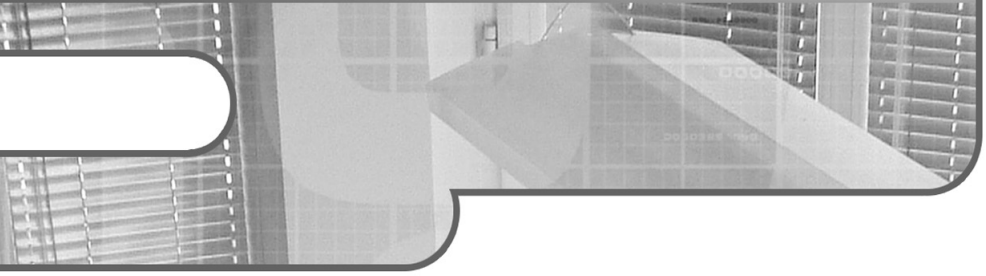


Chapitre 3
Du Lakehouse à la première analyse

1. Évolution des architectures de données

Les données sont désormais vitales pour toutes les organisations qui souhaitent prendre des décisions averties et améliorer ses performances. Cependant, les données ne sont pas statiques, mais dynamiques et en constante évolution. Par conséquent, la manière dont les données sont stockées et traitées doit également s'adapter à l'évolution des besoins des utilisateurs et des défis des organisations de plus en plus axées sur les données. On évoque souvent le terme de « Data Compagnies » (https://go.fabricbook.fr/ch3-1) pour les entreprises pilotées par les données.

Les bases de données de traitement transactionnel en ligne (*Online Transactional Processing*, **OLTP**) sont souvent la source des systèmes de traitement analytique en ligne (*Online Analytical Processing*, **OLAP**). OLTP stocke et met à jour les données transactionnelles à volume élevé de manière fiable et efficace tandis que OLAP combine et regroupe les données afin que vous puissiez les analyser de différents points de vue.

Examinons l'évolution du stockage et du traitement des données, depuis l'entrepôt de données jusqu'au lac de données, et comparons les propriétés, les avantages et les inconvénients de chaque méthode, ainsi que la façon dont ils répondent aux différents besoins de l'analyse des données. Voyons aussi quelques évolutions plus récentes du stockage et du traitement des données dans le concept de Lakehouse.

1.1 Entrepôt de données

Un entrepôt de données, aussi appelé *Data Warehouse* en anglais, est un dépôt centralisé de données structurées, nettoyées, vérifiées… ayant été extraites, transformées et chargées à partir de diverses sources. Ces étapes sont communément appelées **ETL**, ce qui signifie *Extract, Transform, Load* (Extraction, Transformation, Chargement). Cette méthodologie de traitement des données consiste à extraire des données de sources multiples, les transformer pour répondre aux besoins de l'entreprise, puis les charger dans une destination pour l'analyse et la consultation.

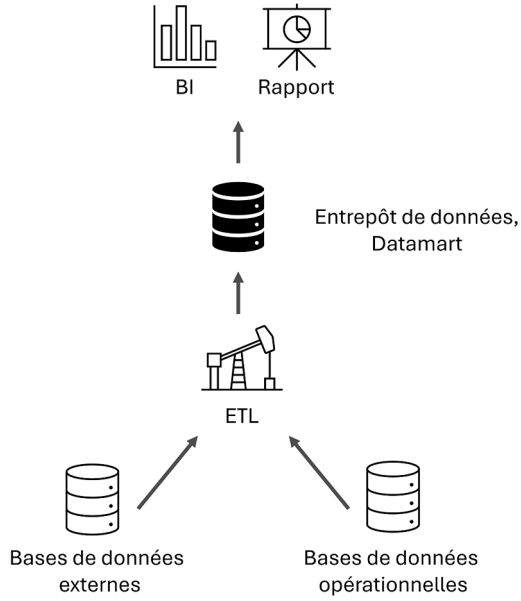
Un entrepôt de données est conçu pour prendre en charge l'informatique décisionnelle (Business Intelligence) et les requêtes analytiques, les rapports, les tableaux de bord et les analyses ad hoc. Les données dans un entrepôt de données sont généralement sauvegardées dans un schéma en étoile (https://go.fabricbook.fr/ch3-2) ou en flocon (https://go.fabricbook.fr/ch3-3) organisant les données en « Faits » et en « Dimensions ». Un entrepôt de données assure également la qualité, la cohérence et la sécurité des données par le biais de divers mécanismes, tels que le nettoyage des données, la validation et le contrôle d'accès.

Le concept d'entrepôt de données a été proposé pour la première fois par Bill Inmon (https://go.fabricbook.fr/ch3-4) dans les années 1980 et popularisé par Ralph Kimball (https://go.fabricbook.fr/ch3-5) dans les années 1990. L'entrepôt de données est devenu le paradigme dominant pour le stockage et le traitement des données pendant plusieurs décennies, car il offre plusieurs avantages, notamment :

- **Une facilité d'analyse** : un entrepôt de données fournit une source unique de vérité pour l'analyse des données, avec un schéma et une terminologie cohérente et normalisée. Pour les bases de données relationnelles, une forme normale caractérise le fait que sa structure respecte certaines contraintes de modélisation, vous retrouverez plus d'informations à l'adresse https://go.fabricbook.fr/ch3-6.
- **Une haute performance** : un entrepôt de données est optimisé pour un traitement rapide et efficace des requêtes, en utilisant des techniques telles que l'indexation, le partitionnement, l'agrégation et la mise en cache.
- **Des données fiables** : un entrepôt de données garantit la qualité et l'exactitude des données en appliquant des règles de nettoyage, de validation et de gouvernance des données.
- **Des données sécurisées** : un entrepôt de données protège les données contre l'accès, la modification ou la suppression non autorisés, en mettant en œuvre des politiques de chiffrement, d'authentification et d'autorisation fine d'accès aux données.

Cependant, l'entrepôt de données a ou est également confronté à plusieurs défis et limitations, tels que :

- **Un coût élevé** : un entrepôt de données nécessite une quantité importante de matériel, de logiciels et de ressources humaines pour être construit, entretenu et mis à l'échelle.
- **Une faible flexibilité** : un entrepôt de données est rigide et inflexible, car il suit un schéma prédéfini qu'il est difficile de modifier ou de mettre à jour.
- **Une faible évolutivité** : un entrepôt de données est limité par la capacité et les performances du matériel et des logiciels sous-jacents, et ne peut pas facilement gérer le volume, la variété et la vitesse croissantes des données.
- **Une portée limitée** : un entrepôt de données ne prend en charge que les données structurées et ne peut pas prendre en charge les données non structurées ou semi-structurées, telles que le texte, les images, l'audio, la vidéo ou les données des médias sociaux.



Les Datamarts et les entrepôts de données sont deux types de référentiels, alors que les entrepôts de données sont prévus pour contenir l'intégralité des données d'une entreprise, les Datamarts répondront seulement aux besoins d'un département donné ou d'une fonction commerciale spécifique.

1.2 Lac de données

Un lac de données, ou *Data Lake* en anglais, est un référentiel distribué de données brutes et non traitées, stockées dans leur format d'origine, sans schéma ni structure prédéfinis. Un lac de données est conçu pour prendre en charge un large éventail de types de données, de sources et de cas d'utilisation, tels que l'exploration, la découverte et l'expérimentation de données. Un lac de données suit une approche dite de « schéma à la lecture ». Les données sont structurées et traitées que lorsqu'elles sont consultées ou consommées par un utilisateur ou une application (ELT - *Extract-Load-Transform*). Un lac de données permet également la démocratisation des données, ce qui signifie que les données sont accessibles et disponibles pour quiconque en a besoin, sans aucune barrière ou restriction.

Remarque

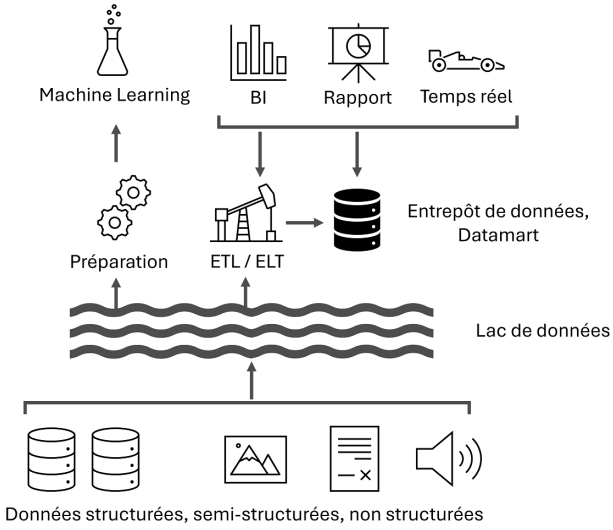
Les termes « barrière » ou « restrictions » sont à prendre ici dans le sens « difficultés techniques ou organisationnelles ». Les lacs de données sont là pour mettre à disposition de toute l'organisation l'ensemble des données, quel que soit leur format d'origine ou l'usage que l'on en a. Dans Microsoft Fabric, il est bien évidemment possible de mettre en place des droits d'accès, ou des limites d'utilisations sur vos lacs de données. Le chapitre Sécuriser sa plateforme de données vous donnera les clés pour les mettre en place.

Le concept de lac de données a été introduit pour la première fois par James Dixon (https://go.fabricbook.fr/ch3-7) en 2010 et a gagné en popularité dans les années 2010 avec l'émergence des technologies Big Data, telles que Hadoop, Spark et NoSQL. Le lac de données offre un nouveau paradigme pour le stockage et le traitement des données, car il présente plusieurs avantages, notamment :

- **Une grande flexibilité** : un lac de données est agile et adaptable, car il n'impose aucun schéma ou structure aux données, ce qui permet aux données d'évoluer et de changer au fil du temps.
- **Un faible coût** : un lac de données exploite du matériel de base et des logiciels open source pour stocker et traiter les données, réduisant ainsi le coût et la complexité de la gestion des données.
- **Une grande évolutivité** : un lac de données est évolutif et élastique, car il peut traiter n'importe quels volumes, types et vitesses de données, en utilisant des techniques de calcul distribué et parallèle.
- **Un vaste champ d'application** : un lac de données prend en charge tout type de données, qu'elles soient structurées, non structurées ou semi-structurées, et tout type de cas d'utilisation, qu'il soit analytique, opérationnel ou expérimental.

Cependant, le lac de données a ou est également confronté à plusieurs défis et limitations, tels que :

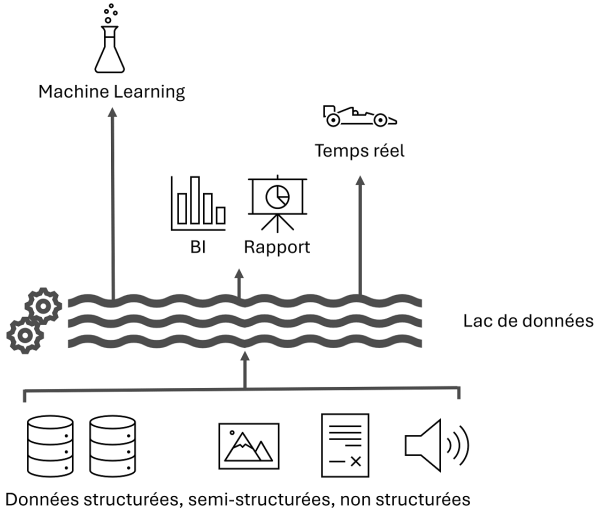
- **De faibles performances d'accès** : un lac de données n'est pas optimisé pour le traitement des requêtes, car il nécessite beaucoup de calculs et de ressources pour structurer et traiter les données à la volée.
- **Une analyse difficile** : un lac de données ne fournit pas une source unique de vérité pour l'analyse des données, car il manque un schéma et une terminologie cohérents et normalisés.
- **Des données peu fiables** : un lac de données ne garantit pas la qualité et l'exactitude des données, car il n'applique aucune règle de nettoyage, de validation ou de gouvernance des données.
- **Des données non sécurisées** : un lac de données ne protège pas les données contre l'accès, la modification ou la suppression non autorisée, car il ne met pas en œuvre de politiques de chiffrement, d'authentification ou d'autorisation des données.



1.3 Data Lakehouse

Le concept de Data Lakehouse a été proposé pour la première fois par Databricks (https://go.fabricbook.fr/ch3-8) en 2019, et s'est démocratisé dans les années 2020, avec les progrès du Cloud Computing, de l'ingénierie des données et de l'apprentissage automatisé. Le Data Lakehouse est une architecture de gestion des données innovante qui combine l'agilité, le faible coût et la scalabilité d'un lac de données avec les fonctionnalités éprouvées, comme les transactions ACID (Atomicité, Cohérence, Isolation et Durabilité : https://go.fabricbook.fr/ch3-14) des entrepôts de données traditionnels. Si une transaction respecte les propriétés ACID, alors chaque instruction d'une transaction (lecture, écriture, mise à jour ou suppression de données) est traitée comme une unité indivisible. Cela signifie que l'instruction est soit entièrement exécutée, soit pas du tout. Ce nouveau paradigme présente de nouveaux avantages, tels que :

- **Des données fiables et flexibles** : un Data Lakehouse garantit la qualité et l'exactitude des données, en appliquant des règles de nettoyage via des traitements ELT, de validation et de gouvernance des données, ainsi que l'agilité et l'adaptabilité des données, en permettant aux données d'évoluer et de changer au fil du temps.
- **Des performances optimales** : un Data Lakehouse est optimisé pour un traitement rapide et efficace des requêtes, ainsi que pour un traitement évolutif et parallèle des données, à l'aide de techniques telles que Delta Lake, parquet et Spark.
- **Une analyse facile et diversifiée** : un Data Lakehouse fournit une source unique de vérité pour l'analyse des données, avec un schéma et une terminologie cohérente et standardisée, ainsi qu'un large éventail de types de données et de cas d'utilisation.
- **Des données sécurisées et démocratisées** : un Data Lakehouse protège les données contre l'accès, la modification ou la suppression non autorisée, en mettant en œuvre des politiques de chiffrement, d'authentification et d'autorisation des données, ainsi que l'accessibilité et la disponibilité des données, en supprimant toutes les barrières ou restrictions.

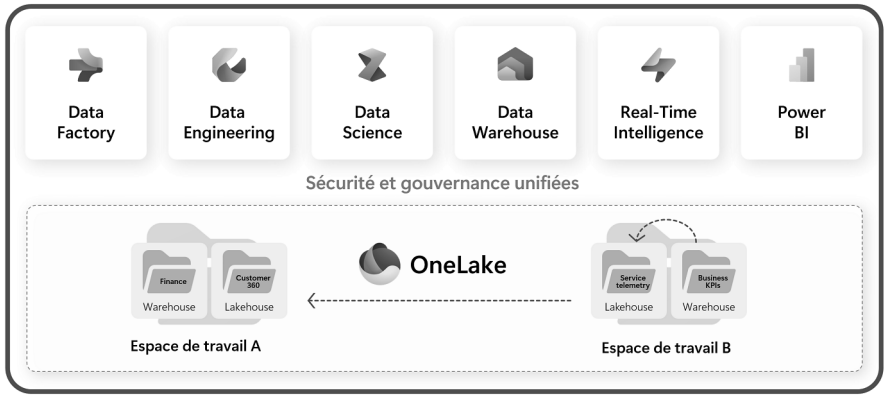


Cependant, le Data Lakehouse est également confronté à certains défis et limitations, tels que :

- **Une faible maturité** : le Data Lakehouse est encore un concept relativement nouveau et émergent, et ne dispose pas ou peu encore de bonnes pratiques, normes et références établies comme pour l'entrepôt de données et le lac de données.
- **Une grande complexité** : la construction, la maintenance et l'exploitation d'un Data Lakehouse requièrent un niveau élevé d'expertise et de compétences techniques, car il fait appel à une variété de technologies, d'outils et de plateformes.
- **Des compromis potentiels** : un Data Lakehouse peut ne pas être en mesure d'atteindre l'équilibre optimal entre les objectifs et les exigences contradictoires du Data Warehouse et du Data Lake, tel que la performance par rapport à l'évolutivité, l'analyse par rapport à l'exploration, et la fiabilité par rapport à la flexibilité.

Le stockage et le traitement des données ont évolué, passant de l'entrepôt de données traditionnelles au lac de données modernes, au fur et à mesure que le paysage des données s'est modifié et diversifié. Chaque approche a ses propres caractéristiques, avantages et inconvénients, et répond aux différents besoins et défis de l'analyse des données. Le Data Lakehouse est le paradigme le plus récent et le plus prometteur pour le stockage et le traitement des données, car il combine le meilleur des deux mondes que sont l'entrepôt de données et le lac de données. Cependant, le Data Lakehouse n'est pas une solution miracle et doit encore faire face à certains problèmes et limitations. Par conséquent, le Data Lakehouse n'est pas une solution universelle mais plutôt une solution qui dépend du contexte spécifique, des objectifs et des préférences de chaque organisation et de chaque utilisateur.

Dans Microsoft Fabric les utilisateurs peuvent interagir et construire aussi bien des Warehouse que des Lakehouse et même composer des architectures utilisant les deux composant, la sauvegarde centralisée des données dans un même format ouvert facilitant leurs interactions.

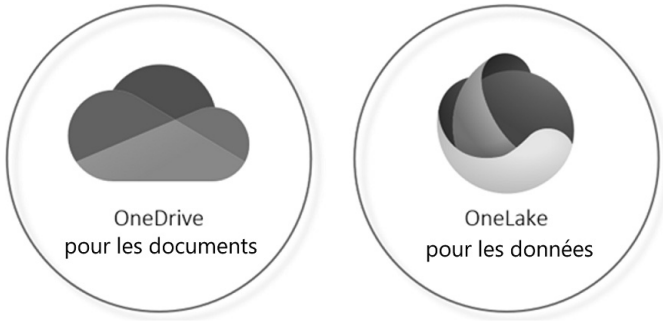


2. Le stockage des données dans Fabric

2.1 OneLake

On observe dans les organisations déployant des lacs de données (Data Lake) qu'ils ont tendance à se multiplier. Cela a pour effet de créer une fragmentation, des silos ou encore la duplication de données et des problèmes liés à leurs maintenances. OneLake se concentre sur l'élimination des défis précédemment énoncés tout en améliorant la collaboration, avec une idée qui s'apparente à une révolution : avoir un seul lac de données pour toute l'organisation. Chaque Tenant a donc exactement un OneLake.

Nous pouvons considérer le OneLake comme le OneDrive pour les données : alors que le OneDrive permet la sauvegarde de vos fichiers d'entreprise, le OneLake permet la sauvegarde des données de l'entreprise.



Toutes les données de OneLake sont stockées dans Azure Data Lake Storage (ADLS, https://go.fabricbook.fr/ch3-9), chaque fichier OneLake est un fichier ADLS. Ainsi, le coût et les performances du stockage des données dans OneLake suivent de près ceux du stockage des données dans ADLS supportant n'importe quel type de fichier, structuré ou non. OneLake supporte aussi les mêmes API et SDK ADLS Gen2 pour être compatible avec les applications existantes, y compris Azure Databricks.

OneLake ajoute une couche d'abstraction pour mieux prendre en charge le modèle SaaS de Microsoft Fabric. Les utilisateurs n'ont pas besoin de créer de comptes ADLS, le stockage est géré en mode SaaS par OneLake, qui prend également en charge des abstractions telles que les raccourcis. En plus des Listes de contrôle d'accès (ACL) de fichiers et de dossiers similaires à ADLS, délimités par les espaces de travail Microsoft Fabric, OneLake supporte aussi certaines règles d'accès aux données plus granulaires telles que le Row Level Security (RLS), Column Level Security (CLS) ainsi que le masquage dynamique de données sur les tables.

Toutes les charges de travail dans Microsoft Fabric ont la capacité d'écrire, de lire et d'interagir avec OneLake. Les entrepôts de données et les Lakehouse stockent automatiquement leurs données dans OneLake au format Delta Parquet. Lorsqu'un Data Engineer utilise un Notebook Spark pour intégrer et transformer des données dans un Lakehouse, ou qu'un développeur utilise le langage T-SQL pour travailler dans un entrepôt de données transactionnelles ils utilisent, travaillent et partagent le même OneLake.

Au sein d'un Tenant Microsoft Fabric, nous collaborons dans des espaces de travail, ces espaces de travail peuvent être répartis dans différentes régions du monde via l'utilisation de capacités. Chaque espace de travail (*Workspace*) apparaît comme un conteneur dans le OneLake, et les différents éléments de données apparaissent comme des dossiers dans ces conteneurs.