# La Virtualisation des Données avec Dremio : Révolutionner l'Accès et l'Analyse des Données à l'Ère du Big Data

[La Virtualisation des Données avec Dremio : Révolutionner l'Accès et l'Analyse des Données à l'Ère du Big Data 1](#_Toc1912130759)

[I. Introduction à la Virtualisation des Données 4](#_Toc1822729821)

[- Définition de la virtualisation des données 4](#_Toc1828718017)

[Comment effectuer la virtualisation des données 4](#_Toc549603612)

[- Importance et avantages de la virtualisation des données 4](#_Toc886962830)

[La virtualisation des données offre les capacités clés suivantes 5](#_Toc2022769835)

[Virtualisation des données vs. ETL 18](#_Toc170956297)

[Exploration des cas d'utilisation de la virtualisation des données 23](#_Toc70748228)

[Fournisseurs de virtualisation des données 34](#_Toc1810865788)

[Dremio - Déplacez vos analyses de données vers la gauche 37](#_Toc1731896728)

[The Dremio Difference 38](#_Toc1044776437)

[Cas d'utilisation de Dremio 39](#_Toc184991864)

[Hadoop Migration 41](#_Toc720989085)

[Dremio & VAST 43](#_Toc848616630)

[Reflections - Transparent Query Acceleration 44](#_Toc1040349432)

[Denodo et Dremio basée sur différents aspects : 45](#_Toc2134064335)

[What is a Data Lakehouse? 49](#_Toc2002815305)

[Pourquoi Dremio 51](#_Toc511467177)

[Data Reflections 57](#_Toc141749198)

[Couche Sémantique en Libre-Service 63](#_Toc842911936)

[SÉCURITÉ ET GOUVERNANCE DES DONNÉES 64](#_Toc551217554)

[PROVENANCE ET TRAÇABILITÉ DES DONNÉES 65](#_Toc1028705218)

[AUTHENTIFICATION 65](#_Toc1818619971)

[Virtualisation des Données avec Dremio 73](#_Toc1760230980)

[Solution Future pour Attijariwafa Bank avec Dremio 74](#_Toc242960258)

[Résultats 74](#_Toc1578656418)

[Summary of Exercises 76](#_Toc980215961)

[Environment Setup 76](#_Toc1006443190)

[Breakdown of the docker-compose file 77](#_Toc199495878)

[Populating the Postgres Database 77](#_Toc1842587840)

[1. Spin up the Postgres Service: 77](#_Toc919150499)

[2. Access the Postgres Shell: 77](#_Toc139511335)

[3. Create a Table and Add Data: 77](#_Toc1959361841)

[Moving the Data to the Data Lake with Spark 77](#_Toc1122882405)

[1. Starting Up Our Data Lake 77](#_Toc2104515829)

[2. Creating a Bucket in Our Data Lake 77](#_Toc1317991724)

[3. Running the PySpark Script 77](#_Toc1652026605)

[Breakdown of the PySpark Code 77](#_Toc457745826)

[Can This Be Easier? 77](#_Toc1620354673)

[Connecting Our Data to Dremio 77](#_Toc548108304)

[Building our BI Dashboard 77](#_Toc601054165)

[Conclusion 77](#_Toc1234483957)

# Abstrac

L'implémentation de la virtualisation des données avec Dremio à la Banque AttijaraWafa vise à révolutionner les capacités d'accès et d'analyse des données à l'ère du Big Data. Comment les organisations peuvent-elles optimiser l'intégration et l'analyse de sources de données diverses pour améliorer l'agilité et la précision des prises de décision ?

Dans le paysage intensif en données d'aujourd'hui, la gestion efficace des données et leur utilisation sont des impératifs cruciaux. Dremio joue un rôle central en permettant le stockage, la gestion et l'analyse fluides de vastes ensembles de données provenant à la fois des systèmes opérationnels et analytiques.

Au cœur de cette initiative se trouve le rôle de Dremio dans l'amélioration de l'accessibilité aux données, l'agilité dans l'analyse et le soutien aux processus décisionnels rapides. Comment l'adoption d'une architecture Lakehouse avec les formats de table ouverte Apache Iceberg peut-elle renforcer la gouvernance des données, la scalabilité et l'efficacité opérationnelle au sein des institutions financières comme la Banque AttijaraWafa ?

Grâce au déploiement de la solution avancée de virtualisation des données de Dremio, cette initiative vise à rationaliser les flux de données, accélérer le temps d'accès aux insights et instaurer une culture d'entreprise centrée sur les données. Comment ces avancées permettent-elles à la Banque AttijaraWafa de naviguer et d'exceller dans un environnement concurrentiel où les données jouent un rôle clé ?

En définitive, ce projet vise à équiper la Banque AttijaraWafa d'outils et de connaissances robustes pour exploiter efficacement les données, propulsant ainsi les initiatives stratégiques et améliorant les capacités de prise de décision face aux défis évolutifs du monde des affaires.

Les banques sont confrontées à un défi majeur avec la gestion de sources de données disparates. L'intégration de ces données nécessite un processu s rigoureux d'extraction, transformation et chargement (ETL) pour assurer leur qualité et cohérence. Une fois traitées, ces données sont essentielles pour les utilisateurs en science des données et en intelligence d'affaires (BI), qui dépendent d'une visualisation claire et précise pour prendre des décisions stratégiques éclairées.La virtualisation des données émerge comme une solution essentielle, permettant aux banques de consolider ces sources disparates en une vue unifiée. Cette vue unifiée facilite l'accès rapide et efficace aux informations critiques pour les analyses avancées et les rapports décisionnels. Grâce à cette architecture centralisée, les données peuvent être utilisées de manière plus agile, accélérant ainsi la capacité des utilisateurs finaux à obtenir des insights pertinents et exploitables pour améliorer les opérations et répondre aux exigences réglementaire

# I. Introduction à la Virtualisation des Données

## - Définition de la virtualisation des données

La virtualisation des données est une méthode de traitement des données qui consiste à ajouter une couche d’extraction au niveau logique. Les utilisateurs peuvent ainsi accéder à des ensembles de données disparates et les modifier sans se soucier de détails techniques tels que le format d’origine des données ou leur lieu de stockage.Les utilisateurs peuvent accéder à toutes leurs données via une interface unique. Il n’est pas nécessaire de déplacer physiquement de grands blocs de données. Au lieu de cela, il utilise des pointeurs vers les données réelles. Il est ainsi plus facile de stocker les données et plus rapide d’y accéde

## Comment effectuer la virtualisation des données

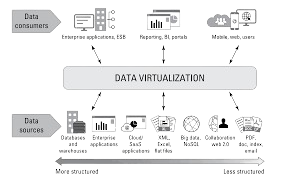
La virtualisation des données est une technique qui permet d’accéder à des données provenant de diverses sources et de les manipuler sans avoir à les déplacer ou à les copier physiquement dans un référentiel centralisé. Il offre une vue unifiée de vos données, ce qui facilite le travail avec des sources de données distribuées et hétérogènes. Voici les étapes à suivre pour réaliser et mettre en œuvre la virtualisation des données

### - Importance et avantages de la virtualisation des données

Alors que la plupart des solutions d'intégration de données déplacent une copie des données vers une nouvelle source consolidée, la virtualisation des données offre une approche totalement différente. Plutôt que de déplacer les données, la virtualisation des données fournit une vue des données intégrées, laissant les données sources exactement là où elles se trouvent. Les entreprises n'ont pas à supporter les coûts liés au déplacement et à l'hébergement des données, tout en bénéficiant des avantages de l'intégration des données.

La virtualisation des données effectue de nombreuses fonctions de transformation et de qualité similaires à celles de l'intégration traditionnelle des données — telles que l'ETL, la réplication des données, la fédération des données, le bus de services d'entreprise (ESB) et d'autres — mais elle exploite la technologie moderne pour offrir une intégration des données en temps réel à moindre coût, avec plus de rapidité et d'agilité. Elle peut remplacer l'intégration traditionnelle Des données et réduire le besoin de data marts et d'entrepôts de données répliqués dans de nombreux cas.

La virtualisation des données est également une couche d'abstraction et une couche de services de données. À cet égard, elle est très complémentaire pour une utilisation entre les sources de données originales et dérivées, l'ETL, l'ESB et autres middleware, les applications et les appareils, qu'ils soient sur site ou basés sur le cloud.



La virtualisation des données offre les capacités clés suivantes :

* - Abstraction logique et découplage : Les sources de données disparates, les middleware et les applications consommatrices qui utilisent ou nécessitent des plateformes et interfaces spécifiques, des formats, des schémas, des protocoles de sécurité, des paradigmes de requête et d'autres particularités peuvent désormais interagir facilement grâce à la virtualisation des données.
* - Fédération des données avancée : La fédération des données est une sous-partie de la virtualisation des données, mais elle est enrichie d'une optimisation de requête en temps réel plus intelligente, de mise en cache, de traitement en mémoire et de stratégies hybrides qui sont automatiquement (ou manuellement) sélectionnées en fonction des contraintes de la source, des besoins de l'application et de la prise de conscience du réseau.
* - Intégration sémantique des données structurées et non structurées : La virtualisation des données est l'une des rares technologies à combler la compréhension sémantique des données non structurées et des données web avec la compréhension basée sur le schéma des données structurées pour permettre l'intégration et l'amélioration de la qualité des données.
* - Approvisionnement agile de services de données : La virtualisation des données promeut l'économie des interfaces de programmation d'applications (API). Toute source de données principale, dérivée, intégrée ou virtuelle peut être rendue accessible dans un format ou un protocole différent de l'original, avec un accès contrôlé en quelques minutes.
* - Gouvernance unifiée des données et sécurité fine avec auditabilité complète : La virtualisation des données permet un contrôle précis sur les informations sensibles des clients stockées dans plusieurs systèmes en établissant une couche d'accès unifiée à travers des systèmes sur site et hors site. Toutes les données sont rendues découvrables et peuvent être facilement intégrées via une seule couche virtuelle qui expose plus rapidement les redondances et les problèmes de qualité. La virtualisation des données impose la gouvernance du modèle de données et la sécurité du source aux services de données de sortie, ainsi que la cohérence dans les règles d'intégration et de qualité des données. Lorsque les consommateurs de données doivent accéder à une source, ils le font via la couche de virtualisation des données
* La virtualisation des données élimine le déplacement inutile des données : Avec une couche de virtualisation des données en place, aucune réplication des données n'est nécessaire à des fins de reporting, et aucun script d'Extraction, Transformation et Chargement (ETL) ne doit être réécrit. Une couche de virtualisation des données fonctionne avec l'infrastructure existante de l'entreprise, configurée exactement comme elle est. La couche de virtualisation des données abstrait simplement les fonctions d'accès de sorte que les utilisateurs perçoivent les données comme existant dans un seul référentiel virtuel. Cependant, si, pour des raisons de performance, les données doivent être persistées, les outils de virtualisation des données offrent également des moyens simples de persister un ensemble de données en activant simplement certains paramètres dans le modèle. La réplication devient simplement une autre option, pas une nécessité.
* Lignage complet des données et règles métier agiles : À tout moment, les entreprises peuvent comprendre et rendre compte du lignage complet de tout ensemble de données sensibles, y compris sa source d'origine, toutes les vues et toutes les modifications. De plus, grâce à la couche de virtualisation des données, les entreprises peuvent établir des règles sophistiquées pour automatiser la conformité, telles que le masquage des données à la volée, afin qu'elles ne puissent pas être consultées par des utilisateurs qui ne possèdent pas les droits nécessaires. Ces règles peuvent être appliquées rapidement et efficacement à travers divers systèmes car elles sont appliquées dans la couche de virtualisation des données.
* Sécurisation des données au repos et en mouvement : La couche de virtualisation des données peut effectuer une authentification basée sur les rôles à n'importe quel niveau, tel que invité, employé ou entreprise ; appliquer des permissions spécifiques aux données, y compris le masquage au niveau des lignes et des colonnes ; et définir des permissions à l'échelle du schéma et une sécurité basée sur des politiques. La couche de virtualisation sécurise les données en transit via des protocoles Secure Sockets Layer/Transport Layer Security (SSL/TLS) et authentifie les utilisateurs via des protocoles éprouvés par l'industrie tels que Lightweight Directory Access Protocol (LDAP), le passage avec Kerberos, Windows Single Sign-On (SSO), Open Authorization (OAuth), Simple and Protected GSS-API Negotiation Mechanism (SPNEGO) authentication, OAuth et SAML authentication, ainsi que Java Database Connectivity/Open Database Connectivity (JDBC/ODBC) Security.

* La virtualisation des données fournit des informations abstraites et intégrées en temps réel à partir de sources disparates à plusieurs applications et utilisateurs. Elle est également facile à construire, consommer et maintenir. Pour construire des services de données virtuelles, l'utilisateur suit trois étapes simples
* - Connecter et virtualiser n'importe quelle source : Accédez rapidement à des sources de données structurées et non structurées disparates à l'aide de connecteurs inclus. Introspectez leurs métadonnées et exposez-les sous forme de vues sources normalisées dans la couche de virtualisation des données.
* - Combiner et intégrer dans des vues de données métier : Combinez, intégrez, transformez et nettoyez les vues sources en vues métier basées sur un modèle canonique dans une interface utilisateur graphique (GUI) ou via des scripts documentés.
* - Connecter et sécuriser les services de données : Chacune des vues de données virtuelles peut être sécurisée et publiée sous forme de vues SQL ou des dizaines d'autres formats de services de données.

Concepts Avancés dans la Virtualisation des Données

En nous appuyant sur les bases de notre introduction, explorons plus en détail certains concepts avancés de la virtualisation des données :

Architecture de la Virtualisation des Données

Les logiciels de virtualisation des données suivent généralement une architecture en couches :

- Couche de Présentation: Cette couche interagit avec les utilisateurs et les applications, offrant une interface conviviale pour accéder et interroger les données via la couche virtuelle.

- Couche de Logique Métier : Cette couche contient les règles et la logique métier qui peuvent être appliquées aux données avant leur présentation aux utilisateurs. Cela peut impliquer le filtrage des données, l'agrégation ou les contrôles de sécurité.

- Couche de Services de Données : Cette couche centrale gère les connexions à diverses sources de données et effectue des transformations de données pour assurer un format cohérent pour la couche virtuelle. Elle utilise des connecteurs de données, le référentiel de métadonnées et le moteur de transformation.

- \*\*Couche de Sources de Données :\*\* Cette couche représente les sources de données physiques sous-jacentes, telles que les bases de données, les applications et les systèmes de stockage en nuage.

Techniques de Virtualisation des Données

- Vues Matérialisées : Ce sont des résumés pré-calculés de données provenant de sources sous-jacentes, stockés localement pour un accès plus rapide et une meilleure performance des requêtes.

- Fédération : Cette technique permet aux utilisateurs d'interroger simultanément des données provenant de plusieurs sources, les traitant comme une source de données unifiée.

- Mashup de Données : La virtualisation des données facilite la création de mashups de données, qui combinent des données de différentes sources pour créer une nouvelle vue intégrée pour des objectifs spécifiques.

Considérations de Sécurité dans la Virtualisation des Données

La sécurité des données est primordiale lors de la mise en œuvre de la virtualisation des données. Voici quelques considérations clés :

- Contrôle d'Accès : L'accès des utilisateurs aux données via la couche virtuelle doit être strictement contrôlé en fonction des permissions et des rôles.

- Masquage et Chiffrement des Données : Les données sensibles peuvent être masquées ou chiffrées au niveau de la couche virtuelle pour minimiser les risques de sécurité.

- Audit et Journalisation : Toutes les activités d'accès et de modification des données doivent être auditées et journalisées à des fins de conformité.

Virtualisation des Données vs. Intégration des Données

La virtualisation des données et l'intégration des données sont souvent utilisées de manière interchangeable, mais elles ont des objectifs distincts :

- Intégration des Données : Se concentre sur le déplacement ou la réplication physique des données de diverses sources vers un emplacement central, comme un entrepôt de données.

- Virtualisation des Données :Crée une vue logique unifiée des données sans déplacer physiquement les données. Elle fournit une couche d'abstraction entre les utilisateurs et les sources de données sous-jacentes.

La sélection de la solution de virtualisation des données appropriée dépend de plusieurs facteurs, y compris :

- Votre Paysage de Données : La complexité et le nombre de vos sources de données.

- Exigences Métier : Vos besoins spécifiques en matière d'accès aux données, d'intégration et de reporting.

- Budget et Évolutivité : Considérez le coût de la solution et sa capacité à évoluer en fonction de vos besoins croissants en matière de données.

Exemples Concrets de Virtualisation des Données

Voici quelques exemples pratiques de la manière dont les organisations utilisent la virtualisation des données :

- Une entreprise de vente au détail utilise la virtualisation des données pour intégrer les données des clients provenant des systèmes de point de vente, des programmes de fidélité et des bases de données marketing. Cela leur permet d'obtenir une vue à 360 degrés des clients et de créer des campagnes marketing ciblées.

- Une organisation de soins de santé utilise la virtualisation des données pour fournir aux médecins une vue unifiée des données des patients provenant de divers systèmes, tels que les dossiers médicaux électroniques, les résultats de laboratoire et les données d'imagerie. Cela améliore les soins aux patients en fournissant aux professionnels de la santé une vue complète de l'historique médical de leurs patients.

- Une entreprise manufacturière utilise la virtualisation des données pour intégrer les données des lignes de production, des systèmes d'inventaire et des logiciels de gestion de la relation client (CRM). Cela leur permet d'optimiser les processus de production et d'améliorer la gestion de la chaîne d'approvisionnement.

L'Avenir de la Virtualisation des Données

La virtualisation des données est une technologie en évolution rapide. Voici quelques tendances à surveiller :

- Virtualisation des Données Cloud-Native: Des solutions conçues spécifiquement pour les environnements cloud afin de gérer les données à travers des déploiements sur site et dans le cloud.

- Virtualisation des Données en Libre-Service : Des outils qui permettent aux utilisateurs métier d'accéder et d'intégrer des données sans dépendre du personnel informatique.

- Intégration de l'Intelligence Artificielle et du Machine Learning : Exploiter l'IA pour automatiser les tâches de virtualisation des données et optimiser les performances d'accès aux données.

En comprenant ces concepts avancés et en restant informé des dernières tendances, vous pouvez utiliser efficacement la virtualisation des données pour libérer tout le potentiel de vos données et favoriser le succès de votre entreprise

La Couche Sémantique dans la Virtualisation des Donnée

Les couches sémantiques jouent un rôle crucial dans la gestion des données. Elles servent d'intermédiaire entre les données brutes stockées dans les bases de données et les utilisateurs finaux qui analysent ces données. Elles traduisent des structures de données complexes en un format plus convivial. Cette traduction implique la conversion de la terminologie technique des bases de données, des relations et des structures en termes et formats compréhensibles et pertinents pour les utilisateurs métier.

Couche Sémantique et Virtualisation des Données

Une couche sémantique est une couche d'abstraction qui se situe entre les données brutes dans les bases de données et les utilisateurs finaux des outils de BI. Elle fournit une vue conviviale des données, traduisant des structures de données techniques complexes en termes métiers significatifs. Par exemple, un champ nommé "cust\_id" dans la base de données pourrait être présenté comme "ID Client" dans la couche sémantique. La couche sémantique peut également gérer les permissions d'accès aux données, garantissant que les utilisateurs ne voient que les données qu'ils sont autorisés à voir.

Cependant, si elle n'est pas bien conçue, la couche sémantique peut introduire des goulots d'étranglement en termes de performance, surtout si elle gère de grands volumes de données. De plus, les organisations peuvent devenir dépendantes de la couche sémantique pour leurs activités de BI, et par conséquent, si des erreurs ou des problèmes surviennent dans cette couche, cela peut impacter un large éventail de rapports et d'analyses.

Data Mesh

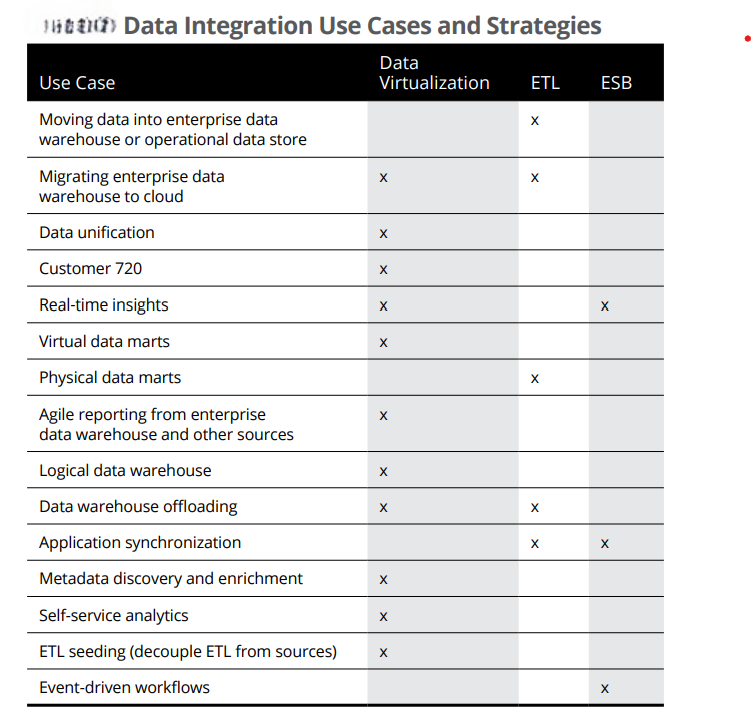
Le Data Mesh est une architecture de données décentralisée qui organise les données par domaines métiers spécifiques, tels que le marketing ou les ventes, et transfère la propriété des données aux producteurs de ces données. Il permet une utilisation en libre-service des données dans toute l'organisation et soutient les politiques de gouvernance des données axées sur la documentation, la qualité et l'accès. Le Data Mesh favorise les technologies cloud-native et peut être particulièrement bénéfique pour les grandes organisations ayant des besoins de données complexes.

Le Data Mesh démocratise les données grâce à l'auto-service et rapproche les informations des métiers. Résultat : décisions plus rapides et agilité accrue. Il capte les données en direct (streaming) via le cloud, permettant des décisions ultra-rapides. Le stockage est flexible et le paiement à l'usage : vous ne payez que pour ce que vous utilisez. Le Data Mesh répartit les données par domaine, facilitant ainsi la maintenance et les mises à jour, réduisant ainsi la complexité et la dette technique.

Self-Service Analytics

L'Analytics en Libre-Service permet aux utilisateurs métier sans expertise en informatique ou en science des données d'analyser des données et d'obtenir des insights. Cela implique l'utilisation d'outils d'analytique qui se connectent directement aux sources de données, permettant aux utilisateurs de sélectionner des données, de réaliser des analyses et de créer des visualisations par eux-mêmes. Cette approche favorise une culture axée sur les données en permettant des insights pertinents et opportuns pour la prise de décision.

Ces concepts font partie d'une stratégie de données moderne qui vise à démocratiser l'accès et l'analyse des données, facilitant ainsi pour les organisations l'exploitation de leurs données pour obtenir un avantage concurrentiel. Si vous avez besoin d'informations plus détaillées ou avez des questions spécifiques



Complementing ETL and ESB

La virtualisation des données prend en charge une grande variété de sources et de cibles, ce qui en fait une stratégie idéale d'intégration des données pour compléter les processus ETL et les ESB.

Les processus ETL ont été conçus pour déplacer des données vers des entrepôts de données et des environnements similaires, et ils sont particulièrement adaptés à cette tâche. Cependant, les processus ETL ne peuvent pas facilement prendre en charge les sources basées sur le cloud. La virtualisation des données peut compléter les processus ETL de plusieurs manières :

- Connexion transparente des sources de données sur site avec les sources de données basées sur le cloud sans avoir besoin de consolider les données dans un seul référentiel.

- Migration des systèmes sur site vers des systèmes basés sur le cloud sans interrompre la continuité des activités.

- Déchargement des entrepôts de données où la virtualisation des données aide non seulement dans le processus de déchargement, mais unifie également les données à travers l'entrepôt de données traditionnel et le nouveau référentiel tel que Hadoop, Amazon Web Services (AWS) S3, ou un magasin de données basé sur le cloud.

- Intégration en temps réel de sources de données disparates.

- Remplacement des processus ETL par la virtualisation des données lorsque l'accès plus rapide aux données est nécessaire.

La virtualisation des données peut également compléter un ESB et améliorer ses performances. L'ajout de nouvelles sources à un ESB peut être complexe ; des sources telles que les bases de données relationnelles, les sources web ou basées sur le cloud, les fichiers plats ou les messages électroniques ne sont pas immédiatement activées pour le SOA que l'ESB prend en charge. Pour simplifier ce processus, toutes les sources que l'ESB ne peut pas gérer peuvent être unifiées par la couche de virtualisation des données avant d'être transmises à l'ESB. Cette architecture tire parti des meilleures qualités des deux technologies : la virtualisation des données unifie les sources disparates, et les ESB délivrent les messages critiques pour soutenir le processus métier.

Delivering Faster

Les entreprises modernes nécessitent un accès rapide aux données les plus récentes et précises pour prendre des décisions stratégiques, anticiper les besoins des clients et rester en tête de leurs concurrents.

Le déplacement physique, l'entreposage et le stockage des données à plusieurs reprises représentent un coût et ralentissent les processus commerciaux lorsqu'il est nécessaire d'apporter des modifications. Les processus ETL (discutés précédemment dans ce chapitre), souvent manuels, présentent des risques d'erreurs coûteuses et retardent l'accès aux données. La réplication entre plusieurs sources de données via des liens réseau ajoute davantage de délais, de coûts et de complexité.

La virtualisation des données permet de mettre en place des architectures de données rapides telles que les entrepôts de données logiques, les data marts virtuels, la BI en libre-service et l'analyse opérationnelle. En agrégeant les données les plus récentes provenant de divers systèmes sources sans avoir besoin de les déplacer physiquement, la virtualisation des données permet aux services informatiques de livrer rapidement les données aux utilisateurs métier pour les intégrer dans leurs systèmes de BI. Le besoin de déplacer physiquement les données est l'une des principales causes de retard dans les architectures de données traditionnelles.

Le paysage informatique devient également plus complexe et distribué avec la création de silos de données supplémentaires grâce à l'adoption de magasins de données volumineuses et de solutions infonuagiques. La virtualisation des données contre cette vague de données en offrant une architecture de données rapide pour l'analyse avancée et le délestage des entrepôts de données. Cette architecture permet aux services informatiques de tirer parti des coûts réduits offerts par les solutions de big data et de cloud, tout en améliorant considérablement le délai de livraison des données avec un accès en temps réel.

Les services de données sont de plus en plus cruciaux pour le développement d'applications. La virtualisation des données permet un développement rapide des applications avec une couche unifiée de services de données créant une abstraction logique des données structurées et non structurées issues des sources sous-jacentes. En utilisant la virtualisation des données, les services informatiques peuvent développer des services de données en moins d'une demi-journée, tandis que les méthodes traditionnelles d'intégration de données telles que l'ETL peuvent souvent prendre une à deux semaines.

Les applications à vue unique telles qu'une vue unique des clients, des produits, des stocks, etc., améliorent l'efficacité des agents de centres d'appels grâce à des temps de réponse rapides, ainsi que celle des équipes de vente et de marketing avec des campagnes ciblées. La virtualisation des données peut sous-tendre ces architectures à vue unique en agrégeant virtuellement les différents données maîtres en temps réel, sans avoir à les répliquer, les stocker de manière centralisée et les gérer.

Pour soutenir à la fois les utilisateurs orientés données et les utilisateurs orientés métier, la virtualisation des données peut fournir un environnement de modélisation des données facile à utiliser mais sophistiqué pour offrir l'accès aux données, gérer les données et fournir des services de données. La virtualisation des données aide les équipes informatiques et les utilisateurs métier à répondre rapidement aux exigences changeantes de la manière suivante :

- Un environnement de développement intégré répondant aux besoins à la fois des équipes informatiques et des utilisateurs métier, avec un studio de développement convivial, basé sur le glisser-déposer et le low-code, conçu pour les équipes orientées données telles que les ingénieurs de données, les utilisateurs avancés et les intégrateurs, qui peuvent publier des services de données en quelques clics seulement.

- Un catalogue complet de vues métier classées et étiquetées selon les catégories métier, facilitant l'accès aux utilisateurs pour parcourir, naviguer et découvrir les relations dans les ensembles de données. Ils peuvent effectuer des recherches sur les métadonnées et les données elles-mêmes pour valider correctement les jeux de données, inspecter les informations détaillées sur la lignée et exporter les résultats des requêtes.

- Des connecteurs préconstruits vers les systèmes sources et cibles, permettant à l'informatique de se connecter rapidement à des sources disparates et garantir des performances maximales.

- Une livraison flexible des données avec la publication en un clic de puissants services web REST et OData.

To make self-service a reality:

Rendre Self-Service une réalité

L'analyse en libre-service a longtemps été recherchée comme une panacée, promettant de libérer les utilisateurs métier pour effectuer des analyses sans l'assistance de l'informatique, permettant ainsi à l'informatique de se concentrer sur d'autres initiatives stratégiques.

Aujourd'hui, de nombreux outils d'analyse de bureau permettent aux utilisateurs de découper et d'analyser les données, et de les présenter sous forme de rapports complets et de tableaux de bord variés. Cependant, plusieurs défis clés au niveau des données ont empêché l'analyse en libre-service de devenir une réalité :

- Données fragmentées : Les données sont dispersées dans plusieurs bases de données hétérogènes, entrepôts de données, systèmes infonuagiques et de big data, sources NoSQL et fichiers plats.

- Initiatives d'intégration de données multiples et à forte maintenance : Lorsqu'un utilisateur métier doit interroger plusieurs sources hétérogènes, les entreprises chargent souvent l'informatique de créer des intégrations ad hoc point-à-point en utilisant des processus ETL. Si une source doit être modifiée, ces processus doivent être réécrits, ce qui est coûteux et prend du temps.

- Retards dans les données : Il peut souvent falloir des mois pour livrer les données demandées en utilisant les processus traditionnels d'intégration de données, augmentant ainsi le risque que les données soient inexactes ou non pertinentes.

- Intégrité des données médiocre : Sans une "source unique de vérité", les utilisateurs métier peuvent involontairement utiliser des sources moins fiables, ce qui entraîne des données de qualité douteuse.

- Ligne de données non traçable : Si les utilisateurs collectent directement des données à partir de sources, ils peuvent ne pas conserver un enregistrement précis de l'origine des données, ce qui entrave la capacité à déterminer la qualité des données et mine davantage la confiance envers les données.

Ces défis doivent être surmontés pour rendre l'analyse en libre-service une réalité efficace et fiable pour les entreprises modernes.

Data virtualization increases operational efficiencies, reduces costs and complexity, minimizes data replication, and fosters data reusability and collaboration. By enabling self-service for business users, data virtualization further accelerates decision-making and time to market by reducing reliance on scarce IT resources for data access and integration.

Data virtualization allows for replication, but only when necessary.

### Virtualisation des données vs. ETL

Qu'est-ce que l'ETL ?

Avec l'ETL, les données sont extraites d'un ou plusieurs systèmes sources ou flux de données, transformées à l'aide de diverses règles et logiques, avant d'être répliquées et chargées dans le stockage cible.

Ce type d'intégration de données est idéal pour le déplacement de grandes quantités de données et la migration des données des systèmes hérités vers une nouvelle application cible ou une structure de stockage. Un exemple classique d'architecture de données d'entreprise utiliserait l'ETL pour déplacer des données de différentes

La virtualisation des données offre de nombreux avantages par rapport aux approches ETL traditionnelles, avec des bénéfices clés tels que l'accès en temps réel aux données et une intégration simplifiée des données.

* Avantages de la virtualisation des données

Accès en temps réel aux données :

- Contrairement à l'ETL qui nécessite le déplacement des données, la virtualisation des données permet un accès direct et en temps réel aux données.

- Cela conduit à une livraison plus rapide des données et à une réduction des besoins en infrastructure.

Intégration rapide :

- Les nouvelles sources de données peuvent être intégrées rapidement dans la couche virtuelle avec un effort minimal.

- Cela favorise le prototypage rapide et le développement agile.

Évolutivité et mise en cache :

- La virtualisation des données peut rencontrer des problèmes d'évolutivité ; cependant, certains produits offrent des fonctionnalités de mise en cache pour atténuer ce problème.

- La mise en cache peut aider dans certains cas d'utilisation mais pourrait ne pas être idéale pour tous les scénarios, comme l'analyse historique.

Architecture de données moderne

Les architectures de données modernes sont conçues pour gérer la complexité et la diversité des données d'aujourd'hui, en assurant flexibilité et agilité. Ces architectures intègrent plusieurs composants clés pour améliorer les performances et la fonctionnalité :

Composants unifiés :

- Virtualisation des données : Fournit un accès en temps réel aux données et à leur intégration.

- Mise en cache des données : Améliore les performances en stockant temporairement les données fréquemment consultées.

- Matérialisation des données : Crée des copies physiques des données pour un accès et un traitement plus rapides.

Langage de requête :

- Utilise généralement SQL pour des requêtes rapides et pour construire divers modèles et vues de données.

Fonctionnalités innovantes :

- ntelligence artificielle (IA) : Améliore le traitement des données et les analyses.

- Déplacement et optimisation des requêtes : Améliore les performances des requêtes.

- Catalogues de métadonnées : Offre une vue complète des actifs de données.

- Automatisation des tâches : Rationalise les flux de travail de données.

- Traçabilité des données : Suit le flux des données de la source à la destination.

- \*\*Mise en cache des données : Améliore les performances en réduisant les temps de requête.

Cas d'utilisation de l'architecture de données moderne

Entrepôt de données logique :

- Combine la réplication des données (ETL/ELT) et la virtualisation des données pour répondre à divers cas d'utilisation.

- Améliore l'intégration des données, la transparence, la gouvernance et la sécurité.

Tissu et maillage de données :

- Ces tendances intègrent des architectures distribuées à travers les clouds pour répondre aux exigences modernes en matière de données.

Big Data, analyses prédictives et apprentissage automatique :

- Combine les entrepôts de données traditionnels avec des outils de big data pour une fonctionnalité en temps réel.

- Fournit une vue unifiée pour l'application de modèles d'apprentissage automatique.

Intégration des données clients :

- Centralise les données de divers systèmes pour comprendre le comportement des clients et améliorer les efforts de marketing.

- Permet une meilleure segmentation des audiences, des performances de campagne et des analyses.

Services financiers :

- Élimine les silos de données, fournissant une source unique de vérité.

- Permet une livraison de données en temps réel pour des processus tels que les approbations de prêt, renforçant ainsi la compétitivité.

Santé :

- Agrège les données de différents systèmes pour améliorer la transparence des flux de travail et leur automatisation.

- Les avantages incluent des coûts réduits, un turnover du personnel plus faible et une gestion efficace des ressources.

E-commerce :

- Centralise les données des systèmes ERP, CRM, applications web/mobile, analyses d'emails et plateformes de marketing.

- Fournit des vues complètes des clients et des produits pour une meilleure prise de décision.

Prochaines étapes

Pour découvrir toute la puissance de la virtualisation des données, essayez un essai gratuit de CData Virtuality et observez comment cela peut améliorer votre architecture de données moderne. Cette approche pratique démontrera comment la virtualisation des données peut transformer vos processus de gestion et d'intégration des données.

la complexité de cette première génération de technologies Big Data a été l'une des raisons qui ont motivé la création de Dremio. La réalisation que les entreprises recherchaient quelque chose de simple, une solution efficace qui offre systématiquement les performances et les fonctionnalités nécessaires sans les complexités inutiles. L'objectif n'était pas de proposer une multitude d'options, mais plutôt de fournir des capacités essentielles. En profitant de l'émergence du cloud public en 2015, tant pour les startups que pour les entreprises, a permis de concevoir une nouvelle architecture qui simplifie considérablement les interactions utilisateur, réduisant au minimum le nombre de composants à gérer.

Bien sûr ! Voici le paragraphe sans utiliser de termes comme "nous" ou "notre" :

Au début de Dremio, l'accent était mis sur la possibilité pour les utilisateurs d'exécuter des requêtes SQL sur des bases de données NoSQL telles qu'Elasticsearch, MongoDB et autres. Cependant, environ un an et demi après la fondation de Dremio, il est devenu évident que l'opportunité réelle résidait dans l'autonomisation des entreprises pour qu'elles puissent réaliser des analyses en libre-service et construire des data lakes basés sur le cloud, ou lakehouses.

Cette prise de conscience a entraîné une réorientation de notre feuille de route et a nécessité de communiquer ce changement à nos premiers clients. Bien que ce changement ait déçu quelques clients initiaux, il constituait une étape cruciale pour la croissance de notre entreprise.

Ce passage à une nouvelle orientation a également représenté un défi pour nos employés, en particulier pour les ingénieurs logiciels qui devaient abandonner certains de leurs projets en cours pour se concentrer sur de nouvelles capacités. Néanmoins, en tant que startup visant à se développer et éventuellement à devenir une entreprise publique, ces décisions difficiles étaient essentielles pour saisir les opportunités du marché et nécessitaient un leadership fort pour faire avancer l'entreprise.

Pour assurer la disponibilité des données pour les analystes et les rapports, les entreprises agrègent souvent des sources de données. La virtualisation des données et les lacs de données sont deux méthodes populaires pour y parvenir, chacune offrant des avantages distinct

## Exploration des cas d'utilisation de la virtualisation des données

1. Rendre la BI Agile une Réalité

L'un des usages les plus courants de la virtualisation des données est pour le reporting agile, la business intelligence opérationnelle (BI) et les tableaux de bord en temps réel qui nécessitent une agrégation, une analyse et une présentation opportunes des données les plus pertinentes provenant de plusieurs sources. Les individus ainsi que les gestionnaires doivent surveiller la performance pour prendre des décisions opérationnelles quotidiennes dans des processus métier clés tels que les ventes, le support, la production, la logistique, les finances, le juridique et la conformité.

La virtualisation des données permet à l'informatique d'être plus agile en répondant à cette demande quasi insatiable d'informations exploitables. Avec la virtualisation des données, les données restent dans les magasins de données sources. La réplication, avec ses outils et processus de mise en scène, de transformation et de copie par lots, n'est pas nécessaire. L'accès aux données se fait via des vues virtuelles qui peuvent être créées rapidement (et jetées, si nécessaire). Les changements sont également rapides,

Les avantages de la virtualisation des données pour les projets de BI agile incluent :

» Réduction significative de la réplication : Depuis la couche de virtualisation des données, l'utilisateur peut pointer vers les sources de données d'origine, mettre en cache partiellement certaines données, et construire des "data marts virtuels" définis dans la couche de virtualisation des données, évitant ainsi la création de nouveaux référentiels (et de copies supplémentaires des données).

» Facilite les changements de vues et permet une réponse plus rapide aux demandes métier : Les modifications sont effectuées dans la couche de virtualisation des données, évitant ainsi de modifier les scripts ETL dans toute la chaîne de réplication.

» Permet l'accès aux données en temps réel pour la BI opérationnelle : En permettant un accès direct aux systèmes opérationnels, la virtualisation des données fournit les données aux applications consommatrices beaucoup plus rapidement que les approches traditionnelles d'intégration de données

» Intègre tous les types de données: De semi-structurées à structurées et non structurées, la virtualisation des données intègre de manière transparente les données de toutes les sources de données.

» Permet la BI en libre-service: Les utilisateurs peuvent exécuter tout rapport sur des services de données "prêts pour l'auto-service" créés et gérés par l'informatique, et atténuer les éventuels problèmes d'administration et de gouvernance grâce à des mesures de sécurité et de gouvernance fournissant un "auto-service encadré".

1. Obtaining a comprehensive view of your customers

Trop souvent, les informations sur un client, un fournisseur, un produit ou un projet sont dispersées parmi divers systèmes et sources. Ces informations doivent être regroupées dans une "vue holistique" pour mener les affaires de manière plus efficace et stimuler l'innovation. L'implémentation la plus courante de ce cas d'utilisation est les services de données de "vue unique du client" pour les centres de contact et les portails libre-service des clients.

La connaissance et l'insight sur les clients conduisent à des expériences client positives et sont le moteur de toute entreprise. Le succès de la relation entre une organisation et ses clients est déterminé par la qualité de leurs interactions tout au long du cycle de vie du client. Pendant ces moments cruciaux, les organisations peuvent créer la fidélité, se différencier de leurs concurrents et augmenter la valeur de la relation avec leurs clients. Si elles ne sont pas gérées correctement, ces interactions peuvent compromettre le succès de cette relation. L'un des principaux défis auxquels les organisations sont confrontées aujourd'hui est la nécessité de créer une vue unique, précise, cohérente et opportune de leurs clients - une vue qui traverse toutes les applications, systèmes, unités commerciales et points de contact client.

Le centre de contact d'une organisation est un point de contact client clé. Pour beaucoup, le centre de contact est le canal principal pour fournir des expériences enrichissantes de service client. Pour d'autres, le centre de contact n'est pas seulement une partie du service client mais aussi un canal de vente et de marketing. Les deux types d'organisations nécessitent un accès rapide à une gamme diversifiée d'informations sur le client. L'interaction positive et correcte avec les clients dans le centre de contact nécessite non seulement une bonne infrastructure de communication, mais surtout la mise en œuvre de processus internes et d'une infrastructure d'information étroitement alignés sur les besoins des clients.

Malheureusement, les outils disponibles pour le personnel des centres de contact sont trop souvent insuffisants. La réalité est que, pour de nombreuses organisations, le manque de données complètes, cohérentes et opportunes nécessaires à leurs processus orientés vers le client est un obstacle majeur à la maximisation de la valeur commerciale dans les interactions avec les clients. Les données fragmentées et le manque d'intégration dans les processus commerciaux créent des obstacles qui réduisent la qualité du service client, entraînent la perte d'opportunités commerciales et conduisent à une satisfaction client médiocre. Ces obstacles contribuent également

Les solutions de virtualisation des données centrées sur le client comprennent :

- Bureaux unifiés pour les centres de contact

- Portails web de libre-service améliorés

- Vue unique du client

- Surveillance des performances en temps réel

- Accès automatisé aux informations web externes

Les avantages de la virtualisation des données incluent :

- Amélioration du service client à moindre coût

- Amélioration de la motivation et de l'efficacité du personnel

- Réduction du taux de désabonnement des clients

- Différenciation concurrentielle accrue

- Amélioration des capacités de cross-selling et up-selling

VIRTUALISATION DES DONNÉES : Accéder et intégrer des sources de données distribuées.

Cependant, la compréhension de la virtualisation des données ne sera pas complète si nous ne savons pas ce qu'est la consolidation des données.

Consolidation des Données

La consolidation des données consiste à combiner des données disparates dans un seul dépôt, généralement un entrepôt de données, via le processus d'extraction, transformation et chargement (ETL). Que ce soit ETL ou ELT, le concept de base reste le même : des volumes massifs de données provenant de nombreuses sources disjointes sont copiés vers un nouveau système consolidé, subissant des transformations en cours de route.

Avantages et limitations de la virtualization des données

Accès en temps réel

Accès en temps réel : Au lieu de déplacer physiquement les données vers un nouvel emplacement, la virtualisation des données permet aux utilisateurs d'accéder aux données et de les manipuler depuis la source via la couche virtuelle/middleware. L'ETL n'est généralement pas nécessaire.

Les informations sont instantanément disponibles pour une grande variété de fonctions de reporting et d'analyse, accélérant et améliorant considérablement le processus de prise de décision.

Faibles coût: La mise en œuvre de la virtualisation des données nécessite moins de ressources et d'investissements par rapport à la construction d'un stockage de données consolidé.

Sécurité améliorée:Les données n'ont pas besoin d'être déplacées et le niveau d'accès peut être géré, éliminant ainsi les risques de fuite ou de perte de données lors de l'extraction ou de la transformation physique.

Agilité:Toutes les données de l'entreprise sont disponibles via une seule couche virtuelle pour différents utilisateurs et divers cas d'utilisation.

La conception et la réalisation d'analyses et de rapports BI peuvent être effectuées facilement sans se soucier des formats de données et/ou de l'emplacement des données.

Gouvernance des données cohérente et sécurisée:Avoir un point d'accès unique aux données, au lieu de multiples points pour chaque département, permet une gestion simple des utilisateurs et des permissions avec une conformité totale au RGPD.

Les KPI et les règles sont définis de manière centralisée pour garantir une compréhension et une gestion uniformes des métriques critiques à l'échelle de l'entreprise

La métadonnée globale améliore la gouvernance des données de haute qualité et offre une meilleure compréhension des données d'entreprise grâce à la traçabilité des données et aux catalogues de métadonnées (en fonction de l'outil).

Les erreurs sont détectées et résolues plus rapidement avec la virtualisation des données par rapport à d'autres approches d'intégration des données en raison de l'accès en temps réel aux données.

Voici quelques limitations de la virtualisation des données:

Point de défaillance unique:Étant donné que le serveur fournit un point d'accès unique à toutes les sources de données, cela pourrait entraîner un point de défaillance unique. Si le serveur est en panne, le risque que tous les systèmes opérationnels soient en panne sans flux de données est élevé.

Les données en batch ne sont pas prises en charge:Les données en batch sont une collecte de données programmée ou planifiée. La virtualisation des données ne prend pas en charge ce type de mouvement de données en vrac qui peut être nécessaire dans un certain nombre de cas. Un exemple est une organisation financière qui doit analyser de grands volumes de données une fois par semaine.Alors, quand la virtualisation des données est-elle une bonne idée ? Quand une organisation devrait-elle envisager d'utiliser la virtualisation des données par rapport à l'ETL.La virtualisation des données peut être une bonne alternative à l'ETL dans un certain nombre de situations différentes.

Le déplacement physique des données, principalement via ETL, est inefficace, difficile ou trop coûteux.

Un environnement flexible est nécessaire pour prototyper, tester et mettre en œuvre de nouvelles initiatives.

Lorsque les données doivent être disponibles en temps réel ou quasi temps réel à des fins analytiques diverses, la DV peut être considérée comme une alternative à l'ETL.

Plusieurs outils BI nécessitent un accès aux mêmes sources de données.

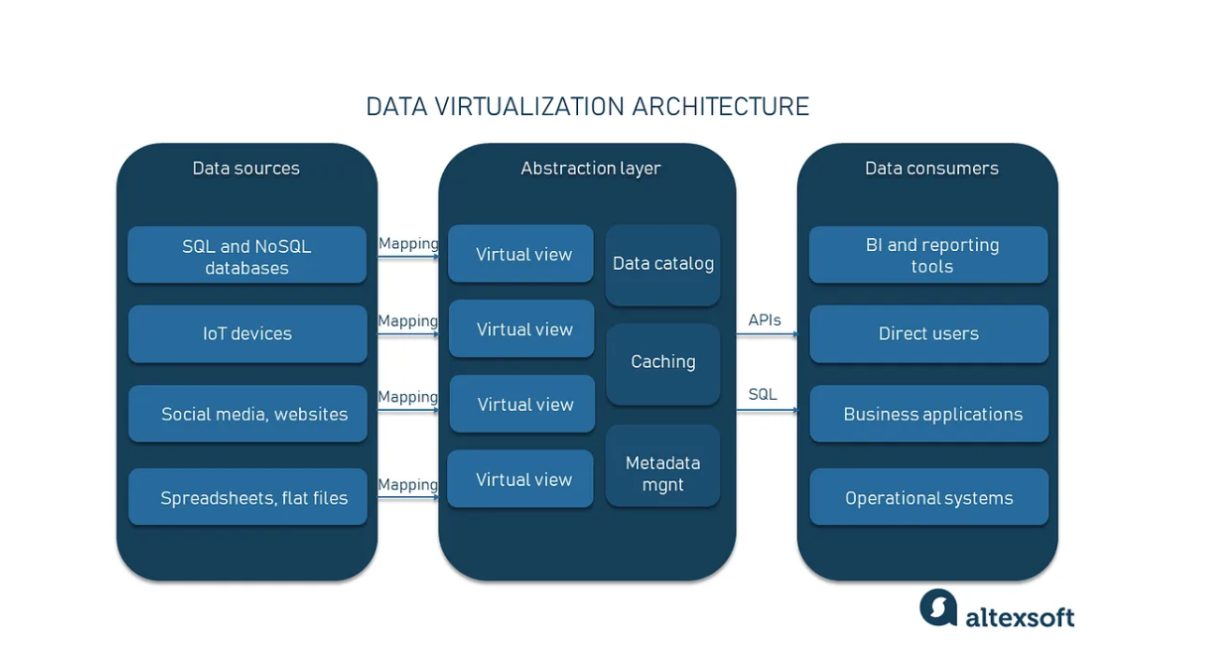
Lorsque les données sont fournies pour analyse, la virtualisation des données peut aider à résoudre les problèmes liés à la confidentialité. La virtualisation permet de combiner des données personnelles provenant de différentes sources sans les copier physiquement à un autre endroit tout en limitant la vue à toutes les autres variables collectées.

Il y a trois blocs de construction qui composent la structure de virtualisation, à savoir

Couche de connexion — un ensemble de connecteurs aux sources de données en temps réel ;

Couche d'abstraction — des services pour présenter, gérer et utiliser des vues logiques des données, parfois appelées couche virtuelle ou sémantique ;

Couche de consommation — une gamme d'outils et d'applications consommateurs demandant des données abstraites.

 Figure :

Comment fonctionne la virtualisation des données

La couche de données virtuelle/couche sémantique

Le cœur d'une application de virtualisation des données est la couche virtuelle ou sémantique, qui permet aux utilisateurs de données ou aux utilisateurs métiers de manipuler, de joindre et de calculer les données indépendamment de leur format source et de leur emplacement physique, qu'il s'agisse de stockage dans le cloud ou sur site.

Bien que toutes les sources de données connectées et les métadonnées associées apparaissent dans une seule interface utilisateur, la couche virtuelle permet à l'utilisateur d'organiser davantage ses données dans différents schémas virtuels et vues virtuelles. Les utilisateurs peuvent facilement enrichir les données brutes des systèmes sources avec une logique métier simple et préparer les données pour les processus d'analytique, de reporting et d'automatisation.

Comment fonctionne la virtualisation des données

La couche de données virtuelle/couche sémantique

Le cœur d'une application de virtualisation des données est la couche virtuelle ou sémantique, qui permet aux utilisateurs de données ou aux utilisateurs métiers de manipuler, de joindre et de calculer les données indépendamment de leur format source et de leur emplacement physique, qu'il s'agisse de stockage dans le cloud ou sur site.

Bien que toutes les sources de données connectées et les métadonnées associées apparaissent dans une seule interface utilisateur, la couche virtuelle permet à l'utilisateur d'organiser davantage ses données dans différents schémas virtuels et vues virtuelles. Les utilisateurs peuvent facilement enrichir les données brutes des systèmes sources avec une logique métier simple et préparer les données pour les processus d'analytique, de reporting et d'automatisation.

Produits de virtualisation des données

Plusieurs produits de virtualisation des données étendent cette couche virtuelle avec des capacités de gouvernance des données et d'exploration des métadonnées, bien que cette fonctionnalité ne soit pas incluse dans tous les outils.

Gestion des permissions

Avec une gestion sophistiquée des permissions basée sur les utilisateurs, la couche virtuelle crée une source unique de vérité pour toute l'organisation de manière totalement conforme et sécurisée. Les utilisateurs autorisés peuvent accéder aux données dont ils ont besoin depuis un seul point dans un seul outil, aidant à éliminer les silos de données supplémentaires et à simplifier l'architecture des données.

La virtualisation des données ne persiste généralement pas les données des systèmes sources, contrairement à la réplication simple des données comme les outils ETL traditionnels.

Au lieu de cela, la virtualisation des données stocke les métadonnées pour alimenter les vues virtuelles et permettre la création de logique d'intégration qui aide à fournir les données intégrées des systèmes sources, en temps réel, à toute application frontale telle que :

- Outils de Business Intelligence (BI) et plateformes d'analyse de données

- Programmes et outils personnalisés

- Microservices

Principaux avantages de la virtualisation des données

Utiliser la virtualisation des données pour intégrer des données métiers provenant de sources disparates présente plusieurs avantages :

Temps de solution plus rapide

- Grâce à l'accès immédiat aux données, toutes les données peuvent être intégrées en temps réel sans connaissances techniques approfondies ou effort de codage manuel.

- L'accès en temps réel différencie la virtualisation des données des approches d'intégration plus lentes de type batch, où les données peuvent souffrir d'un manque de fraîcheur et de précision.

- La virtualisation des données assure une conception plus rapide et un prototypage rapide, créant un retour sur investissement (ROI) beaucoup plus rapide.

- Les informations sont instantanément disponibles pour une grande variété de fonctions de reporting et d'analyse, accélérant et améliorant considérablement le processus de prise de décision.

Flexibilité et simplicité

- Le prototypage rapide permet des cycles de test plus rapides avant de passer aux environnements de production.

- Les sources de données apparaissent dans une interface unifiée, ce qui signifie que la virtualisation des données masque la complexité sous-jacente d'un paysage de données hétérogène.

- La couche virtuelle permet aux utilisateurs d'adapter rapidement la logique métier aux demandes changeantes.

Rentabilité

- Contrairement aux entrepôts de données traditionnels, aucune infrastructure supplémentaire n'est requise car les données restent dans les systèmes sources. Cette approche est moins coûteuse que les solutions ETL traditionnelles où les données sont transformées en différents formats avant d'être déplacées vers une zone de stockage cible.

- Un changement dans les sources de données ou les solutions frontales ne nécessite pas de restructuration coûteuse et d'effort complexe.

- La virtualisation des données agit comme un middleware, permettant d'intégrer sans effort l'infrastructure existante avec de nouvelles applications tout en éliminant le besoin de silos de données inutiles et coûteux.

Gouvernance des données cohérente et sécurisée

- Avoir un point d'accès unique aux données, au lieu de multiples points pour chaque département, permet une gestion simple des utilisateurs et des permissions avec une conformité totale au RGPD.

- Les KPI et les règles sont définis de manière centralisée pour garantir une compréhension et une gestion uniformes des métriques critiques à l'échelle de l'entreprise.

- La métadonnée globale améliore la gouvernance des données de haute qualité et offre une meilleure compréhension des données d'entreprise grâce à la traçabilité des données et aux catalogues de métadonnées (en fonction de l'outil).

- Les erreurs sont détectées et résolues plus rapidement avec la virtualisation des données par rapport à d'autres approches d'intégration des données en raison de l'accès en temps réel aux données.

Considérations de conception de la virtualisation des données

Les plateformes de virtualisation des données offrent de nombreux avantages par rapport aux solutions de données traditionnelles.

Cependant, il existe certaines contraintes à prendre en compte lors de la conception de votre solution :

- La technologie de virtualisation des données fonctionne en accédant aux données sources en temps réel via les systèmes de production. Cela contraste avec un entrepôt de données ou une solution de gestion des données maîtres où les données sont souvent stockées dans un stockage pré-agrégé/classé et offrent donc généralement des temps de réponse plus rapides.

- La virtualisation des données n'est pas capable de fournir une analyse historique des données. Un entrepôt de données ou une base de données analytique est généralement requis, ce qui n'est pas le concept original de la virtualisation des données.

- Le nettoyage des données et/ou la transformation des données peut encore être une tâche complexe dans la couche virtuelle.

- Les modifications du modèle de données virtuel peuvent parfois nécessiter un effort accru car pour qu'une modification soit pleinement appliquée, elle doit être acceptée par toutes les applications et utilisateurs consommateurs.

- L'exigence initiale de la virtualisation des données était de récupérer des données en utilisant un seul langage de requête pour fournir une réponse rapide aux requêtes et pour assembler rapidement différents modèles de données ou vues des données pour répondre à des besoins spécifiques. Cependant, en réalité, cet objectif n'a pas encore été réalisé dans chaque produit.

Cas d'utilisation de la virtualisation des données

Mart de données virtuel Un mart de données fournit une vue agrégée des données, généralement extraites d'un entrepôt de données traditionnel. La virtualisation des données facilite la création d'un mart de données virtuel pour plus de rapidité.

En combinant l'infrastructure de données principale de l'organisation avec des sources de données auxiliaires pertinentes pour des unités commerciales spécifiques basées sur les données, les initiatives peuvent avancer plus rapidement que si les données devaient être intégrées à un entrepôt de données traditionnel.

Prototypage rapide

Les entreprises modernes et agiles aiment expérimenter de nouvelles idées et modèles commerciaux – principalement soutenus par des données pour à la fois mettre en œuvre l'initiative et mesurer le succès. Par conséquent, un système flexible est nécessaire pour tester, ajuster et mettre en œuvre de nouvelles idées.

Avec l'entrepôt de données logique, le composant de virtualisation des données peut être utilisé pour une configuration rapide, des itérations plus rapides et des capacités de matérialisation des données pour facilement passer les données en production au besoin. Le moteur de recommandation intégré analyse l'utilisation des données prototypes et fait des suggestions sur la meilleure façon de stocker les données pour une utilisation en production, y compris la création automatique d'index de base de données et d'autres optimisations.

La virtualisation des données dans les lacs de données offre une solution robuste en créant une couche virtuelle sur diverses sources de données IoT. Voici comment cela fonctionne et pourquoi c'est crucial :

1. Accès unifié aux données : Le personnel de la banque peut accéder aux données IoT en temps réel à partir de plusieurs sources via une interface unique, sans déplacer physiquement ou dupliquer les données.

2. Intégration en temps réel : La virtualisation des données permet l'intégration et l'analyse en temps réel des données IoT, facilitant ainsi les insights immédiats et la prise de décision proactive.

3. Flexibilité et agilité : Les banques peuvent rapidement adapter leurs analyses et leurs rapports aux besoins de données changeants ou aux exigences réglementaires, sans nécessiter de restructuration importante des données.

4. Efficacité des coûts : En éliminant le besoin de duplication des données et en réduisant les coûts de stockage, la virtualisation des données optimise les dépenses globales de gestion des données.

5. Amélioration de la gouvernance des données : L'accès centralisé aux données via une couche virtuelle renforce la sécurité, la confidentialité et la conformité aux normes réglementaires.

Étape 1 : Intégration des appareils IoT

Les appareils IoT tels que les DAB et les terminaux de transaction sont intégrés dans le lac de données, où les données sont stockées de manière distribuée.

Étape 2 : Création de la couche de virtualisation

Une couche de virtualisation des données est mise en place pour abstraire les emplacements physiques et les formats des données IoT, offrant ainsi une vue unifiée.

Étape 3 : Configuration des requêtes en temps réel

Des requêtes et analyses en temps réel sont configurées pour accéder et intégrer les données IoT de manière transparente via la couche de virtualisation.

Étape 4 : Surveillance et alertes en temps réel

Les données sont surveillées en temps réel pour détecter les anomalies, déclencher des alertes et faciliter les actions correctives immédiates

Exemple d'application pratique

Une banque utilise des capteurs IoT installés dans les DAB pour surveiller les volumes de transactions, l'état des appareils et les interactions avec les clients en temps réel. Avec la virtualisation des données, la banque peut :

- Surveiller en temps réel les métriques de performance des DAB.

- Détecter immédiatement les modèles de transactions inhabituels ou les problèmes opérationnels.

- Déclencher des alertes en temps réel aux directeurs de succursale ou aux équipes de maintenance pour une intervention rapide.

- Analyser les données historiques des DAB pour optimiser la gestion des liquidités et l'efficacité opérationnelle.

Conclusion

La virtualisation des données dans les lacs de données révolutionne la manière dont les banques accèdent et analysent les données IoT en temps réel. Elle offre une solution flexible, agile et rentable pour intégrer et exploiter les données IoT, permettant aux banques d'améliorer l'efficacité opérationnelle, la sécurité et le service client. En adoptant la virtualisation des données, les banques renforcent leur capacité à surveiller, analyser et optimiser leurs opérations en temps réel, assurant ainsi des performances robustes et une conformité réglementaire

Les organisations reconnaissent que pour prendre des décisions plus intelligentes, ravir leurs clients et surpasser leurs concurrents, elles doivent exploiter leurs actifs de données plus efficacement.

### Fournisseurs de virtualisation des données

- CData Virtuality: CData Virtuality est une plateforme d'intégration de données pour un accès instantané aux données, une centralisation facile des données et une gouvernance des données d'entreprise. CData Virtuality combine deux technologies distinctes, la virtualisation des données et la réplication des données, pour une architecture haute performance et une livraison de données flexible.

- IBM Cloud Pak for Data: IBM Cloud Pak for Data, anciennement connu sous le nom de IBM Cloud Private for Data, est une plateforme de données et d'IA qui aide à collecter, organiser et analyser les données, tout en utilisant la virtualisation des données.

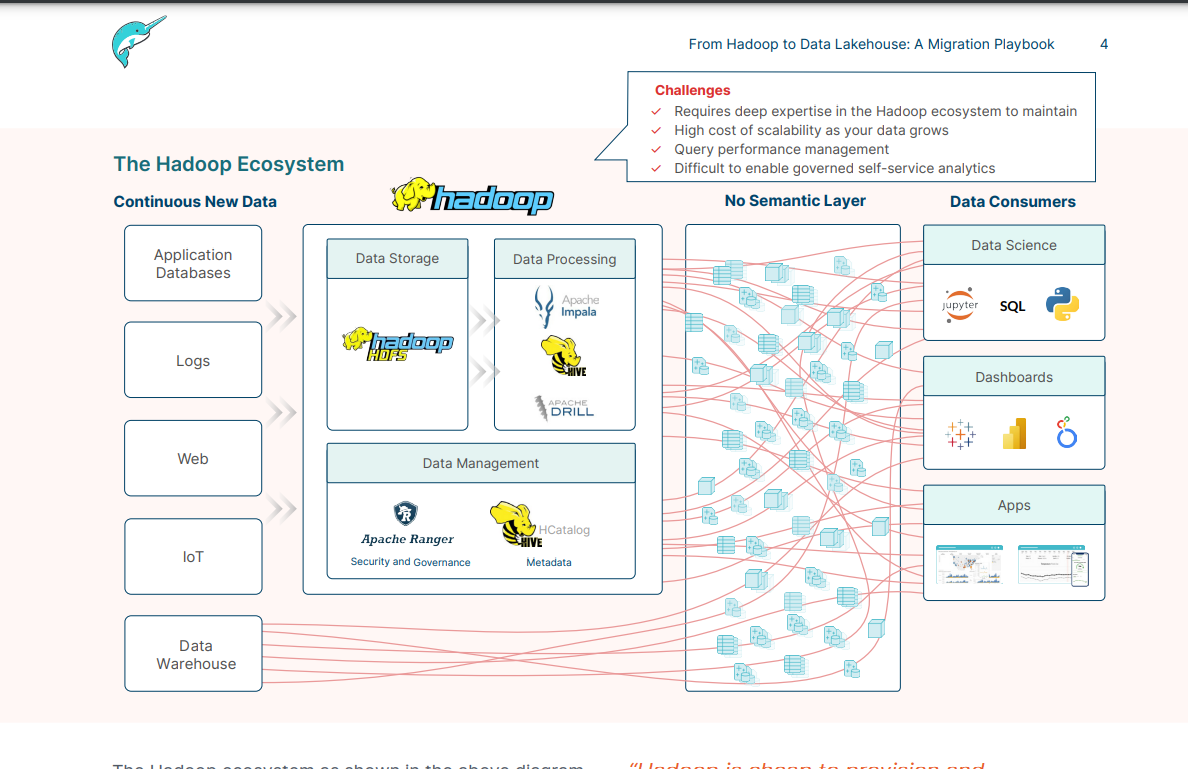
- Denodo: Denodo propose une plateforme de virtualisation des données avec une fonctionnalité de catalogue de données associée, permettant aux utilisateurs de non seulement combiner mais aussi identifier et structurer les données existantes.

- Informatica PowerCenter : Informatica PowerCenter est une plateforme d'intégration de données d'entreprise avec des fonctionnalités telles que l'archivage des données des anciennes applications ou une analyse d'impact pour examiner les modifications structurelles avant leur mise en œuvre.

- TIBCO : Le produit de virtualisation des données de TIBCO contient un annuaire des données d'entreprise pour aider les utilisateurs dans leurs analyses et un moteur de transformation intégré pour les sources de données non structurées.

# Starting our data virtualisation with dremio

Dans le cadre de sa transformation digitale visant à optimiser l'efficacité opérationnelle et à valoriser pleinement son écosystème de données, AttijaraWafa Bank a opté stratégiquement pour la virtualisation des données. Parmi les technologies évaluées à cette fin, Dremio s'est imposé comme la solution optimale, répondant aux besoins complexes d'intégration et d'analyse de données de la banque.



L'architecture et les capacités de Dremio sont étroitement alignées sur les objectifs d'AttijaraWafa Bank, offrant une plateforme unifiée pour un accès fluide à des sources de données diverses. Grâce à son moteur avancé d'optimisation de requêtes basé sur Apache Arrow, Dremio permet aux analystes et décideurs de la banque d'interroger et d'analyser les données avec une rapidité et une agilité inégalées, même à travers des environnements distribués.

Cette section du rapport explore la logique derrière le choix de Dremio, mettant en lumière sa puissance technologique, ses avantages de performance et son adéquation stratégique au cadre de gestion des données de la banque. En détaillant les critères ayant conduit à l'adoption de Dremio et en explorant ses implications pour la gouvernance des données, l'efficacité opérationnelle et l'agilité métier, cette étude vise à souligner l'impact transformateur de Dremio en tant que composant essentiel de la stratégie de virtualisation des données d'AttijaraWafa Bank.

Dremio est considéré comme une solution solide pour la visualisation des données pour plusieurs raisons :

- Analytique Unifiée : Dremio fournit une plateforme analytique unifiée qui permet une analyse en libre-service et gouvernée sur l'ensemble de vos données. Cela signifie que les analystes de données et les scientifiques peuvent accéder et analyser les données facilement, favorisant ainsi un processus décisionnel plus orienté données.

- Performance: Il offre des temps de réponse aux requêtes de moins d'une seconde, ce qui est crucial pour l'analyse interactive et la prise de décision en temps réel. Cet aspect de performance est particulièrement important lors du traitement de grands ensembles de données et de requêtes complexes.

- Gestion des Data Lakehouse : Dremio supporte l'architecture moderne des data lakehouse, combinant les avantages des data lakes et des data warehouses. Cela permet aux organisations de gérer leurs données plus efficacement et de les exploiter à des fins analytiques.

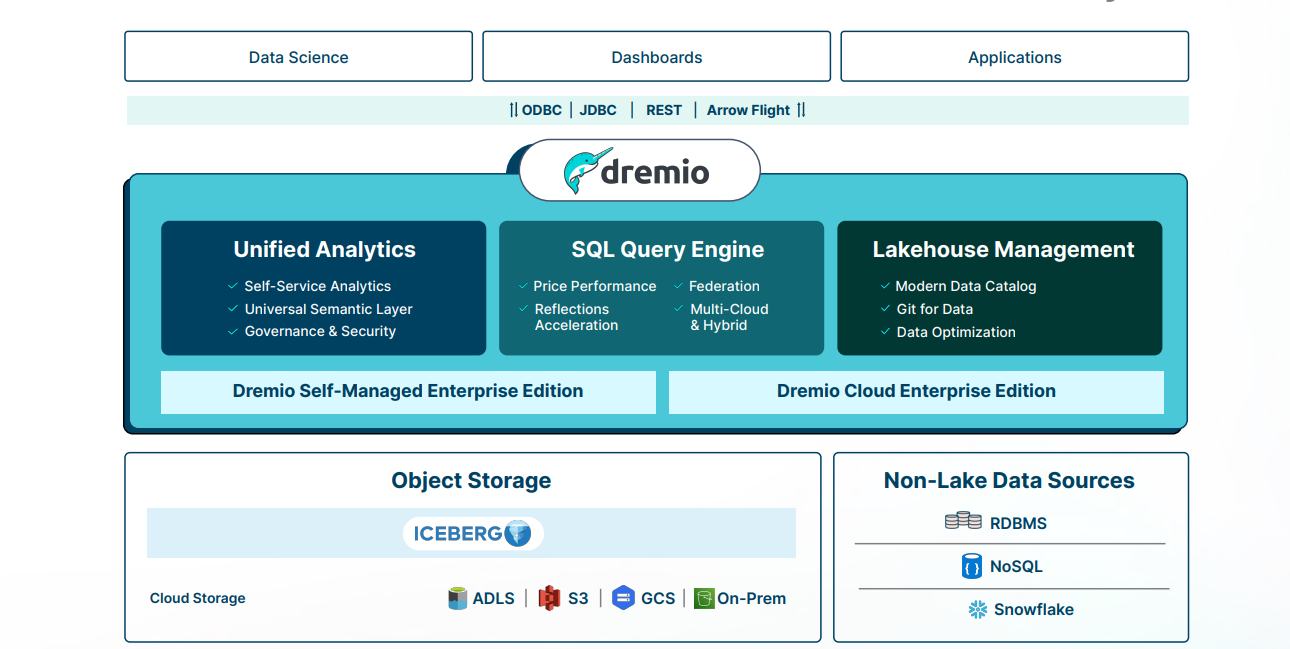
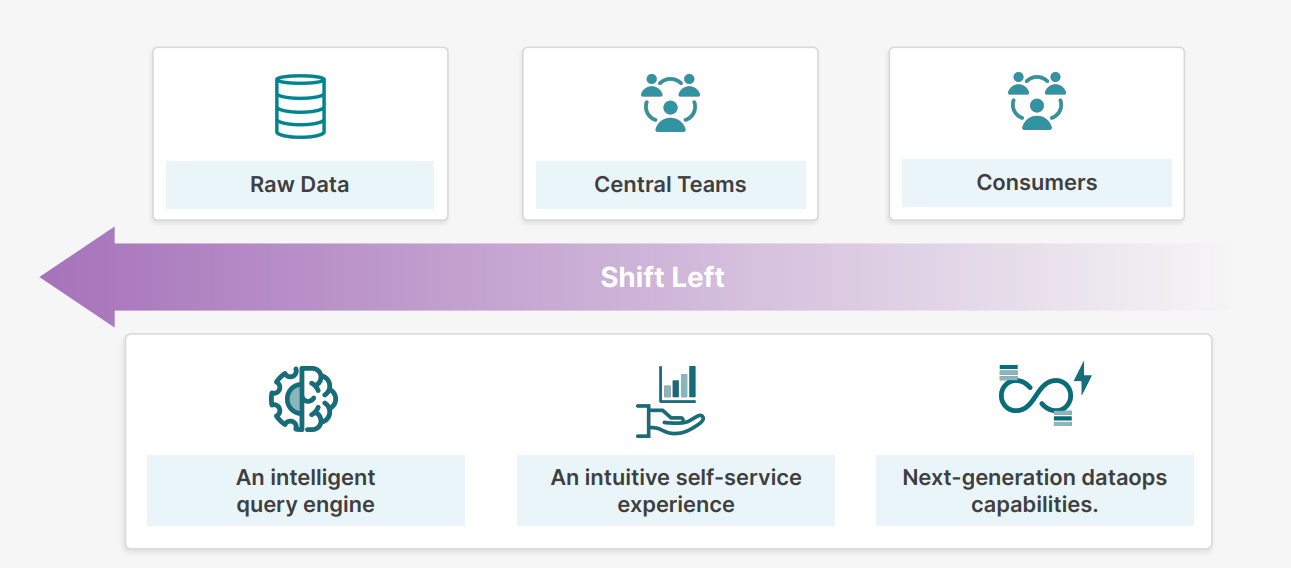
- Architecture de Données Ouverte : Construit sur des projets open-source clés, y compris les contributions dirigées par Dremio, il favorise la flexibilité et la scalabilité. Cela est bénéfique pour les organisations qui souhaitent éviter le verrouillage par un fournisseur et maintenir le contrôle sur leur infrastructure de données.

- Rentabilité : Dremio vise à offrir le coût total de possession le plus bas pour l'analyse. En réduisant le besoin de réplication des données et les coûts de stockage, il peut considérablement diminuer les dépenses liées à la gestion des données.

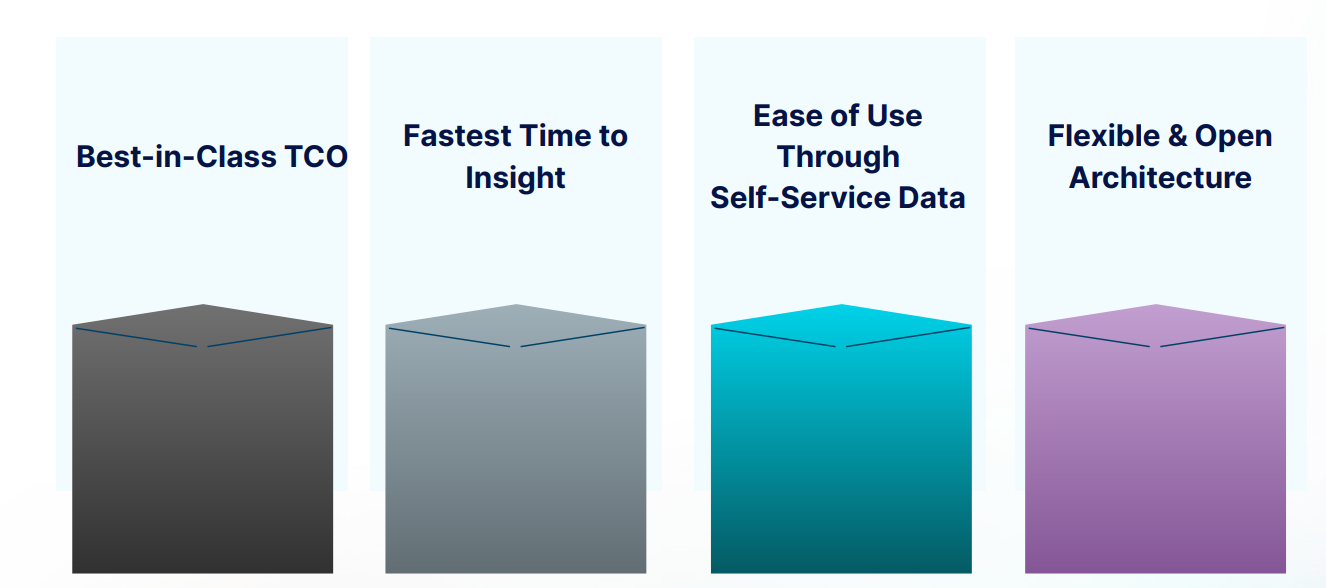
- Analytique en Libre-Service : La plateforme permet aux équipes de données d'explorer et de visualiser les données sans avoir besoin d'un support informatique étendu. Cela permet à davantage d'utilisateurs au sein d'une organisation de travailler avec les données et de générer des insights.

L'approche de Dremio pour le traitement des requêtes est conçue pour être conviviale, permettant une exploration et une manipulation faciles des données tout en assurant des performances optimales et une utilisation efficace des ressources.

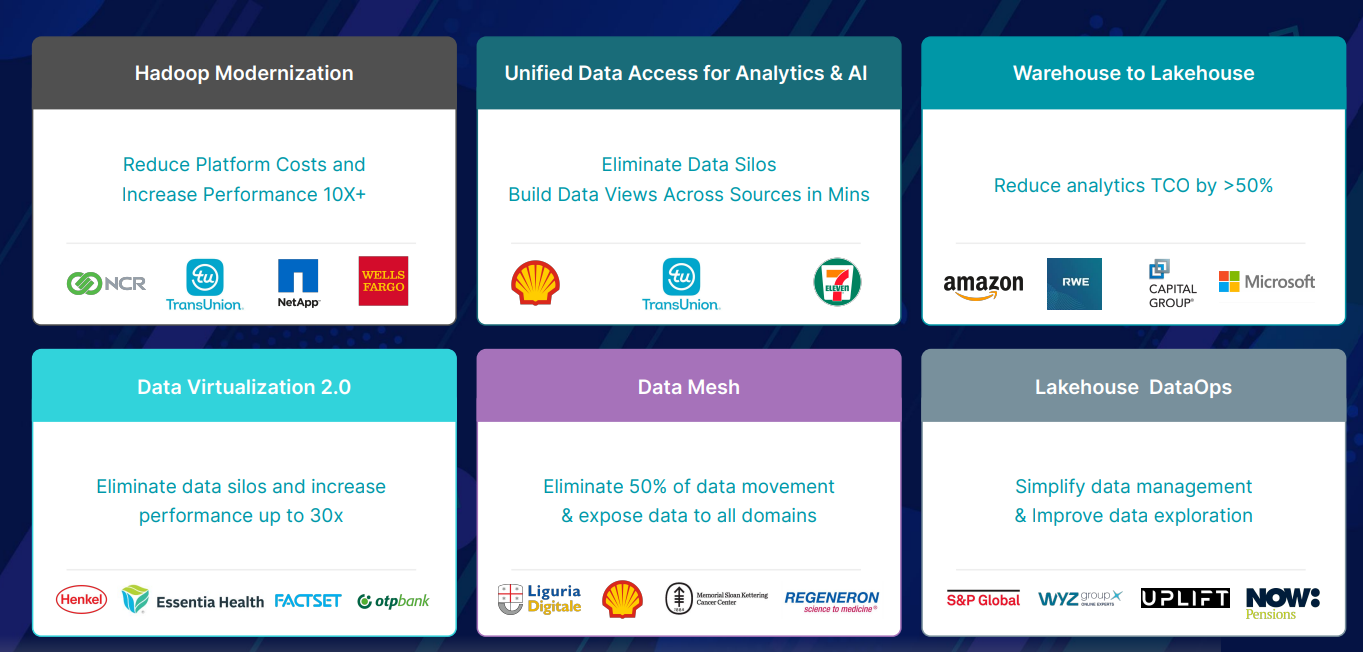
##### Dremio - Déplacez vos analyses de données vers la gauche

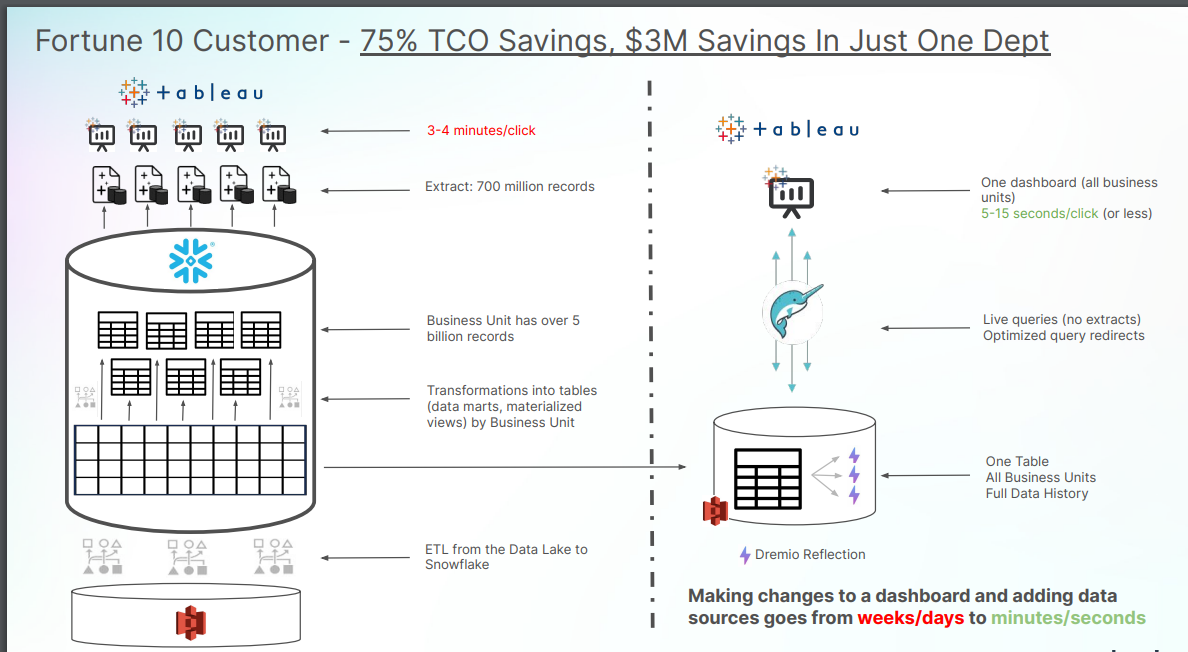


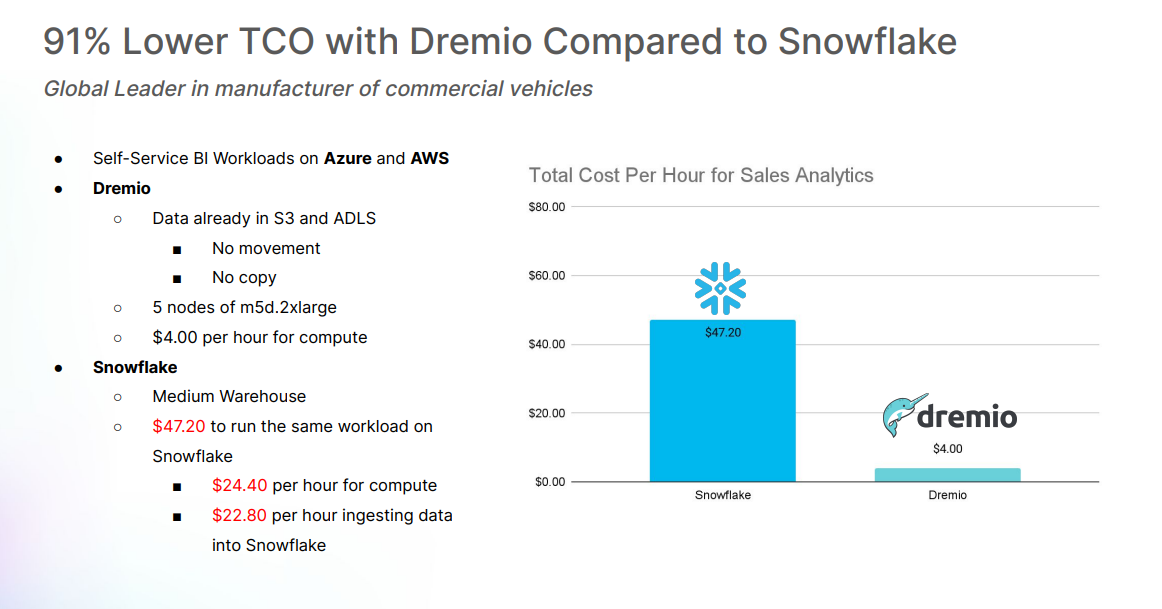
##### The Dremio Difference



##### Cas d'utilisation de Dremio

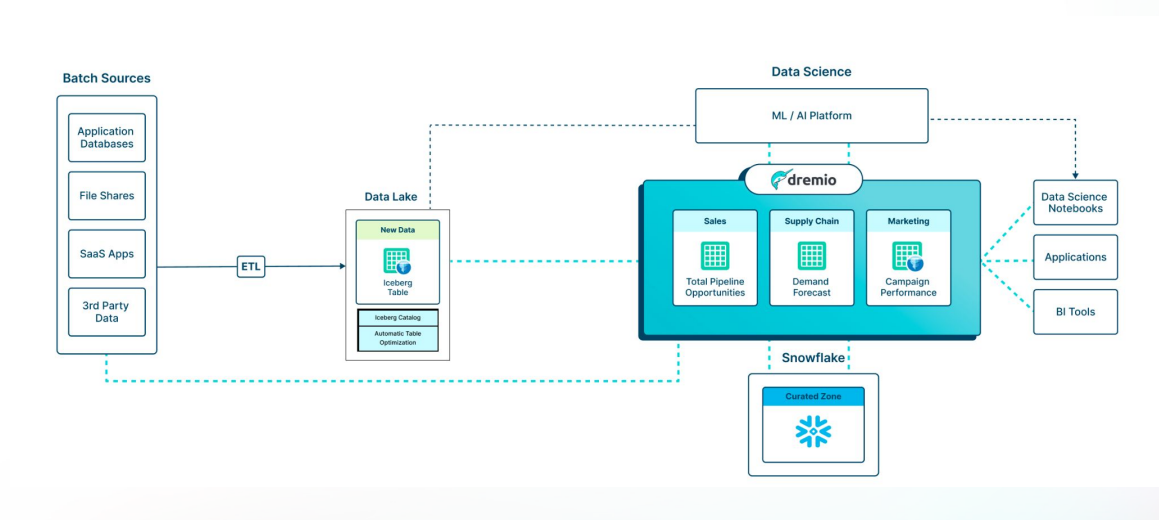






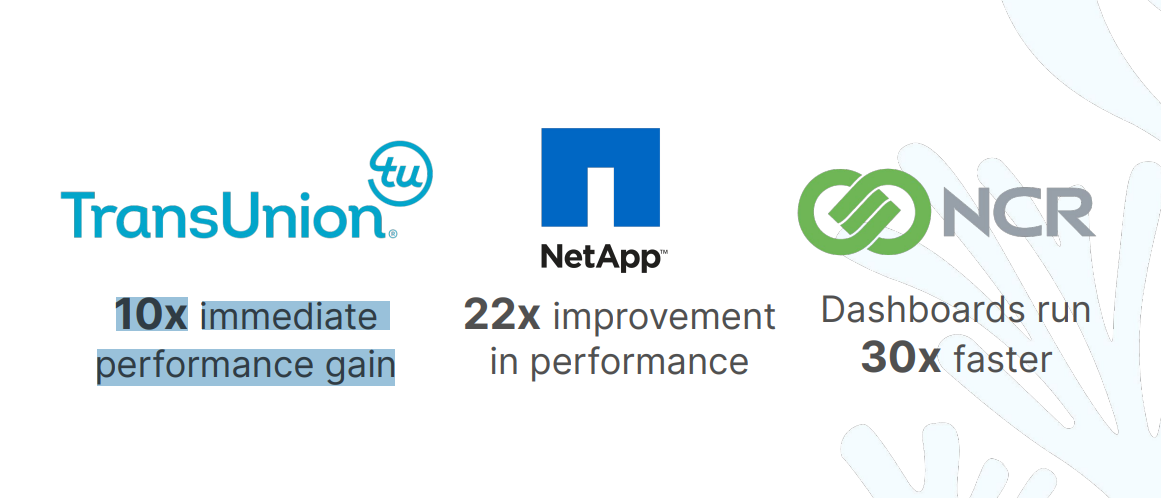
**Warehouse et Lakehouse - Mieux Ensemble**

Optimisation des Coûts | Fournir une Analyse Unifiée et Autonome | Architecture Ouverte

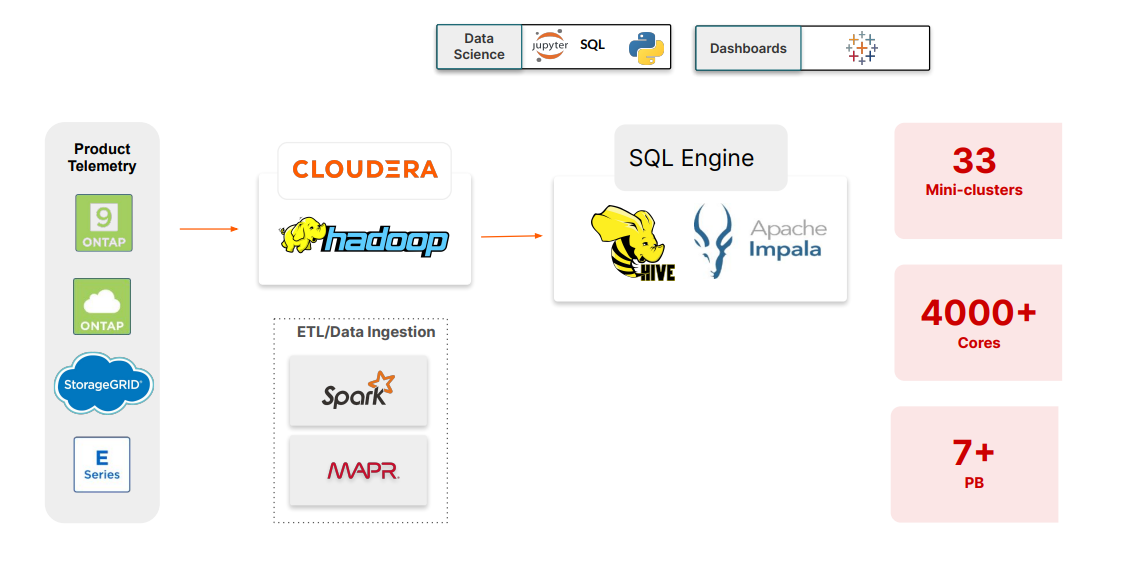


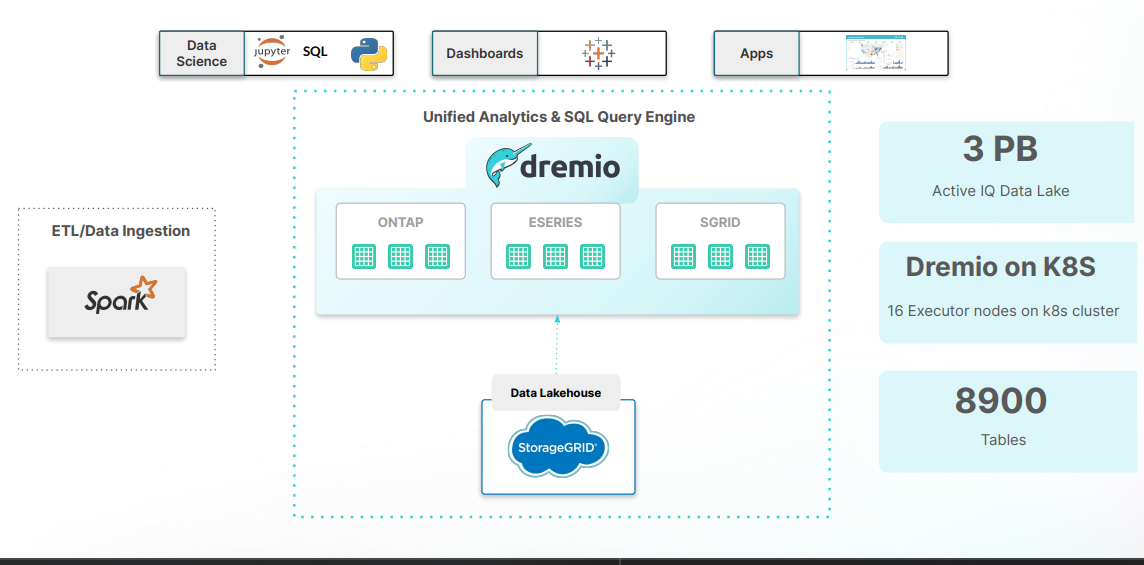
#### Hadoop Migration

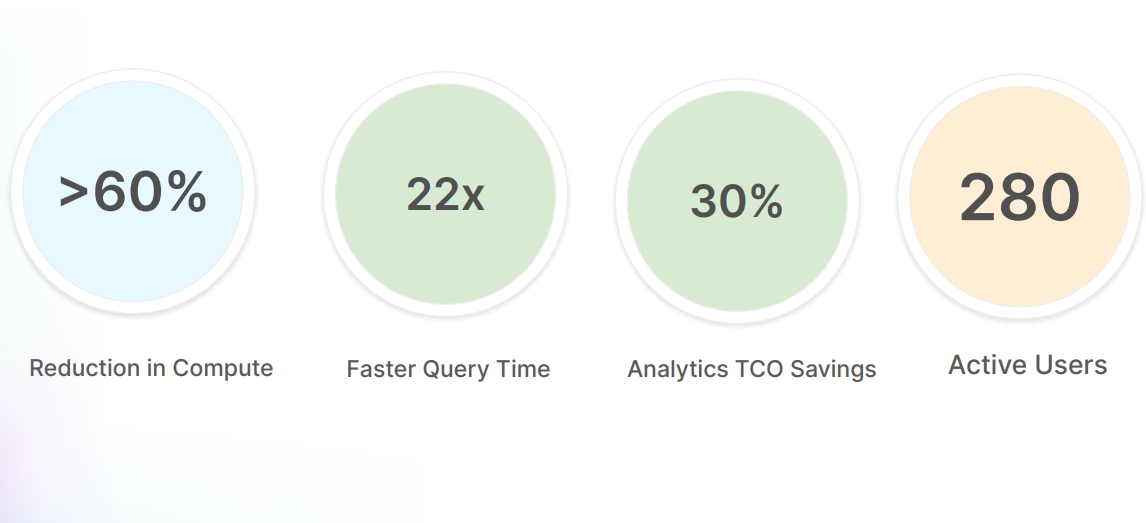
Rapid Hadoop Modernization to Drive Value 10x immediate performance gain Self-fund Dremio in 3-months cost savings

Hadoop Modernization: Install, Query, Save

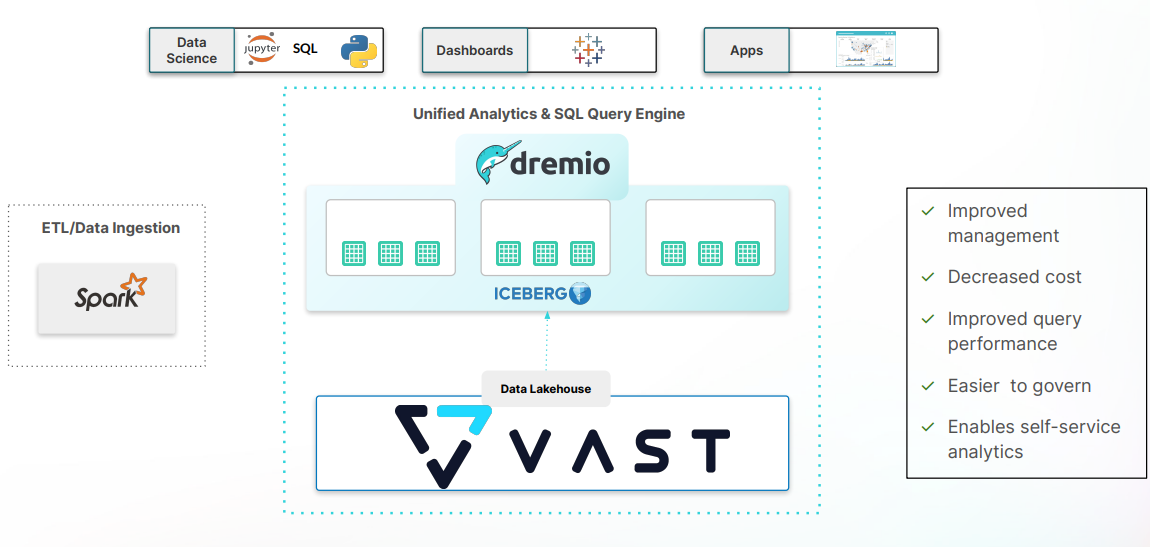
Architecture Before Dremio

Architecture After Dremio

Business Outcomes

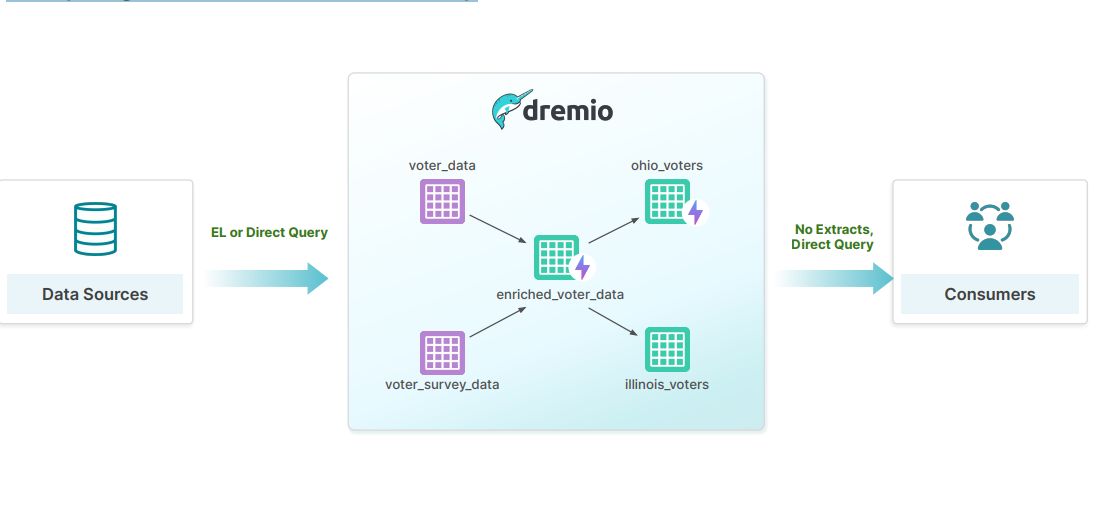


##### Dremio & VAST

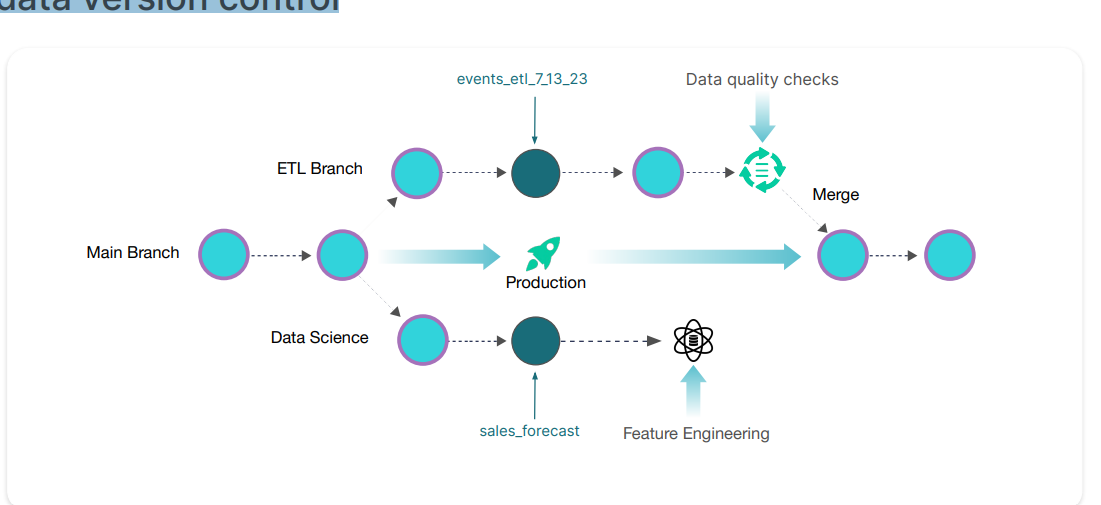
Reflections Intelligent Query Engin

##### Reflections - Transparent Query Acceleration

Query Engine Rewrites SQL on the Fly

Git-Inspired Data Versioning Next-Gen DataOps

To simplify pipeline management, we’ve created git-inspired data version control



#### Denodo et Dremio basée sur différents aspects :

Facilité d'utilisation

- Denodo : Les utilisateurs le trouvent relativement facile à utiliser, mais il peut y avoir une courbe d'apprentissage pour ceux qui découvrent la virtualisation des données.

- Dremio : Généralement considéré comme plus facile à utiliser, à configurer et à administrer par rapport à Denodo. Il est apprécié pour son interface conviviale et sa simplicité.

Performance

- Denodo : Offre des mécanismes de mise en cache des données et optimise l'exécution grâce à la mise en cache des ensembles de résultats et à la parallélisation des requêtes pour améliorer les performances des requêtes.

- Dremio : Se concentre sur l'accélération des requêtes pour les charges de travail de big data, en utilisant des technologies telles que Apache Arrow et le stockage colonnaire pour des performances élevées.

Support des Sources de Données

- Denodo : Supporte une large gamme de sources de données, y compris les bases de données relationnelles, les sources de big data, les plateformes cloud et les services web.

- Dremio : Plus axé sur les sources de big data telles que Hadoop, les bases de données NoSQL et les systèmes de stockage cloud, avec un support natif et une optimisation pour ces types de données.

Gouvernance des Données et Sécurité

- Denodo : Propose des fonctionnalités robustes de gouvernance des données et de sécurité

incluant un contrôle d'accès détaillé, le masquage des données, le chiffrement et des capacités d'audit.

- Dremio : Offre des fonctionnalités de sécurité de base comme l'authentification des utilisateurs et l'autorisation, mais peut ne pas avoir autant de capacités avancées de gouvernance des données que Denodo.

Transformation et Intégration des Données

- Denodo : Propose des outils complets pour l'intégration et la transformation des données, y compris une interface ETL visuelle et des capacités d'automatisation des pipelines de données.

- Dremio : Se concentre principalement sur la virtualisation et l'exploration des données, avec des capacités de transformation des données plus limitées.

Options de Déploiement

- Denodo : Peut être déployé sur site, dans le cloud ou dans un environnement hybride, prenant en charge diverses plateformes cloud et options de mise à l'échelle.

- Dremio : Conçu principalement pour les environnements cloud, offrant une mise à l'échelle automatique et une optimisation de l'infrastructure pour les déploiements basés sur le cloud.

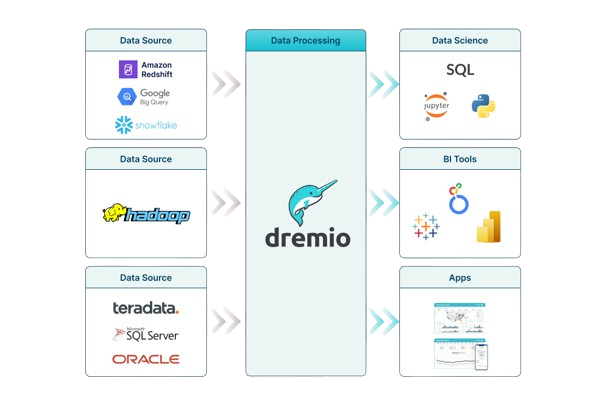
Communauté et Support

- Denodo : Dispose d'une communauté large et active avec un support technique complet.

- Dremio : Dispose également d'une communauté solide et est reconnu pour la qualité de son support produit continu.

Cette comparaison devrait mettre en lumière les principales différences et forces de chaque plateforme en fonction de leurs fonctionnalités et capacités respectives.

Cas d'utilisation : Les organisations adoptent la virtualisation des données pour fournir un accès aux rapports BI et analytiques aux consommateurs de données. Elle abstrait la complexité des systèmes sous-jacents et de l'infrastructure, promettant un accès rapide aux données. Cependant, à mesure que la solution se développe, les performances peuvent se dégrader et les coûts augmenter en raison de la nécessité d'une infrastructure de virtualisation robuste.



Dremio est souvent préféré pour sa facilité d'utilisation, ses performances et son support, notamment dans les environnements cloud et big data. Denodo, quant à lui, est reconnu pour ses capacités complètes de gouvernance des données, de sécurité et d'intégration. Le choix entre les deux doit se baser sur vos besoins spécifiques en virtualisation des données, votre infrastructure existante et vos objectifs stratégiques

Pourquoi j'ai choisi Dremio

J'ai sélectionné Dremio principalement pour sa capacité à exécuter des requêtes ultra-rapides sur vos données, quelle que soit la source, en faisant un outil puissant pour l'accélération analytique. Sa capacité à organiser toutes les données dans un état plus gérable, performant et fiable le distingue des autres. Ma conviction que Dremio est le meilleur pour accélérer l'analytique repose sur son architecture efficace qui élimine les barrières de données, permettant une analyse des données plus rapide et plus précise

Définition : Dremio permet l'analyse en libre-service pour les équipes de données. Les analystes de données peuvent explorer et visualiser les données avec des temps de réponse quasi instantanés, tandis que les ingénieurs de données peuvent ingérer et transformer les données directement dans le data lake, avec un support complet pour les opérations DML. De plus, les analystes peuvent joindre les données du lac avec celles des bases de données externes, évitant ainsi de déplacer les données dans le stockage d'objets pour en tirer de la valeur. La plateforme open lakehouse de Dremio, basée sur des standards communautaires comme Apache Iceberg et Apache Arrow, permet aux organisations d'utiliser les meilleurs moteurs de traitement et d'éviter l'enfermement avec un fournisseur spécifique.

Architecture : Les lacs de données sur site sont souvent associés à Hadoop, tandis que des solutions basées sur le cloud comme Amazon S3 ou Microsoft ADLS offrent des options de stockage évolutives et rentables.

Avantages : Les lacs de données prennent en charge de grands volumes de données et des types de données diversifiés. Ils permettent l'interrogation directe des données dans leur format d'origine à partir de divers systèmes sources, favorisant la flexibilité et la scalabilité. Le concept émergent de "lakehouses" combine ces avantages avec les capacités traditionnelles des entrepôts de données pour une performance accrue et une gestion des données améliorée.

Data engineering aims to make data accessible and usable for data analytics and data science purposes. This involves several key aspects:

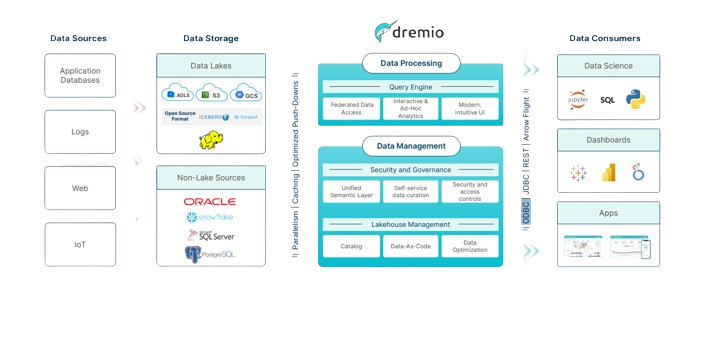
* Transferring data from operational systems like databases to systems optimized for analytical access.
* Modeling and optimizing data for improved accessibility and performance.
* Governing data access to ensure that only authorized individuals can access specific data.
* Creating abstractions to simplify data access.

This tutorial focuses on the initial step of moving data between systems, introducing various systems commonly used in modern data platforms. Specifically, [**we'll explore a "Data Lakehouse" architecture**](https://bit.ly/dremio-blog-why-lakehouse).

### What is a Data Lakehouse?

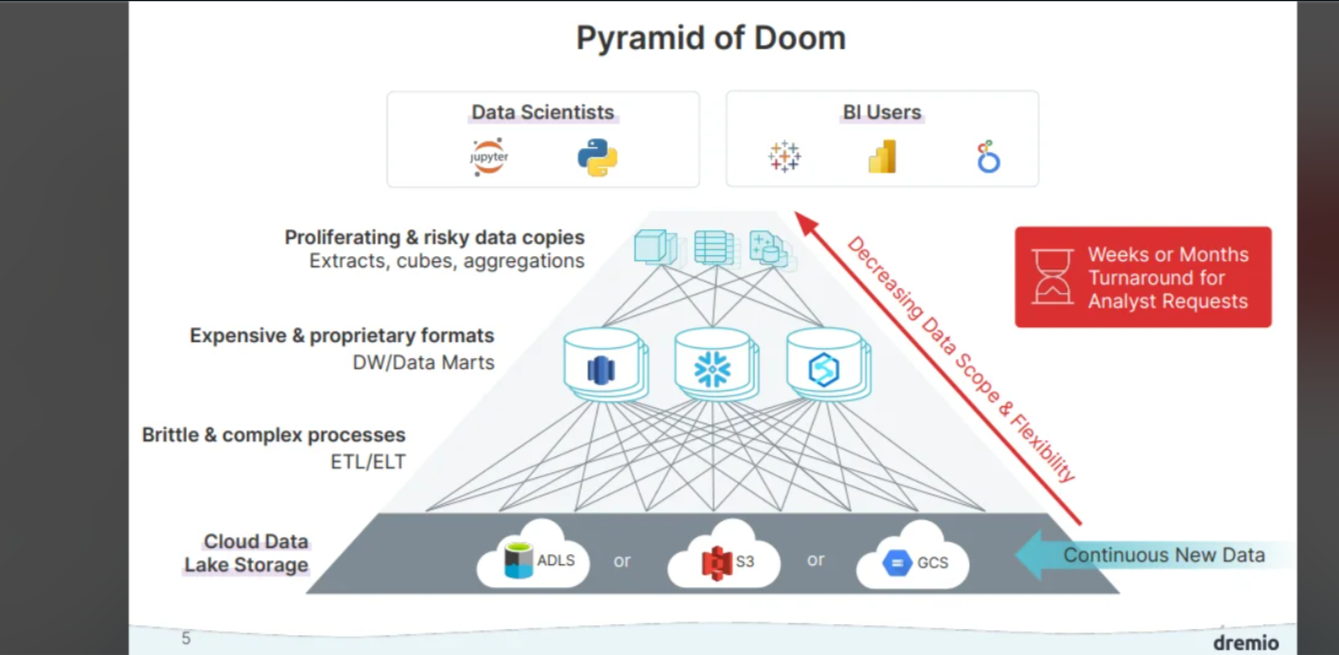
In many data systems, there are two primary hubs for data:

* **Data Lake:** A storage system like Hadoop or Object Storage (ADLS/S3) that stores structured and unstructured data.
* **Data Warehouses:** These systems store structured data optimized for analytical workloads, unlike databases designed for transactional taska

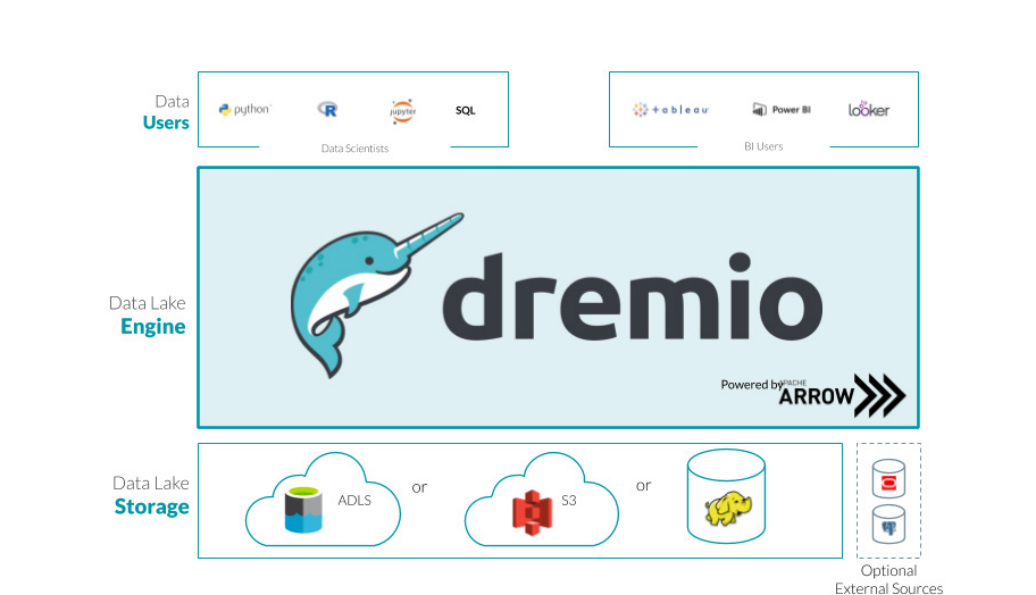
 figure (transform Hadoop into an open architecture for self- service analytics(DREMIO)

### Pourquoi Dremio

Le marché des lacs de données connaît une croissance rapide. De plus en plus de données sont stockées dans des solutions de stockage en cloud pour lacs de données, telles que AWS S3 et Microsoft ADLS. Par nature, les lacs de données diffèrent des entrepôts de données car ils ne nécessitent pas un schéma rigide pour définir les données avant leur chargement, ce qui facilite la collecte de données brutes provenant de nombreux systèmes pour l'analyse. Cependant, l'adoption des lacs de données entraîne des défis : les moteurs SQL traditionnels comme Presto et Athena sont souvent trop lents pour interroger directement les données des lacs de données pour la plupart des cas d'utilisation analytique, à l'exception de certaines requêtes ad-hoc spécifiques. Par conséquent, la plupart des organisations préfèrent utiliser une combinaison d'entrepôts de données, d'extraits, de cubes et d'ETL pour rendre les données des lacs de données plus utilisables et accessibles pour les requêtes analytiques. Cependant, lorsque les systèmes d'entrepôts de données sont utilisés à grande échelle, le coût de possession devient extrêmement élevé. De plus, il existe un certain verrouillage fournisseur en raison du caractère propriétaire des solutions d'entreposage de données.



Cette superposition des data warehouses et des technologies et processus associés crée une infrastructure complexe et coûteuse pour l'analyse des données, comme le montre la Figure 1 ci-dessus. C'est regrettable, car le stockage en cloud pour les data lakes est peu coûteux, facile à provisionner et à faire évoluer, et possède généralement déjà un écosystème d'outils et de solutions associés. Dans de nombreuses organisations, c'est également le premier endroit où les données atterrissent après leur génération par les systèmes opérationnels. Si le stockage en data lake pouvait être interrogé directement de manière performante, de nombreuses solutions supplémentaires mises en place pour contourner ces problèmes de performance pourraient être simplifiées ou complètement éliminées.

 figure2 : dremio data lake engine

Comme le montre la Figure 2 ci-dessus, le moteur de data lake de Dremio offre cette performance, ainsi qu'un couche sémantique en libre-service qui facilite l'accès et l'analyse des données même pour les utilisateurs non techniques. Cela se fait tout en maintenant la flexibilité inhérente au stockage en data lake ; vous n'avez pas besoin de copier et déplacer vos données vers un tiers ou de les mettre dans un format propriétaire. Dremio offre

- Des requêtes ultrarapides directement sur le stockage en data lake\*\*. Les technologies de Dremio, telles que les Data Reflections, le Columnar Cloud Cache (C3) et le Predictive Pipelining, fonctionnent avec Apache Arrow alimenté par Gandiva pour exécuter des requêtes directement sur le stockage en data lake à une vitesse interactive. Ces technologies d'accélération des requêtes combinées offrent une performance de 4 à 100 fois plus rapide par rapport aux moteurs de data lake traditionnels comme Presto.

- Une couche sémantique en libre-service. Cette couche d'abstraction permet aux ingénieurs de données d'appliquer des politiques de sécurité et des significations métier tout en permettant aux analystes et aux scientifiques des données d'explorer les données et de dériver de nouveaux ensembles de données virtuels.

- Flexibilité et ouverture. Dremio vous permet d'éviter le verrouillage fournisseur, de requêter les données directement à travers les clouds ou en local, et de conserver vos données dans un stockage que vous possédez et contrôlez. De cette manière, vous maximisez votre flexibilité et votre liberté d'utiliser vos données comme vous le souhaitez.

- Des économies importantes sur les coûts d'infrastructure. Dremio combine l'accélération des requêtes et des ressources de calcul hautement élastiques pour offrir des avantages significatifs en termes de performance et de coûts. Par exemple, une augmentation moyenne de la vitesse de 4x entraîne une réduction de 75 % ou plus des coûts d'infrastructure par rapport à Presto au même niveau de performance.

Technologies Clés de Dremio

Dremio, dans son essence, utilise un stockage et une exécution en colonnes haute performance, alimentés par Apache Arrow (en mémoire colonne) avec Gandiva (noyau d'exécution basé sur LLVM), Apache Arrow Flight (protocole distribué haute vitesse) et Apache Parquet (colonnes sur disque). Dremio possède une connaissance approfondie et une expérience en analytique haute performance, et est le co-créateur et le mainteneur actuel des projets Apache Arrow, Gandiva et Arrow Flight.

APACHE ARROW, ARROW FLIGHT ET GANDIVA

Apache Arrow est un projet open source qui permet le traitement et l'échange de données en mémoire colonne. Il utilise également le noyau d'exécution Gandiva pour compiler les requêtes en code vectorisé optimisé pour les CPU modernes. Dremio faisait partie de l'équipe fondatrice derrière Arrow, qui comprend désormais des contributeurs de diverses organisations telles qu'IBM, Cloudera, Databricks, Hortonworks, Intel, MapR et Two Sigma.

Dremio est le premier moteur de data lake construit de zéro sur Apache Arrow. En interne, les données en mémoire sont conservées hors tas dans le format Arrow, et Arrow Flight a introduit une API RPC qui retourne les résultats des requêtes sous forme de buffers mémoire Arrow.

Divers autres projets ont également adopté Arrow. Python (Pandas) et R en font partie, permettant aux scientifiques des données de travailler plus efficacement avec les données. Par exemple, Wes McKinney, créateur de la bibliothèque Pandas populaire, a récemment démontré comment Arrow permet aux utilisateurs de Python de lire des données dans Pandas à plus de 10 Go/s.

Dremio comprend un lecteur Parquet ultra-performant et vectorisé qui lit les données au format Parquet depuis le disque en données au format Arrow en mémoire. Le lecteur Parquet permet un traitement rapide des données brutes ainsi que des Data Reflections. Il comprend des capacités de pointe telles que :

- Optimisation des prédicats intelligents et taille des pages.

- Opérations en place sans décompression des données.

- Traitement vectorisé.

- Aucune copie mémoire.

Une Instance Dremio

Dremio dispose d'une architecture élastique évolutive. Elle est conçue pour évoluer de un à plusieurs milliers de nœuds au sein d'une seule instance. Les modèles de déploiement courants incluent :

Environnement Kubernetes hébergé :

- Amazon Elastic Container Service for Kubernetes (Amazon EKS)

- Azure Kubernetes Service (AKS)

Environnement hébergé via des modèles de provisionnement :

- Modèle AWS Cloud Formation

- Modèle Azure Resource Management

Environnement multi-tenant partagé :

- Hadoop utilisant YARN

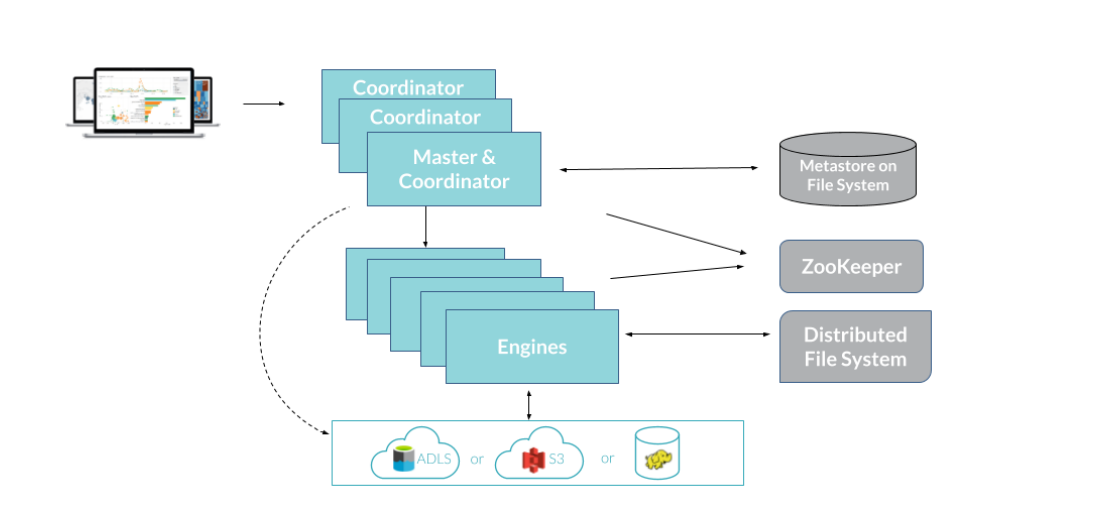
- MapR utilisant YARn

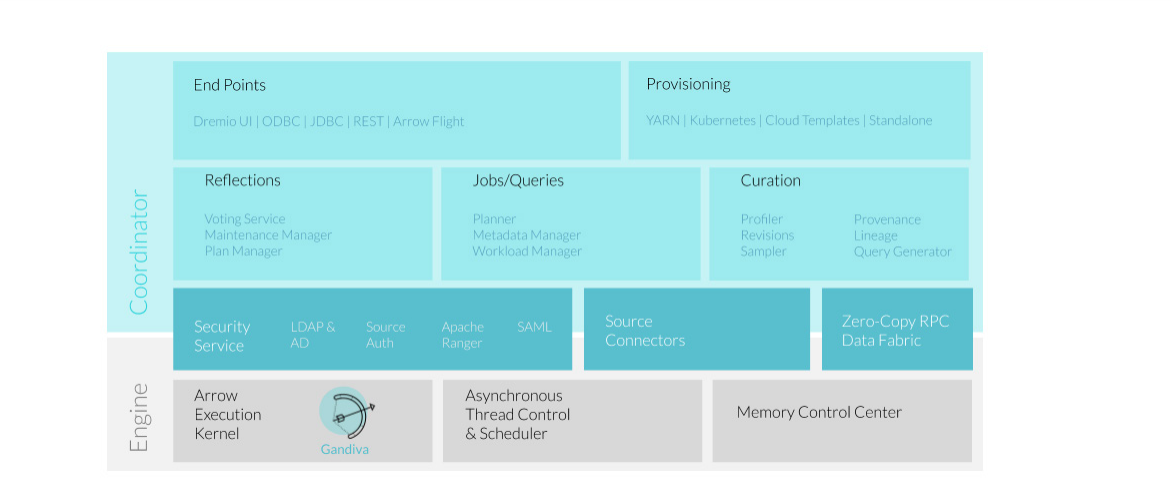
TYPES DE NŒUDS D'INSTANCE

Il existe deux types de nœuds distincts dans une instance Dremio, et la propriété des services détermine si le nœud est activé avec le rôle de master-coordinator ou d'engine. Chaque type de nœud peut être évolué indépendamment ; voir la Figure 3 ci-dessous pour l'architecture de déploiement de Dremio et la Figure 4 pour l'architecture fonctionnelle de Dremio. Les deux types de nœuds sont :

- Coordinators. Ces nœuds sont responsables de la planification des requêtes, de la gestion des métadonnées, de la gestion de l'interface utilisateur de Dremio et de la gestion des connexions des clients. Les applications clientes, telles que les outils BI ou de science des données, se connectent et communiquent avec les coordinators. Les coordinators sont hautement disponibles et peuvent être augmentés pour traiter un plus grand nombre de clients simultanés. Notez que tous les nœuds d'instance Dremio avec des services de coordinator doivent avoir le rôle de master-coordinator activé. Un nœud avec uniquement le rôle de coordinator activé n'est pas pris en charge.

- Engines. Ces nœuds sont responsables de l'exécution des requêtes. Les applications clientes ne se connectent pas aux engines. Les engines peuvent être augmentés pour traiter de grands volumes de données et un plus grand nombre de requêtezsimultanées. Étant donné que les engines sont sans état, les déploiements peuvent traiter ces nœuds comme des ressources élastiques et faire évoluer le système de manière dynamique.

 figure4: Dremio deployment architecture

 figure 4: Dremio functional architecture

##### Data Reflections

Une Data Reflection est représentée par un plan logique et la matérialisation physique correspondante. Le plan logique peut être basé sur un ou plusieurs ensembles de données physiques (par exemple, un compartiment S3, un fichier journal, un répertoire journal, une table Hive, une table Oracle). Du point de vue de la gestion, une Data Reflection est ancrée à un ensemble de données physique ou virtuel unique, bien qu'au moment de la requête, elle puisse également accélérer les requêtes sur d'autres ensembles de données.

La matérialisation physique d'une Data Reflection est basée sur Apache Parquet, avec une variété d'optimisations environnantes telles que les statistiques au niveau des colonnes. Les Data Reflections peuvent être triées, partitionnées, agrégées et distribuées par des colonnes spécifiques.

Pour comprendre comment fonctionnent les Data Reflections, nous devons répondre aux questions indépendantes suivantes :

1. Comment les Data Reflections sont-elles gérées ?Comment sont-elles créées et actualisées ?

2. Comment les Data Reflections sont-elles utilisées pour accélérer les requêtes ?

3. Comment les Data Reflections sont-elles stockées et optimisées physiquement ?

Gestion des Data Reflections

Du point de vue de la gestion, les Data Reflections sont toujours ancrées à un ensemble de données physique ou virtuel spécifique dans le système. Bien que les requêtes SQL sur l'ensemble de données X soient souvent accélérées via des Data Reflections ancrées à l'ensemble de données X, l'optimiseur est libre d'utiliser toute autre Data Reflection pour accélérer ces requêtes. De plus, les Data Reflections ancrées à l'ensemble de données X peuvent être utilisées par l'optimiseur pour accélérer les requêtes sur l'ensemble de données Y.

Il existe deux types de Data Reflections :

- Raw Reflections : une projection d'une ou plusieurs colonnes de l'ensemble de données d'ancrage. Les données peuvent être triées, partitionnées et distribuées par différentes colonnes de l'ensemble de données d'ancrage.

- Aggregation Reflections : une agrégation sur une ou plusieurs colonnes de l'ensemble de données d'ancrage. La Data Reflection est définie par des dimensions et des mesures et contient des données au niveau d'agrégation pour chaque mesure telle que le nombre, la somme, le minimum et le maximum. Les données peuvent être triées, partitionnées et distribuées par différentes colonnes de l'ensemble de données d'ancrage.

Il peut y avoir des circonstances uniques où il est souhaitable de créer une Data Reflection personnalisée, similaire à une vue matérialisée dans une base de données relationnelle. Dans ces cas, il suffit de créer un nouvel ensemble de données virtuel avec la requête SQL qui définit la matérialisation souhaitée, puis de créer une seule Raw Reflection qui inclut toutes les colonnes dans l'ensemble de données virtuel.

Actualisation des Data Reflections

Dremio actualise automatiquement les Data Reflections pour garantir la fraîcheur des données. Il existe trois méthodes d'actualisation :

- Actualisation complète : Cette méthode est la plus adaptée pour les ensembles de données en mutation.

- Actualisation incrémentielle pour les sources basées sur des fichiers :\*\* Cette méthode est la plus adaptée pour les grands ensembles de données en append-only représentés comme une collection de fichiers (par exemple, S3). Le système identifie automatiquement les nouveaux fichiers dans le répertoire.

- Actualisation incrémentielle pour les sources basées sur des tables :\*\* Cette méthode est la plus adaptée pour les grands ensembles de données en append-only représentés sous forme de tables ou de collections (par exemple, Oracle, Elasticsearch, MongoDB). Une colonne croissante de manière monotone, telle qu'un horodatage `created\_at` ou une clé primaire séquentielle, est requise.

En interne, Dremio maintient un graphe de dépendances (DAG) qui définit l'ordre dans lequel les Data Reflections sont actualisées. Les dépendances sont calculées en fonction de l'algèbre relationnelle, et le temps de début réel de l'actualisation prend en compte le temps attendu pour compléter tout le cycle d'actualisation.

Notez que cette approche basée sur un graphe réduit le temps de cycle de bout en bout, ainsi que les ressources informatiques nécessaires pour compléter le cycle. De plus, en utilisant une Data Reflection pour actualiser une autre Data Reflection, le système peut éviter les lectures gourmandes en ressources sur les bases de données opérationnelles plus d'une fois.

Data Reflections Externe

Pour des cas uniques, Dremio prend en charge la notion de Data Reflections externes. Les utilisateurs peuvent créer et maintenir des Data Reflections en utilisant un processus externe tel qu'Apache Spark, puis enregistrer la Data Reflection externe dans Dremio. Dremio considère ces Data Reflections dans son analyse basée sur les coûts pour accélérer les requêtes. Toute source de données supportée par Dremio peut être utilisée pour les Data Reflections externes.

Les Data Reflections externes sont utiles dans les cas où il est judicieux de gérer la création en dehors du processus Dremio (par exemple, un traitement qui dure plusieurs heures ou jours, ou lorsque des processus existants sont déjà en place pour créer des représentations optimisées des données pour des modèles de requêtes spécifiques). L'avantage d'enregistrer ces ressources dans Dremio est qu'il simplifie l'expérience pour les consommateurs de données, et fournit des capacités supplémentaires pour sécuriser et gouverner l'accès aux données, ainsi que pour suivre la lignée des données.

Utilisation des Data Reflections pour Accélérer les Requêtes

Dremio inclut un optimiseur basé sur les coûts qui non seulement planifie et optimise les requêtes, mais explore également les opportunités d'utiliser les Data Reflections pour réduire le coût des requêtes (c'est-à-dire, accélérer la requête). Lorsqu'une nouvelle requête arrive, l'optimiseur prend en compte toutes les Data Reflections et réécrit automatiquement le plan de la requête pour utiliser les Data Reflections lorsque cela est possible (voir Figure 10). L'optimiseur utilise un algorithme en deux phases :

1. Élagage des Data Reflections : L'optimiseur écarte les Data Reflections qui sont non pertinentes car leurs plans logiques n'ont aucun ensemble de données physique en commun avec le plan logique de la requête.

2. Correspondance des sous-graphes : L'optimiseur utilise un algorithme hiérarchique innovant pour faire correspondre les sous-graphes du plan logique de la requête avec les plans logiques des Data Reflections.

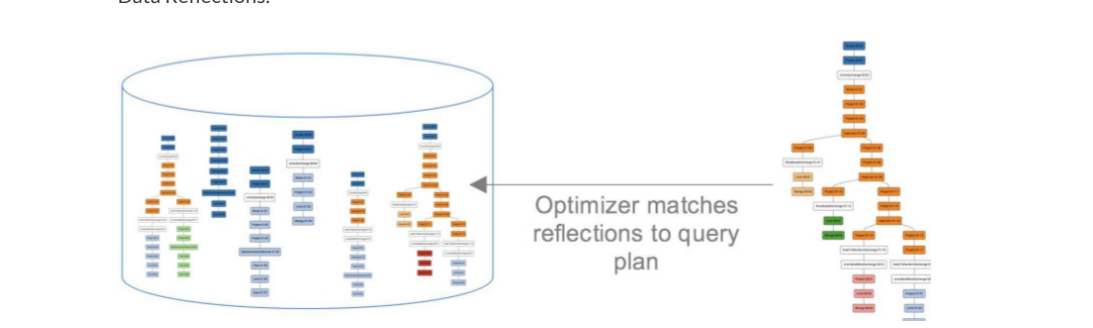


Figure 10 - Subgraph matching

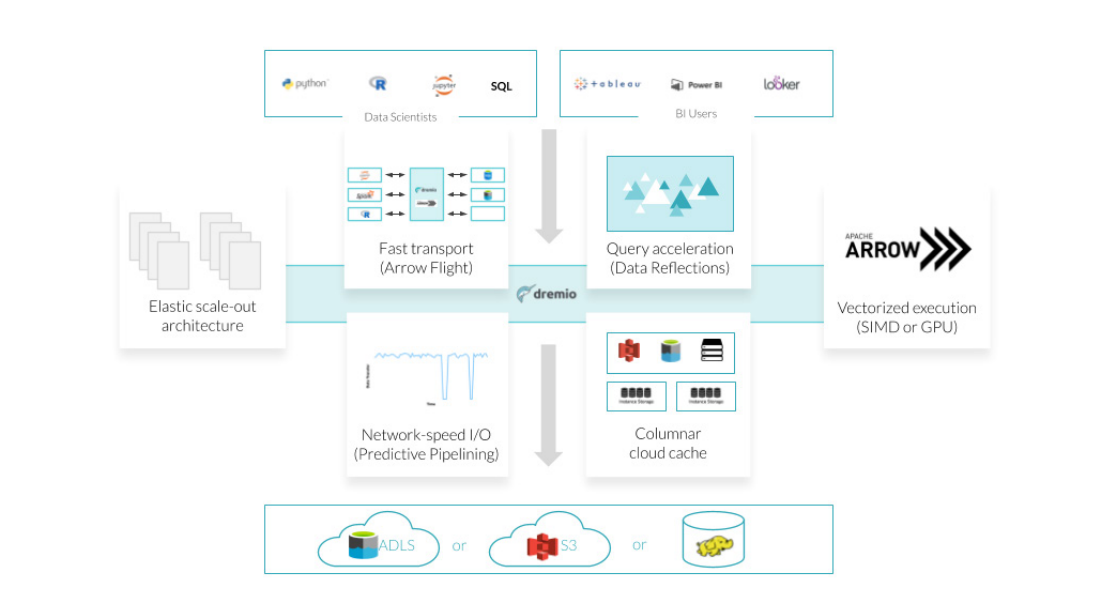
Accélération des Requêtes

Il existe deux types généraux de requêtes analytiques : ad-hoc et tableau de bord/rapports. Le moteur de data lake de Dremio dispose de plusieurs technologies d'accélération, lui permettant d'exécuter ces requêtes de manière extrêmement rapide et efficace. Les requêtes ad-hoc sont dynamiques et interactives par nature, nécessitant un accès direct et à la demande aux données sources. Le moteur basé sur Arrow de Dremio, alimenté par Gandiva, C3 et le pipeline prédictif, fournit des résultats de requêtes ad-hoc jusqu'à 4 fois plus rapidement que les moteurs SQL traditionnels (voir Figure 5).

Les requêtes de tableau de bord/rapports sont bien définies et sont généralement exécutées contre des ensembles de données curatés. Pour ces requêtes, Dremio offre des technologies d'accélération supplémentaires telles que les Data Reflections, permettant de rendre ces requêtes jusqu'à 100 fois plus rapides par rapport aux moteurs SQL traditionnels.

THE LIFE OF A QUERY

Client applications can issue queries to Dremio over ODBC, JDBC or REST. A query might involve one or more datasets, mostly residing in data lake storage but potentially in optional, relatively small external data sources. Dremio utilizes the following primary techniques to reduce the amount of processing required for a query: Direct data lake storage queries. The optimizer uses Apache Arrow to read data directly from data lake storage into the Arrow buffer with massively parallel readers and predictive pipelining at high speed and extreme concurrency. Also, the optimizer interrogates C3 to fetch automatically cached data, achieving NVMe-level performance. Push-downs into optional external data sources. The optimizer considesr the capabilities of the underlying external data source and the relative costs. It then generates a plan that performs stages of the query, either in the external source or in Dremio’s distributed execution environment, to achieve the most efficient overall plan possible. Optional acceleration via Data Reflections. The optimizer uses Data Reflections, a highly optimized physical representation of source data, if available for all or portions of the query when this produces the most efficient overall plan. In many cases, the entire query can be serviced from Data Reflections

 figure5: Dremio query acceleration

Poussée des Requêtes

Bien que la plupart des sources de données résident dans le stockage en data lake, lorsque nécessaire, Dremio peut pousser le traitement dans des sources de données relationnelles et non relationnelles externes. Les sources de données non relationnelles ne supportent généralement pas SQL et ont des capacités d'exécution limitées. Par exemple, un système de fichiers ne peut pas appliquer de prédicats ou d'agrégations. L'optimiseur de Dremio comprend les capacités de chaque source de données. Lorsque cela est le plus efficace, Dremio pousse autant de la requête que possible vers la source sous-jacente et effectue le reste dans son propre moteur d'exécution distribué.

Data Reflections

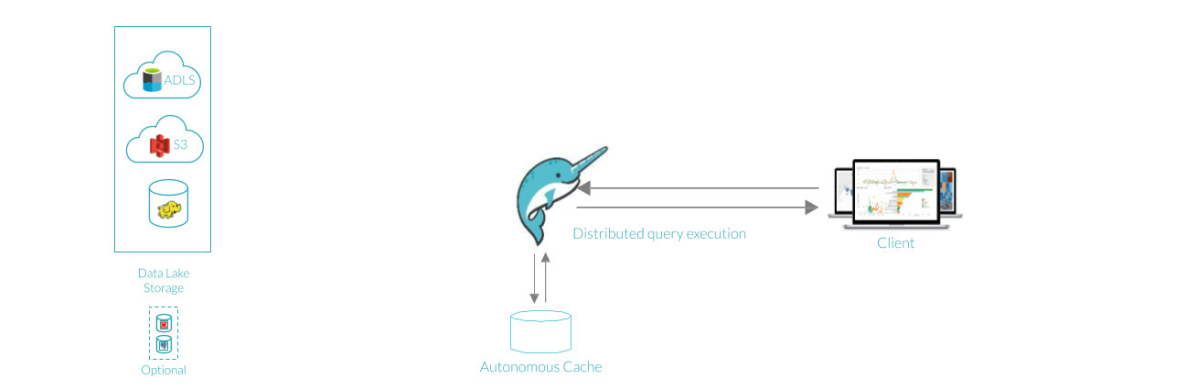
Dremio accélère les requêtes en utilisant des représentations physiques hautement optimisées des données sources appelées Data Reflections. Le Data Reflection Store peut résider dans un stockage en cloud tel qu'Amazon S3, Azure ADLS, HDFS ou dans un stockage attaché directement (DAS). La taille du Data Reflection Store peut dépasser celle de la mémoire physique. Cette architecture permet à Dremio d'accélérer davantage de données à un coût plus bas, entraînant un taux de réussite de cache beaucoup plus élevé par rapport aux architectures traditionnelles basées uniquement sur la mémoire. Les Data Reflections sont automatiquement utilisées par l'optimiseur basé sur les coûts de Dremio au moment de la requête.

Notez que les Data Reflections sont invisibles pour les utilisateurs finaux. Contrairement aux cubes OLAP, aux tables d'agrégation et aux extraits BI, l'utilisateur ne se connecte pas explicitement à une Data Reflection. Au lieu de cela, les utilisateurs se connectent à des ensembles de données virtuels (c'est-à-dire des vues de données) dans la couche sémantique de Dremio, et émettent des requêtes contre le modèle logique de Dremio. L'optimiseur de requêtes de Dremio accélère automatiquement la requête en tirant parti des Data Reflections qui sont appropriées pour la requête en fonction du graphe de dépendance et de l'analyse de coûts de l'optimiseur.

Par défaut, Dremio utilise son moteur d'exécution distribué haute performance, tirant parti du traitement en mémoire colonne (via Apache Arrow) et des poussées avancées dans les sources de données sous-jacentes lorsque cela est disponible (voir Figure 7)

Figure 7 - Default query execution path

Dans certains cas où l'ensemble de la requête peut être satisfait par les Data Reflections sans consulter la source de données, le plan est réécrit par l'optimiseur pour utiliser les Data Reflections (voir Figure 8).

 Figure 8 - Query execution with Data Reflection acceleration

Bien sûr, il existe des situations où le plan optimal inclut une combinaison de données provenant de la source de données et d'une ou plusieurs Data Reflections (voir Figure 9).

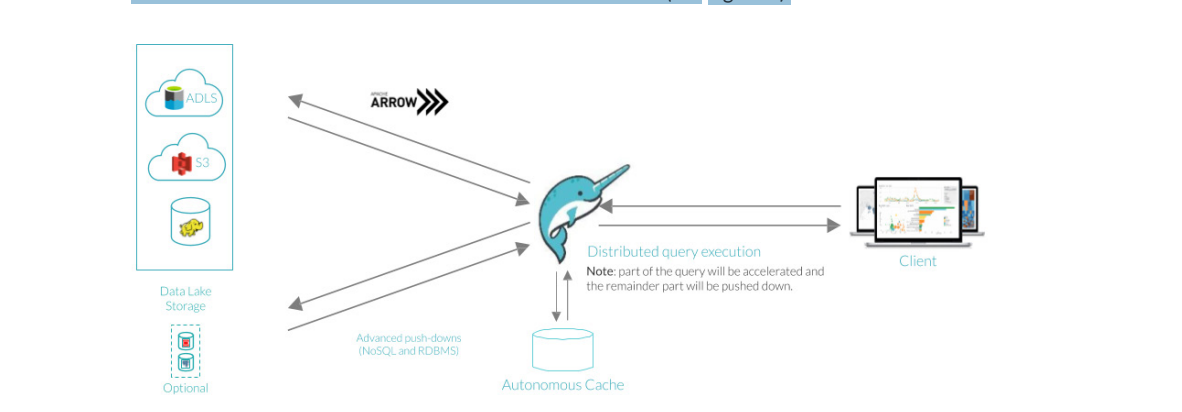
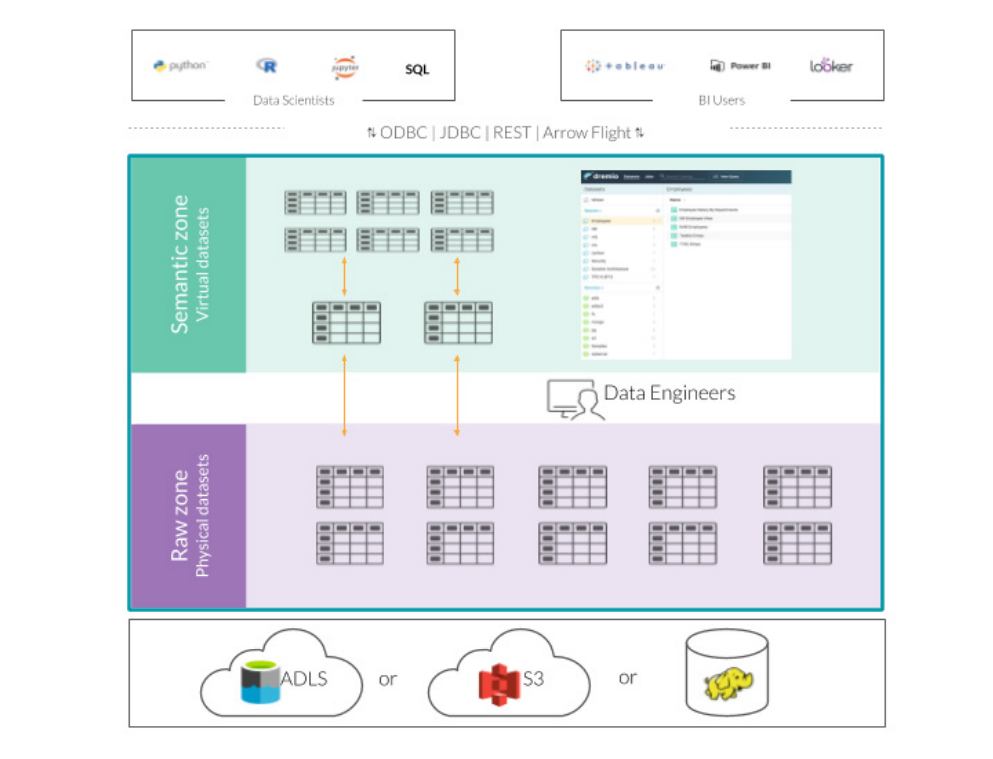


Figure 9 - Query execution with a partial Data Reflection acceleration

##### Couche Sémantique en Libre-Service

Dremio établit des vues sur les données (appelées ensembles de données virtuels) dans une couche sémantique au-dessus de vos données physiques (voir Figure 11), permettant ainsi aux analystes et ingénieurs de données de gérer, curer et partager les données tout en maintenant la gouvernance et la sécurité—mais sans la surcharge et la complexité liées à la copie des données. Connectez n'importe quel outil de BI ou de science des données, y compris Tableau, Power BI, Looker et Jupyter Notebooks, à Dremio et commencez à explorer et exploiter votre data lake pour en extraire de la valeur.

La couche sémantique de Dremio est entièrement virtuelle, indexée et interrogeable, et les relations entre vos sources de données, ensembles de données virtuels et transformations, ainsi que toutes vos requêtes, sont maintenues dans le graphe de données de Dremio, vous permettant de savoir exactement d'où provient chaque ensemble de données virtuel. Le contrôle d'accès basé sur les rôles garantit que chacun a accès exactement à ce dont il a besoin (et rien de plus), et l'authentification unique (SSO) permet une expérience d'authentification fluide. Figure 11 - Dremio semantic layer architecture

### **SÉCURITÉ ET GOUVERNANCE DES DONNÉES**

La sécurité et la gouvernance des données sont cruciales pour toute entreprise. Cependant, la complexité croissante et la demande de données entraînent souvent une dispersion des données, ce qui augmente les risques. Dremio, grâce à sa couche sémantique en libre-service, permet aux analystes d'affaires et aux scientifiques des données de découvrir, curer, accélérer et partager des données sans avoir à exporter des copies dans des systèmes non régulés, tels que des feuilles de calcul déconnectées, des serveurs BI non gouvernés et des bases de données privées. Cela réduit le risque d'accès non autorisé aux données ainsi que le vol de données.

Dremio offre une variété de capacités en matière de sécurité et de gouvernance, notamment :

* **Contrôle d'accès par lignes et colonnes sur toute source**
* **Masquage des données**
* **Provenance et traçabilité des données**
* **Authentification basée sur LDAP/AD**
* **Contrôle d'accès basé sur les rôles**
* **Audit**
* **Chiffrement**

#### PROVENANCE ET TRAÇABILITÉ DES DONNÉES

Le Graph de Données de Dremio montre la provenance et la traçabilité de chaque ensemble de données dans le système, y compris ceux dans l'espace personnel d'un utilisateur. Le service informatique peut facilement comprendre comment un ensemble de données a été créé, transformé, joint et partagé, ainsi que la pleine traçabilité de ces étapes entre les ensembles de données. Ces informations sont automatiquement capturées et suivies par Dremio.

#### **AUTHENTIFICATION**

Dremio prend en charge les modes d'authentification suivants :

* **Local** : Dremio gère les utilisateurs en interne.
* **LDAP** : Dremio se connecte à un service d'annuaire LDAP existant tel qu'Active Directory. Dremio se fie au service d'annuaire pour vérifier les identifiants et vérifier l'appartenance aux groupes.
* **Azure AD** : Dremio authentifie les utilisateurs via Azure Active Directory Single Sign On (SSO).
* **OpenID** : Dremio authentifie les utilisateurs via des fournisseurs d'identité tiers utilisant le protocole OpenID.

CONTRÔLE D'ACCÈS

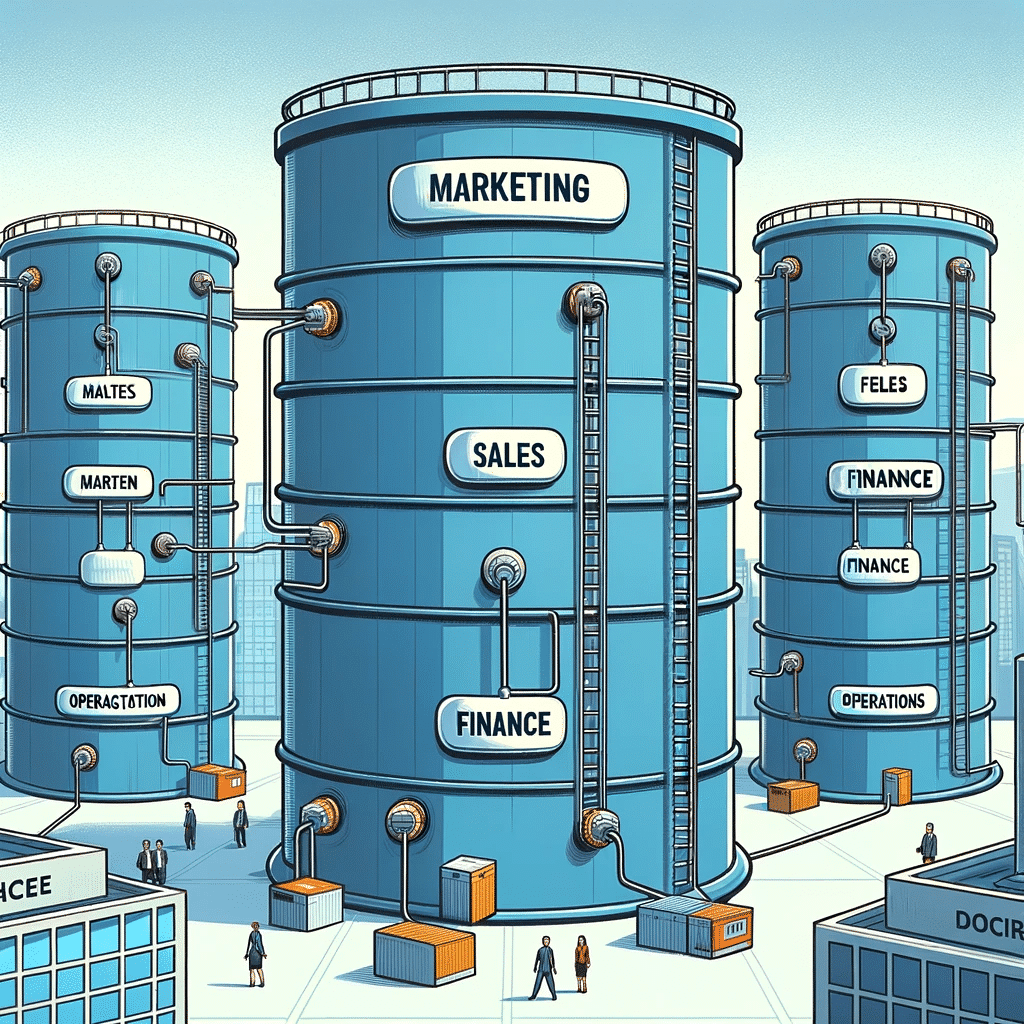
Dremio prend en charge un contrôle d'accès granulaire :

* **Autorisations sur les ensembles de données physiques** : Contrôle quels utilisateurs et/ou groupes peuvent interroger un ensemble de données physique spécifique (par exemple, répertoire HDFS, table Hive, type Elasticsearch).
* **Autorisations sur les ensembles de données virtuels** : Contrôle quels utilisateurs et/ou groupes peuvent interroger un ensemble de données virtuel spécifique.
* **Autorisations au niveau des colonnes** : Peuvent être utilisées pour restreindre l'accès aux colonnes sensibles dans un ensemble de données pour des utilisateurs spécifiques.
* **Autorisations au niveau des lignes** : Peuvent être utilisées pour restreindre l'accès à un sous-ensemble des enregistrements dans un ensemble de données pour des utilisateurs spécifiques.
* **Masquage des données sensibles** : Appliqué dynamiquement au moment de la requête en fonction de l'appartenance aux groupes LDAP/AD et d'autres règles définies par l'utilisateur.

Les ensembles de données virtuels sont le mécanisme recommandé pour gérer l'accès aux données. Un utilisateur possédant un ensemble de données spécifique peut créer un ensemble de données virtuel dérivé, incluant uniquement un sous-ensemble des colonnes ou des enregistrements de l'ensemble de données original. L'utilisateur peut ensuite refuser l'accès à l'ensemble de données original tout en permettant l'accès à l'ensemble de données virtuel dérivé

La gestion efficace et l'utilisation des données sont cruciales pour le succès de toute entreprise. Cependant, un obstacle majeur auquel de nombreuses organisations sont confrontées dans leur quête de devenir orientées vers les données est la prévalence des silos de données. Les silos de données se produisent lorsque l'information est isolée dans des départements ou des systèmes distincts au sein d'une organisation, la rendant inaccessible ou invisible aux autres parties de l'entreprise. Cette compartimentation des données entrave la collaboration et le partage des connaissances, conduisant à des inefficacités et à un manque de vision cohérente.C'est là qu'intervient Dremio, une solution conçue pour éliminer ces silos et unifier des sources de données disparates pour des analyses transparentes. Elle permet à l'entreprise de combiner diverses sources de données, des bases de données traditionnelles aux systèmes de stockage cloud modernes, dans un environnement analytique unique, cohérent et accessible. Ce blog explore comment Dremio réalise cet exploit et pourquoi il révolutionne l'approche des es plu efficacement entreprises souhaitant exploiter leurs données plus efficacement.



Data silos sont essentiellement des poches isolées de données stockées dans des systèmes ou départements distincts au sein d'une organisation. Ces silos peuvent résulter de structures organisationnelles historiques, de systèmes technologiques disparates ou même de barrières culturelles au sein de l'entreprise. Les caractéristiques clés des silos de données comprennent un accès limité, où seuls des groupes ou départements spécifiques peuvent accéder aux données, et un manque d'intégration, où les données ne sont pas reliées à d'autres données pertinentes au sein de l'organisation

Les Causes Fondamentales

Plusieurs facteurs contribuent à la formation des silos de données :

1. Structure organisationnelle : Les entreprises aux divisions départementales rigides finissent souvent par créer des silos de données, chaque département générant et stockant ses propres données indépendamment.

2. Technologies diverses : L'utilisation de différentes technologies et plateformes au sein des départements peut entraîner des problèmes de compatibilité, rendant l'intégration des données difficile.

3. Barrières culturelles: Parfois, le problème est moins technologique et plus lié à l'état d'esprit, où les départements hésitent à partager l'information en raison de la concurrence ou du manque de confiance.

L'Impact des Silos de Données

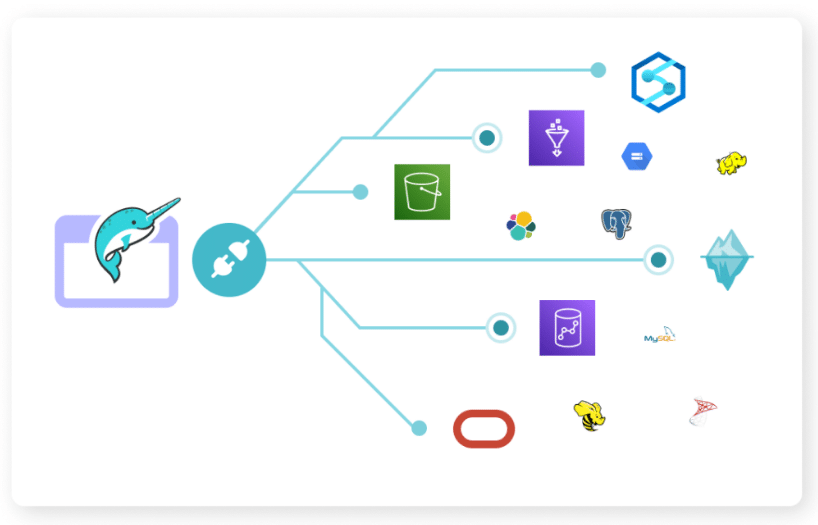
Les silos de données peuvent avoir un impact néfaste sur la capacité d'une organisation à prendre des décisions éclairées. Les principaux problèmes incluent :

1. Inefficacité et duplication : Lorsque les données sont cloisonnées, les mêmes données peuvent être collectées et stockées plusieurs fois dans l'organisation, entraînant des inefficacités et des coûts accrus.

2. Prise de décision altérée : Les silos empêchent une vue d'ensemble des données de l'organisation, conduisant à des décisions basées sur des informations incomplètes.

3. Réduction de l'agilité : Dans l'environnement commercial actuel rapide, l'incapacité à accéder et analyser rapidement les données peut entraver la capacité d'une organisation à répondre aux changements du marché.

Dremio est conçu pour alléger les défis posés par les silos de données. C'est une solution d'unification des données, permettant un accès et une manipulation des données à grande vitesse à travers diverses sources de données, tout en transformant votre data lake en un véritable lakehouse Apache Iceberg. Dremio se distingue par sa capacité à intégrer des référentiels de données disparates sans nécessiter de déplacement ou de duplication des données, favorisant ainsi une approche plus efficace et agile de l'analyse des données.

Principales Caractéristiques de Dremio

1. Virtualisation des données : Dremio permet la virtualisation des données, ce qui signifie qu'il peut se connecter à diverses sources de données et fournir une vue unifiée sans déplacer physiquement les données.

2. Accès aux données en libre-service : Il démocratise l'accès aux données, permettant aux utilisateurs finaux et aux analystes de récupérer et d'analyser les données de manière indépendante, réduisant ainsi la dépendance aux équipes informatiques. Dremio dispose également d'une couche sémantique robuste.

3. Accélération avancée des requêtes : Dremio intègre des techniques sophistiquées d'accélération des requêtes, réduisant considérablement le temps nécessaire pour traiter des requêtes complexes.

Fonctionnalités Avancées de Dremio pour une Analyse Transparente

Dremio est doté de plusieurs fonctionnalités avancées :

1. Accélération des requêtes avec les reflections de données : Les reflections de données créent des représentations optimisées des données, accélérant ainsi les performances des requêtes.

2. Fonctionnalités de sécurité robustes : Dremio garantit la sécurité des données grâce à des fonctionnalités telles que le chiffrement, les contrôles d'accès et la conformité aux normes industrielles.

3. Scalabilité et flexibilité : Conçu pour évoluer avec les besoins croissants en données d'une organisation, Dremio prend en charge les déploiements sur site et dans le cloud.

#### Virtualisation des Données avec Dremio(use for attijariwafa)

Dremio offre une solution de virtualisation des données qui permet à Attijariwafa Bank d'accéder et d'analyser ses données de manière unifiée et efficace, sans nécessiter de déplacement physique des données. Voici comment Dremio va transformer la gestion des données à l'avenir :

1. Accès Unifié aux Données : Dremio créera une couche d'accès unique qui connecte diverses sources de données, permettant aux utilisateurs de découvrir et d'explorer les données de manière transparente, sans avoir à les déplacer ou les dupliquer.

2. Optimisation des Performances : Grâce à des techniques avancées de virtualisation, Dremio facilitera une gestion efficace des données en offrant une vue cohérente et intégrée, tout en améliorant les performances des requêtes par l'optimisation des représentations de données.

3. Flexibilité et Scalabilité : La solution permettra à Attijariwafa Bank de s'adapter facilement à l'évolution des besoins en données. Dremio offrira une flexibilité maximale en supportant à la fois les déploiements sur site et dans le cloud, ce qui permettra d'ajuster l'infrastructure en fonction de la croissance des volumes de données.

4. Réduction des Coûts et Complexité : En éliminant la nécessité de déplacer ou de dupliquer les données, Dremio réduira les coûts associés à la gestion des données et simplifiera l'infrastructure IT. Cela se traduira par une meilleure efficacité opérationnelle et une réduction des dépenses liées au stockage et à la gestion des données.

5. Amélioration de la Sécurité et de la Conformité : Avec des fonctionnalités robustes de sécurité et de gouvernance des données, Dremio assurera que les données sont protégées tout en respectant les normes et réglementations en vigueur, garantissant ainsi la conformité et la sécurité des informations sensibles.

En résumé, la virtualisation des données avec Dremio permettra à Attijariwafa Bank de bénéficier d'une approche moderne et agile pour la gestion et l'analyse des données, tout en optimisant les coûts et en améliorant la sécurité et l'efficacité de l'infrastructure de données.

### Solution Future pour Attijariwafa Bank avec Dremio

Dremio proposera à Attijariwafa Bank une feuille de route pour la transition vers une analyse unifiée grâce à une approche progressive pour moderniser son infrastructure de données basée sur Hadoop. À l'avenir, Dremio exigera peu de modifications des pipelines existants, réduisant ainsi les coûts et les perturbations associées à la migration des données.

En adoptant Dremio et un environnement entièrement conteneurisé, Attijariwafa Bank pourra considérablement réduire ses frais de gestion tout en améliorant la sécurité et la résilience. Dremio, en intégrant l'écosystème ouvert autour d'Apache Iceberg et d'Arrow, garantira une solution évolutive, transparente et extensible. En remplaçant leur infrastructure Hadoop/Hive, Dremio offrira des fonctionnalités adaptées à divers cas d'utilisation secondaires grâce à sa couche sémantique.

Les mécanismes existants d'ETL et d'ingestion de données basés sur Spark resteront en place, mais Dremio introduira une couche d'accès unifiée qui facilitera la découverte et l'exploration des données pour les utilisateurs finaux sans duplication. Cela permettra une réduction significative du facteur de réplication des données et un découplage efficace entre le stockage et le calcul, optimisant ainsi les performances et l'efficacité de l'infrastructure de données d'Attijariwafa Bank.

#### Résultats

Avec Dremio en place, Attijariwafa Bank pourra réaliser des économies significatives en réduisant de manière drastique à la fois la consommation de calcul et l'espace disque nécessaire dans ses environnements de données. La nouvelle infrastructure de données pourra gérer des volumes importants tout en optimisant les coûts. Par exemple, la banque pourrait traiter des milliers de tables contenant plusieurs pétaoctets de données, tout en réduisant l'empreinte physique des données par rapport aux anciennes infrastructures.

La nouvelle infrastructure, soutenue par des nœuds exécutants sur des clusters Kubernetes, permettra de remplacer un grand nombre de mini-clusters et de cœurs nécessaires précédemment. Cette réduction des ressources de calcul, associée à l'amélioration des performances, offrira une accélération significative des temps de réponse des requêtes, même avec une diminution des ressources.

En accédant directement aux données via leur lakehouse avec Dremio, le temps de traitement des requêtes sera considérablement réduit, par exemple de 45 minutes à seulement 2 minutes, permettant ainsi une amélioration de 95 % du temps nécessaire pour obtenir des insights pour la maintenance prédictive et l'optimisation des données. La migration devrait aboutir à une réduction de plus de 60 % des coûts de calcul par rapport à l'infrastructure de données précédente, à des requêtes plus de 20 fois plus rapides, et à des économies de plus de 30 % sur le coût total de possession (TCO).

Résultats

Avec la plateforme Unified Lakehouse de Dremio, Attijariwafa Bank pourra atteindre un temps d'analyse 95 % plus rapide tout en simplifiant la gestion proactive de la relation client. La solution permettra aux consommateurs de données de gérer eux-mêmes les données, de gérer proactivement l'expérience client, et d'optimiser et identifier les problèmes avant qu'ils ne surviennent. Le lakehouse de données de Dremio permettra à l'équipe de la banque d'exploiter les données de télémétrie, de réduire les risques de désabonnement des clients, et d'assurer une meilleure disponibilité des produits tout au long du parcours client.

### Summary of Exercises

This exercise assumes that our operational applications use Postgres as a database. Our goal is to migrate this data to our data lakehouse, specifically into [**Apache Iceberg tables**](https://bit.ly/am-iceberg-101) managed and stored in Minio as our object storage, these tables will be tracked by a [**Nessie catalog**](https://bit.ly/am-nessie-101). We'll utilize Apache Spark as the data movement tool to the data lake and Dremio as the query engine powering our business intelligence (BI) dashboards through [**Apache Superset**](https://www.dremio.com/blog/bi-dashboards-101-with-dremio-and-superset/).

### Environment Setup

You will need a docker desktop installed on your machine to set up our environment. Then in an empty folder, create a docker-compose.yml file and include the following:

version: "3"  
  
services:  
 # Nessie Catalog Server Using In-Memory Store  
 nessie:  
 image: projectnessie/nessie:latest  
 container\_name: nessie  
 networks:  
 de-end-to-end:  
 ports:  
 - 19120:19120  
 # Minio Storage Server  
 minio:  
 image: minio/minio:latest  
 container\_name: minio  
 environment  
 - MINIO\_ROOT\_USER=admin  
 - MINIO\_ROOT\_PASSWORD=password  
 - MINIO\_DOMAIN=storage  
 - MINIO\_REGION\_NAME=us-east-1  
 - MINIO\_REGION=us-east-1  
 networks:  
 de-end-to-end:  
 ports:  
 - 9001:9001  
 - 9000:9000  
 command: ["server", "/data", "--console-address", ":9001"]  
 # Dremio  
 dremio:  
 platform: linux/x86\_64  
 image: dremio/dremio-oss:latest  
 ports:  
 - 9047:9047  
 - 31010:31010  
 - 32010:32010  
 container\_name: dremio  
 networks:  
 de-end-to-end:  
 # Spark  
 spark:  
 platform: linux/x86\_64  
 image: alexmerced/spark35notebook:latest  
 ports:   
 - 8080:8080 # Master Web UI  
 - 7077:7077 # Master Port  
 - 8888:8888 # Notebook  
 environment:  
 - AWS\_REGION=us-east-1  
 - AWS\_ACCESS\_KEY\_ID=admin #minio username  
 - AWS\_SECRET\_ACCESS\_KEY=password #minio password  
  
 container\_name: spark  
 networks:  
 de-end-to-end:  
 # Postgres  
 postgres:  
 image: postgres:latest  
 container\_name: postgres  
 environment:  
 POSTGRES\_DB: mydb  
 POSTGRES\_USER: myuser  
 POSTGRES\_PASSWORD: mypassword  
 ports:  
 - "5435:5432"  
 networks:  
 de-end-to-end:  
 #Superset  
 superset:  
 image: alexmerced/dremio-superset  
 container\_name: superset  
 networks:  
 de-end-to-end:  
 ports:  
 - 8088:8088  
networks:  
 de-end-to-end:

### Breakdown of the docker-compose file

This Docker Compose file defines a set of services that work together to create a data engineering environment. Let's break down each service and its purpose:

1. **Nessie Catalog Server (nessie):**

- Image: projectnessie/nessie:latest

- Purpose: This service sets up a Nessie catalog server using an in-memory store.

- Ports: Exposes port 19120 for external communication.

2. **Minio Storage Server (minio):**

- Image: minio/minio:latest

- Environment Variables:

- MINIO\_ROOT\_USER=admin

- MINIO\_ROOT\_PASSWORD=password

- MINIO\_DOMAIN=storage

- MINIO\_REGION\_NAME=us-east-1

- MINIO\_REGION=us-east-1

- Purpose: Sets up a Minio storage server for object storage.

- Ports: Exposes ports 9001 and 9000 for external access and uses port 9001 for the Minio console.

- Command: Starts the server with the specified parameters.

3. **Dremio (dremio):**

- Platform: linux/x86\_64

- Image: dremio/dremio-oss:latest

- Ports: Exposes ports 9047, 31010, and 32010 for Dremio communication.

- Purpose: Sets up Dremio, a data lakehouse platform, for data processing and analytics.

4. **Spark (spark):**

- Platform: linux/x86\_64

- Image: alexmerced/spark35notebook:latest

- Ports: Exposes ports 8080, 7077, and 8888 for Spark services, including the web UI, master port, and notebook.

- Purpose: Sets up Apache Spark for distributed data processing and analytics.

5. **Postgres (postgres):**

- Image: postgres:latest

- Environment Variables:

- POSTGRES\_DB=mydb

- POSTGRES\_USER=myuser

- POSTGRES\_PASSWORD=mypassword

- Ports: Exposes port 5435 for external access.

- Purpose: Sets up a Postgres database with a specified database name, username, and password.

6. **Superset (superset):**

- Image: alexmerced/dremio-superset

- Ports: Exposes port 8080 for Superset access.

- Purpose: Sets up Apache Superset, a data visualization and exploration platform, for creating BI dashboards.

Additionally, the file defines a network called de-end-to-end that connects all the services, allowing them to communicate with each other within the Docker environment.

This Docker Compose file creates a comprehensive data engineering environment with services for data storage, processing, analytics, and visualization.

### Populating the Postgres Database

The first step is to populate our Postgres database with some data to represent operational data.

### 1. Spin up the Postgres Service:

Open a terminal, navigate to the directory containing the Docker Compose file, and run the following command to start the Postgres service:

docker-compose up postgres

### 2. Access the Postgres Shell:

After the Postgres service is running, you can access the Postgres shell using the following command in another terminal:

docker exec -it postgres psql -U myuser mydb

Enter the password when prompted (use mypassword in this example).

### 3. Create a Table and Add Data:

Once in the Postgres shell, you can create a table and add data. Here's an example SQL script:

-- Create a table for a mock BI dashboard dataset  
CREATE TABLE sales\_data (  
 id SERIAL PRIMARY KEY,  
 product\_name VARCHAR(255),  
 category VARCHAR(50),  
 sales\_amount DECIMAL(10, 2),  
 sales\_date DATE  
);  
  
-- Insert sample data into the table  
INSERT INTO sales\_data (product\_name, category, sales\_amount, sales\_date)  
VALUES  
 ('Product A', 'Electronics', 1000.50, '2024-03-01'),  
 ('Product B', 'Clothing', 750.25, '2024-03-02'),  
 ('Product C', 'Home Goods', 1200.75, '2024-03-03'),  
 ('Product D', 'Electronics', 900.00, '2024-03-04'),  
 ('Product E', 'Clothing', 600.50, '2024-03-05');

Run the above SQL script in the Postgres shell to create the sales\_data table and populate it with sample data that is ideal for a mock BI dashboard. Leave the Postgres shell with the command:

\q

### Moving the Data to the Data Lake with Spark

Next, we must move the data to our data lake. Let's spin up the following services.

* minio: This will be our storage layer, an object storage service for holding all our files.
* nessie: This will be our Apache iceberg catalog, tracking our different tables and the location of their latest metadata file in our storage.
* spark: This will have Apache Spark, a data processing framework, running along with a Python notebook server for writing code to send Spark instructions for processing data.

### 1. Starting Up Our Data Lake

To run these services in an available terminal, run the following command:

docker compose up spark nessie minio dremio

Keep an eye out. The URL to access the Python notebook server will appear in the terminal output, and you will need this to access the server running on localhost:8888.

spark | [I 2024-04-01 15:02:50.052 ServerApp] <http://127.0.0.1:8888/lab?token=bdc8479a80be54e723eb636e1b62de141a553b75e984a9da>

Put the URL in the browser, and you can create a new notebook. We'll add some code to it later.

### 2. Creating a Bucket in Our Data Lake

Head over to localhost:9001 and enter in the username admin and password password to get access to the minio console, where you can create a new bucket called "warehouse".

### 3. Running the PySpark Script

with the following code:

import pyspark  
from pyspark.sql import SparkSession  
import os  
  
  
## DEFINE SENSITIVE VARIABLES  
CATALOG\_URI = "<http://nessie:19120/api/v1>" ## Nessie Server URI  
WAREHOUSE = "s3://warehouse/" ## S3 Address to Write to  
STORAGE\_URI = "[http://minio:9000](http://minio:9000/)"  
  
  
conf = (  
 pyspark.SparkConf()  
 .setAppName('app\_name')  
 #packages  
 .set('spark.jars.packages', 'org.postgresql:postgresql:42.7.3,org.apache.iceberg:iceberg-spark-runtime-3.5\_2.12:1.5.0,org.projectnessie.nessie-integrations:nessie-spark-extensions-3.5\_2.12:0.77.1,software.amazon.awssdk:bundle:2.24.8,software.amazon.awssdk:url-connection-client:2.24.8')  
 #SQL Extensions  
 .set('spark.sql.extensions', 'org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions,org.projectnessie.spark.extensions.NessieSparkSessionExtensions')  
 #Configuring Catalog  
 .set('spark.sql.catalog.nessie', 'org.apache.iceberg.spark.SparkCatalog')  
 .set('spark.sql.catalog.nessie.uri', CATALOG\_URI)  
 .set('spark.sql.catalog.nessie.ref', 'main')  
 .set('spark.sql.catalog.nessie.authentication.type', 'NONE')  
 .set('spark.sql.catalog.nessie.catalog-impl', 'org.apache.iceberg.nessie.NessieCatalog')  
 .set('spark.sql.catalog.nessie.s3.endpoint', STORAGE\_URI)  
 .set('spark.sql.catalog.nessie.warehouse', WAREHOUSE)  
 .set('spark.sql.catalog.nessie.io-impl', 'org.apache.iceberg.aws.s3.S3FileIO')  
  
)  
  
## Start Spark Session  
spark = SparkSession.builder.config(conf=conf).getOrCreate()  
print("Spark Running")  
  
# Define the JDBC URL for the Postgres database  
jdbc\_url = "jdbc:postgresql://postgres:5432/mydb"  
properties = {  
 "user": "myuser",  
 "password": "mypassword",  
 "driver": "org.postgresql.Driver"  
}  
  
# Load the table from Postgres  
postgres\_df = spark.read.jdbc(url=jdbc\_url, table="sales\_data", properties=properties)  
  
# Write the DataFrame to an Iceberg table  
postgres\_df.writeTo("nessie.sales\_data").createOrReplace()  
  
# Show the contents of the Iceberg table  
spark.read.table("nessie.sales\_data").show()  
  
# Stop the Spark session  
spark.stop()

***If you run into a "Unknown Host" issue using*** [http://minio:9000](http://minio:9000/) then there may be an issue with the DNS in your Docker network that watches the name minio with the ip address of the image on the docker network. In this situation replace minio with the containers ip address. You can look up the ip address of the container with docker inspect minio and look for the ip address in the network section and update the STORAGE\_URI variable for example STORAGE\_URI = "[http://172.18.0.6:9000](http://172.18.0.6:9000/)"

### Breakdown of the PySpark Code

This PySpark script demonstrates how to configure a Spark session to integrate with Apache Iceberg, and Nessie. Apache Spark reads data from a PostgreSQL database and writes it to an Iceberg table Nessie manages.

1. **Import necessary modules:**

- pyspark: The main PySpark library.

- SparkSession: The entry point to programming Spark with the Dataset and DataFrame API.

2. **Define sensitive variables:**

- CATALOG\_URI: The URI for the Nessie server.

- WAREHOUSE: The S3 bucket URI where the Iceberg tables will be stored.

- STORAGE\_URI: The URI of the S3-compatible storage, in this case, a MinIO instance running at 172.18.0.6:9000.

3. **Configure Spark session:**

- Set the application name.

- Specify necessary packages (`spark.jars.packages`) including PostgreSQL JDBC driver, Iceberg, Nessie, and AWS SDK.

- Enable required SQL extensions for Iceberg and Nessie (`spark.sql.extensions`).

- Configure Nessie catalog settings such as URI, reference branch, authentication type, and implementation class.

- Set the S3 endpoint for Nessie to communicate with the S3-compatible storage (MinIO).

4. **Start the Spark session:**

- The SparkSession is initialized with the above configuration.

5. **Database connection setup:**

- Define the JDBC URL for the PostgreSQL database.

- Set connection properties including user, password, and driver.

6. **Data ingestion from PostgreSQL:**

- Read data from the sales\_data table in PostgreSQL into a DataFrame (`postgres\_df`).

7. **Write data to an Iceberg table:**

- Write the DataFrame to an Iceberg table named sales\_data in the Nessie catalog.

8. **Read and display the Iceberg table:**

- Read the newly created Iceberg table from the Nessie catalog and display its contents.

9. **Stop the Spark session:**

- Terminate the Spark session to release resources.

### Can This Be Easier?

Configuring Apache Spark, a standard tool for the Data Engineer can be tedious to configure and troubleshoot. We could use our Data Lakehouse Platform, Dremio, to handle the ingestion of the data with simple SQL statements. To see an example of this, check out the following tutorials:

* [**From Postgres -> Dremio -> Dashboards**](https://www.dremio.com/blog/from-postgres-to-dashboards-with-dremio-and-apache-iceberg/)
* [**From SQLServer -> Dremio -> Dashboards**](https://www.dremio.com/blog/from-sqlserver-to-dashboards-with-dremio-and-apache-iceberg/)
* [**From MongoDB -> Dremio -> Dashboards**](https://www.dremio.com/blog/from-mongodb-to-dashboards-with-dremio-and-apache-iceberg/)
* [**From AWS Glue -> Dremio -> Dashboard**](https://www.dremio.com/blog/bi-dashboards-with-apache-iceberg-using-aws-glue-and-apache-superset/)

### Connecting Our Data to Dremio

[**Dremio is a powerful data lakehouse platform**](https://www.dremio.com/solutions/data-lakehouse/) that can connect several data sources across cloud and on-prem sources and deliver them anywhere you need them, like [**BI Dashboards**](https://www.dremio.com/blog/bi-dashboards-101-with-dremio-and-superset/) and [**Python notebooks**](https://www.dremio.com/blog/connecting-to-dremio-using-apache-arrow-flight-in-python/). We will use Dremio to process queries that power our BI Dashboards.

Now, head to localhost:9047 in your browser to set up your Dremio admin account. Once set up, click “add a Source” and select a “Nessie” as the source. Enter in the following settings:

* General settings tabSource Name: nessieNessie Endpoint URL: [**nessie:19120/api/v2**](http://nessie:19120/api/v2)Auth Type: None
* Storage settings tabAWS Root Path: warehouseAWS Access Key: adminAWS Secret Key: passwordUncheck “Encrypt Connection” Box (since we aren’t using SSL)Connection PropertiesKey: fs.s3a.path.style.access | Value: trueKey: fs.s3a.endpoint | Value: minio:9000Key: dremio.s3.compat | Value: true

Click on “Save,” and the source will be added to Dremio. You can then run full DDL and DML SQL against it. Dremio turns your data lake into a data warehouse—a data lakehouse!

Now we can connect superset and build BI dashboards over any data we have connected to Dremio which can not only include our data lake but many sources like Postgres, SQLServer, Mongo, ElasticSearch, Snowflake, Hadoop, ADLS, S3, AWS Glue, Hive and much more!

### Building our BI Dashboard

Dremio can be used with most existing BI tools, with one-click integrations in the user interface for tools like Tableau and Power BI. We will use an open-source option in Superset for this exercise, but any BI tool would have a similar experience. Let's run the Superset service:

docker compose up superset

We need to initialize Superset, so open another terminal and run this command:

docker exec -it superset superset init

This may take a few minutes to finish initializing, but once it is done, you can head over to localhost:8080 and log in to Superset with the username “admin” and password “admin”. Once you are in, click on “Settings” and select “Database Connections”.

* Add a New Database
* Select “Other”
* Use the following connection string (make sure to include Dremio username and password in URL):

dremio+flight://USERNAME:PASSWORD@dremio:32010/?UseEncryption=false

* Test connection
* Save connection

The next step is to add a dataset by clicking on the + icon in the upper right corner and selecting “create dataset”. From here, choose the table you want to add to Superset, which is, in this case, our sales\_data table.

We can then click the + to add charts based on the datasets we’ve added. Once we create the charts we want, we can add them to a dashboard, and that’s it! You’ve now taken data from an operational database, ingested it into your data lake, and served a BI dashboard using the data.

### Conclusion

In conclusion, this comprehensive guide has journeyed through the critical steps of data engineering, from moving data between operational systems and analytical platforms to leveraging modern data architectures like the Data Lakehouse. Using tools like Apache Iceberg, Nessie, Minio, Apache Spark, and Dremio, we've demonstrated how to efficiently migrate data from a traditional database like Postgres into a scalable and manageable data lakehouse environment. Furthermore, integrating Apache Superset for BI dashboarding illustrates the seamless end-to-end data workflow.

Many other tutorials and resources help you learn more about data engineering.