Le regard des Kommunautés

Google Cloud

Utilisez les services managés dans le Cloud pour booster votre plateforme big data !

Pour commencer, et pour les plus novices d'entre nous, qu'est-ce que le cloud ?

C'est tout simplement le fait de pouvoir accéder à des ressources (puissance de calcul ou de stockage) de manière très simple, sans vraiment les avoir à portée de main. Elles sont quelque part dans les « nuages ».

Ces ressources, mises à disposition par un fournisseur, seront adaptées à nos demandes et facilement évolutives. Aujourd'hui, beaucoup d'entreprises font le choix de déplacer leurs infrastructures dans le cloud, afin de bénéficier de plus de flexibilité mais aussi pour réduire les coûts, souvent très élevés.

Lorsqu'on parle de Cloud, on entend très souvent parler de termes tels que IaaS, PaaS et SaaS (et bien d'autres « as-a-Service ») dont voici les définitions :

IaaS

(Infrastructure-as-a-Service)

Pour résumer/simplifier, il s'agit de recourir à du matériel virtualisé (par exemple, une machine virtuelle). Une fois une machine virtuelle disponible, le client devra se charger d'installer les logiciels/composants adéquats. Le fournisseur gère donc la partie infrastructure pour le compte des clients.

PaaS

(Platform-as-a-Service)

Pour ce type d'offre dans un contexte Data, le fournisseur met à disposition une plateforme qu'il va administrer et les clients se chargeront d'intégrer et gérer leurs données. A titre d'exemple, les fournisseurs proposent des bases de données managées tel qu'Azure SQL Database chez Microsoft ou encore Cloud SQL chez Google.

SaaS

(Software-as-a-Service)

Ce type d'offre a pour objectif de mettre à disposition un logiciel clé-en-main intégralement géré par le fournisseur. Un des exemples connus sur le marché est la solution CRM Salesforce.

Dans cet article, nous nous concentrerons sur Google Cloud Platform (GCP) et nous expliquerons comment cette plateforme aboutie permet de moderniser les plateformes Big Data basées sur les technologies Hadoop/Spark.

Les services Data proposés par Google Cloud

Présentation Cloud, GCP et des services Data

Google Cloud propose à ses clients une plateforme de données unifiées entièrement gérées permettant d'obtenir des insights métier tout en innovant rapidement et en maîtrisant les coûts. En utilisant les services **Google Cloud**, les clients peuvent :

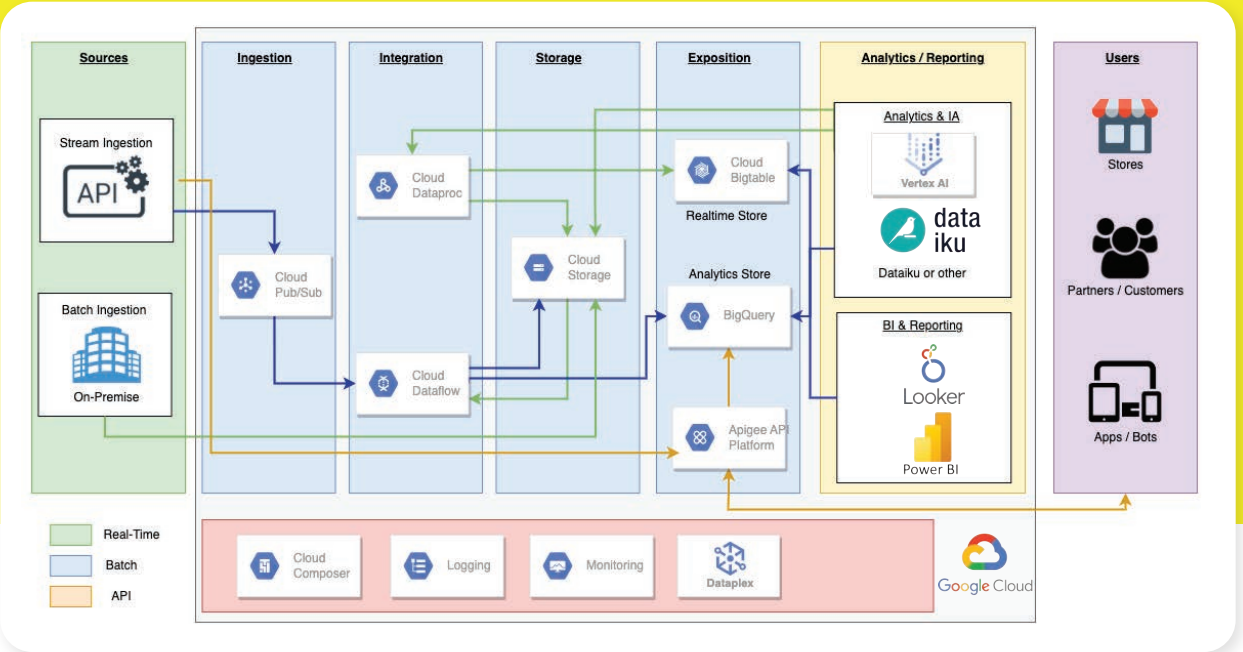
- Accélérer le processus de déploiement ;
- Optimiser les coûts en ne payant que la consommation des services utilisés ;
- Innover en choisissant à la carte des services managés (PaaS) ou encore même « serverless » ;
- Ne plus avoir à gérer l'infrastructure sous-jacente puisqu'entièrement à la charge de **Google Cloud** ;
- S'adapter à l'évolution des usages à la hausse comme à la baisse et dans une logique de maîtrise des coûts.

Google Cloud propose un ensemble de services permettant de répondre aux besoins des clients sur l'ensemble des thématiques autour de la Data.

Le tableau ci-dessous représente un ensemble non-exhaustif de services **Google Cloud** :

STOCKAGE	
Cloud Storage	Service de stockage objet, multi-class et multi-région
Cloud Firestore	Serveur NFS managé (système de fichiers réseaux)
DATABASE	
Cloud BigTable	Base de données non relationnelle, à très faible latence et scalable avec des pétabytes de données
Cloud Spanner	Base de données relationnelle scalable horizontalement
Cloud SQL	Base de données managées pour MySQL, PostGre, SQL Server
DATA & ANALYTICS	
Big Query	Datawarehouse analytics
Cloud Dataflow	Service de traitement des données, en batch ou en streaming
Cloud Dataproc	Service Hadoop/Spark managé
Cloud Data Fusion	Gérer graphiquement les pipelines de données
Cloud Composer	Service managé d'orchestration des workflows
Cloud Pub/Sub	Service de messaging temps réel
DataPlex	Service de gestion de métadonnées et de découverte de données entièrement géré et hautement évolutif
Data Studio	Exploration/Dashboarding des données collaboratives
Looker	BI et Analytics d'entreprise

Schéma d'architecture cible d'une plateforme data dans le cloud.



Source Keyrus : Exemple de schéma d'architecture cible d'une plateforme Data

Pour expliquer l'architecture dans les grandes lignes, vous trouverez ci-dessous des explications sur les étapes majeures du cycle de vie des données :

1 Les données sont intégrées en mode batch (fichiers dans **Google Cloud** Storage) ou en Streaming via Pub/Sub. La brique Apigee nous servira pour l'exposition des données à des tiers (partenaires, clients, etc), tout en pouvant les monétiser.

2 Elles sont ensuite transformées à l'aide d'outils tels que Cloud DataProc ou Cloud Dataflow

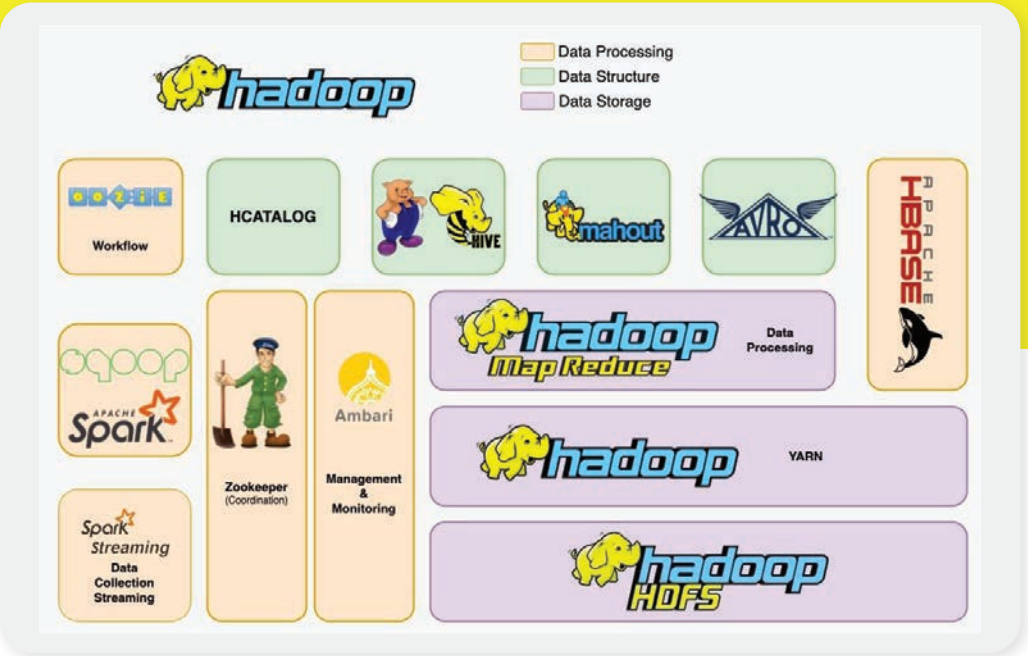
3 Les données sont ensuite chargées dans l'entrepôt de données BigQuery (serverless) pour des usages analytiques ou bien dans BigTable pour des usages nécessitant l'ingestion de gros volumes de données avec des latences inférieures à 10ms

4 Enfin, les données sont disponibles pour des usages BI/Reporting avec Looker (ou avec des produits équivalents : PowerBI, Qlik, Tableau) et Data Science/IA avec la plateforme Vertex AI (ou autres : Dataiku, Databricks, Open Source,etc...)

Les plateformes Big Data basées sur Hadoop : Quelle situation à date et pourquoi moderniser ?

Technologies & Architecture

Le schéma ci-dessous décrit une architecture Big Data basée sur la pile technologique Hadoop/Spark.



Cette architecture est dite « classique » puisqu'elle est très courante au sein des Systèmes d'Information Data (SID) des entreprises qui ont pris très tôt le virage du Big Data (période 2014/2015 jusqu'en 2018/2019). Dans la majorité des cas, elle repose sur le déploiement d'une plateforme proposée par des éditeurs spécialisés se chargeant de consolider les différents projets Open Source et surtout de simplifier son administration et son exploitation. Historiquement, le marché était « trusté » par 2 voire 3 éditeurs : Cloudera, Hortonworks et MapR.

Constats

Depuis plusieurs mois et encore plus aujourd'hui, plusieurs « points durs » amènent les entreprises à repenser leur stratégie technologique orientée Big Data et de manière générale de gestion et de valorisation des données :

- La complexité élevée de mise en place et d'exploitation d'une architecture Big Data « on-premise » basée sur des distributions Hadoop ;
- Une capacité à monter en charge rapidement (scalabilité) complexe et limitée liée à l'infrastructure « on-premise » et aux activités de déploiement/configuration requises pour déployer des capacités additionnelles ;
- Des difficultés à obtenir un niveau de performance acceptable pour les utilisateurs sur les usages analytiques ;
- Une consolidation du marché des éditeurs de distribution Hadoop (abandon par IBM de sa distribution « maison », rachat de MapR – alors en grande difficulté financière – par HPE, fusion de Hortonworks et de Cloudera avec une prise de pouvoir de la part de ce dernier,etc...) laissant les entreprises avec Cloudera comme quasi seul acteur spécialisé de ce marché. Les clients historiques se retrouvent sans alternative et dépendent de la stratégie technologique et tarifaire de Cloudera ;
- Une innovation bridée par les principes technologiques des architectures Hadoop et cela y compris pour les plateformes portées dans le Cloud ;
- Des savoir-faire et une expertise technique de plus en plus complexes à maintenir et à recruter du fait de la transformation du marché des compétences liées à l'avènement du Cloud.

Dans un contexte Data, pourquoi aller vers le cloud et quels sont les apports ?

Là où les entreprises investissaient auparavant sur des infrastructures physiques, elles s'orientent maintenant majoritairement vers le Cloud. Les offres proposées permettront aux entreprises d'avoir des technologies et des solutions plus modernes, plus simples et surtout à la carte avec les bénéfices suivants :

- Disposer de gains sur les coûts à iso périmètre technologique (1er niveau de gain en mode « as-is » avant d'éventuels chantiers de modernisation) ;
- Accéder à un catalogue de services proposant une Data Platform complète et/ou des solutions spécialisées en gestion et valorisation de données ;
- Être en capacité de traiter des cas d'usages analytiques de manière plus rapide et efficace avec une meilleure performance, le tout pour un coût optimisé ;
- Minimiser l'impact de la migration sur les métiers en réduisant au maximum les interruptions des services ;
- Minimiser le coût de la migration en utilisant des services compatibles avec une approche « Lift and Shift ».

Pourquoi GCP ?

Une des principales raisons de choisir **Google Cloud** comme fournisseur de services Data est le fait que de nombreux services proposés reposent sur des projets Open Source qui sont optimisés et managés par Google.
Dans un contexte de « Move-to-Cloud » d'existants Big Data/Hadoop « on-premise », les services proposés par GCP permettent de sécuriser et de minimiser l'effort d'une migration reposant sur une logique « Lift and Shift ». Une pile technologique adaptée à ce contexte pourrait par exemple reposer sur les services suivants : Dataproc (Hadoop/Spark), Google Composer (Apache Airflow), Google DataFlow (Apache Beam), etc...

Google Cloud offre des avantages permettant de répondre aux enjeux exprimés précédemment dont l'optimisation du temps et des coûts, à savoir :

- Aucune gestion de serveurs physiques (pour les services managés) ;
- Gestion des versions des systèmes et des outils à jour supervisées par Google ;
- Réduction de l'effort d'administration des clusters distribués ;
- Réduction de la complexité liée à la maintenance de la plateforme par les équipes **Google Cloud** ;
- Réduction du délai d'approvisionnement d'environnements techniques ;
- Haut niveau de compatibilité entre les produits Hadoop et les services GCP ;
- Reprise simplifiée d'une architecture de type Data Lake ;
- Facilité et simplicité de mise à disposition de nouveaux services ;
- Réduction du « Time-To-Market » ;
- Capacité à monter en charge et amélioration des performances des environnements ;
- Réduction de l'impact sur la conduite du changement.

Stratégie de « move-to-cloud » et de modernisation

Pour réussir une migration vers le Cloud, un plan de migration doit être défini afin de lister les tâches devant être déroulées dans le temps et de manière progressive, en commençant par une approche « Lift and Shift » puis en modernisant l'existant migré.

Mapping des services « on-premise » VS Google Cloud

Catégorie	Type	Hadoop On-Premise	Google Cloud
Data Integration	Batch	Scoop	Google Dataproc
	Streaming	Kafka	Google Pub/Sub
Data Storage	Objet	HDFS	Google Cloud Storage
	NoSQL	HBase	Google BigTable
Data Processing	Batch	Spark	Google Dataproc Google Dataflow
	Streaming	Spark Streaming	Google Dataproc Google Dataflow
Servir/Exposer	Datawarehouse	Hive	Google BigQuery
Administration	Catalogue de données	Data Catalogue Atlas	Google DataPlex
	Sécurité	Ranger	IAM Google Dataproc Ranger
	Orchestration	Oozie	Google Composer

Migration en mode « Lift and Shift »

Le chemin le plus rapide et avec le moins d'impact sur l'architecture existante est de recourir à une approche dite « Lift and Shift ». Elle consiste à porter l'existant sur **Google Cloud** en utilisant les services avec le plus haut niveau de compatibilité (cf. tableau de mapping présenté précédemment) afin de limiter au maximum les activités de refonte/réécriture.

Quelques recommandations :

- Les données hébergées sur HDFS peuvent techniquement être migrées sur le composant HDFS de DataProc. Néanmoins, cette solution n'est ni efficace ni flexible et, de surcroît, potentiellement très coûteuse. C'est pourquoi Google recommande une stratégie de stockage « off-cluster », c'est-à-dire stocker les données sur **Google Cloud** Storage. Au-delà de l'optimisation financière, un des objectifs sera de décorréler la capacité à monter en charge entre le stockage (GCS) et le processing (Dataproc) ;

- Les processus Hadoop/Spark sont à porter sur Dataproc qui est, pour rappel, la distribution Hadoop managée proposée par GCP. L'effort de portage devrait être minime dans un contexte classique d'usage Spark.

- Le service BigQuery est le service à privilégier pour remplacer Hive (ou d'autres services de DataWareHouse). Il est flexible, serverless (c'est-à-dire totalement géré par Google) et supporte le langage SQL.
Totalement adapté pour des usages analytiques, il est simple d'utilisation, rapide et surtout très fiable. Le service BigTable est quant à lui, le service idéal pour remplacer HBase au sein de **Google Cloud**. Il fournit les mêmes fonctionnalités et expose les mêmes API que HBase ;

- Pour des besoins de temps réel reposant sur Kafka, le service PubSub est la solution recommandée par Google ;

- En termes d'administration et d'exploitation, les services de management « on-premise » (Oozie, Ranger,...) peuvent être remplacés par des services managés tels que IAM et Composer (en rajoutant aussi Google Data Catalog).

Modernisation et perspectives d'innovation

Une fois l'existant migré sur GCP en mode « Lift and Shift », il est primordial de penser à faire évoluer l'architecture. L'objectif est d'améliorer la qualité de services, simplifier les nouveaux développements et leur maintenance et cela sans oublier l'optimisation de la consommation et donc améliorer le coût de possession de l'architecture globale.

A titre d'exemple, il pourrait être pertinent de réécrire tout ou partie des « Data Pipelines » pour s'adapter à l'architecture de référence proposée par Google (Figure 1).

Cela reviendrait à utiliser les services suivants :

- Cloud Storage pour le stockage des données non structurées ;
- Dataflow (qui est un service serverless) pour le traitement des données en mode batch et/ou streaming ;
- BigQuery comme socle de données orienté valorisation.

L'effort principal consistera à réécrire les processus Hadoop/Spark en processus Dataflow.

En conclusion, **Google Cloud** est, de par son éventail de service data, un parfait candidat pour la modernisation de vos plateformes traditionnelles Data.

Il permettra notamment de larges possibilités (lift&shift, refactoring, etc.) pour permettre les migrations / modernisations de vos plateformes traditionnelles. La plupart des services Data de **Google Cloud** sont basées sur de l'open source, ce qui permettra de gérer plus facilement les migrations dans le cloud, mais aussi la montée en compétence des data engineers.

Le regard des kommunautés

Google Cloud