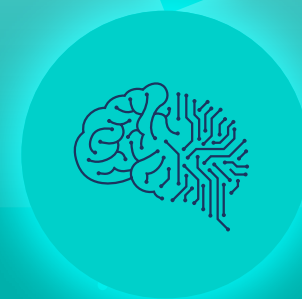
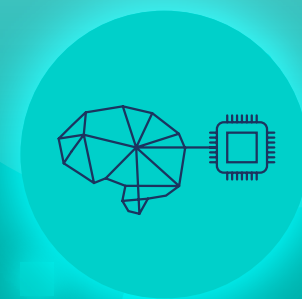


Guide comparatif des Data Lakes dans le Cloud

AWS, Azure, Google, Cloudera, Databricks et Snowflake



Les Data Lakes migrent vers le Cloud.

Initialement créés pour surmonter les limites des data warehouses traditionnels, les data lakes optent pour l'évolutivité, la rapidité et la rentabilité afin de vous aider à gérer d'importants volumes de données hétérogènes dans le cadre de multiples initiatives d'analytique : IA, machine learning, analytique en streaming, BI, et bien plus encore.

Si les premières solutions sur site, telles que Hadoop, ont ouvert la voie aux data lakes et créé un cadre d'architecture initial, les fournisseurs de plateformes cloud et autres pionniers du cloud natif ont largement élargi le spectre des possibilités. Aujourd'hui, avec l'envolée des data lakes dans le cloud, le moment n'a jamais été aussi propice pour tirer parti d'avantages exclusifs : vous pouvez à présent éviter les coûts élevés en amont pour configurer et assurer la maintenance de votre data lake, et vous concentrer sur l'exploitation de vos données pour en extraire la plus grande valeur.

Pourtant, le choix de la solution idéale peut s'avérer plus que difficile, en raison du nombre important de fournisseurs de data lakes dans le cloud présents sur le marché, et du large éventail de solutions disponibles, des plus classiques aux plus récentes, telle que le lakehouse. C'est pourquoi cet e-book vous propose de mieux comprendre les principales différences qui rendent ces plateformes uniques, afin de vous aider à prendre la décision la plus pertinente en fonction de vos besoins.

LIMITES DES DATA LAKES SUR SITE

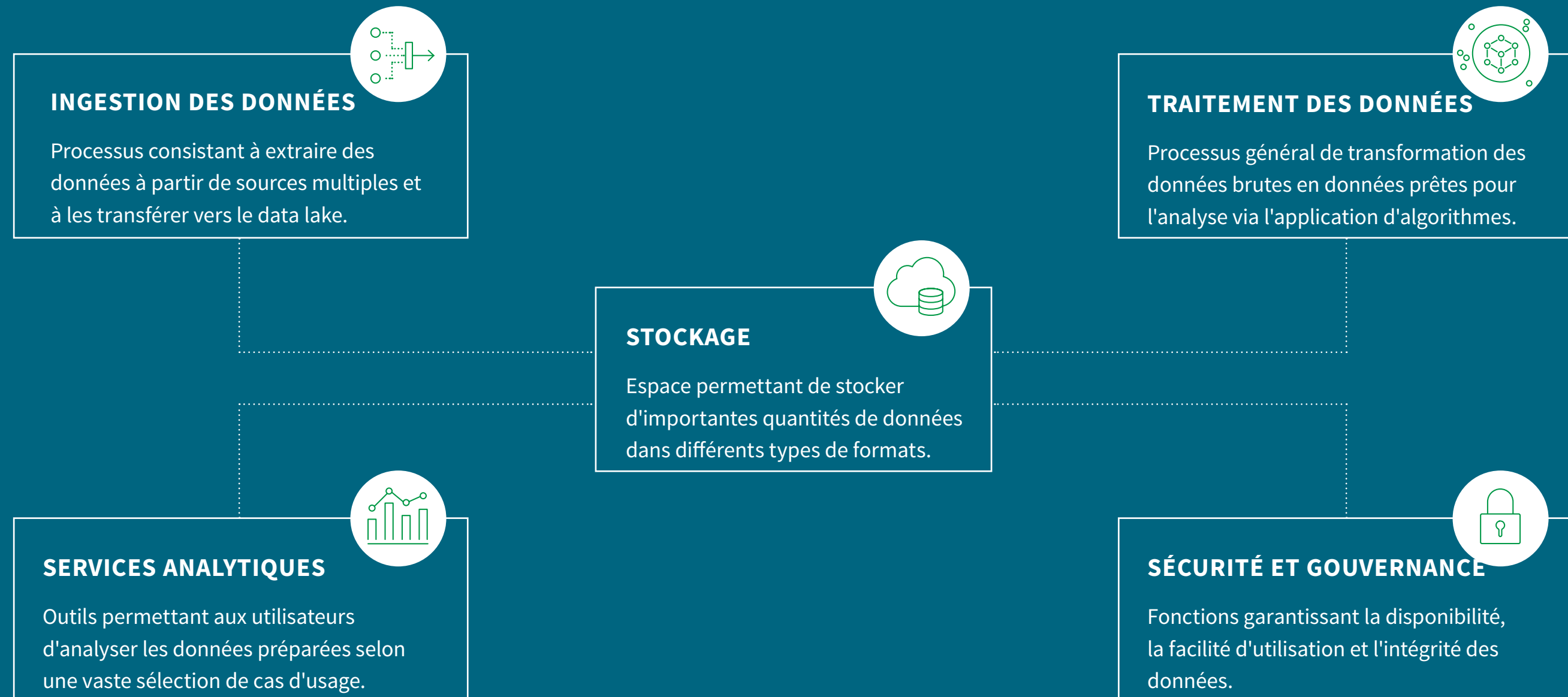
- ✗ Manque de flexibilité
- ✗ Sécurité et gouvernance insuffisantes
- ✗ Coûts de maintenance élevés

AVANTAGES DES DATA LAKES DANS LE CLOUD

- ✓ Stockage et calcul dissociés
- ✓ Sécurité et chiffage intégrés
- ✓ Mise à l'échelle transparente
- ✓ Infrastructure flexible à la demande
- ✓ Tarification basée sur la consommation

Capacités incontournables du Data Lake dans le Cloud

Chaque fournisseur de data lake dans le cloud propose des caractéristiques, des fonctionnalités et des capacités uniques à votre entreprise. Par essence, tous les data lakes offrent plusieurs éléments clés, dont les attributs varient en fonction des fournisseurs.



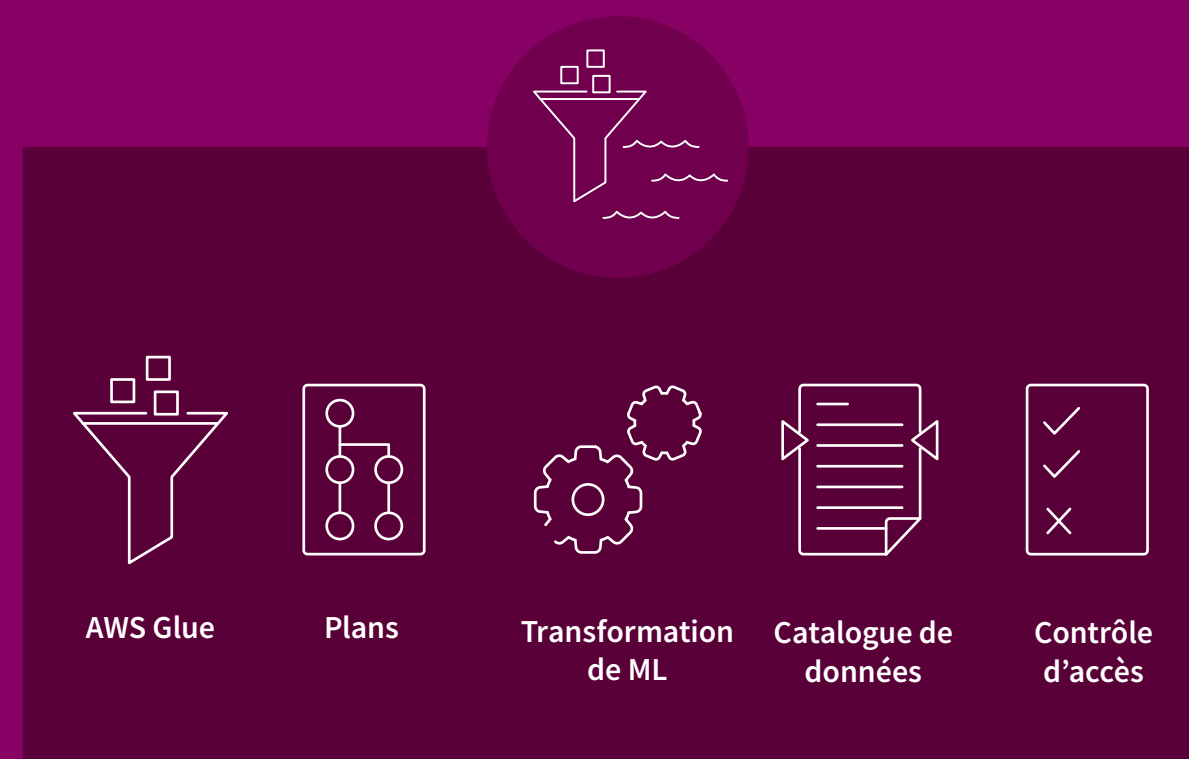
AWS Data Lake

Amazon Web Services (AWS) offre de multiples services pour développer des data lakes sécurisés, flexibles et économiques. Voici les principaux services facilitant la création de data lakes sur AWS :

- Amazon Simple Storage Service (S3), qui fournit un stockage à usage général. Parfois, Amazon DynamoDB, une base de données NoSQL, peut également être utilisée pour stocker des données à faible latence, telles que les flux de clics ou les données IoT.
- Amazon Elastic MapReduce (EMR), le moteur de traitement basé sur des outils open source (par ex. Apache Spark, Apache Hive, Presto), qui automatise le traitement par lots ou en streaming des données

AWS fournit de nombreux services Web (par ex. Kinesis Stream, Kinesis Firehose, Database Migration Service [DMS]), de même que des solutions partenaires qui facilitent l'ingestion et la migration des données du cloud et des sources sur site vers le service de stockage Amazon S3.

Par ailleurs, AWS offre plusieurs services d'analytique entièrement gérés tels qu'Elasticsearch et Athena pour simplifier l'analyse des données du journal et exécuter des requêtes interactives.



AWS Lake Formation

Pour vous aider à bâtir des data lakes plus facilement, le service d'Amazon, AWS Lake Formation, a été spécialement conçu pour automatiser leur création et leur configuration dans Amazon S3.

Malgré ses multiples capacités, AWS Lake Formation est essentiellement constitué d'AWS Glue, un service de catalogage et ETL sans serveur d'Amazon, qui aide les utilisateurs à rechercher, enregistrer et fusionner les données.

Principalement axé sur l'accès aux données et la sécurité, Lake Formation propose une gestion des autorisations précise en plus du service AWS Identity and Access Management (IAM) d'Amazon S3.

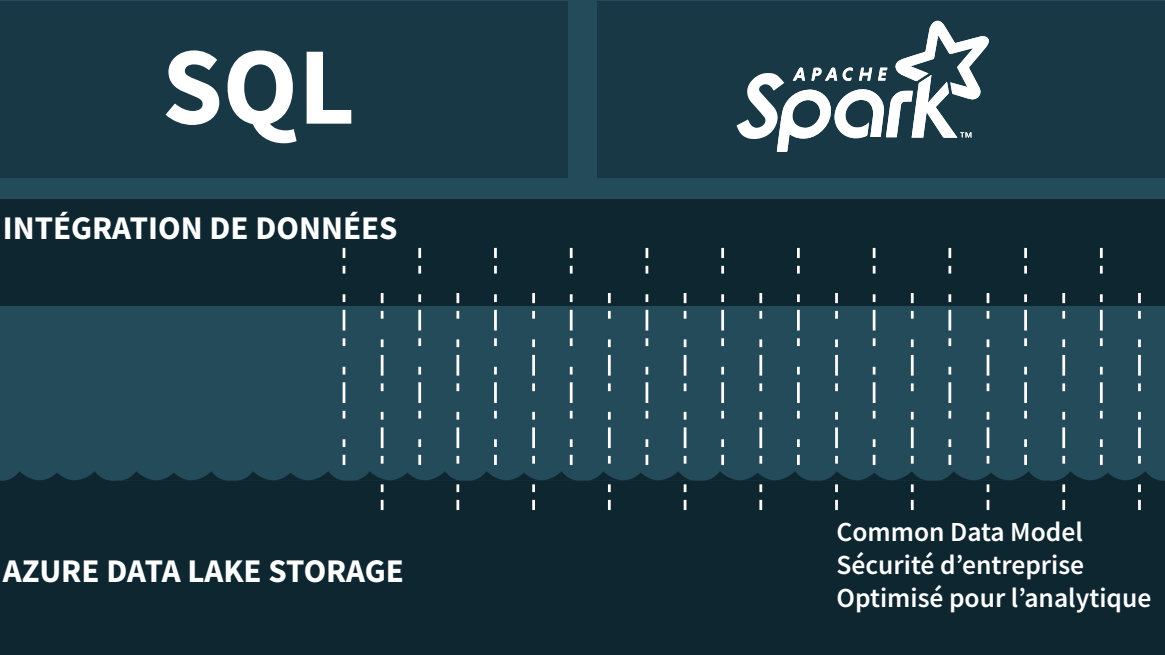
Azure Data Lake

Partie intégrante de la plateforme cloud Microsoft Azure, Azure Data Lake fournit un stockage évolutif et permet d'effectuer tous types de traitements et d'analyses sur de multiples plateformes, dans divers langages de programmation.

Voici ses principaux services :

- Azure Data Lake Storage (ADLS) Gen2, qui combine le stockage du système de fichiers d'ADLS Gen 1 avec le stockage BLOB (Binary Large Object), afin de fournir une évolutivité accrue, des performances optimisées pour gérer les charges de travail d'analyses, et des coûts avantageux.
- Azure HDInsight, un service géré basé sur des outils open source et Azure Synapse, qui combine les requêtes SQL et le traitement de données à grande échelle basé sur Apache Spark.
- Azure Data Lake Analytics, une plateforme à la demande qui vous permet de développer votre propre code tout en garantissant une prise en charge multi-langage, y compris U-SQL, R, Python et .NET

Azure Data Lake comprend des fonctionnalités de récupération d'urgence et s'intègre à d'autres services Azure, tels qu'Azure Active Directory, afin de fournir des contrôles d'accès basés sur les rôles et des capacités d'authentification unique. Vous pouvez également étendre vos contrôles de sécurité sur site avec l'environnement cloud Azure.



Azure Synapse Analytics

Azure Synapse Analytics est le service de Microsoft qui fournit une architecture lakehouse, autrement dit une approche hybride de plus en plus populaire qui rassemble les constructions de data lake et de data warehouse.

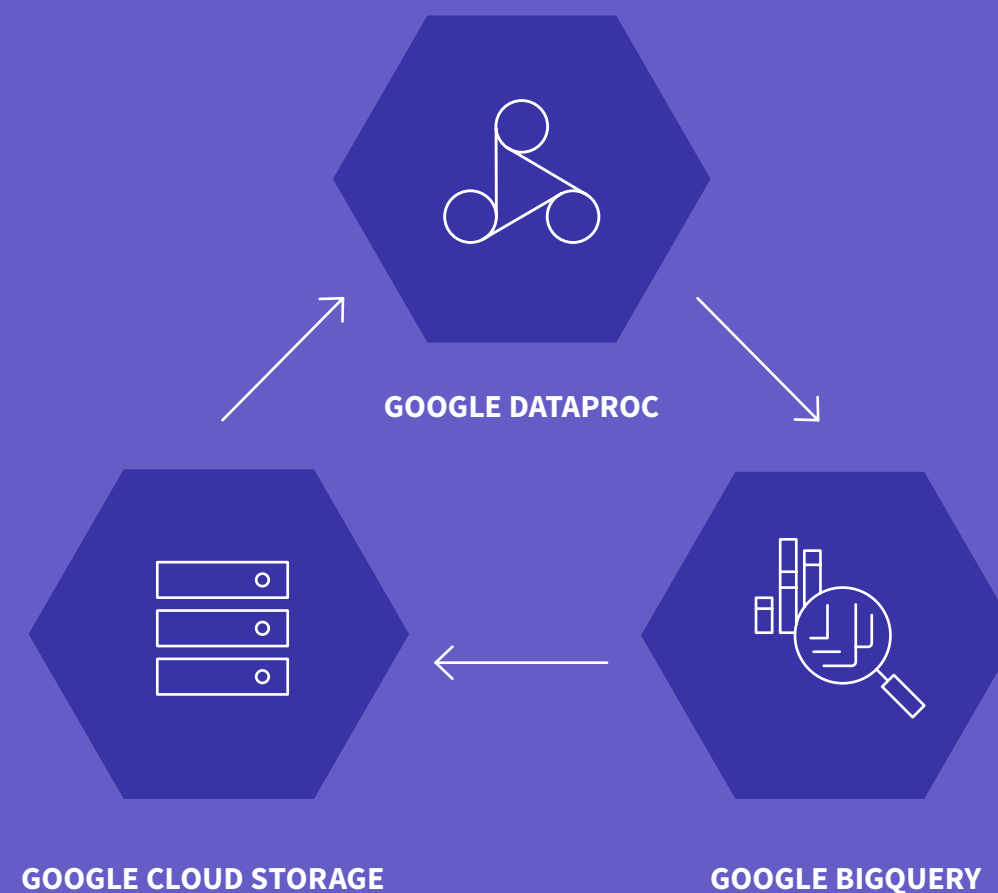
Basé sur ADLS Gen 2, Azure Synapse réunit un moteur SQL et Apache Spark au sein d'une seule et même plateforme afin de vous aider à traiter et à interroger d'importants volumes de données.

Google Data Lake

Google Cloud Platform (GCP) propose son propre data lake pour vous aider à ingérer, stocker et analyser d'importants volumes de données hétérogènes en toute sécurité. Parfaitement intégré aux autres services GCP, Google Data Lake vous propose les éléments clés suivants :

- Google Cloud Storage (GCS), un service de stockage à usage général qui livre une option économique pour des entreprises de toutes tailles
- Google Dataproc, un service entièrement géré basé sur des outils open source (par ex. Apache Hive, Apache Spark), qui traite et analyse des jeux de données dans le cloud. En outre, le service de data warehouse sans serveur de Google, Google BigQuery, permet aux utilisateurs d'exécuter des requêtes natives sur les données GCS pour profiter de fonctionnalités semblables à celles du lakehouse.

Google fournit de nombreux autres services qui s'intègrent nativement à GCS. Pour l'ingestion et la migration de données en temps réel et stockées, Google offre des outils tels que Pub/Sub, Transfer Services, et Transfer Appliance. Pour le traitement et l'analyse de données, Dataflow facilite le traitement par lots ou en temps réel des données, tandis que Cloud Datalab, permet d'explorer, d'analyser et de visualiser les données, et d'utiliser le machine learning.



Google BigQuery

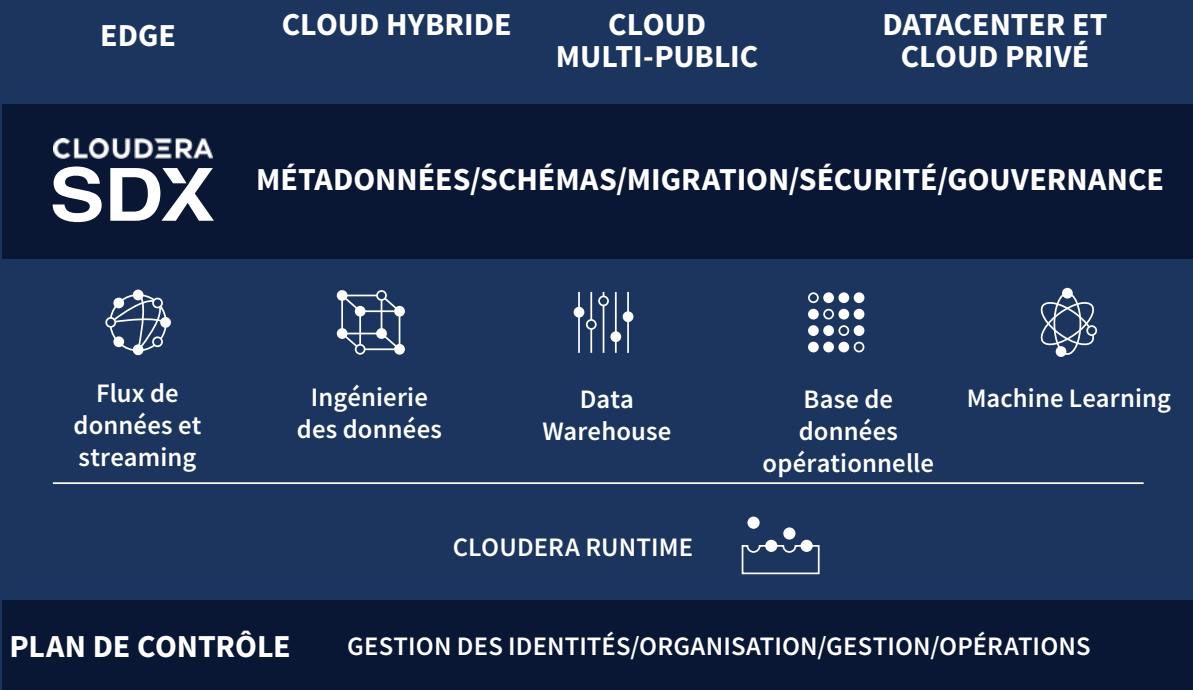
Non seulement Google BigQuery offre aux utilisateurs de SQL la possibilité d'effectuer des requêtes natives de hautes performances pour les données stockées dans GCS, mais il se marie également à la perfection avec Google Data Lake.

Grâce à un transfert gratuit des données entre GCS et Google BigQuery, et à un modèle de sécurité compatible, Google livre aux utilisateurs un accès cohérent aux deux services. Sans compter que les données transférées de GCS vers BigQuery sont directement enregistrées au sein d'un catalogue de données, automatisant ainsi une étape auparavant manuelle.

Cloudera Data Platform (CDP)

Cloudera Data Platform (CDP) est une plateforme de données garantissant une interopérabilité entre les clouds qui vous permet de gérer votre infrastructure, vos données et vos charges de travail d'analyses quel que soit l'environnement que votre entreprise utilise : public, privé, hybride et multi-cloud. CDP rassemble les capacités de Cloudera et Hortonworks, en positionnant Cloudera sur le marché du lakehouse en proposant des services de data lake et de data warehouse au sein d'une seule et même plateforme. Voici les principaux éléments et services de CDP :

- Data Hub, un service gérant les charges de travail qui vous permet de déployer un cluster complet dans le cloud en quelques clics seulement, sans intervention manuelle
- Shared Data Experience (SDX), qui vous aide à consolider toutes vos données dans un emplacement unique afin de les partager en toute sécurité parmi les différentes équipes et multiples services
- Services d'analytique en libre-service couvrant les cas d'usage des data warehouses et du machine learning
- Une console de gestion qui vous permet de gérer, surveiller et organiser les utilisateurs et les services de manière centralisée au sein de multiples environnements grâce à une interface unique



Cloudera Data Platform

CDP comprend des services de data lakes et de data warehouses, de même que des outils d'analytique qui vous permettent de gérer simultanément plusieurs charges de travail d'analyses.

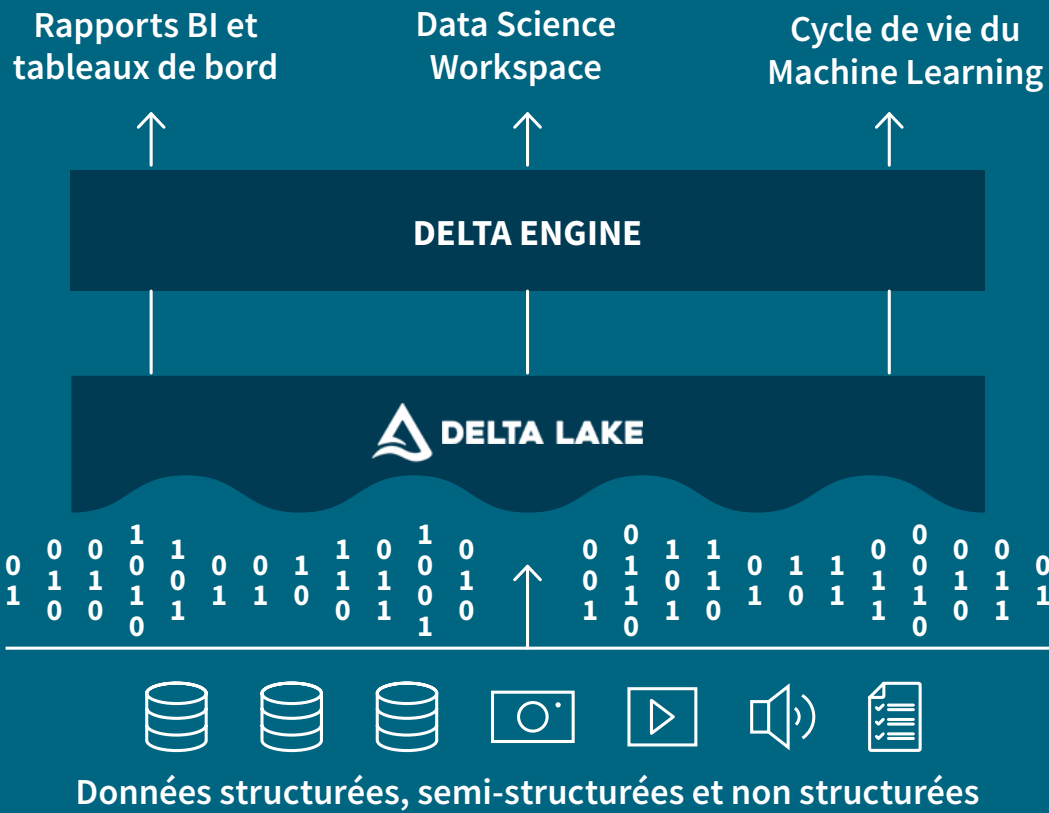
Si grâce au service de data lakes de CDP vous pouvez créer des data lakes sécurisés et gouvernés, et partager ces données dans tous les services et pour toutes les charges de travail, le service de data warehouses est quant à lui automatiquement isolé, pour que les utilisateurs n'aient accès qu'aux données appropriées.

Databricks Unified Analytics Platform

Initialement centrée sur la modernisation des data lakes, Databricks se positionne désormais comme un lakehouse : une plateforme unifiée et ouverte conçue pour stocker et gérer toutes vos données afin de répondre à l'ensemble des besoins de votre entreprise en matière d'analytique. La plateforme multi-cloud, disponible sur AWS, Azure, et GCP, comprend les éléments clés suivants :

- Delta Lake, un centre de stockage open source qui complète votre data lake existant sur votre plateforme cloud préférée, éliminant ainsi le besoin de modifier votre architecture actuelle. Axé sur la fiabilité des données, Delta Lake propose des fonctionnalités de transactions ACID et de déclaration du schéma, ainsi qu'une compression et une optimisation automatiques afin d'améliorer l'exactitude et la performance du data lake.
- Delta Engine, un moteur de requêtes compatible avec Apache Spark qui traite les données dans Delta Lake. Contrairement à Spark, Delta Engine est optimisé pour les données du lakehouse et prend en charge une variété de charges de travail, du traitement ETL à grande échelle aux requêtes interactives ad-hoc.

Par ailleurs, Databricks offre une prise en charge native pour une variété de langages de programmation courants, y compris R et Python, de même qu'une plateforme collaborative de science des données et de machine learning.



SQL Analytics de Databricks

Le service SQL Analytics de Databricks est l'ultime outil proposé par l'entreprise pour s'affirmer en tant que fournisseur de lakehouse : une seule et même plateforme pour toutes les initiatives d'analytique. Conçu pour prendre en charge vos besoins uniques de BI et de reporting, ce service offre aux utilisateurs de SQL une interface familière pour interroger les données et créer des tableaux de bord en toute simplicité.

Snowflake Cloud Data Platform

Mieux connu comme le data warehouse cloud, Snowflake a progressivement brouillé les lignes entre le data lake et le data warehouse. Bâti sur une plateforme flexible, Snowflake fournit l'évolutivité, la flexibilité et le stockage économique d'un lac de données avec la sécurité, la gouvernance et les performances d'un data warehouse.

Disponible sur AWS, Azure et GCP, Snowflake vous permet de transférer un vaste ensemble de données natives, sans avoir à les transformer pour bénéficier de la flexibilité et de l'agilité d'un data lake. Les utilisateurs peuvent également tirer parti de l'architecture MPP de Snowflake pour activer de multiples data warehouses virtuels et exécuter plusieurs requêtes en même temps.

Snowflake vous permet également de partager des données avec des outils partenaires tels qu'Apache Spark, avec les pilotes ODBC et JDBC pour un traitement des données à grande échelle en temps réel.



Snowpark

Snowpark, le tout nouvel outil pour développeurs de Snowflake, s'inscrit dans la continuité de leur approche orientée lakehouse pour permettre aux data scientists, ingénieurs, et programmeurs de créer et déployer un code personnalisé sur Snowflake dans divers langages de programmation, y compris Java, Scala, et Python.

Aperçu des Data Lakes dans le Cloud

CAPACITÉS		 Microsoft Azure	 Google Cloud Platform	CLLOUDERA	 databricks	 snowflake
Service de stockage principal	Amazon S3	ADLS Gen 2	Google Cloud Storage	Cloudera Data Platform	Data Lake sur AWS, GCS ou ADLS	Snowflake Cloud Data Platform
Moteur de traitement	Amazon EMR	Azure HDInsight, Azure Synapse	Google Dataproc, Data Flow	CDP Data Engineering	Delta Engine	Snowflake
Support Apache Hive et Spark	Oui	Oui	Oui	Oui	Spark	N/A
Stockage et calcul dissociés	Oui	Oui	Oui	Oui	Oui	Oui
Service de pipeline	AWS Glue	Azure Data Factory	Cloud Data Fusion, Data Flow	Cloudera Data Engineering	Delta Engine	Snowpark
Support SQL	Amazon Athena, Redshift, Spectrum	Azure Synapse	Google BigQuery	Services d'analytique en libre-service pour Data Warehouse	Service SQL Analytics	Snowflake
Prise en charge de multiples langages de programmation	Oui	Oui	Oui	Oui	Oui	Oui
Catalogue	AWS Glue	Azure Data Catalog	Google Data Catalog	Cloudera Data Platform	Solutions partenaires	Solutions partenaires
Architecture Lakehouse	Oui	Oui	Oui	Oui	Oui	Oui
Multi-cloud	Non	Non	Non	Oui	Oui	Oui



Les meilleurs data lakes s'appuient sur une intégration optimale des données.

À mesure que vous commencez à migrer vos données dans le cloud, de plus en plus de fournisseurs sont prêts à vous proposer leurs services, avec différentes solutions spécialement conçues pour répondre à vos besoins uniques. Cependant, quelle que soit la plateforme que vous choisissiez, vous devez absolument compter sur une intégration des données robuste, pour acheminer les données partout où vous en avez besoin.

Que ce soit via un lac de données ou un lakehouse, l'intégration des données est nécessaire non seulement pour ingérer et migrer un ensemble de données issues de sources multiples, mais également pour traiter et raffiner ces données de sorte qu'elles soient prêtes à l'emploi pour tous vos cas d'usage d'analytique.

La plateforme d'intégration des données Qlik® automatise le pipeline du data lake pour accélérer et optimiser la livraison de données prêtes pour l'analyse, tout en permettant à vos équipes d'ingénierie de fournir en continu des données actualisées aux data scientists, aux analystes métier et aux autres utilisateurs au sein de votre organisation.

Qlik® pour les Data Lakes dans le Cloud.

La plateforme d'intégration des données de Qlik peut vous aider à accélérer le retour sur investissement de votre data lake dans le cloud en fournissant, en continu, les données précises, rapides et fiables dont vous avez besoin. La plateforme offre une capacité inégalée : l'automatisation des flux de données issues de toutes sources, y compris des mainframes, des applications d'entreprises telles que SAP, des bases de données, des data warehouses et autres, vers votre data lake. Qlik livre également des jeux de données prêts pour l'analyse, sans programmation requise.

Exploitez toute la valeur de vos données en accélérant la livraison de jeux de données prêts pour l'analyse, afin d'optimiser l'efficacité, l'agilité et la flexibilité de vos programmes d'analytique.

Vous voulez voir comment Qlik peut optimiser l'exploitation des données au sein de votre entreprise ?

Commencez par un essai gratuit.

Essai

Qlik automatise le pipeline complet du data lake :



Capture et fournit en temps réel des données actualisées dans le data lake de votre choix



Catalogue et sélectionne les données pour faciliter la recherche et le libre approvisionnement, tout en garantissant leur sécurité et leur gouvernance



Est compatible avec tous les principaux fournisseurs cloud, les points de terminaison source et cible, et les outils d'analyse de votre choix



Normalise, fusionne et raffine les données, avant de les subdiviser en jeux de données prêts pour l'analyse



Organise et simplifie la configuration, le monitoring et la gestion des pipelines de données

Qlik s'est donnée pour objectif la création d'un monde « data literate », où chacun peut exploiter la data et l'analytique pour résoudre les défis les plus complexes. Qlik offre une plateforme cloud de bout en bout d'intégration des données et d'analytique en temps réel, afin de combler l'écart entre les données, les informations et les actions. En transformant les données en Intelligence Active, les entreprises peuvent s'orienter vers de meilleures décisions, améliorer leur chiffre d'affaires et leur rentabilité, et optimiser la relation client. Qlik exerce ses activités dans plus de 100 pays et offre ses services à plus de 50 000 clients à travers le monde.



qlik.com