de grandes quantités de données.

2.1. Stockage

- Stockage actif : 0,02 € par Go/mois.
- Stockage à long terme : 0,01 € par Go/mois (les données sont considérées comme étant à long terme si elles n'ont pas été modifiées depuis 90 jours).

Selon les tarifs que j'ai mentionnés précédemment, le stockage actif dans Google BigQuery coûte 0,02 € par Go/mois et c’est sur celui-ci que l’on va s’appuyer :

Le stockage actif dans des systèmes comme Google BigQuery est intéressant pour plusieurs raisons:

1. **Accès rapide aux données :** Les données en stockage actif sont immédiatement disponibles pour l'analyse et le traitement. Il n'y a pas de délai pour accéder à ces données, ce qui est essentiel pour de nombreuses applications en temps réel ou interactives.
2. **Performance d'analyse :** Les données stockées activement sont prêtes à être analysées à tout moment. Cela signifie que vous pouvez exécuter des requêtes et obtenir des résultats plus rapidement, ce qui est essentiel pour les applications d'analyse de données de grande taille et de haute performance.
3. **Gestion des données :** Le stockage actif peut offrir des fonctionnalités plus avancées, comme la possibilité d'organiser les données en partitions, d'optimiser les requêtes, d'ajouter et de supprimer des données en fonction des besoins, etc. Cela peut faciliter la gestion des données et améliorer les performances des requêtes.
4. **Fiabilité et sécurité :** Les systèmes de stockage actif offrent généralement une haute fiabilité et des garanties de durabilité. Cela signifie que vous pouvez compter

sur eux pour conserver vos données en toute sécurité. De plus, ils peuvent offrir des fonctionnalités de sécurité avancées, comme le chiffrement des données en repos et en transit.

Cependant, le stockage actif peut coûter plus cher que les autres types de stockage, comme le stockage à long terme ou le stockage d'archives. Il est donc important d'équilibrer les besoins en termes de performance et d'accès aux données avec les coûts lorsque vous choisissez le type de stockage à utiliser.

💶 Par conséquent, l'estimation du coût sera la suivante :

25 782 948 Go * 0,02 € par Go = 515 658,96 € par mois

Donc, le coût estimé du stockage de 25 782 948 Go de données par mois dans Google BigQuery serait d'environ 515 658,96 €.

2.3. Analyse

- Analyse sur demande : 5,00 € par To analysée.
- Analyse en mode flat-rate : les prix commencent à 10 000 € par mois.

L'analyse à la demande (on-demand analysis) et le forfait à tarif fixe (flat-rate) sont deux modèles de tarification proposés par Google BigQuery. Chacun a ses avantages et le choix entre les deux dépend principalement de vos besoins spécifiques.

L'analyse à la demande est généralement plus intéressante pour les raisons suivantes :

1. **Flexibilité :** Avec l'analyse à la demande, vous payez uniquement pour les données que vous analysez. Cela peut être plus rentable si vos besoins d'analyse sont irréguliers ou si le volume de données que vous analysez varie beaucoup d'un mois à l'autre.
2. **Pas de frais fixes :** Il n'y a pas de frais mensuels fixes avec l'analyse à la demande. Vous ne payez que pour ce que vous utilisez. Cela peut être un avantage si vous n'avez pas besoin d'analyser de grandes quantités de données tous les mois.

3. **Pas de gestion de capacité :** Avec l'analyse à la demande, vous n'avez pas à vous soucier de la gestion de la capacité. Google BigQuery alloue automatiquement les ressources nécessaires pour exécuter vos requêtes.

Cependant, le forfait à tarif fixe peut être plus rentable si vous analysez constamment de grandes quantités de données. Avec ce modèle, vous payez un tarif mensuel fixe et vous pouvez analyser autant de données que vous le souhaitez sans frais supplémentaires. Cela peut simplifier la budgétisation et peut être moins cher si vos besoins d'analyse sont importants et constants.

D'autre part, le tarif du forfait à tarif fixe commence à 10 000 € par mois, mais cela ne garantit pas nécessairement la capacité d'analyser autant de données que vous en avez. La tarification à tarif fixe de BigQuery est basée sur des "emplacements de slots" réservés pour l'exécution des requêtes. Le coût de ces slots peut augmenter en fonction de la quantité de données que vous devez analyser.

Pour déterminer si le forfait à tarif fixe serait plus rentable, vous devriez contacter un représentant de Google Cloud pour discuter de vos besoins spécifiques en matière d'analyse de données et obtenir une estimation précise du coût du forfait à tarif fixe adapté à vos besoins.

Cependant, sur la base des chiffres que nous avons ici, si vos besoins d'analyse sont constants et proches de 25 782,948 To par mois, l'option à tarif fixe pourrait potentiellement être plus rentable, à condition que le coût des slots nécessaires ne dépasse pas le coût estimé pour l'analyse à la demande.

Dans un premier temps nous allons estimer le prix pour le tarifs d’analyse à la demande.

Selon les tarifs que j'ai mentionnés précédemment, l'analyse à la demande dans Google BigQuery coûte 5,00 € par To analysée.

💶 Par conséquent, l'estimation du coût sera la suivante :

25 782,948 To * 5,00 € par To = 128 914,74 € par mois

Donc, le coût estimé de l'analyse de 25 782 948 Go de données par mois dans Google BigQuery serait d'environ 128 914,74 €.

III. Google Dataflow

Description : Google Dataflow est un service de traitement des données entièrement géré pour le streaming en temps réel et les tâches de traitement par lots.

Coût :

1. **Traitement :**
   - Traitement par lots : 0,01 € par vCPU-heure.
   - Streaming en temps réel : 0,02 € par vCPU-heure.

   Coût prévu : (Coût par vCPU-heure * Heures de traitement)
2. **Stockage des données :**
   - Stockage des données temporaires : 0,02 € par Go/mois.

   Coût prévu : (Coût par Go * Volume de données temporaires en Go)

En accord avec l'équipe de l'IA en quantité de 158,1 Go/h soit 117 626,4 Go/mois à été décidé.

Pour estimer le coût du traitement par lots de 158,1 Go/h pendant 744 heures (ce qui représente environ un mois) dans Google Dataflow, nous allons procéder comme suit :

1. **Traitement :**

Pour le traitement par lots, Google facture en fonction des vCPU-heures utilisées. Supposons que nous utilisions une machine standard à 4 vCPU.

- Coût par vCPU-heure pour le traitement par lots : 0,01 €.

Donc, pour une machine de 4 vCPU fonctionnant pendant 744 heures, le coût serait de :

- 4 vCPU * 0,01 €/vCPU-heure * 744 heures = 29,76 € pour le mois.

2. **Stockage des données :**

   Google facture également pour le stockage des données temporaires. Supposons que chaque heure de traitement génère 158,1 Go de données temporaires qui sont stockées pendant une heure.

   - Coût de stockage des données temporaires : 0,02 € par Go/mois.

   Pour 158,1 Go de données stockées pendant 744 heures, le coût serait de :

   - 158,1 Go * 0,02 €/Go * 744/720 (pour convertir les heures en mois) = 3277,37 € pour le mois.

Donc, le coût total pour un mois de traitement par lots de 158,1 Go/h avec Google Dataflow serait d'environ 3307,13 € (29,76 € pour le traitement + 3277,37 € pour le stockage temporaire).

💶 Donc, le coût total pour un mois de traitement par lots de 158,1 Go/h avec Google Dataflow serait d'environ 3307,13 € (29,76 € pour le traitement + 3277,37 € pour le stockage temporaire).

Ces coûts sont des estimations basées sur les taux que j'ai mentionnés et peuvent varier. Les coûts réels dépendent de nombreux facteurs, y compris la complexité du travail de traitement, la configuration de la machine, le volume réel de données temporaires stockées, et d'autres facteurs. De plus, ce calcul ne tient pas compte d'éventuels coûts de réseau ou autres coûts associés.

IV. Ressources Humaines

En tant que développeur de Big Data, mon expertise vous aidera à mettre en place et à gérer votre infrastructure de Big Data sur Google Cloud, comprenant Google Cloud Storage, Google BigQuery et Google Dataflow.

💶 Le coût de mes services est d'environ 47,50 € par heure, ce qui se traduit par un coût total de 7600 € pour un mois de travail à temps plein (soit environ 160 heures).

Je suis conscient que cela peut sembler être un coût significatif, mais voici quelques raisons pour lesquelles ce coût est justifié :

1. **Expertise technique :** J'ai une connaissance approfondie des technologies de Big Data, y compris Google Cloud Storage, Google BigQuery, et Google Dataflow. Je peux non seulement mettre en place votre infrastructure de Big Data, mais aussi l'optimiser pour assurer une performance maximale et minimiser les coûts.
2. **Expérience :** Avec plusieurs années d'expérience dans le développement de Big Data, j'ai la capacité de prévoir et de résoudre les problèmes avant qu'ils ne se produisent. Cela peut éviter des retards et des coûts supplémentaires dans votre projet.
3. **Gain de temps :** En me confiant la mise en place de votre infrastructure de Big Data, votre équipe peut se concentrer sur ce qu'elle fait le mieux, sans avoir à se soucier des détails techniques. Cela peut accélérer le développement de votre projet et améliorer l'efficacité de votre équipe.
4. **Support continu :** Mon travail ne s'arrête pas une fois que votre infrastructure de Big Data est mise en place. Je peux fournir un soutien continu pour garantir que

tout fonctionne comme prévu et pour effectuer les mises à jour et les modifications nécessaires.

Ces facteurs justifient le coût de mes services et garantissent que vous obtiendrez une valeur significative en retour pour votre investissement.

V. Total

Bien sûr, voici un résumé sous forme de tableau des coûts que nous avons discutés :

Service / Ressource	Quantité / Durée	Coût unitaire (€)	Coût total (€)
Stockage Google BigQuery	25 782 948 Go / mois	0,02 / Go	515 658,96
Analyse Google BigQuery	25 782 948 Go / mois	0,005 / Go	128 914,74
Stockage Google Cloud Storage	25 782 948 Go / mois	0,02 / Go	515 658,96
Google Dataflow - Traitement	4 vCPU * 744 h	0,01 / vCPU-h	29,76
Google Dataflow - Stockage temp.	158,1 Go * 744 h	0,02 / Go	3277,37
Main-d'œuvre - Développeur Big Data	160 h	47,5 / h	7600,00
Total			1 170 120,79

Ces chiffres sont des estimations basées sur les informations que vous avez fournies et sur les tarifs que j'ai mentionnés, et peuvent varier en fonction de facteurs tels que le volume réel des données, la complexité du travail de traitement, les fluctuations des tarifs, et d'autres facteurs.

De plus, veuillez noter que ces coûts ne prennent pas en compte d'autres frais potentiels tels que les coûts de réseau, les coûts de récupération de données, les coûts

de transfert de données, les coûts de support technique, et autres.

Sources:

Budgétisation : Google Cloud Plateforme

Statistiques TFl :

- https://www.gov.uk/
- https://tfl.gov.uk/