Éliminez les silos de données : Interrogez plusieurs systèmes en une seule requête

La virtualisation des données dans IBM Cloud Pak for Data

Points saillants

- Interroger de multiples bases de données et référentiels de mégadonnées, individuellement ou collectivement.
- Centraliser le contrôle et la gouvernance des accès.
- Faire apparaître de nombreuses bases de données — distribuées partout dans le monde — comme une seule dans une application.
- Simplifier l'analytique des données avec une plateforme puissante et évolutive.

Contexte

Les données sont partout et les meilleures entreprises au monde sont aujourd'hui axées sur les données. Les entreprises recueillent des données provenant de sources de plus en plus nombreuses et diverses pour les analyser et effectuer leurs opérations; ces sources de données peuvent se compter par milliers ou par millions. La complexité, les coûts, le temps et les risques d'erreur dans la collecte, la gouvernance, le stockage, le traitement et l'analyse centralisés de ces données augmentent de façon exponentielle. Les bases de données et les référentiels dont proviennent toutes ces données sont également plus puissants et dotés de capacités considérables de traitement et de stockage de données, pour que celles-ci soient toujours immédiatement disponibles.

Vue d'ensemble de la virtualisation des données

La virtualisation des données dans IBM Cloud Pak for Data (auparavant appelé IBM Cloud Private for Data) est une nouvelle technologie très puissante qui connecte toutes ces sources de données en un seul ensemble de sources ou de bases de données à équilibrage automatique appelé une *constellation*. (Voir la Figure 1.) Les requêtes d'analytique ne sont plus réalisées sur des données copiées et stockées dans un emplacement centralisé. L'application d'analytique soumet une requête qui est traitée sur le serveur où se trouve la source de données. Les résultats de la requête sont regroupés dans la constellation et retournés à l'application. Aucune donnée n'est copiée, les données n'existent que dans la source.

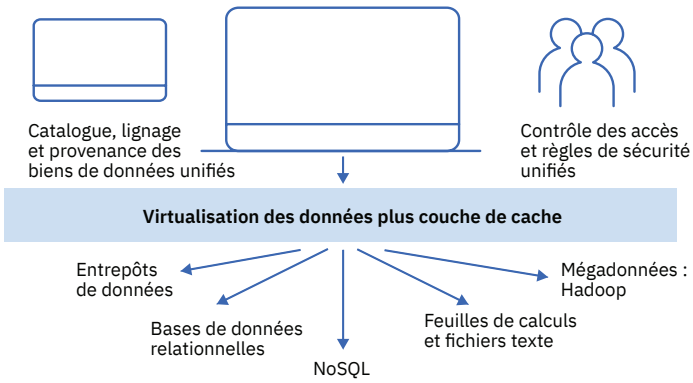

Figure 1 : Virtualisation des données dans Cloud Pak for Data.

Comment fonctionne la virtualisation des données

Les applications se connectent à IBM Data Virtualization comme elles le feraient avec une base de données IBM Db2 et soumettent des requêtes au système comme elles le feraient avec une seule base de sources de données. La charge de travail est distribuée de façon collaborative et traitée par toutes les sources de données participantes.

Des fonctions qui comptent

Plusieurs importantes fonctions d'IBM Data Virtualization permettent aux entreprises d'exploiter plus efficacement leurs données.

Informatique coopérative

En exploitant la puissance de traitement de chaque source de données et en accédant aux données que chaque source de données a stockées, on évite tout temps d'attente. Toutes les données des référentiels sont accessibles en temps réel et les problèmes de gouvernance et de données erronées sont éliminés. Nul besoin d'effectuer l'extraction, la transformation et le chargement (ETL) ni de dupliquer le stockage des données, ce qui accélère le temps de traitement. On fournit donc des connaissances en temps réel aux applications de prise de décision et aux analystes plus rapidement qu'avec les méthodes existantes. Toutefois, s'il est encore nécessaire de copier et de déplacer certaines données à des fins d'historique, d'archivage ou de réglementation, le processus peut fonctionner et coexister avec les méthodes existantes.

Repliement de schémas

Il arrive dans les systèmes de données distribués que plusieurs bases de données stockent des données selon un schéma commun. Par exemple plusieurs bases de données qui stockent des données sur les ventes ou sur les transactions, chacune pour un ensemble de titulaires ou pour une région donnée. IBM Data Virtualization détecte automatiquement les schémas communs entre les systèmes et peut les faire apparaître comme un seul schéma dans la virtualisation des données : c'est le repliement de schémas. Ainsi, un tableau appelé VENTES qui existe dans 20 bases de données peut apparaître comme un seul tableau VENTES, qu'on peut interroger comme un seul tableau virtuel au moyen du langage SQL.

Outils simples pour joindre des vues

Des outils en ligne permettent de définir des vues de tableaux entre différentes bases de données de différents types, parfois même géographiquement dispersées (voir Figure 2).

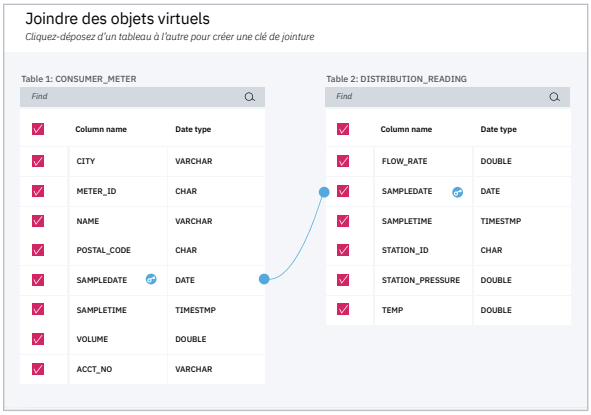

Figure 2 : Une interface intuitive simplifie la jointure de vues de tableaux.

Sécurité

Toutes les communications au sein de la constellation et pour le retour vers l'application sont chiffrées au moyen d'une technologie IBM hautement sécurisée, robuste et puissante, et du chiffrage SSL (Secure Sockets Layer) et TLS (Transport Layer Security) à l'aide de protocoles standards.

Performance

La conception et l'architecture en maillage de calcul d'égal à égal d'IBM Data Virtualization confèrent un important avantage par rapport à une architecture de fédération traditionnelle. Grâce aux avancées de l'équipe Recherche IBM, le moteur de virtualisation des données est capable de fournir rapidement les résultats de requêtes provenant de multiples sources de données, grâce au traitement parallèle et aux optimisations. Les modèles de traitement collaboratif hautement parallèle offrent des performances supérieures pour les requêtes par rapport aux systèmes de fédération, avec des résultats jusqu'à 430 % plus rapides avec des ensembles de données[1] de 100 To. IBM Data Virtualization offre des fonctions sans pareilles de mise à l'échelle de requêtes complexes, avec jointures et agrégats entre des douzaines de systèmes en direct.

IBM Data Virtualization est non seulement rapide, il trouve aussi automatiquement les bases de données et les tableaux, ce qui simplifie la recherche d'informations provenant de multiples sources de données. Les requêtes peuvent facilement combiner des données de sources multiples, notamment de bases de données relationnelles, de sources NoSQL, de feuilles de calcul et de fichiers plats.

Prise en charge des plateformes

Pour les applications, IBM Data Virtualization apparaît comme une instance unique d'une base de données Db2. Par conséquent, les clients et applications de connexion Db2 les plus répandus peuvent être rattachés à IBM Data Virtualization et fonctionner sans modification. C'est le cas même si l'ensemble de sources de données interrogées comprend un mélange de plusieurs types de sources de données, par exemple :

- PostgreSQL
- Oracle
- Netezza
- Microsoft SQL Server

La technologie IBM Data Virtualization est capable de convertir les données vers et en provenance de tous les dialectes SQL. Vos applications peuvent donc librement coder avec SQL, Procedural Language/SQL (PL/SQL) et SQL PL comme si elles travaillaient directement sur la base de données Db2, sans essayer de déterminer si la syntaxe est prise en charge par le système de données cible. Ainsi, les outils les plus répandus peuvent se connecter à IBM Data Virtualization sans aucune modification ni mise à niveau, notamment :

- Logiciel IBM Cognos Business Intelligence (BI)
- Tableau
- MicroStrategy
- Looker
- Plotly
- R
- Jupyter

Le nœud de service de virtualisation des données auquel les applications se connectent est un microservice qui fait partie de Cloud Pak for Data.

Apache Hive	Serveur de base de données Informix
Cloudera Impala	MariaDB
Logiciel Db2	MySQL
IBM Db2 Big SQL	Netezza
IBM Db2 Event Store	Oracle
DerbyDB	PostgreSQL
Fichier Excel et CSV (séparé par une virgule)	SQL Server
Hortonworks Data Platform (HDP) avec Apache Hive	

Tableau 1 : Sources de données prises en charge.

Exigences matérielles minimales

Data Virtualization in Cloud Pak for Data nécessite la configuration suivante :

- Processeur 16 cœurs (v)
- Au moins 64 gigaoctets de mémoire vive (RAM) physique
- 200 gigaoctets d'espace disque (recommandé)

Scénarios courants avec IBM Data Virtualization

IBM Data Virtualization convient bien à l'analyse d'ensembles de données hautement distribuées où les données et les résultats d'analytique sont sensibles au facteur temps. Il fonctionne aussi très bien dans les cas où l'analytique peut constituer une opération unique sur de tels ensembles de données. Il convient aussi dans les cas où le temps d'attente pour la copie de lots à partir de certaines sources de données est trop grand par rapport au besoin commercial d'obtenir les résultats d'analytique.

De nombreuses organisations dupliquent les données et créent de nouveaux référentiels de données pour satisfaire les besoins en analytique des secteurs d'activité. Ce processus nécessite la configuration des biens physiques ainsi que la création et la maintenance de nouveaux processus ETL pour charger et transformer les données pour ces référentiels. Souvent, les données sont désuètes avant d'être rendues disponibles aux secteurs d'activité.

Pour beaucoup d'organisations TI, les approches existantes ont presque atteint leur point de saturation. Étant donné l'augmentation du nombre et de la diversité des sources de données et des besoins en analytique, le modèle ne peut plus évoluer. IBM Data Virtualization peut augmenter la productivité des organisations TI et offrir aux secteurs d'activité une approche évolutive de l'accès aux données, à l'échelle de l'entreprise.

Dans plusieurs cas, les entreprises se heurtent à des problèmes de politiques ou légaux pour la copie ou le déplacement de données, par exemple dans le cas des informations personnelles. Ces restrictions peuvent entraver les efforts d'une entreprise pour combler ses besoins commerciaux en résultats d'analytique démographique. IBM Data Virtualization contribue à résoudre ces problèmes en laissant les données protégées à leur source et en ne retournant que les résultats des requêtes démographiques.

Aujourd'hui, pour pouvoir tester des hypothèses au moyen de l'analytique, un analyste scientifique des données est obligé de créer un lac de données, de copier les données à partir des sources d'intérêt et de les intégrer. IBM Data Virtualization élimine le besoin d'un lac de données, ce qui permet à l'analyste de fédérer les données dont il a besoin pour tester des hypothèses en connectant des outils comme IBM Watson Studio directement aux sources de données.

Plus d'agilité pour les grands projets analytiques

La simplicité offerte par IBM Data Virtualization permet aux utilisateurs d'acquérir des données unifiées exploitables quand ils le veulent, comme ils le veulent et à un rythme qui correspond à leurs besoins en analytique. C'est une technologie qui permet d'augmenter la vitesse et l'efficacité de l'intégration ainsi que la prise de décision, ce qui permet de s'adapter aux besoins en constante évolution du marché.

IBM Data Virtualization in Cloud Pak for Data prend en charge de nombreuses initiatives, notamment :

- La modernisation pour une livraison plus rapide et facile de systèmes d'interaction.
- L'analytique en temps réel, qui répond aux besoins immédiats de l'entreprise.
- L'optimisation pour réduire les coûts et la complexité de l'accès aux données organisationnelles.

IBM Data Virtualization permet d'exploiter l'intelligence d'entreprise en libre-service. Les biens de données virtuels réutilisables offrent une représentation des données adaptée à l'entreprise, permettant à l'utilisateur d'interagir avec les données sans devoir connaître les complexités de la couche de données physiques ou l'endroit où les données sont stockées. Ils permettent également à de nombreux outils d'intelligence d'entreprise et de production de rapports d'acquérir des données en provenance d'une couche de virtualisation de données.

IBM Data Virtualization offre une vue unifiée à 360 degrés. Le bien de données virtualisées présente une vue complète des données en temps réel. La couche de données virtuelles procure une vue unifiée et intégrée des informations commerciales qui permet à l'utilisateur de mieux comprendre et utiliser les données organisationnelles.

IBM Data Virtualization fournit des services de données SOA agiles. Une couche de virtualisation des données fournit la couche de services de données aux applications SOA (architecture orientée services). Elle accélère la création de biens virtuels sans qu'on doive toucher aux sources sous-jacentes, grâce à des fonctions d'autodétection et de mappage des métadonnées qui encapsulent la logique d'accès aux données. La virtualisation des données permet également à de nombreux services de l'entreprise d'acquérir des données à partir d'un emplacement centralisé et réalise un couplage lâche entre les services de l'entreprise et les sources de données physiques.

IBM Data Virtualization permet un meilleur contrôle de l'information. Il améliore la qualité des données grâce à un contrôle centralisé des accès, à une infrastructure de sécurité robuste et à la réduction des copies physiques des données, ce qui réduit aussi les risques. Le référentiel de métadonnées catalogue les magasins de données de l'organisation et les relations entre les données dans différents magasins de données, pour plus de transparence et de visibilité.

Objectif atteint : Transformer et accélérer la prise de décision

Data Virtualization in Cloud Pak for Data est idéal pour les organisations qui recherchent :

- la rentabilité, la croissance et la réduction des risques;
- une agilité et une productivité accrues;
- l'optimisation des investissements TI existants.

Il améliore l'exploitation des investissements existants dans les serveurs et le stockage tout en réduisant la réplication inutile de données et les coûts associés en duplication et en gestion d'infrastructure. Grâce à une administration simplifiée et à un ensemble d'interfaces de programme d'application (API) SQL, il permet à l'entreprise de tirer des bénéfices de l'analytique en temps réel.

Pour de plus amples renseignements, faites l'essai de Cloud Pak for Data sans frais, ou programmez une rencontre de consultation. Nous vous invitons également à explorer plus en profondeur les détails de ce produit en nous visitant sur le Web.

1. Les mesures de performance ont été recueillies dans un environnement de test contrôlé aux laboratoires d'IBM dans la Silicon Valley, à l'aide de la solution de virtualisation des données d'IBM avec différentes sources de données totalisant 100 To. Les mesures ont été réalisées en mai 2019 et les gains de performance ont été comparés aux résultats avec des systèmes IBM de fédération.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produit au Canada
20-02