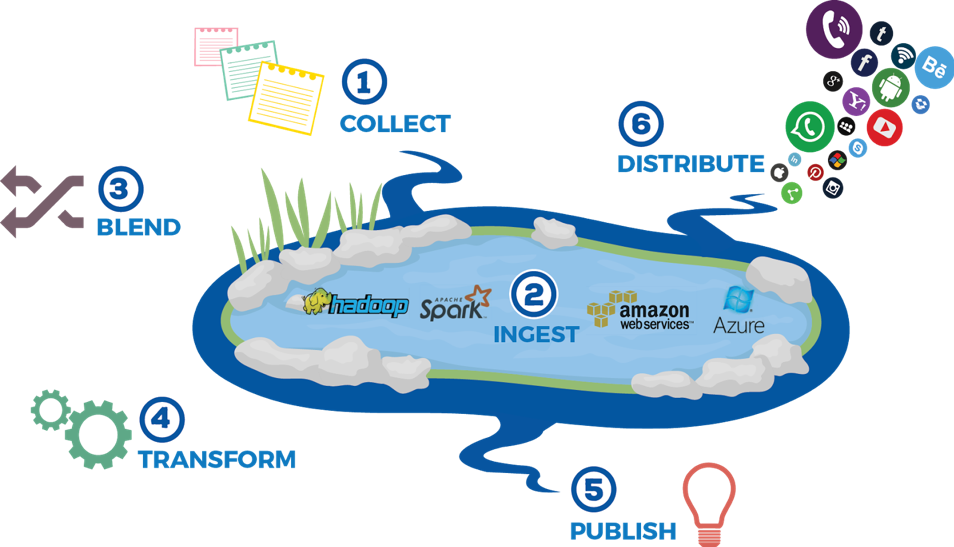
Une image contenant texte

Description générée automatiquement

**DATA LAKE**

**MASTER 2 BI & BD – Réalisé par Justin BENDE**

© 2021



**SOMMAIRE**

**INTRODUCTION**

**PREMIERE PARTIE : Comprendre le Data Lake**

1. Qu’est-ce qu’un Data Lake ?
2. Pourquoi un Data Lake ?
3. Les usages d’un Data Lake
4. Quel coût ?
5. Les avantages et inconvénients du Data Lake
6. Distinguer Data Lake de Datawarehouse
7. Le binôme Data Lake – Big Data
8. Les métadonnées dans les Data Lake
9. La gouvernance des données dans les Data Lake

**DEUXIEME PARTIE : Architecture, composants et couches**

1. Architecture d’un Data Lake
2. Les composants et couche d’un Data Lake
3. Les principes de base à respecter lors de la mise en place d’un Data Lake

### TROISIEME PARTIE : Mise en place d’un data Lake COVID 19 (Azure Data Lake Storage Gen2) avec Azure Data Factory

1. Architecture de la solution Cloud
2. Qu’est-ce qu’Azure Data Factory ?
3. Les solutions de stockages utilisées pour la solution
4. Les sources de données
5. Ingestion des données
6. Transformation des données
7. Publication des données

**INTRODUCTION**

Avec l’avènement des mégadonnées (*big data*), l’informatique décisionnelle a dû trouver des solutions pour gérer des volumes et une variété de données plus grands encore que dans les entrepôts de données, qui se sont révélés mal adaptés. Les lacs de données répondent à ces besoins du point de vue du stockage.

Le concept de **Data Lake** où lac de données a été évoqué pour la première fois, en 2010, par James Dixon, CTO de Penthao, comme une solution pour le stockage de données sans pré-traitement et sans connaître précisément l’usage futur qu’il en sera fait. L'image du lac, permettant d'expliquer, que différentes sources peuvent l'alimenter de manière naturelle et brute, et que les utilisateurs peuvent y plonger pour l'explorer (reporting, visualisation, fouille de données) et en rapporter des échantillons à examiner.

C’est un concept qui s’oppose au processus d’entreposage de données, qui aboutit à une base de données décisionnelles en silos étanches (Stein et Morrison, 2014)

Toutefois, un lac de données doit être bien conçu sous peine de devenir un marécage de données (*data swamp*) inexploitable ; c’est-à-dire qu’il doit permettre le requêtage des données (sélection/restriction) avec un bon temps de réponse et pas seulement leur stockage et leur accès « clé-valeur ».

**PREMIERE PARTIE : Comprendre le Data Lake**

1. **Qu’est-ce qu’un Data Lake ?**

Un Data Lake, ou lac de données, est un dépôt de stockage permettant de stocker de larges quantités de données structurées, non structurées ou semi-structurées. Il est possible d’y entreposer tous les différents types de données dans leur format natif. Comme dans un véritable lac, les données découlent de différentes sources en temps réel.

1. **Pourquoi un Data Lake ?**

On dispose souvent d’une myriade de sources de données applicatives, que l’on souhaite consolider au sein d’une structure plus globale. Cette structure doit permettre de garantir une agilité business, c’est-à-dire de donner aux équipes la possibilité de croiser des données issues de déférentes sources métier pour en tirer de la valeur pour l’entreprise. Le data Lake doit en outre être capable d’ingérer de la donnée avec une composante temps-réel (logs applicatifs, IoT), de stocker cette donnée en vue d’être traitée par la suite pour des besoins analytiques, du reporting, ou tout autre recherche de valeur (machine learning…)

1. **Les usages d’un Data Lake**

D’une manière générale, de nombreux Data Lake voient le jour par des projets de remplacement et d’amélioration des infrastructures data existantes. Les organisations sont motivées par le besoin d’améliorer leur utilisation des données, de centraliser toutes les sources en un seul point et d’accélérer les cycles d’innovation. Les secteurs du marketing et des médias ont été évidement les premiers à saisir cette opportunité bien avant que le terme de data Lake ne popularise cette pratique.

Dans une démarche de Data Management Platform, le data Lake permet de collecter toutes les données issues des interactions avec les clients de raffiner celles-ci pour offrir une vision à 360° sur les clients. Très souvent, ces projets ont vocation à appliquer sur ces données des algorithmes de segmentation, ou de prédiction pour anticiper les comportements des consommateurs. Ils mettent aussi en avant les capacités à assembler et valoriser une grande variété de données. Thomas Faivre-Duboz relève que c’est un réel avantage pour mettre en œuvre un marketing plus rapide, en mode test and learn.

Les usages sont nombreux. Ainsi, l’enseigne Picard a sauté de pas du data lake pour identifier les appétences canal-magasins versus web afin d’optimiser ses campagnes de couponing. La filiale du Crédit Agricole, Sofinco a œuvré à la personnalisation des campagnes RTB pour optimiser les investissement médias

1. **Quel coût ?**

Le coût de la mise en œuvre d’un Data Lake dépend du volume de données à stocker, auquel s’ajoute le coût de setup du flux de données. Il ne faut pas oublier le coût d’une réflexion autour de l’inventaire de données disponibles et l’ordre d’intégration des données à l’outil.

1. **Les avantages et inconvénients du Data Lake**

Le Data Lake est avant tout un moyen de stockage dont les avantages sont :

* Une capacité de stockage de gros volume de données
* Une rapidité de stockage sans pré-traitement des données brutes
* Une souplesse et polyvalence pour stocker différents formats et sources de données.
* Réduction des coûts de stockage
* Réduire le temps passé et les coûts liés à préparer les données avant leur stockage, puisque leur format d’origine est conservé.
* Garder un historique des données anciennes sur de longue durée.
* Rechercher, analyser et corréler l’ensemble des données de l’entreprise pour une meilleur analyse et création de valeur.
* Facilite l’accès pour l’analyse et la prise de décisions d’une façon globale.

Malgré tout, la solution présente aussi certains inconvénients :

* La difficulté à conserver un lac de données propre et organisé
* La difficulté à organiser et maintenir une gouvernance de données efficace.
* Le temps nécessaire à organiser et traiter et analyser les données stockées à l’état brut.
* L’expertise requise pour rechercher, analyser et traiter les données de manière pertinente et créatrice de valeur, souvent confiées aux Data Scientists.
* La sécurité, la confidentialité, problématiques liées aux données personnelles et au respect des réglementations, inhérente aux données, est toujours présente et d’autant plus importante que le lac de données regroupe et conserve des données massives.

1. **Distinguer Data Lake de Datawarehouse**

****

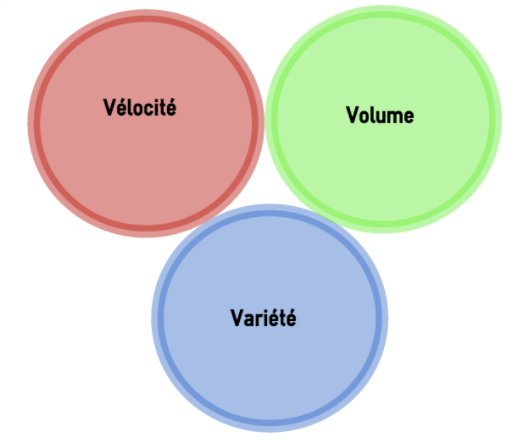
Un « Lac de données » représente un maillon essentiel de la connaissance du client et de son secteur d’activité. En ce sens, il apporte des fonctions complémentaires à celles d’un Datawarehouse qui est un simple entrepôt de données organisées par thèmes, horodatées et structurées. Il est donc parfaitement adapté aux analyses répétitives.

A l’inverse, un Data Lake analyse les données en fonction de besoins exprimés par un service de l’entreprise. Il est en effet possible de charger des informations brutes, et de leur conférer une forme et une structure uniquement lorsque le moment est venu de les exploiter.

|  |  |  |
| --- | --- | --- |
| Caractéristiques | Data Lake | Data Warehouse |
| Typologie de donnée | Intégralité de donnée brute | Donnée préstructurée pour un besoin métier précis |
| Processing | Données brutes et non traitées. Ingestion rapide de nouvelles données | Donnée nettoyée et prétraitée. Difficcultés potentielles à inclure de nouvelles données |
| Type de donnée | Structurée, semi-structurée et non structurée | Structurée (tabulaire) |
| Stockage | Stockage bas-coût | Stockage plus cher mais permettant d’excellents temps de chargement et de reponse |
| Schéma | Schéma inféré à la lecture | Schéma défini à l’écriture |
| Outillage | Majorité d’outils open-source | Majorité d’outils commerciaux |
| Utilisateurs | Les spécialistes des données, les développeurs de bases de données et analystes commerciaux | Analystes métier |
| Analyse | Machine Learning, analyse prédictive, découverte de données et profilage | Rapport de production par lot, BI et visualisation |

1. **Le binôme Data Lake – Big Data**

Les [big data](https://www.futura-sciences.com/tech/dossiers/internet-big-data-boom-donnees-numeriques-1936/page/2/) ou mégadonnées désignent l'ensemble des [données numériques](https://www.futura-sciences.com/tech/dossiers/internet-big-data-boom-donnees-numeriques-1936/) produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Cela regroupe les données d'entreprise (courriels, documents, bases de données, historiques de processeurs métiers...) aussi bien que des données issues de [capteurs](https://www.futura-sciences.com/tech/definitions/technologie-senseur-8460/), des contenus publiés sur le [web](https://www.futura-sciences.com/tech/definitions/internet-internet-3983/) (images, vidéos, sons, textes), des transactions de commerce électronique, des échanges sur les [réseaux sociaux](https://www.futura-sciences.com/tech/definitions/informatique-reseau-social-10255/), des données transmises par les objets connectés (étiquettes électroniques, [compteurs intelligents](https://www.futura-sciences.com/planete/definitions/developpement-durable-compteur-intelligent-6952/), smartphones...), des données géolocalisées, etc. ce concept regroupe un famille d’outils qui répondent à une triple problématique dite **règle des 3V**. il s’agit notamment d’un **Volume** de données considérable à traiter, un grande **Variété** d’information (venant de diverses sources, non structurées, organisées, Open…), et un certain niveau de **Vélocité** à atteindre, autrement dit fréquence de création , collecte et partage de ces données.



Certaines entreprises ajoutent deux autres principes clés pour former les « **5V**». Ces 2 principes sont la **Véracité** qui invoque la nécessité de vérifier la crédibilité de la source et la qualité du contenu afin de pourvoir exploiter ces données et la **Valeur** qui désigne le fait que chaque donnée doit apporter une valeur ajoutée à l’entreprise.

Le domaine des Big data a engendré celui des Data Lake.

1. **Les métadonnées dans les Data Lake**

Les métadonnées sont de l’information structurée qui décrit, explique, localise ou facilite autrement l’obtention, l’utilisation ou la gestion d’une ressource d’information. Les métadonnées sont souvent appelées données des données ou information sur l’information.

Aussi, les vertus de la métadonnée sont multiples, dont entre autres :

***•* L’accès à l’information par des utilisateurs sans culture technique**

**• La qualité des données :**

L’évaluation de la qualité de la donnée est plus facile. Une fois qualifiées, les données n’ont plus besoin d’être passées au tamis pour savoir si elles sont utiles, d’actualité et pertinentes.

***•* La protection des données sensible :**Le règlement RGPD impose un nouveau cadre pour l’utilisation des données personnelles. La cartographie des données personnelles (DCP) et des données sensibles, facilite le travail de protection (chiffrement, gestion des accès…) de ces données et la mise à jour du registre des traitements, demandé par le régulateur.

***•* Une exploitation et une collaboration facilitées :**La traçabilité procurée par les métadonnées apporte une transparence sur les traitements réalisés.

**• La mise au jour de données cachées :** L’entreprise dispose souvent d’un réservoir de données cachées, donc gâchées. Une fois cartographiés avec les métadonnées, ces jeux d’informations retrouvent la lumière et peuvent être exploités par les métiers dans l’entreprise.

1. **La gouvernance des données dans les data Lake**

La gouvernance des données est un ensemble de processus, rôles, règles, normes et métriques permettant d’assurer une utilisation efficace et efficiente des informations, dans le but d’aider les entreprises à atteindre leurs objectifs. Elle définit également qui peut effectuer quelle action, sur quelles données, dans quelle situation et selon quelle méthode.

Une stratégie de gouvernance des données efficace offre de nombreux avantages à une organisation, notamment :

* **Une compréhension commune des données** : la gouvernance des données offre une vue cohérente des données et une terminologie commune, tout en laissant aux unités opérationnelles la flexibilité dont elles ont besoin.
* **Une meilleure qualité des données** : la gouvernance des données crée un plan assurant l’exactitude des données, leur exhaustivité et leur cohérence.
* **Une carte des données** : la gouvernance des données offre une fonctionnalité avancée de compréhension de l’emplacement de toutes les données liées aux entités clés, facteur indispensable de l’intégration des données. Tout comme un GPS permet de représenter une zone géographique et de guider les utilisateurs, la gouvernance des données rend les éléments de données plus faciles à utiliser et à connecter avec des résultats commerciaux.
* **Une vue à 360 degrés de chaque client et des autres entités de l’entreprise** : la gouvernance des données établit un cadre dans lequel l’organisation partage « une seule version de la vérité » et assure la cohérence nécessaire entre toutes les entités et activités de l’entreprise.
* **Une conformité à tous les niveaux** : la gouvernance des données offre une plate-forme permettant de répondre aux exigences réglementaires, telles que le RGPD (Règlement général sur la protection des données) de l’Union européenne ou la loi HIPAA (Health Insurance Portability and Accountability Act) des États-Unis, et aux autres exigences du secteur, telles que les normes PCI DSS (Payment Card Industry Data Security Standards).
* **Une meilleure gestion des données** : la gouvernance des données apporte une dimension humaine dans ce monde des données fortement automatisé. Elle établit des codes de conduite et des bonnes pratiques en matière de gestion des données, afin de répondre aux besoins et inquiétudes dans tous les domaines : données et technologies bien sûr, mais aussi sécurité, conformité et autres aspects juridiques.

**DEUXIEME PARTIE : Architecture, composants et couches**

1. **Architecture du Data Lake**

Une image contenant texte, carte de visite, capture d’écran

Description générée automatiquement

1. **Les composants et couche d’un Data Lake**

Les composants et couche d’un Data Lake sont les suivantes :

* **Ingestion** : représente les connexions aux différentes sources externes et les processus de déversement vers notre Data Lake. Les données peuvent être déversées de manière « batch » ou en « temps réel »
* **Stockage** : un stockage de type S3 ou HDFS, infiniment scalable, est de manière générale très peu couteux et bien adapté à ce besoin tant en termes de données structurées que de données non structurées. C’est un point de chute pour l’ensemble des données au repos
* **Exploration** : Les données stockées doivent pouvoir être requêtées facilement et de manière transparente, de préférence avec des langages déclaratifs usuels de type SQL.
* **ETD (Extract, transform, load)** : l’étape appelée ici ETL est celle qui consiste à convertir de la donnée brute depuis l’étape de stockage vers un état structuré, et ce des fins de traitement ultérieur.
* **Processing** : on souhaite évidemment pouvoir manipuler, requêter et croiser les données, et ce quel que soit la fréquence d’input (batch, live), à des fins d’analyse et d’extraction de valeur ajoutée.
* **Opérations** : on regroupera sous cette appellation tous les services ayant trait à la sécurité, aux performances globales, au monitoring, à l’audit, ou encore au management du catalogue de données. Un DataLake correctement architecturé doit garantir une bonne disponibilité de la donnée – y compris sous la charge – ainsi qu’une bonne sécurité et intégrité via un monitoring et une restriction des accès à un gain très fin. Les opérations regroupent pour cela :
* **La gouvernance** : la découverte de données est une étape importante de l’organisation de notre DataLake et sa structuration générale. Cette étape est cruciale car elle va servir à constituer notre catalogue de données ; on y trouvera des informations aussi cruciales que la source initiale, le format, la sémantique…
* **La Sécurité**: elle doit être implémentée dans chaque couche du DataLake, du stockage(résilience et réplication), à la lecture des données et leur consommation (Authentification et contrôle des accès au datasets par les différentes équipes). Enfin chaque accès doit pouvoir être loggé à des fins d’audit ; un audit du catalogue de données permet d’évaluer la compliance aux différentes régulations ainsi que les risques liés à la manipulation de données sensibles.
* **La qualité :** elleest une composante essentielle de l’architecture de notre datalake. En effet, la donnée sera la source de la valeur métier. Une donnée de mauvaise qualité entrainera à coup sûr une valeur métier moindre.
* **Le lineage** : la couche de Data Lineage doit permettre de « retracer » les évolutions apportées à la donnée par rapport à son ingestion initiale depuis une « gloden source ». En effet, lors d’audit, on cherchera à traquer les changements apportés à la donnée par les différentes applications, et les auteurs de ces changements.

1. **Les principes de base à respecter lors de la mise en place d’un Data Lake**

De manière générales, pour piloter aux mieux les coûts de la mise en place ainsi que les projections de coût à long terme, il convient de se poser la question de la structuration de la donnée, de sa fréquence de mise à jour, mais aussi et surtout ce que l’on souhaite en faire.

Modulariser son Data Lake de manière à séparer les composantes listées ci-dessus. Stockage, ingestion, administration, qualité, visualisation… doivent pouvoir être gérées indépendamment.

Un Data Lake sera différent d’une entreprise à l’autre par la variété des métiers, des approches et des besoins. Les spécificités de l’entreprise et du métier doivent être prises en compte dès la conception.

Un data Lake correctement architecturé doit pouvoir intégrer facilement et rapidement de nouvelles sources de données.

### TROISIEME PARTIE : Mise en place d’un data Lake COVID 19 (Azure Data Lake Storage Gen2) avec Azure Data Factory

1. **Architecture de la solution**

