Pour une formation en personne, le cas échéant, se présenter et demander au participant de faire de même.

Programme

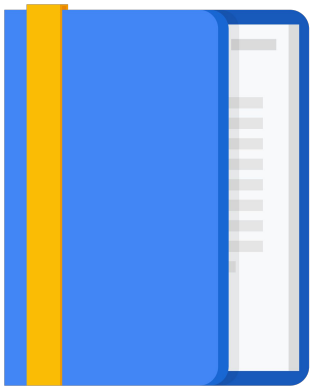
Traiter des flux de données

Cloud Pub/Sub

Fonctionnalités de traitement par flux Cloud Dataflow

Fonctionnalités de traitement par flux BigQuery

Cloud Bigtable



Google Cloud

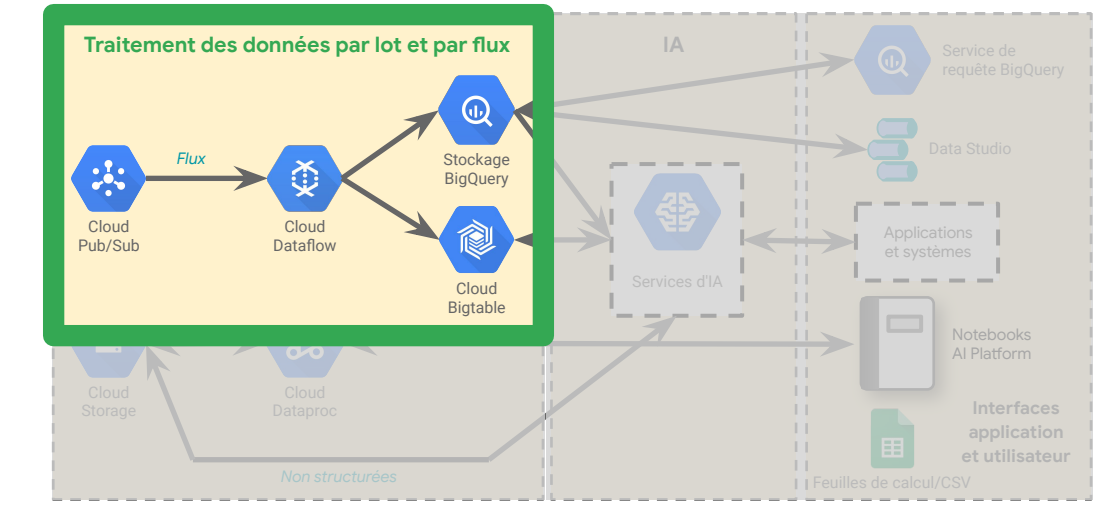
Bonjour et bienvenue dans ce cours intitulé Créer des systèmes de flux de données résilients.

Je m'appelle _____.

Dans ce cours, nous allons parler des sujets suivants :

- Traitement des données par flux
- Cloud Pub/Sub
- Fonctionnalités de traitement par flux Cloud Dataflow
- Fonctionnalités de traitement par flux BigQuery
- Cloud Bigtable

Traitement des flux de données.

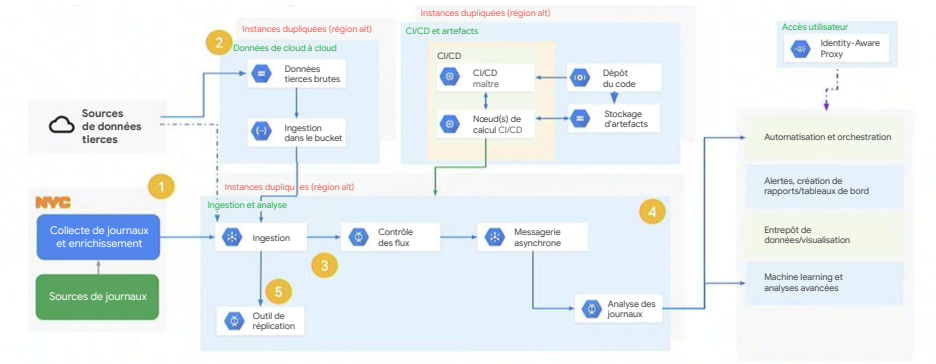


Google Cloud

Ce module est consacré au traitement par flux et nous aborderons l'architecture de référence. Les données sont habituellement transmises par Cloud Pub/Sub, puis passent par Dataflow pour être agrégées et transformées. Vous utilisez ensuite BigQuery ou Cloud Bigtable selon que vous voulez coder des enregistrements agrégés ou individuels provenant de sources de flux.

De nombreuses entreprises veulent donner à leurs analystes la possibilité de prendre des décisions en temps réel. C'est ce que NYC3 a fait.

Schéma général



"Seules les données en temps réel comptent pour nous."

-- Noam Dorogoyer, New York City Cyber Command

Google Cloud

Article dans GCN : https://gcn.com/articles/2019/08/01/nyc-cloud-cyber-pipeline.aspx
Présentation à la conférence GCP NEXT 2019 : https://www.youtube.com/watch?v=x4yQY8yhVJY

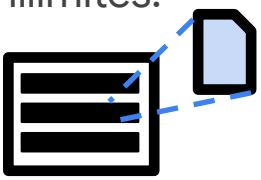
Examinons d'abord des cas de traitement par flux. Quels types de données sont traitées par flux ? Le traitement par flux nous permet d'obtenir des informations en temps réel dans un tableau de bord, mais aussi de connaître l'état de votre activité.

Noam Dorogoyer de New York City Cyber Command déclare : "Nos données proviennent de fournisseurs externes. Elles sont ingérées via Pub/Sub, qui à son tour les envoie vers Dataflow à des fins d'analyse ou d'enrichissement.
Si les données n'arrivent pas en temps et en heure, surtout lorsqu'il s'agit de cybersécurité, elles sont inutiles, en particulier en cas d'urgence. Du point de vue de l'ingénierie des données, le pipeline est construit de manière à minimiser la latence à chaque étape. Dans le cas d'une tâche Dataflow, par exemple, la conception vise à activer autant d'éléments que possible en même temps, de sorte qu'à aucun moment une étape soit bloquée par la précédente."

La quantité de données qui circule par la commande change chaque jour. En semaine, aux heures de pointe, cela peut atteindre 5 ou 6 To, d'après Noam Dorogoyer. Le week-end, le volume passe à 2 ou 3 To. À mesure que NYC Cyber Command étend sa visibilité à d'autres agences, elle devra traiter des pétaoctets de données.

"Les analystes de sécurité peuvent accéder aux données de différentes manières", a déclaré Anthony Bocekci, Community Emergency Response Team Specialist. Ils peuvent exécuter des requêtes ou utiliser d'autres outils de visualisation des données, comme Data Studio, une solution de création de rapports.

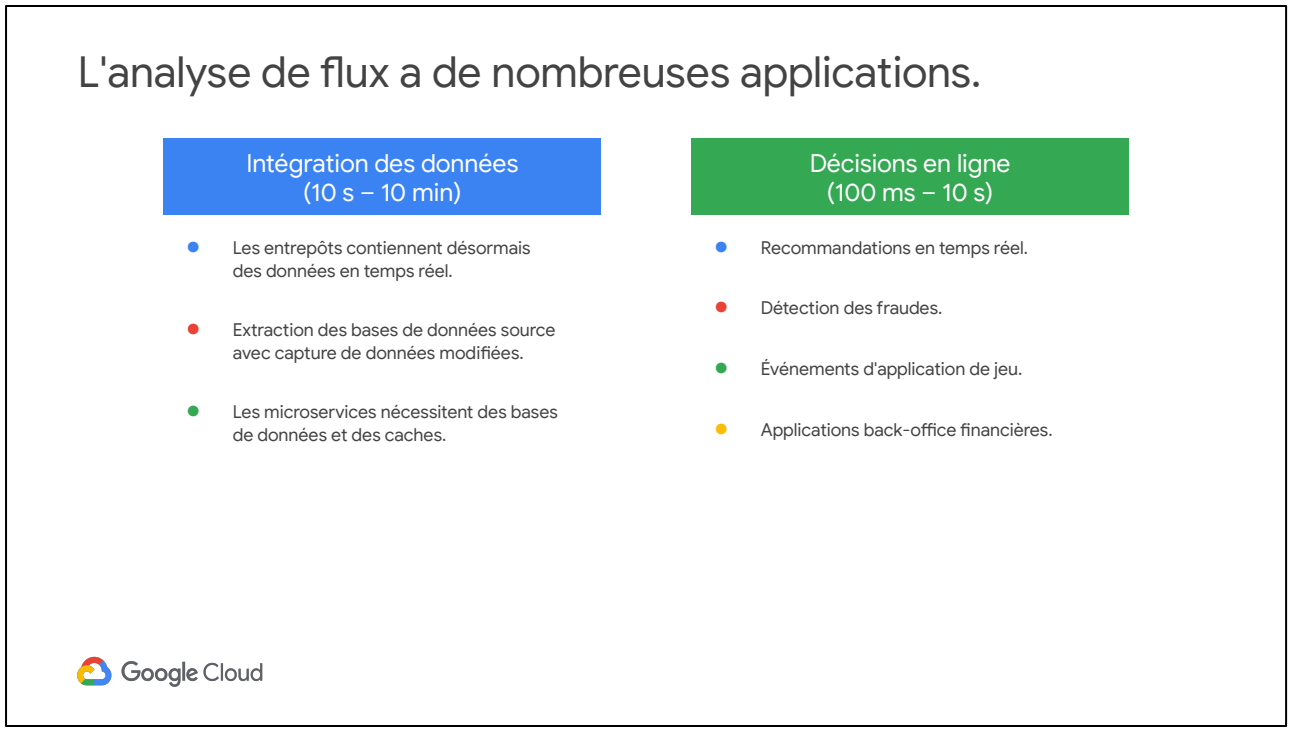
Les données traitées par flux sont des ensembles de données illimités.



Données limitées (lot)	Données illimitées (flux)
Ensemble de données limité Généralement complet Chronologie généralement ignorée Normalement au repos Stockage durable	Ensemble de données illimité Jamais complet Chronologie généralement importante Normalement en transit Stockage temporaire

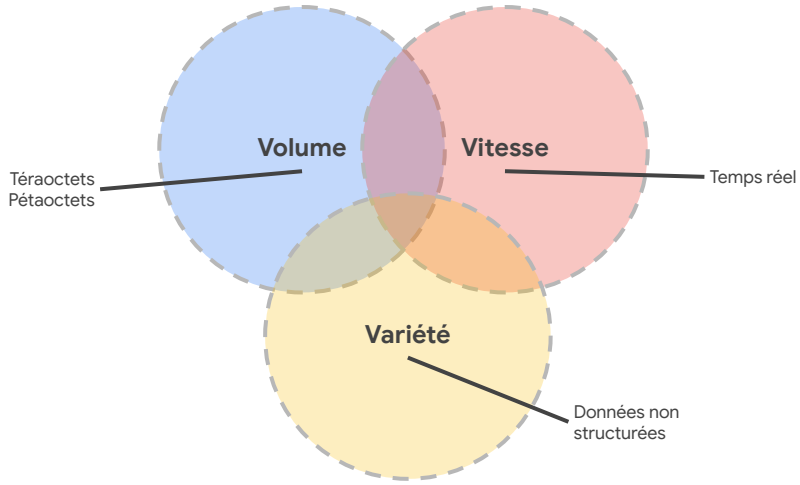
Google Cloud

Les données traitées par flux sont des données illimitées. Les données limitées sont des données au repos. Le traitement par flux fait référence à la manière dont les données illimitées sont traitées.
Un moteur de traitement par flux offre une latence faible, des résultats spéculatifs ou partiels, un raisonnement chronologique flexible, des contrôles qui garantissent l'exactitude des résultats et les performances nécessaires à la réalisation d'analyses complexes.



Un flux de données permet de bénéficier d'un entrepôt de données en temps réel, et ainsi de générer un tableau de bord avec ces informations. Par exemple, vous pouvez afficher en temps réel une comparaison entre les tweets négatifs et les tweets positifs publiés au sujet des produits de votre entreprise. Cela sert aussi à la détection des fraudes, aux événements d'application de jeu ou aux applications back-office financières, comme les opérations de bourse et les ventes sur les marchés.

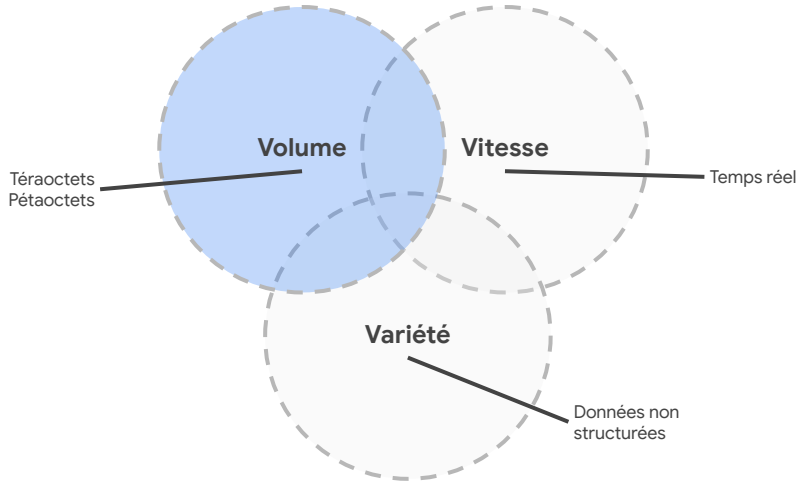
Comment gérer le volume, la vitesse et la variété des données ?



Google Cloud

Les défis associés aux applications de flux tournent autour des trois V, qui sont le volume, la vitesse et la variété des données.

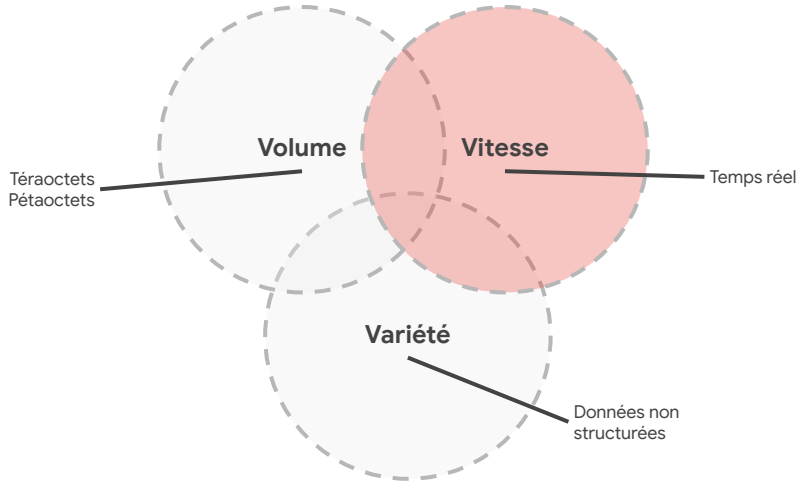
Comment gérer le volume, la vitesse et la variété des données ?



Google Cloud

Le volume est problématique, car les données affluent en continu et leur quantité augmente rapidement.

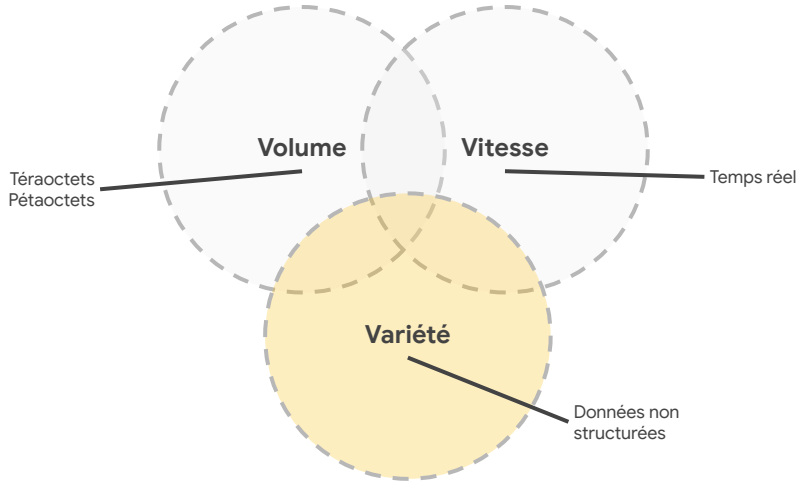
Comment gérer le volume, la vitesse et la variété des données ?



Google Cloud

La vitesse, en fonction de votre activité (opérations boursières, suivi des informations financières, ouverture des portes de métro), l'est aussi, en raison des transferts qui peuvent atteindre des dizaines de milliers d'enregistrements par seconde. La vitesse peut varier considérablement. Par exemple, un commerçant qui conçoit son système de points de vente au niveau national va probablement avoir un volume stable toute l'année jusqu'à la période des soldes où les ventes et les données transférées vont exploser. Il est donc important de concevoir des systèmes en mesure de gérer cette charge supplémentaire.

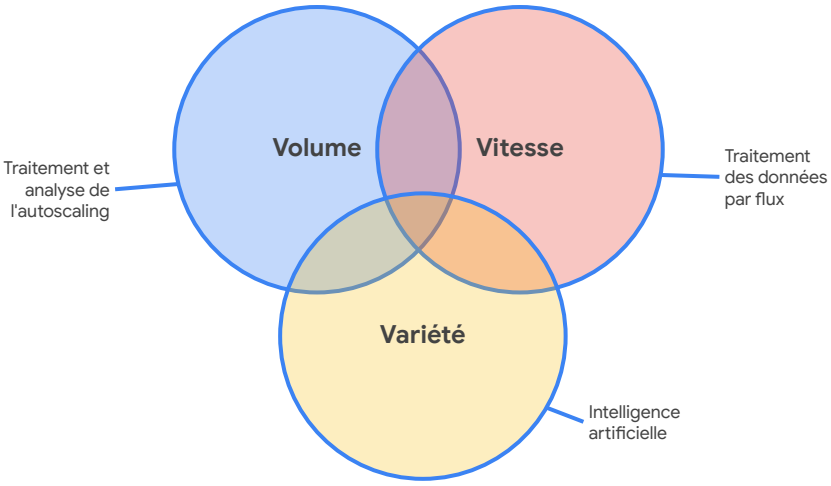
Comment gérer le volume, la vitesse et la variété des données ?



Google Cloud

Le troisième défi réside dans la variété des données. Si vous n'utilisez que des données structurées, qui proviennent d'une application mobile, la gestion est plutôt simple. Mais qu'en est-il des données non structurées, comme les données vocales ou les images ? Ce sont des enregistrements en flux continu pour lesquels il faudra définir une valeur nulle.

Autoscaling, machine learning et traitement par flux.



Google Cloud

Nous allons voir comment nous pouvons simplifier leur traitement dans le cloud. Pour que le système puisse gérer le volume, nous allons utiliser un outil de traitement et d'analyse de l'autoscaling. Pour la vitesse, nous allons utiliser un outil adapté à l'évolutivité du traitement par flux. Pour résoudre le problème de la variété, nous verrons l'utilité de l'intelligence artificielle pour classer les données non structurées.

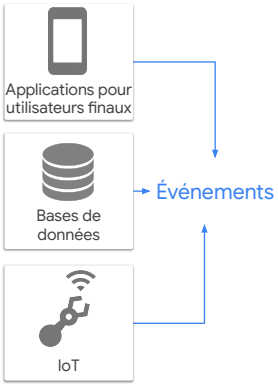
Les produits GCP permettent de relever les défis majeurs posés par le traitement et l'analyse des flux de données.

Volumes de données évolutifs et variables	Traitement des données sans retard inutile	Analyse ad hoc et tendances immédiates indispensables
		
Cloud Pub/Sub	Cloud Dataflow	BigQuery

Google Cloud

Nous aborderons trois outils essentiels : Cloud Pub/Sub qui gère les volumes des données variables et évolutifs, Cloud Dataflow qui assure le traitement des données sans retard inutile, et BigQuery qui sert à créer des rapports ad hoc, même sur des flux de données.

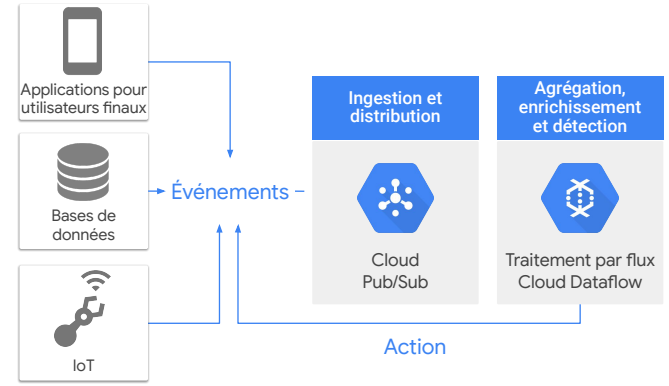
L'analyse du flux de données comprend certaines étapes communes.



Google Cloud

Examinons les différentes étapes. D'abord, un certain type de données est récupéré, en provenance d'une application, d'une base de données ou de l'IoT (Internet des objets). Ce sont des événements de génération.

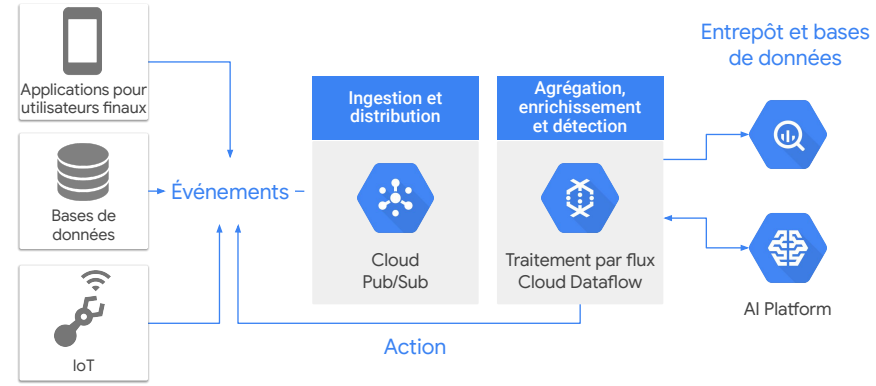
L'analyse du flux de données comprend certaines étapes communes.



Google Cloud

Ils sont suivis d'une action. Nous ingérons ces données et les distribuons à l'aide de Cloud Pub/Sub. Cela permet de s'assurer que les messages sont fiables. Nous bénéficions ainsi d'une mise en mémoire tampon. Dataflow agrège, enrichit et détecte les données.

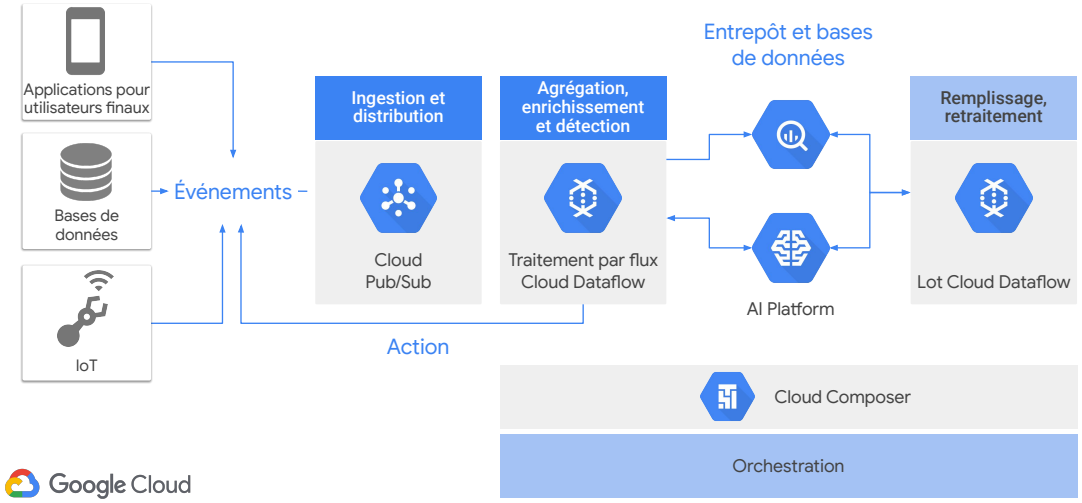
L'analyse du flux de données comprend certaines étapes communes.



Google Cloud

Elles sont ensuite encodées dans un entrepôt, comme BigQuery ou Bigtable, ou exécutées à partir d'un modèle de ML (machine learning). Par exemple, nous pouvons utiliser ce flux de données tel quel pour entraîner un modèle dans Cloud ML Engine.

L'analyse du flux de données comprend certaines étapes communes.



Enfin, vous pouvez utiliser Dataflow ou Dataproc pour effectuer un traitement par lot, un remplissage, etc.

C'est un moyen assez pratique d'effectuer le processus en intégralité dans GCP.