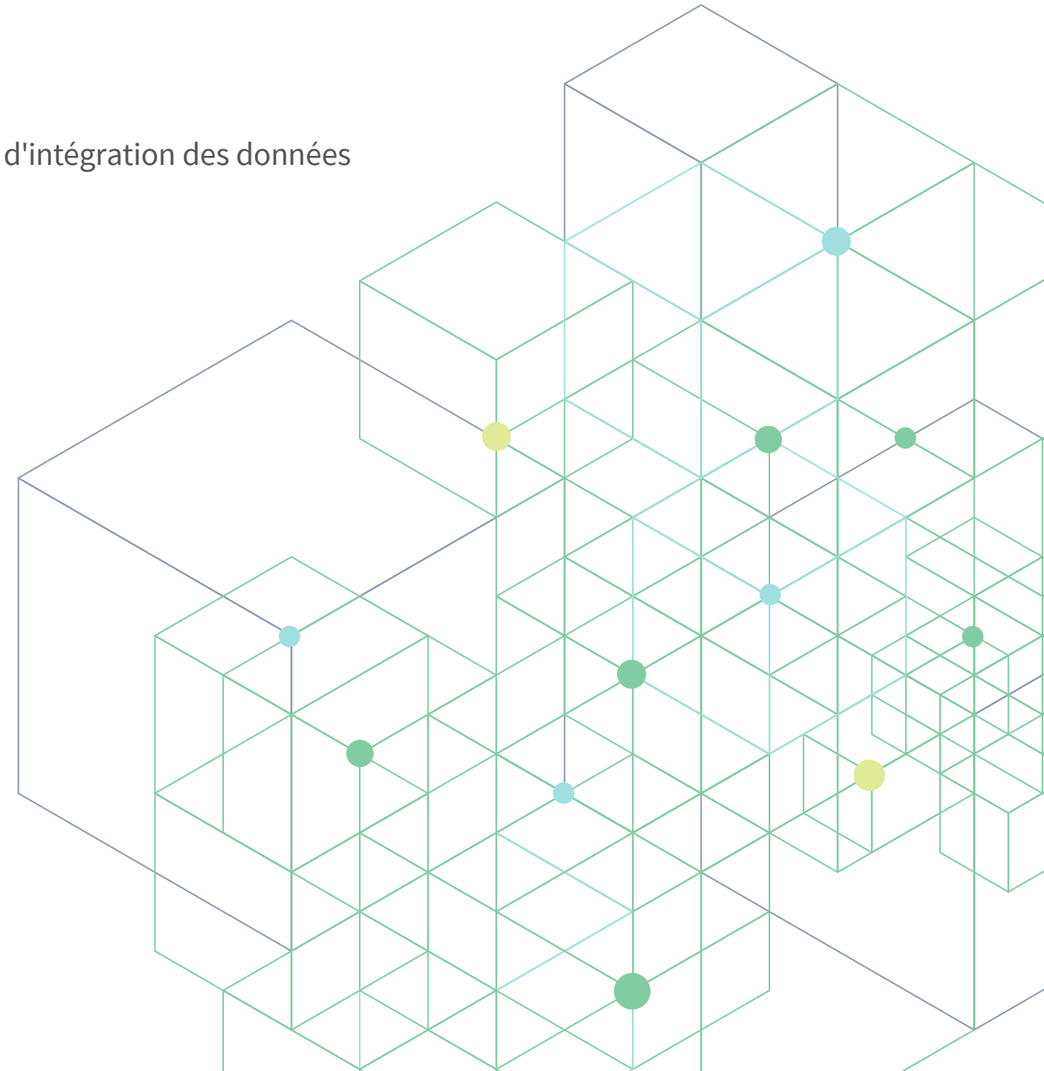
Change Data Capture 101 :

ce qui fonctionne le mieux et pourquoi

LEAD WITH DATA™ Qlik

SOMMAIRE



La nouvelle promesse de l'analytique, et ses exigences sur votre architecture de données.

Des processus métier simplifiés. Des expériences utilisateurs personnalisées. Des approches plus intelligentes en matière de risques. Et de nouvelles sources de revenus. Les initiatives modernes en matière d'analytique ont le potentiel de réinventer votre entreprise de l'intérieur. Mais pour en tirer parti, les services informatiques doivent d'abord réinventer la façon dont ils déplacent, stockent, traitent et analysent les données. Et les défis sont réels.

Les data warehouses dans le Cloud, les data lakes et le streaming de données en temps réel jouent tous un rôle dans les architectures modernes. Ces technologies complètent et, parfois même, remplacent le data warehouse de l'entreprise, le système d'enregistrement structuré traditionnel pour l'analyse de données. Mais n'importe quel data engineer vous le dira : la création d'un pipeline efficace, connecté et rapide entre les données brutes et les données prêtes pour l'analyse est une tâche délicate.

L'un des principaux défis réside dans l'intégration. Les données doivent être répliquées vers des plateformes d'analyse, souvent en continu, sans perturber les applications de production. Et comme les données sont générées à une vitesse vertigineuse, les processus utilisés pour répliquer ces données doivent être évolutifs, efficaces et capables d'absorber de grands volumes de données provenant de nombreuses sources différentes. Pour être pertinents, ils doivent faire tout cela sans augmentation rédhibitoire de la main-d'œuvre ou de la complexité.

Le problème lié à l'intégration traditionnelle des données.

Malheureusement, les exigences d'aujourd'hui liées à l'intégration des données ne peuvent pas être satisfaites avec les processus d'intégration de données d'hier. Les tâches de réplication par lots et les procédures manuelles d'extraction, de transformation et de chargement (ETL) des scripts sont lentes et inefficaces. Elles perturbent la production, contraignent les programmeurs talentueux et créent des goulets d'étranglement au niveau du réseau et du traitement. De plus, elles ne sont pas suffisamment extensibles pour prendre en charge les projets stratégiques. En conséquence, les entreprises passent à côté de certaines opportunités, perdent du terrain sur le plan de la concurrence et dépassent leurs budgets de fonctionnement.

Exemples concrets en entreprise.

- Une entreprise de télécommunications du classement Fortune 25 était incapable d'extraire les données de son ERP SAP et PeopleSoft suffisamment rapidement pour les intégrer dans son data lake. Des processus de chargement laborieux, à plusieurs niveaux, produisaient des retards aussi longs qu'une journée, interférant avec la génération des rapports financiers.

- Une entreprise agro-alimentaire du classement Fortune 100 effectuait, la nuit, des tâches par lots qui ne permettaient pas de rapprocher les commandes et les articles de la chaîne de production dans les délais, ce qui ralentissait les horaires de l'usine et empêchait d'établir des rapports de vente précis.

- L'une des plus grandes entreprises au monde pour le traitement des paiements perdait de la marge sur chaque transaction parce qu'elle n'était pas en mesure d'évaluer assez rapidement la solvabilité de ses clients en interne. Au lieu de cela, elle devait payer une agence extérieure.

- Une grande compagnie d'assurances européenne perdait des clients en raison de retards dans la récupération des informations sur les comptes.

LA SOLUTION CDC

Les avantages de Change Data Capture.

Les processus manuels traditionnels d'intégration des données ne peuvent pas répondre aux demandes actuelles en matière de données. En revanche, les technologies modernes et automatisées en sont parfaitement capables. Une technologie fondamentale pour la modernisation de l'environnement des données est CDC (Change Data Capture), qui identifie et capture en permanence les modifications incrémentielles apportées aux données et aux structures de données d'une source (ou de plusieurs sources) et réplique ces modifications vers une cible (ou plusieurs cibles), où les données peuvent ensuite être transformées et distribuées vers des applications d'analyse des données.

Lorsqu'elle est correctement conçue et mise en œuvre, la technologie CDC permet un transfert de données efficace et à faible latence aux utilisateurs opérationnels et analytiques, répondant ainsi à toutes les exigences actuelles en matière d'évolutivité, de distribution en temps réel et d'impact zéro.

Pourquoi choisir CDC plutôt que la réplication par lots ? La technologie CDC vous permet de :

Prendre des décisions plus rapides et plus précises en permettant aux utilisateurs de tirer parti de données actualisées.

Réduire au minimum les perturbations au niveau des charges de travail en production en envoyant les mises à jour incrémentielles des sources vers les destinations analytiques.

Gagner du temps et réduire les coûts en éliminant la nécessité de transférer des magasins de données massifs sur site vers le Cloud.

Libérer des ressources qualifiées pour les projets à plus grande valeur ajoutée en éliminant la nécessité de scripts manuels.

Modifications, sources et cibles.

Change Data Capture identifie et capture uniquement les modifications les plus récentes au niveau des données de production et des métadonnées, enregistrées par la source pendant une période donnée, généralement un intervalle de quelques secondes ou de quelques minutes. Ensuite, la technologie permet au logiciel de réplication de copier ces modifications dans un référentiel de données séparé.

Types de modifications de données capturés :

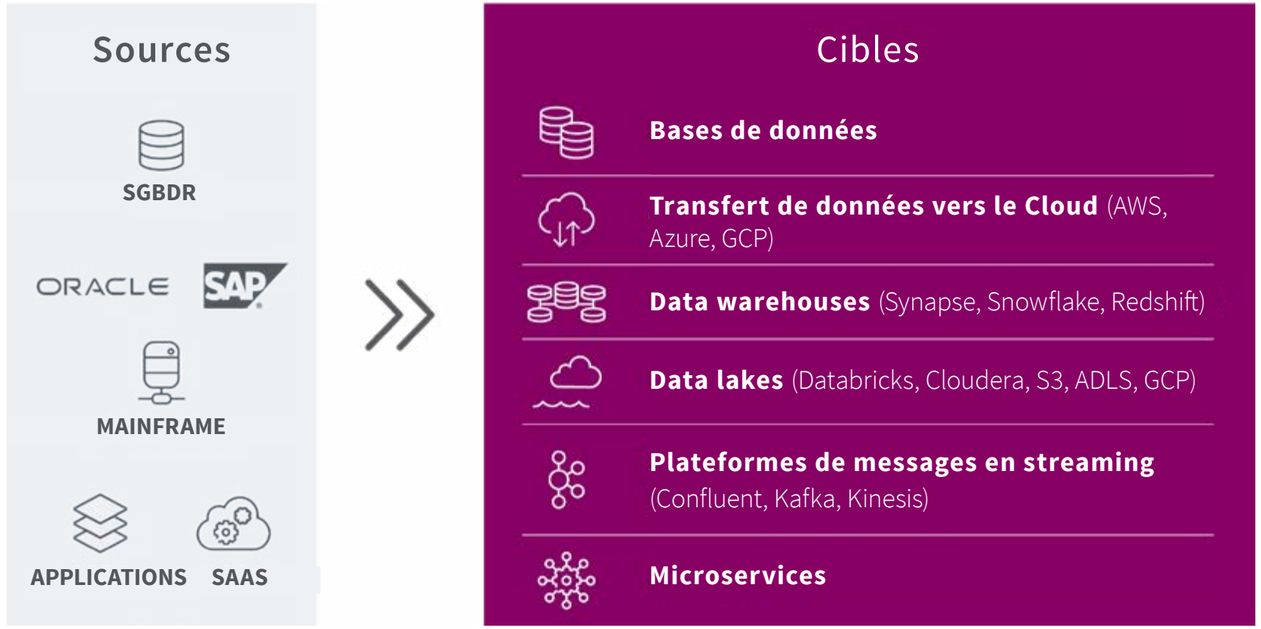
Modifications de lignes

- **INSERTIONS :** lignes ajoutées dans une base de données
- **MISES À JOUR :** champs modifiés dans des lignes existantes
- **SUPPRESSIONS :** lignes supprimées d'une base de données

Modifications de métadonnées

- **DDL :** création, modification et/ou suppression d'objets de base de données comme les colonnes, les tables et les types de données avec le langage de définition de données (DDL).

Sources et cibles concernées :



Méthodes de capture et de distribution.

La capture des modifications dans CDC s'appuie sur trois approches technologiques différentes, certaines plus efficaces que d'autres :

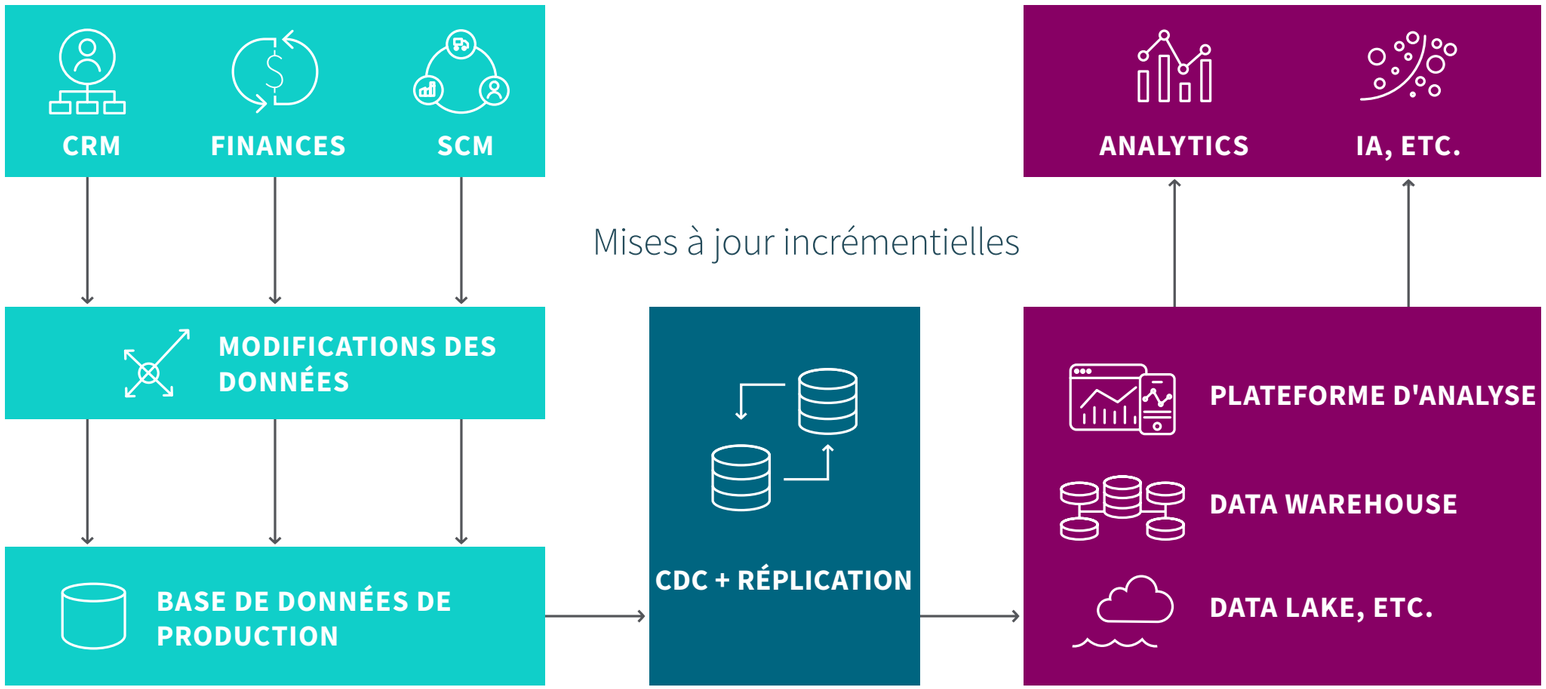
MÉTHODE DE CAPTURE	IMPACT SUR LA PRODUCTION
1. Déclenchement Les transactions au niveau de la source « déclenchent » les copies vers la table de capture des modifications. Méthode privilégiée lorsque les journaux de transactions ne sont pas accessibles.	Moyen
2. Requête Le logiciel signale les nouvelles transactions dans la colonne de la table de production avec les horodatages, les numéros de version, etc. et le moteur CDC demande régulièrement les mises à jour à la base de données de production.	Faible
3. Lecture des journaux Les modifications sont identifiées grâce à l'analyse des journaux de transactions de sauvegarde/ récupération. Méthode privilégiée lorsque les journaux sont accessibles.	Minimal

La distribution des données répliquées vers la cible offre également trois options :

MÉTHODE DE DISTRIBUTION	ÉTUDE DE CAS
1. Transactionnelle CDC copies les mises à jour, ou transactions, dans l'ordre dans lequel elles ont été appliquées à la source. Méthode appropriée lorsque l'intégrité séquentielle est plus importante que des performances ultra-élevées.	Rapports financiers quotidiens, lesquels doivent intégrer l'ensemble des transactions réalisées à partir d'un point spécifique dans le temps.
2. Agrégée (optimisation pour le traitement par lots) CDC regroupe plusieurs mises à jour de la source et les envoie ensemble à la cible. Méthode appropriée lorsque les performances sont plus importantes que l'intégrité séquentielle sur la cible.	Analyse cumulée des tendances sur la base du plus grand nombre possible de points de données.
3. Optimisation pour les flux CDC réplique les mises à jour de la source dans un flux de messages géré par des plateformes telles que Kafka, Azure Event Hub et Amazon Kinesis. Contrairement aux autres méthodes, les cibles gèrent les données en mouvement plutôt que les données au repos.	Série de nouveaux cas d'utilisation, y compris les offres aux clients en temps réel et en fonction de la localisation et l'analyse des données sur les transactions boursières en continu.

Fonctionnement de Change Data Capture avec l'analytics.

CDC a évolué pour devenir un élément essentiel des architectures de données modernes.



Ce schéma présente une vue simplifiée du rôle de CDC.

Avec agents **ou** sans agent

Le logiciel CDC offre deux options architecturales principales :

Avec agents. CDC réside sur le serveur source et interagit directement avec la base de données de production. Les agents de CDC ne sont pas idéaux car ils détournent le processeur, la mémoire et le stockage des charges de production de la source, ce qui dégrade les performances.

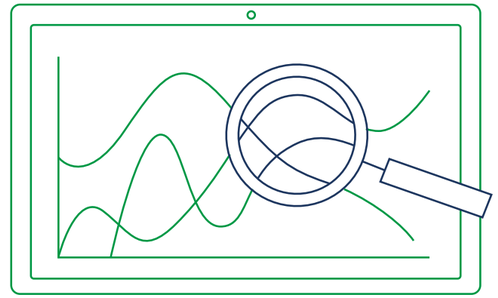
L'architecture sans agent est plus moderne et n'a aucune empreinte sur la source ou la cible. Au lieu de cela, le logiciel interagit avec la source et la cible à partir d'un serveur intermédiaire séparé, ce qui minimise l'impact et améliore la facilité d'utilisation.

Intégration de Change Data Capture dans les architectures modernes.

Les méthodes de transfert des données varient en fonction de la cible.

CIBLE	ACTION ET AVANTAGES DE CDC
Réplication vers des bases de données	Copie les enregistrements nécessaires dans les bases de données de reporting, vous permettant ainsi de délester les requêtes et la charge de travail d'analytics de la production, et profiter de tableaux de bord opérationnels en temps réel.
Publication vers des plateformes de streaming	Convertit les mises à jour de la source en un flux de messages qui peuvent être divisés en thèmes. Ainsi, vous pouvez offrir une latence pratiquement nulle dans les analyses en temps réel.
Microservices	Distribue et synchronise les données sur les différents référentiels de données de microservices spécialisés, ce qui vous permet de fournir des services granulaires à un large éventail de clients.
Transfert de données vers le Cloud	Synchronise en permanence les référentiels de données sur site et dans le Cloud, ce qui vous permet d'éliminer les tâches par lots répétitives et perturbatrices et de n'avoir aucun temps d'arrêt.
ETL et le data warehouse	Étale l'extraction et le chargement des données dans le temps en traitant continuellement les mises à jour incrémentielles, ce qui vous permet de prendre en charge une transformation en temps réel, tout en réduisant l'impact sur les performances.
Ingestion de data lake	Ingère des « Big Data » et des « Wide Data » dans le data lake, pratiquement en temps réel, au fil des modifications, permettant ainsi aux équipes de votre entreprise d'effectuer les analyses rapides et en temps réel qui nécessitent les données les plus récentes.

Présentation de Qlik pour CDC.



La plateforme Qlik d'intégration des données pour le streaming CDC fournit une solution simple et universelle qui permet de convertir les bases de données de production en streaming de données, afin de prendre en charge les microservices et analyses de données modernes. Avec Qlik, les data engineers peuvent entièrement automatiser le transfert des données de bout en bout, en temps réel, entre plusieurs sources et plusieurs destinations. La configuration rationalisée et sans agent, avec une interface graphique simple, facilite le paramétrage, le contrôle et la supervision des pipelines de données.

Qlik permet aux équipes de gestion des données de :

Prendre en charge, avec flexibilité, la publication one-to-many, le mappage automatique des types de données et l'intégration complète des métadonnées, le tout sans codage manuel

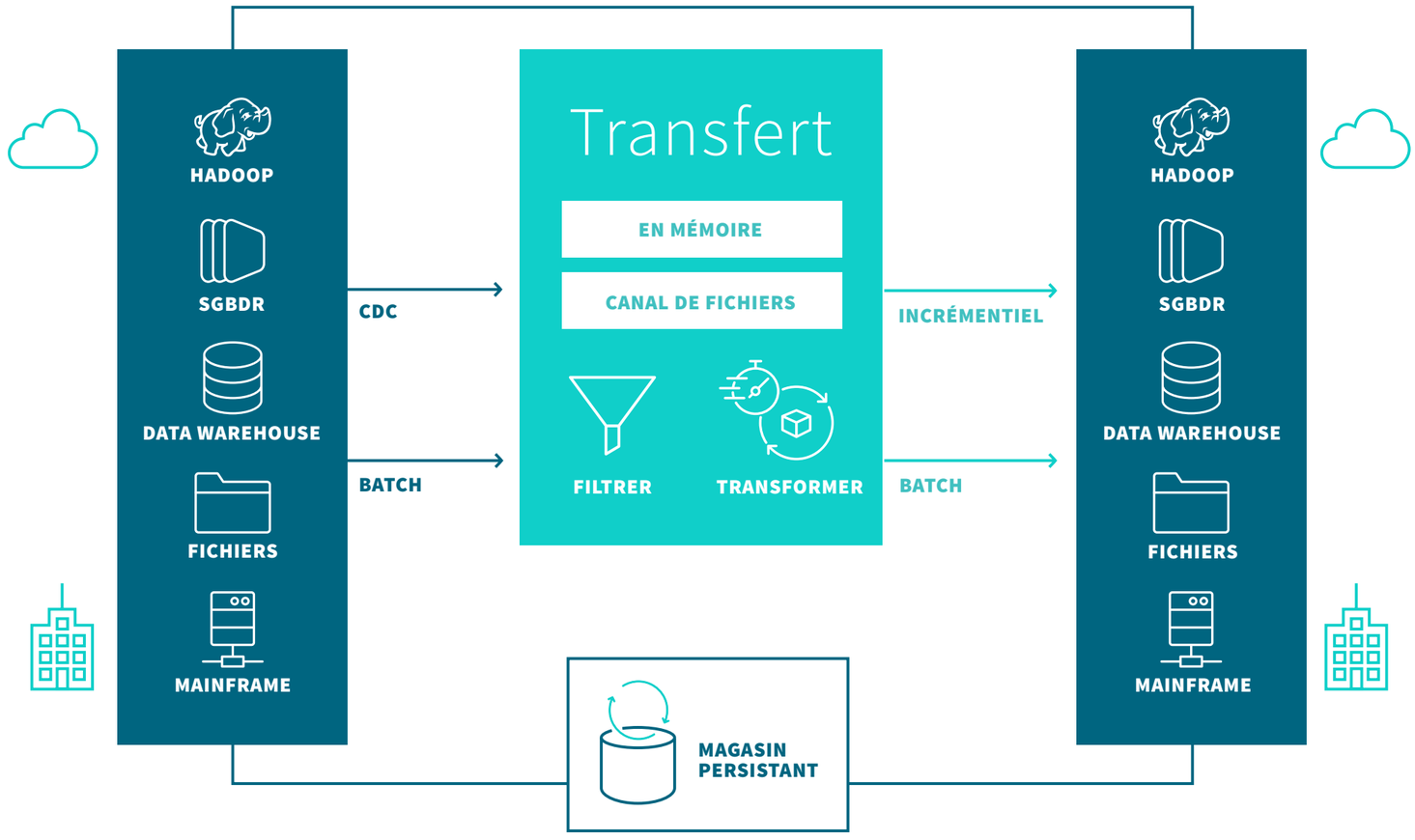
Réduire au minimum l'impact sur les sources, avec la technologie CDC basée sur les journaux et « sans empreinte »

Libérer du temps pour les projets à forte valeur ajoutée en réduisant considérablement votre charge de travail, grâce à un processus automatisé à 100 % et une interface graphique utilisateur intuitive

Tirer parti de l'optimisation du Cloud et de la prise en charge étendue des plateformes, y compris toutes les principales sources de données structurées (Confluent/Apache Kafka, Amazon Kinesis et Azure Event Hub)

CDC pour le transfert de données en temps réel

Qlik pour le streaming CDC réside sur un serveur intermédiaire qui se situe entre une ou plusieurs sources et une ou plusieurs cibles. À l'exception des sources SAP, qui présentent certaines exigences natives particulières, aucun logiciel agent n'est nécessaire, que ce soit sur la source ou la cible. Le mécanisme CDC de Qlik capture les modifications apportées aux données et aux métadonnées, au moyen de la méthode la moins perturbatrice possible, pour chaque source spécifique, généralement son lecteur de journaux.



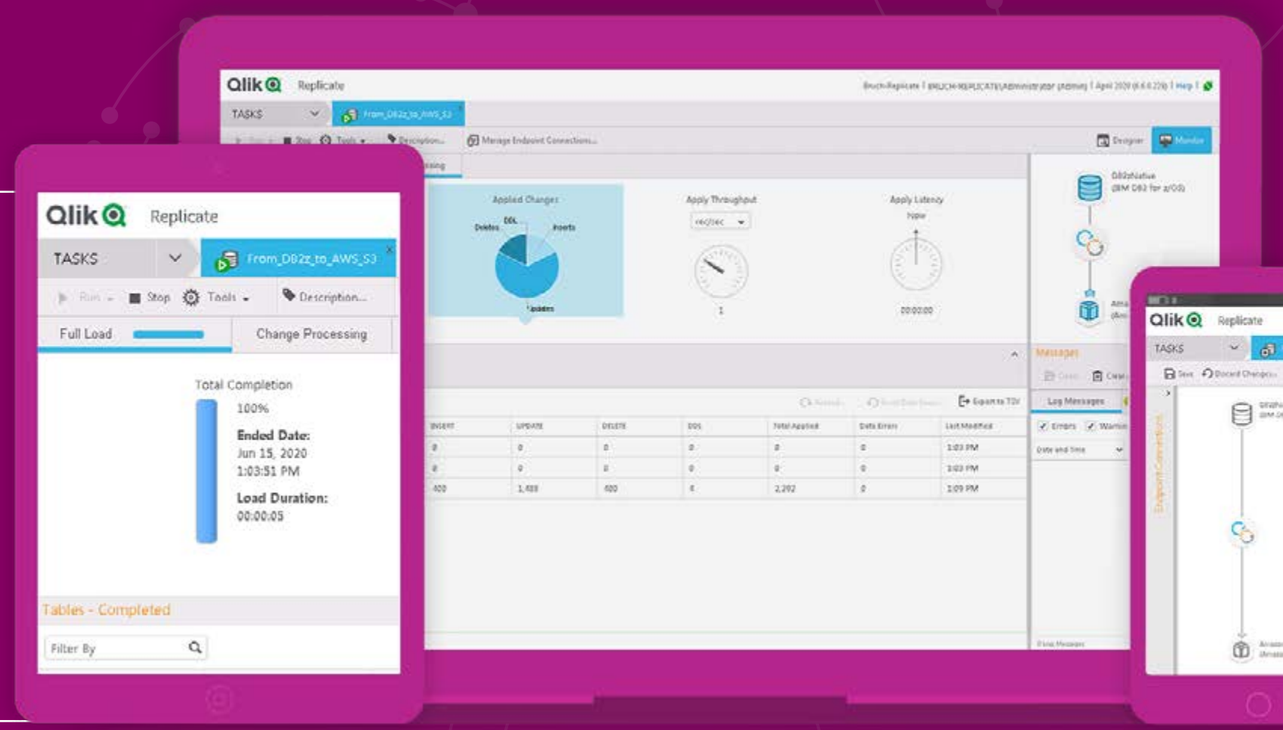
Révolution dans la distribution des données.

Aux quatre coins du monde, des organisations utilisent Qlik pour le streaming CDC afin de réinventer la façon dont elles distribuent leurs données. Plutôt que de se contenter d'un codage manuel, lequel nécessite beaucoup de main-d'œuvre, et des délais inhérents au traitement par lots, elles établissent des flux continus et en temps réel de données fiables et les mettent à la disposition des équipes, à tous les niveaux de l'entreprise, pour qu'elles les utilisent dans les analyses et les microservices. Et les résultats observés sont remarquables.

Dans les pages suivantes, vous trouverez quatre témoignages de réussite avec Change Data Capture.

La plateforme complète

La possibilité de répliquer vos données en continu est l'une des composantes de la plateforme d'intégration de données Qlik. Vous pouvez encore accélérer vos pipelines de données, de façon spectaculaire, avec les solutions Qlik pour l'automatisation du data warehouse et la création de data lakes.



SUCCESS STORIES AVEC CDC

Aggreko stimule sa croissance avec des insights en temps réel.

Aggreko, l'un des principaux fournisseurs mondiaux de solutions d'énergie, de chauffage et de refroidissement, s'était fixé comme mission de s'appuyer de plus en plus sur les insights. Pour ce faire, l'entreprise devait commencer par centraliser l'ensemble de ses données opérationnelles.

DÉFI

Dans sa quête pour établir un processus propre et cohérent d'ingestion des données issues de ses systèmes sources, l'équipe d'Aggreko avait du mal à gérer les différentes exigences de chaque pipeline. Ceci devait vraisemblablement l'amener à développer une solution personnalisée pour chaque source de données, ce qui deviendrait rapidement ingérable.

SOLUTION

La recherche a été lancée afin de trouver un produit unique, capable de gérer tous les pipelines d'ingestion. L'équipe a choisi Qlik pour deux raisons : parce que la solution est extrêmement légère, avec un impact minimal sur les systèmes de production, et parce que sa facilité de mise en œuvre aiderait la petite équipe de data engineers d'Aggreko à gérer de gros volumes avec rapidité.

RÉSULTATS

Qlik CDC Streaming a été mis en œuvre pour le système ERP d'Aggreko en une semaine seulement, sans aucun impact sur les systèmes de production. Aggreko peut désormais fournir des insights en temps réel et prévoit d'étendre la solution à d'autres sources de données.

Generali réduit ses délais « source vers cible » à moins de 10 secondes.

Generali Switzerland, la branche suisse d'un leader mondial de l'assurance, emploie 1 800 personnes sur 56 sites et dessert plus d'un million de clients.

DÉFI

Alors que les applications principales héritées pesaient de plus en plus lourd sur les applications orientées clients et les applications de canaux, l'entreprise a entrepris de moderniser son architecture de données. Ses objectifs : 1) fournir des données actualisées et de haute qualité par l'intermédiaire de différents canaux et 2) améliorer la fourniture des services informatiques.

SOLUTION

Dès le début, les données à répliquer étaient extrêmement volumineuses, ce qui impliquait une intégration importante. L'équipe a développé une plateforme de connexion hybride, alimentée par Qlik et Confluent (Apache Kafka), en partie parce que les deux solutions fonctionnent harmonieusement ensemble. Qlik se situe entre les sources de données et les destinations cibles, lisant et comprenant les informations et envoyant celles-ci vers la cible choisie par Generali.

RÉSULTATS

Les délais « source vers cible » ont été réduits, de plusieurs jours à moins de 10 secondes. Les données sont extraites sans perturber l'activité ou les applications. Une version unique et fiable des données est désormais disponible via un grand nombre d'applications et de canaux. L'accès à des données précises en temps réel a permis d'améliorer le service client et l'engagement. Et enfin, Generali peut prendre en charge le développement de solutions agiles.

« Le streaming de données avec Qlik nous permet d'assurer la réplication et le streaming des données en quelques secondes seulement. Avant, cela pouvait nous prendre plusieurs jours. La valeur ajoutée est considérable pour notre entreprise. »

Christian Nicoll, directeur des opérations et de l'ingénierie des plateformes

GENERALI

Veritix (AXS) accélère l'intégration des données et dope ses performances en analytics.

Veritix (qui fait maintenant partie d'AXS) conçoit des applications de billetterie numérique, de marketing événementiel et de gestion de la relation client pour les équipes sportives professionnelles et les lieux de divertissement dans le monde entier. Grâce à son reporting d'analytics, ses clients peuvent cibler plus efficacement certains groupes démographiques, améliorer les taux de renouvellement des abonnements et attirer de nouveaux fans dans les salles.

DÉFI

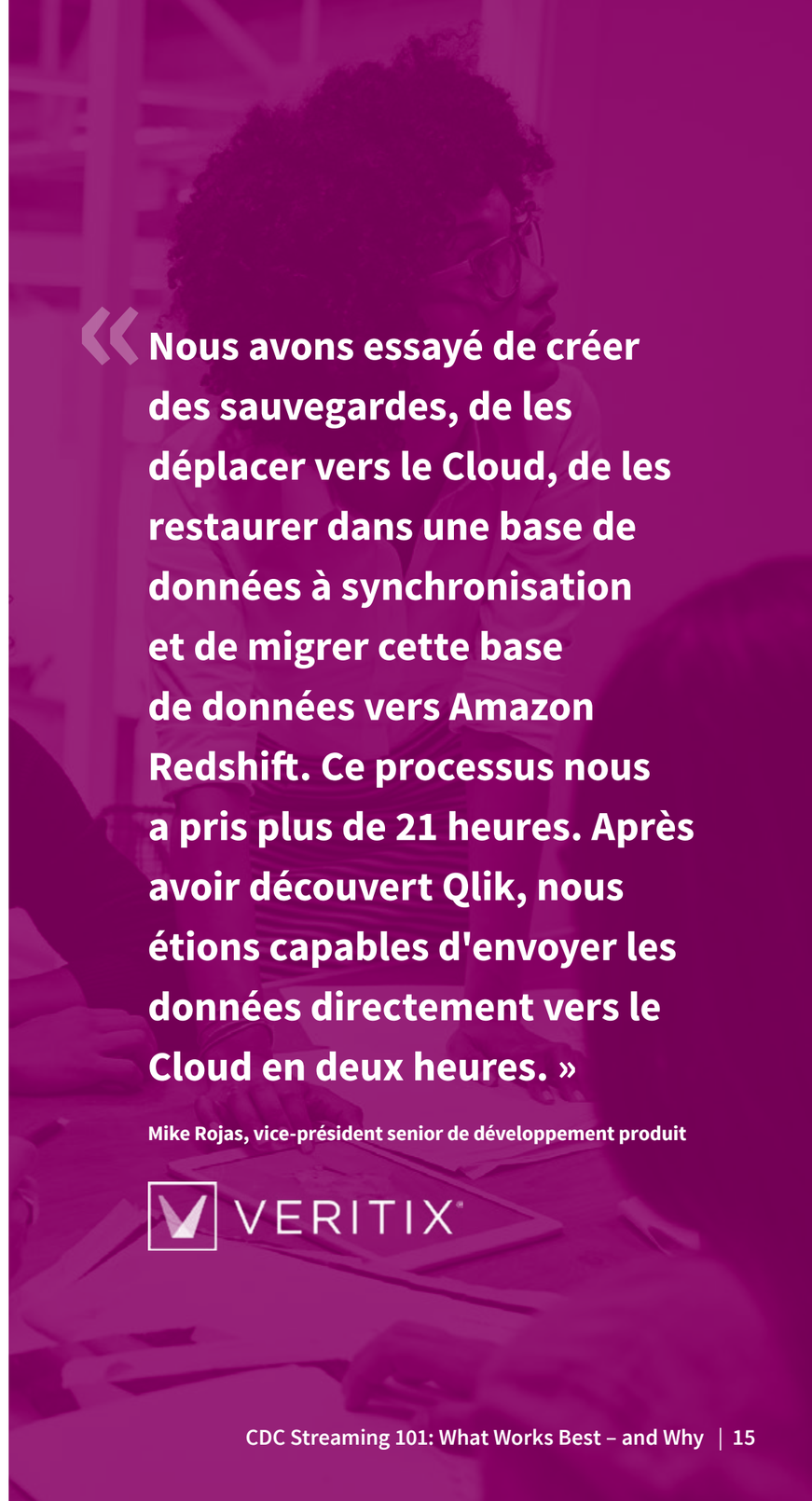
La base de données transactionnelle de Veritix était utilisée à la fois à des fins de production et de reporting. Par conséquent, elle n'était pas optimisée pour l'analyse, et les performances étaient préoccupantes. L'équipe a donc décidé de créer un data warehouse distinct et, pour ce faire, a choisi Amazon Redshift.

SOLUTION

Dans un premier temps, les data engineers de Veritix sont partis du principe qu'ils devaient créer des outils personnalisés ou s'appuyer sur une solution standard pour répliquer les données dans le Cloud. Leur première tentative a pris plus de 21 heures. Puis ils ont découvert Qlik, qui n'a pris que deux heures pour installer, configurer et transférer plusieurs centaines de millions d'enregistrements vers Redshift.

RÉSULTATS

Le data warehouse contient maintenant environ trois téra-octets d'informations, avec des milliards d'enregistrements disponibles pour l'analyse client. Et non seulement Qlik a répondu aux exigences de performance de Veritix, mais a également dépassé les attentes de l'équipe en termes de facilité d'utilisation.



Présentation de la plateforme Qlik d'intégration des données.

Dans la modernisation de votre environnement de données, la réplication de CDC est une pièce du puzzle, ou plus précisément, une pièce du pipeline. Chez Qlik, nous avons conçu une plateforme d'intégration de données de bout en bout, laquelle accélère la découverte et la disponibilité de données prêtes à l'emploi, grâce à l'automatisation du streaming de données en temps réel, mais aussi du perfectionnement, du catalogage et de la publication des données.

Streaming des données en temps réel
Développez le streaming des données d'entreprise pour favoriser la mise en place d'analyses et microservices modernes.

Création de Data Lakes gérés
Automatisez les processus complexes d'ingestion et de transformation des données pour fournir des data lakes prêts à l'emploi.

Automatisation agile du data warehouse
Concevez, créez, déployez et gérez rapidement des data warehouses sur mesure, sans codage manuel.

Catalogue de données d'entreprise
Donnez à l'ensemble des utilisateurs de votre entreprise les outils nécessaires pour trouver, préparer et partager des données prêtes à l'emploi.

Tous les éléments de la plateforme d'intégration de données Qlik fonctionnent ensemble, ce qui vous permet d'établir des pipelines de données automatisés et en temps réel, avec des flux issus des systèmes transactionnels, data warehouses ou data lakes, pour générer des données fiables et exploitables à la demande, à tous les niveaux de votre organisation.

Tirez parti des dernières innovations en matière d'analyse des données, avec les nouvelles technologies d'intégration des données.

La solution CDC de Qlik, qui constitue la base de pipelines de données modernes et efficaces, permet l'intégration automatique, en temps réel et universelle des données dans toutes les principales architectures de données, sur site et dans le Cloud. Elle permet de transférer les données à grande vitesse. Elle est simple à utiliser et à gérer. Et elle vous assure une visibilité globale et un contrôle centralisé de la réplication des données dans votre environnement hybride et distribué.

LAISSEZ-VOUS GUIDER

Libérez le potentiel de vos données dans vos sources héritées. Facilitez la mise en œuvre de grands projets d'intégration des données. Réduisez l'impact sur vos systèmes de production. Éliminez les requêtes directes. Gagnez du temps. Réduisez la charge de travail IT en « back office ». Et accompagnez vos équipes en leur fournissant la flexibilité dont elles ont besoin, tout en préservant la sécurité de vos données.

Vous souhaitez en savoir plus ?

Visiter le site Web

Ou plongez directement dans Qlik CDC Streaming et essayez notre solution par vous-même.

Essai gratuit

À PROPOS DE QLIK

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

Qlik LEAD WITH DATA™

qlik.com