UN APPRENTISSAGE FACILE

2e édition spéciale Snowflake

Le Datawarehouse dans le cloud

pour les nuls®

- Qu'est-ce qu'un datawarehouse dans le cloud ?
- Comparaison des solutions de datawarehouse
- Comment choisir un datawarehouse dans le cloud

Proposé par :

snowflake®

Joe Kraynak
David Baum

À propos de Snowflake

Dès ses débuts, Snowflake avait une vision claire : rendre le datawarehouse moderne efficace, abordable et accessible à tous les utilisateurs de données. Snowflake permet à l'entreprise orientée données de bénéficier d'une élasticité instantanée, d'un partage sécurisé des données et d'une tarification à la seconde, sur plusieurs clouds. Parce que les solutions on-premise et cloud traditionnelles peinent à ce niveau, Snowflake a développé un nouveau produit avec une nouvelle architecture conçue pour le cloud qui combine la puissance du datawarehouse, la flexibilité des plateformes Big Data et l'élasticité du cloud, à une fraction du coût des solutions traditionnelles. Snowflake : vos données, sans limites.

Pour plus d'informations, consultez le site
Snowflake at snowflake.com.

Le Datawarehouse dans le cloud

Une 2e édition spéciale Snowflake

par Joe Kraynak et David Baum

pour les nuls®

Le Datawarehouse dans le cloud pour les Nuls®, 2e édition spéciale Snowflake

Publié par
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

Pour obtenir des renseignements généraux sur nos autres produits et services, ou sur la publication d'un livre *Pour les Nuls* destiné à votre entreprise ou organisation, veuillez contacter notre service de développement commercial aux États-Unis, par téléphone au 877-409-4177, par e-mail à info@dummies.biz, ou consulter notre site www.wiley.com/go/custompub. Pour obtenir des informations sur les licences relatives à la marque *Pour les Nuls* pour des produits et services, veuillez écrire à l'adresse BrandedRights&Licenses@wiley.com.

ISBN 978-1-119-71448-4 (pbk); ISBN 978-1-119-71451-4 (ebk)

Imprimé aux États-Unis d'Amérique

10 9 8 7 6 5 4 3 2 1

Remerciements de l'éditeur

Nous sommes fiers de ce livre et des personnes qui y ont travaillé. Pour savoir comment créer un livre personnalisé *Pour les nuls* pour votre entreprise ou organisation, contactez info@dummies.biz ou visitez www.wiley.com/go/custompub. Pour obtenir des informations sur les licences relatives à la marque *Pour les Nuls* pour des produits et services, veuillez écrire à l'adresse BrandedRights&Licenses@wiley.com.

Cet ouvrage a été réalisé avec la participation des personnes suivantes :

Éditeur de développement : Nicole Sholly

Rédacteur projet : Martin V. Minner

Rédacteur en chef : Steve Hayes

Responsable éditorial : Rev Mengle

Représentante du développement commercial : Karen Hattan

Éditeur de la production : Mohammed Zafar Ali

Équipe des contributeurs de Snowflake : Vincent Morello, Clarke Patterson, Leslie Steere, Kent Graziano

Table des matières

Introduction

En tant que dirigeant, responsable ou analyste, vous êtes bien conscient que la connaissance est synonyme de pouvoir et que des données correctement analysées en temps utile fournissent les informations nécessaires pour prendre des décisions éclairées et obtenir un avantage concurrentiel. Aujourd'hui, les organisations disposent d'un ensemble de données beaucoup plus important et plus pertinent que jamais auparavant. Cela comprend un large éventail de sources, internes et externes, y compris les datamarts, les applications cloud et les données générées par des machines.

Malheureusement, l'architecture des datawarehouses (c'est-à-dire des entrepôts de données) de ces 30 dernières années continue à être mise à rude épreuve sous le poids d'ensembles de données extrêmement vastes et divers. Les analystes attendent souvent 24 heures ou plus pour que les données entrent dans le datawarehouse avant de pouvoir les analyser. Ils doivent quelquefois attendre encore plus longtemps pour exécuter des requêtes complexes sur ces données. Dans de nombreux cas, les ressources de stockage et de calcul nécessaires pour traiter et analyser ces données sont insuffisantes. Cela conduit la suspension ou à l'effondrement des systèmes. Pour éviter cela, les utilisateurs et les charges de travail doivent être mis en file d'attente, ce qui entraîne des délais encore plus longs. Plus récemment, d'autres approches ont vu le jour, telles que différentes formes de data lakes (ou lacs de données). Ces solutions ont cependant apporté leurs propres limites.

Pour rester efficaces et compétitives, les organisations doivent être capables d'exploiter la puissance des vastes quantités de données générées en permanence et de mener des analyses complexes de ces données. Heureusement, la commercialisation du cloud computing est apparue il y a plus de dix ans et offre des avancées en matière de matériel informatique, d'architecture et de logiciels qui peuvent aider votre organisation à relever ce défi et à dépasser vos attentes.

À propos de ce livre

Bienvenue à la deuxième édition de *Datawarehouse dans le cloud pour les nuls*, dans laquelle vous découvrirez comment votre organisation peut exploiter le pouvoir d'énormes quantités de données de manière pratique et abordable pour améliorer son efficacité et transformer des données brutes en précieuses informations commerciales.

Une plus grande quantité de données ouvre la porte à des opportunités plus nombreuses et plus importantes, qui s'accompagnent presque toujours de défis tout aussi importants. Pour tirer parti de ces nombreuses possibilités, vous devez mettre en œuvre une solution de datawarehouse capable de stocker et d'organiser les données sous divers formats, de fournir un accès pratique à celles-ci et d'améliorer la vitesse à laquelle vous pouvez les analyser. Et cela doit être fait de la manière la plus rentable possible. Ce livre vous montre comment.

Icônes employées dans ce livre

Tout au long de ce livre, les icônes suivantes mettent en évidence les conseils, les points importants à retenir, et bien plus encore :

CONSEIL

Les conseils vous guident vers des moyens plus simples d'effectuer une tâche ou vers de meilleures façons pour votre organisation d'utiliser le datawarehouse dans le cloud.

RAPPEL

Cette icône met en évidence les concepts qui méritent d'être rappelés lorsque vous vous plongez dans l'apprentissage et l'application du datawarehouse dans le cloud.

ÉTUDE DE CAS

Les études de cas présentées dans ce livre révèlent comment ces organisations ont appliqué le datawarehouse dans le cloud pour économiser de l'argent et améliorer considérablement la vitesse et les performances de leurs analyses de données.

Au-delà de ce livre

Si vous avez aimé ce que vous avez lu dans ce livre et que vous souhaitez en savoir plus, nous vous invitons à visiter le site `www.snowflake.com`, où vous pourrez en savoir plus sur l'entreprise et ses offres, essayer gratuitement Snowflake, obtenir des détails sur les différents plans et tarifs, visionner des webinaires, accéder aux communiqués de presse, avoir la primeur des évènements à venir, accéder à la documentation et à d'autres supports, et entrer en contact avec elle – elle sera ravie d'avoir de vos nouvelles !

DANS CE CHAPITRE

- Explorer le datawarehouse: d'hier à aujourd'hui
- Comprendre les avantages d'un datawarehouse dans le cloud
- Reconnaître la place du datawarehouse dans le cloud dans l'économie actuelle

Chapitre 1
En savoir plus sur le datawarehouse dans le cloud

Sous une forme ou une autre, le cloud computing et le SaaS (Software as a Service ou logiciel en tant que service) existent depuis des décennies. Mais le DWaaS dans le cloud (data warehouse-as-a-service ou datawarehouse comme service dans le cloud -) n'est apparu que récemment comme une alternative aux solutions traditionnelles de datawarehouse on-premise et autres solutions similaires. Pourquoi aujourd'hui ? Qu'est-ce qui a changé ? Dans ce chapitre, nous répondons à ces questions, et à bien d'autres encore.

Nous commençons par définir ce qu'est un datawarehouse et explorons son évolution pour montrer comment cette technologie a fait son chemin vers le cloud. Nous examinons ensuite comment les organisations peuvent bénéficier du DWaaS dans le cloud et expliquons pourquoi davantage d'entreprises s'appuient sur le datawarehouse dans le cloud pour être compétitives dans l'économie actuelle axée sur les données.

Définition du datawarehouse

Un *datawarehouse* est un système informatique dédié au stockage et à l'analyse des données afin de révéler des tendances, des schémas et des corrélations qui fournissent des informations et des indicateurs. Traditionnellement, les organisations ont utilisé les datawarehouses pour stocker et intégrer les données collectées à partir de leurs sources internes (généralement des bases de données transactionnelles), notamment de marketing, ventes, production et finances. Le

datawarehouse est apparu lorsque les entreprises ont réalisé que l'analyse des données directement à partir de ces bases de données transactionnelles les ralentissait (et les faisait même crasher) sous la pression de leur activité transactionnelle normale et des charges de travail nécessaires pour analyser ces données. Par conséquent, toutes ces données ont été dupliquées dans un datawarehouse pour analyse, laissant la base de données se concentrer sur les transactions.

Au fil des ans, les sources de données se sont étendues au-delà des opérations commerciales internes et des transactions externes. Elles comprennent maintenant des volumes, une variété et une vitesse de transmission de données exponentiellement plus importants provenant de sites web, de téléphones portables et d'applications, de jeux en ligne, d'applications bancaires en ligne et même de machines. Plus récemment, les organisations capturent d'énormes quantités de données sur les périphériques IoT (Internet of things ou Internet des objets).

L'évolution du datawarehouse

Historiquement, les entreprises collectaient des données sous des formes bien définies et très structurées, à un rythme et un volume raisonnablement prévisibles. Même lorsque la vitesse des anciennes technologies a progressé, l'accès aux données et leur utilisation étaient soigneusement contrôlés et limités pour garantir des performances acceptables pour chaque utilisateur, en raison de la rareté de la puissance de calcul et du stockage on premise et de la difficulté d'augmenter ces ressources. Les organisations devaient donc tolérer de très longs cycles d'analyse.

Les temps ont changé (voir figure 1-1). Grâce aux progrès technologiques, les organisations peuvent prendre des décisions commerciales importantes en s'appuyant sur de grandes quantités de données.

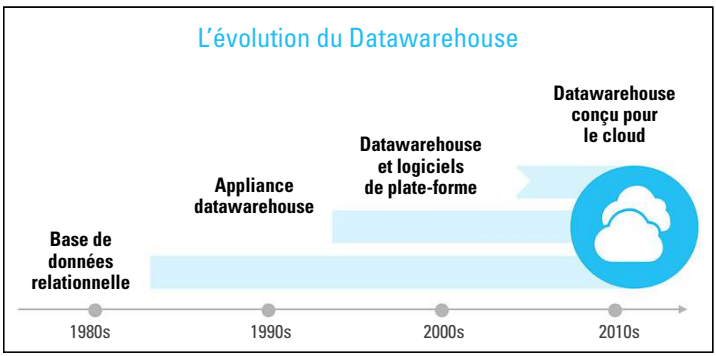

FIGURE 1-1 : Les systèmes traditionnels ont permis l'émergence du datawarehouse dans le cloud.

Il ne s'agit pas seulement des leaders du marché ou des entreprises bien établies. Les petits entrants sur le marché, plus agiles, continuent à transformer des industries bien établies en quelques mois, voire un ou deux ans seulement. Et ce, avec des données pour révéler des opportunités et développer des produits et services qui changent la façon dont les fournisseurs des particuliers et des entreprises éveillent l'intérêt de leurs clients.

Reconnaître les limites du data-warehouse traditionnel

Les datawarehouses traditionnels n'ont pas été conçus pour traiter le volume, la variété et la vitesse des données actuelles. Les systèmes plus récents conçus pour remédier à ces lacunes ont du mal à s'adapter aux exigences actuelles des organisations en matière d'accès et d'analyse des données. Les défis d'aujourd'hui révèlent que :

- Les sources de données sont plus nombreuses et variées, ce qui se traduit par des structures de données plus diverses qui doivent coexister en un seul endroit pour permettre une analyse exhaustive et abordable.
- Les architectures traditionnelles entraînent par nature une concurrence entre les utilisateurs et les activités d'intégration de données, ce qui rend difficile le chargement simultané de nouvelles données dans le datawarehouse et ne fournit pas aux utilisateurs des performances adéquates.
- Le chargement de données par lots à des intervalles spécifiques est encore courant, mais de nombreuses organisations exigent un chargement continu des données (*microbatching*) et des données en continu (*chargement instantané*).
- Développer un datawarehouse traditionnel pour répondre aux demandes croissantes de stockage et de charges de travail d'aujourd'hui, si possible, est coûteux, difficile et lent.
- Les plateformes de données alternatives les plus récentes sont souvent complexes et nécessitent des compétences spécialisées ainsi que de nombreux réglages et configurations. Cela s'aggrave avec le nombre et la diversité croissants des sources de données, des utilisateurs et des requêtes.

La technologie et le design à la rescousse

La bonne nouvelle est que la technologie et l'*architecture* du datawarehouse (les éléments de conception et de construction du datawarehouse moderne) ont évolué pour répondre aux exigences de l'économie basée sur les données grâce aux innovations suivantes :

- **Le cloud :** Le cloud est un facteur clé de l'évolution du datawarehouse moderne. Il permet d'accéder à un stockage quasi illimité et peu coûteux, d'améliorer l'évolutivité, de sous-traiter la gestion et la sécurité du datawarehouse au fournisseur cloud et de ne payer que les ressources de stockage et de calcul réellement utilisées.
- **Traitement massivement parallèle (ou MPP pour Massively parallel processing) :** Le MPP, qui consiste à diviser une seule opération informatique à exécuter simultanément sur un grand nombre de processeurs informatiques distincts, est apparu au début des années 2000. Cette division du travail facilite un stockage et une analyse plus rapides des données lorsque le logiciel est conçu pour capitaliser sur l'approche.
- **Stockage en colonnes :** Traditionnellement, les bases de données stockaient les enregistrements en rangées, de la même manière qu'un tableur. Il pouvait s'agir, par exemple, de toutes les informations concernant un client ou une transaction de détail. Pour récupérer les données de la manière traditionnelle, le système devait lire toute la rangée pour obtenir un élément. C'était laborieux et cela prenait beaucoup de temps. Avec le stockage en colonnes, chaque élément des données d'un enregistrement est stocké dans une colonne. Grâce à cette approche, un utilisateur peut interroger un seul élément de données, par exemple les membres d'un club de gym qui ont payé leur cotisation, sans avoir à lire tout le reste de l'enregistrement, qui peut comprendre le numéro d'identification de chaque membre, son nom, son âge, son adresse, sa ville, son département, les informations de paiement, etc. Cette approche permet une réponse beaucoup plus rapide à ce type de requêtes analytiques.
- **Traitement vectoriel :** Cette forme de traitement pour *l'analyse des données* (la science qui consiste à examiner les données pour en tirer des conclusions) tire profit des conceptions récentes et révolutionnaires des puces électroniques. Cette approche permet des performances bien plus rapides que les anciennes solutions de datawarehouse conçues il y a des décennies pour des technologies de matériel plus ancien et plus lent.
- **Disques à circuits intégrés (Solid State Drive – SSD) :** Contrairement aux lecteurs de disques durs (Hard Disk Drive – HDD), les SSD stockent les données sur des puces de mémoire flash, ce qui accélère le stockage, la récupération et l'analyse des données. Une solution qui tire avantage des SSD peut offrir des performances nettement meilleures.

Pour en savoir plus sur les progrès technologiques et les autres tendances à l'origine de l'évolution du stockage des données, voir le chapitre 2.

Introduction du datawarehouse dans le cloud

Le stockage de données dans le cloud est un moyen rentable pour les entreprises de tirer avantage des technologies et de l'architecture les plus récentes sans avoir à supporter les énormes coûts initiaux d'achat, d'installation et de configuration du matériel, des logiciels et de l'infrastructure nécessaires. Les différentes options de stockage de données dans le cloud sont généralement regroupées en trois catégories.

- **Logiciel de datawarehouse traditionnel déployé sur une infrastructure cloud :** cette option est similaire à un stockage de données traditionnel on premise, car elle réutilise le codebase d'origine. Vous avez encore besoin d'une expertise informatique pour construire et gérer le datawarehouse. Bien qu'il ne soit pas nécessaire d'acheter et d'installer le matériel et les logiciels, il se peut que vous ayez à effectuer des opérations de configuration et de réglage importantes, telles que des sauvegardes régulières.
- **Datawarehouse traditionnel hébergé et géré dans le cloud par une tierce partie en tant que service géré :** avec cette option, le fournisseur tiers fournit l'expertise informatique, mais vous risquez toujours de rencontrer les mêmes limites que celles d'un datawarehouse classique. Le datawarehouse est hébergé sur du matériel installé dans un datacenter géré par le fournisseur. Cela ressemble à ce que l'industrie appelle un *fournisseur de services d'application* (Application Service Provider – ASP). Les clients doivent encore préciser à l'avance la quantité d'espace disque et de ressources de calcul (CPU et mémoire) qu'ils comptent utiliser.
- **Un véritable datawarehouse en mode SaaS :** avec cette option, souvent appelée le *DWaaS*, le fournisseur propose une solution complète de datawarehouse dans le cloud qui inclut tout le matériel et les logiciels, et élimine presque toutes les tâches liées à l'établissement et à la gestion des performances, de la gouvernance et de la sécurité requises par un datawarehouse. Les clients ne paient généralement que pour les ressources de stockage et de calcul qu'ils utilisent, lorsqu'ils les utilisent. Cette option devrait également s'adapter à la demande en ajoutant des quantités illimitées de puissance de calcul dédiée à chaque charge de travail, tandis qu'un nombre illimité de charges de travail fonctionnent simultanément sans incidence sur les performances.

Pour une comparaison plus détaillée de solutions de stockage de données dans le cloud, passer au Chapitre 5.

Pourquoi vous avez besoin d'un datawarehouse dans le cloud

Toute organisation qui dépend des données pour mieux servir ses clients, rationaliser ses opérations et être leader de son industrie trouvera avantage dans un datawarehouse dans le cloud. Contrairement aux énormes datawarehouses traditionnels, le cloud permet aux entreprises, grandes et petites, d'adapter la taille de leur datawarehouse en fonction de leurs besoins et de leur budget, d'augmenter et réduire leur système de manière dynamique au fur et à mesure que les choses changent, de jour en jour et d'année en année.

Voici quelques domaines dans lesquels la technologie de pointe du datawarehouse dans le cloud peut améliorer considérablement les opérations d'une entreprise.

- **Expérience client :** la surveillance du comportement des utilisateurs en temps réel peut aider les organisations à adapter les produits, les services et les offres spéciales en fonction des besoins de clients particuliers. Grâce à l'analyse du ressenti des clients, les entreprises comprennent mieux leurs clients en analysant des quantités massives de messages sur les médias sociaux, de tweets et d'autres activités en ligne.
- **Assurance qualité :** les organisations peuvent également utiliser les données en continu pour surveiller les signes avant-coureurs de problèmes de service client ou de défauts des produits. Elles peuvent agir en quelques minutes ou heures, au lieu de jours ou de semaines, ce qui n'était pas possible lorsque la seule source de données était le journal des plaintes des centres d'appel.
- **Efficacité opérationnelle :** L'*intelligence opérationnelle* (Operational intelligence – OI) consiste à surveiller l'activité et à analyser les évènements afin de déterminer où une organisation peut réduire ses coûts, augmenter ses marges, rationaliser ses processus et répondre plus rapidement aux forces du marché. En déchargeant votre organisation de la gestion d'un datawarehouse, vous pouvez vous concentrer sur l'analyse des données.
- **Innovation :** au lieu de seulement regarder dans le rétroviseur pour comprendre le passé récent d'une industrie, les entreprises peuvent utiliser de nouvelles sources de données et d'analyses de données (prédictives, prescriptives, apprentissage machine) pour repérer et exploiter les tendances, et ainsi perturber leur industrie avant qu'un concurrent inconnu ou imprévu ne le fasse en premier.

RAPPEL

La quasi-totalité des données d'une entreprise est stockée dans une multitude de bases de données disparates. Les principales questions à se poser sont les suivantes : dans quelle mesure ces données sont-elles accessibles ? Combien cela coûtera-t-il de les extraire, les stocker et les analyser ? Que se passera-t-il si vous ne le faites pas ? C'est là qu'entre en jeu le stockage de données dans le cloud.

DANS CE CHAPITRE

- S'adapter aux demandes croissantes d'accès aux données et d'analyse
- S'adapter à la façon dont les données sont créées et utilisées aujourd'hui
- Relever les défis avec de nouvelles et meilleures technologies

Chapitre 2
Apprendre pourquoi le datawarehouse moderne a vu le jour

Le datawarehouse de données dans le cloud est né de la convergence de trois grandes tendances : l'évolution des sources, du volume et de la variété des données ; l'augmentation de la demande d'accès et d'analyse des données ; et les améliorations technologiques qui ont considérablement accru l'efficacité du stockage, de l'accès et de l'analyse des données. Dans ce chapitre, nous décrivons ces tendances de manière plus détaillée et révélons comment un datawarehouse peut tirer profit des avantages du cloud pour y faire face.

Examen des tendances en matière de données : volume, variété et rapidité

Quand on parle de données dans ce livre, on parle de pétaoctets. Un pétaoctet est égal à 1 million de gigaoctets. Cela équivaut à environ 500 milliards de pages de texte imprimé standard ou à 58 333 films en haute définition, chacun d'une durée approximative de deux heures. Les données affluent des opérations quotidiennes d'une entreprise, des personnes qui utilisent des sites web, des applications logicielles sur leurs appareils mobiles et de l'activité quotidienne des appareils numériques et mécaniques.

Dans cette partie, nous nous concentrons sur les changements en matière de données et de leur utilisation, qui ont conduit à la demande du datawarehouse dans le cloud.

Gestion du tsunami de données

Dans un passé pas si lointain, les entreprises géraient généralement des données que des êtres humains saisissaient manuellement dans le système. Elles pouvaient également détenir des données provenant de sources externes, comme de clients, consommateurs ou partenaires. La quantité de données était relativement faible et prévisible et les données étaient stockées, gérées et sécurisées dans un datacenter de l'entreprise, ce qu'on appelle désormais une *méthode « on-premise »cvg*.

Aujourd'hui, le monde des affaires connaît un tsunami de données, disponibles de diverses sources déjà mentionnées dans ce livre et d'autres sources trop nombreuses et variées pour être énumérées. Le volume et la variété de ces données peuvent rapidement submerger un datawarehouse traditionnel on premise, entraînant souvent un blocage de leur traitement et analyse, voire un crash du système, en raison d'une surcharge des utilisateurs et de charges de travail à traiter à tout moment donné.

L'adaptation à l'augmentation exponentielle des données nécessite une nouvelle perspective (voir la figure 2-1). La conversation doit porter non plus sur la taille que doit avoir le datawarehouse d'une organisation, mais sur la question de savoir s'il peut s'étendre de manière rentable, sans friction, et sur l'ordre de grandeur nécessaire pour traiter des volumes énormes de données.

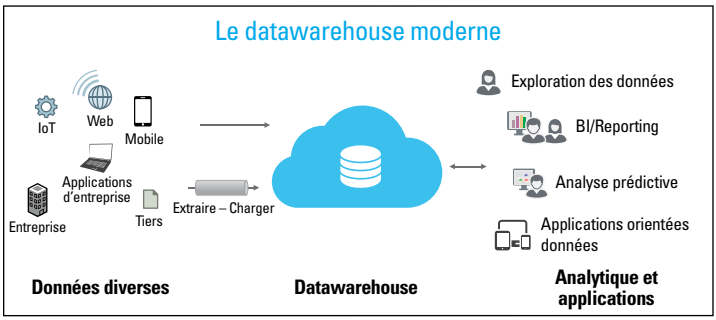

FIGURE 2-1 : Le datawarehouse moderne permet d'analyser toutes les données, par tous les utilisateurs.

RAPPEL

Les cas d'utilisation que le datawarehouse dans le cloud a suscité continuent d'émerger. Par exemple, les entreprises nées du SaaS et les grandes entreprises qui utilisent le cloud pour stocker leurs données monétisent (vendent) ces données. Elles les présentent comme un service et les vendent à d'autres organisations désireuses de prendre des décisions

commerciales encore meilleures en s'appuyant sur les connaissances les plus approfondies possibles.

Bénéfice des données nées dans le cloud

Les organisations ont connu une adoption rapide du SaaS, notamment les logiciels de gestion de la relation client (CRM), les suites logicielles de planification des ressources de l'entreprise (Enterprise Resource Planning – ERP), les plates-formes d'achat de publicité et les outils de marketing en ligne, pour n'en citer que quelques-uns. Grâce au cloud, les nouvelles entreprises de SaaS peuvent s'installer pour le prix d'un ou deux ordinateurs portables. Ces produits SaaS créent d'énormes quantités de données précieuses, stockées dans le cloud. En outre, les organisations réalisent que les fournisseurs de SaaS offrent une meilleure sécurité que ce qui est possible dans leurs propres datacenters on premise.

La demande d'applications SaaS/cloud a également augmenté. La facilité de déploiement n'a rien à voir avec ce qu'exigent les applications on premise pour être opérationnelles. Dans le passé, une entreprise n'exploitait que cinq à dix applications d'entreprise importantes qui généraient des données. Aujourd'hui, il est normal que même les organisations de taille moyenne aient des centaines, voire des milliers d'applications, chacune ayant la possibilité de créer son propre silo de données – données marketing dans un système, finances dans un autre, informations sur les produits dans un autre encore – sans aucune intégration entre elles pour une analyse complète et optimale.

La majorité des données d'une organisation étant désormais dans le cloud, l'endroit naturel pour intégrer ces données est également dans le cloud. Grâce au datawarehouse dans le cloud, vous n'êtes plus obligé de les placer dans votre datacenter : c'est coûteux, cela prend beaucoup de temps et a moins de sens à mesure qu'augmentee la quantité de données natives dans le cloud.

Utilisation de données générées par des machines

Les données générées par des machines sont un sujet clé lié à *l'internet des objets (IoT)* : une collection infinie de périphériques qui communiquent des données via l'internet, y compris les smartphones, les thermostats, les réfrigérateurs, les plates-formes pétrolières, les systèmes de sécurité domestique, les compteurs intelligents, et bien plus encore. Les données recueillies et analysées à partir des périphériques IoT peuvent améliorer les produits et les processus, surveiller les équipements et prévoir la maintenance nécessaire pour éviter les pannes.

Mais beaucoup de données générées par des machines ont un mauvais rapport signal/bruit. Elles contiennent des données précieuses mais aussi beaucoup de bruit. Il faut donc souvent tout stocker pour retrouver les informations utiles. En outre, une part croissante de ces données

provient de l'extérieur de votre datacenter. Cela fait du cloud, grâce à son évolutivité quasi infinie, l'endroit naturel pour stocker et intégrer ces données.

Expérimenter l'exploration des données

L'analyse des données commence par l'exploration des données : en identifiant les liens intéressants et précieux et en les mettant à la disposition des utilisateurs de données sous forme de rapports et d'analyses. Bien que l'exploration des données ne soit pas un concept nouveau, l'augmentation du volume de données en fait un exercice plus exigeant en termes de ressources.

L'exploration des données implique souvent de grands ensembles de données. Elle est également souvent de nature expérimentale, ce qui complique l'évaluation du retour sur investissement nécessaire pour supporter le coût initial important du déploiement d'un datawarehouse traditionnel on premise. En réponse, le cloud peut permettre à un datawarehouse d'évoluer en fonction des besoins et offre un modèle de paiement à l'utilisation qui permet aux organisations d'éviter la question de savoir si elles doivent ou non prendre un engagement initial coûteux.

Introduction aux data lakes

Le besoin croissant de disposer d'énormes quantités de données brutes sous différents formats, le tout en un seul endroit, a donné naissance à ce que l'on considère aujourd'hui comme le data lake (ou lac de données), dit système « legacy ». Les organisations ont rapidement réalisé que ces solutions étaient d'un coût prohibitif, car il était presque impossible de transformer ces données et d'en extraire des informations précieuses.

Mais l'intérêt initial des data lakes a clairement montré que les entreprises voulaient stocker toutes leurs données en un seul endroit à un coût raisonnable. En ajoutant un datawarehouse moderne dans le cloud à votre data lake existant, ou en construisant votre data lake dans le datawarehouse, vous pouvez facilement réaliser cette vision originale du data lake : charger, transformer et analyser de manière rentable des quantités illimitées de données structurées et semi-structurées – avec des ressources de stockage et de calcul quasi illimitées.

Examen des tendances en matière de rapports et d'analyses

La prise de décision basée sur les données n'est plus reléguée à l'équipe de direction ou aux spécialistes des données. Les rapports et analyses sont maintenant utilisés pour améliorer presque tous les aspects opérationnels d'une entreprise. Mais cette demande croissante d'accès aux

données et d'analyse au sein d'une organisation peut ralentir ou faire crasher un système, car les charges de travail se font concurrence pour les ressources de stockage et de calcul des datawarehouses traditionnels. L'efficacité diminue, ce qui oblige les entreprises à investir plus de temps et d'argent dans des infrastructures supplémentaires pour maintenir le système.

Dans cette partie, nous identifions certaines des tendances qui modifient la façon dont on accède aux données et les utilise, et comment ces tendances entraînent le besoin de solutions de datawarehouses modernes, conçus pour le cloud.

L'analyse grâce à l'élasticité

Voici quelques scénarios où le datawarehouse élastique dans le cloud peut permettre de faire plus avec les données :

- L'exploration des données présente de nombreux avantages. Mais personne ne connaît vraiment à l'avance les ressources de calcul nécessaires pour analyser d'énormes ensembles de données, ce qui rend l'extensibilité élastique et à la demande idéale pour ce type d'analyse.
- L'analyse ad hoc des données, qui apparaît tout le temps, répond à une question commerciale unique et spécifique. L'élasticité dynamique et les ressources dédiées pour chaque charge de travail permettent ces requêtes sans ralentir les autres charges de travail.
- L'analyse événementielle exige des données constantes. Elle intègre de nouvelles données pour mettre sans cesse à jour les rapports et les tableaux de bord, afin que les cadres supérieurs puissent suivre l'activité en temps réel ou quasi réel. L'ingestion et le traitement de données en continu nécessitent un datawarehouse élastique pour gérer les variations et les pics de flux de données.

Remplacement de la planification préliminaire exhaustive par une itération rapide

Les entrepreneurs ont généralement deux voies à suivre pour assurer la commercialisation d'un nouveau concept : une planification préliminaire exhaustive ou une itération rapide. La première option est un processus traditionnel, qui prend beaucoup de temps et qui consiste à réfléchir à une opportunité ou à un concept de nouveau produit, à faire du brainstorming et à espérer que cela crée une demande de la part des consommateurs. L'itération rapide consiste à tester rapidement le concept sur le marché afin de l'itérer encore et encore jusqu'à ce qu'une version viable du produit connaisse le succès. A partir de là, le processus recommence.

L'itération rapide est apparue comme le processus le plus efficace pour démanteler des concurrents établis et modifier la façon dont toute une industrie fait des affaires. Mais pour réussir, il faut une collecte et une analyse rapides de grandes quantités de données précises. Les progrès en matière de stockage et d'analyse des données dans le cloud ont rendu l'itération rapide plus pratique, tout en préservant l'exactitude des données.

RÉPONDRE À LA DEMANDE CROISSANTE D'ANALYSE DES DONNÉES

ÉTUDE DE CAS

Jana fournit un accès gratuit et illimité à Internet à plus de 30 millions d'utilisateurs de smartphones dans plus de 15 marchés émergents. Avec son application Android mCent, Jana fait passer le coût de l'Internet mobile des clients à plus de 4 000 marques par le biais de contenus sponsorisés.

Lorsque de nouveaux contenus de marque ou des fonctionnalités mCent sont introduits, Jana analyse et mesure des paramètres clés, notamment l'attention de l'utilisateur, la valeur de l'utilisateur tout au long de sa vie et les indicateurs clés de performance (KPI).

La croissance de Jana et de ses données signifiaient que l'architecture analytique initiale de la société ne pouvait plus servir efficacement ses activités. Les requêtes se sont ralenties et les balayages de tableaux sont devenus impossibles. L'ajout de capacités et de systèmes de sauvegarde, et l'administration du dépôt de données open source de Jana demandaient de plus en plus de temps d'administration.

Comme l'illustre la figure, Jana a mis à niveau la plupart des composants de sa plate-forme de données pour rationaliser son système avec un datawarehouse dans le cloud afin de surmonter ces obstacles et d'obtenir les avantages suivants :

- Suivre le rythme des demandes des entreprises en matière de traitement et d'analyse d'un flux en croissance rapide de données disparates.
- Encourager une utilisation accrue de l'analyse dans toute l'entreprise ; 80 % des employés de Jana ont accès au datawarehouse.
- Réduire significativement les frais administratifs.

La transformation de Jana vers un datawarehouse plus rapide, moins cher et plus efficace.

Intégration de l'analyse

Pour de nombreuses entreprises, l'analyse fonctionne comme un processus commercial séparé et distinct. Mais une tendance croissante consiste à intégrer l'analyse dans les applications commerciales, qui sont de plus en plus souvent établies dans le cloud. Ces applications gèrent une grande variabilité du nombre d'utilisateurs qui les interrogent et du nombre de requêtes (charges de travail) que les utilisateurs exécutent pour analyser ces données. Le cloud facilite les transferts de données des applications dans le cloud vers le datawarehouse dans le cloud de l'organisation, où leur évolutivité et leur élasticité peuvent mieux supporter les fluctuations des utilisateurs et des charges de travail.

La technologie est indispensable à tout datawarehouse moderne

Les innovations technologiques peuvent améliorer le stockage et l'analyse des données en ce qui concerne leur disponibilité, leur simplicité, leur coût et leurs performances. Dans cette section, nous nous concentrons sur les technologies clés qui devraient faire partie de tout datawarehouse moderne.

Le cloud

Les propriétés du cloud le rendent particulièrement adapté au datawarehouse. Nous les avons mentionnées dans d'autres contextes, mais il est important de savoir qu'elles viennent du cloud.

- **Ressources illimitées :** l'infrastructure du cloud fournit des ressources quasi illimitées, à la demande, et en quelques minutes ou secondes. Les organisations ne paient à la seconde que pour ce qu'elles utilisent, ce qui permet de supporter de manière dynamique n'importe quelle quantité d'utilisateurs et de charges de travail sans compromettre les performances.
- **Économiser de l'argent, se concentrer sur les données :** les entreprises qui choisissent une solution dans le cloud évitent les investissements initiaux coûteux en matériel, logiciels et autres infrastructures, ainsi que les coûts de maintenance, de mise à jour et de sécurisation d'un système « on-premise ». Elles se concentrent plutôt sur l'analyse des données.
- **Point d'intégration naturel :** Selon certaines estimations, jusqu'à 80 % des données que vous souhaitez analyser proviennent d'applications extérieures au datacenter de votre entreprise. Rassembler ces données dans le cloud est plus facile et moins cher que d'établir un datacenter interne, car vous n'avez pas à débourser

immédiatement des millions de dollars pour du matériel et des logiciels ni à payer ensuite le personnel technique pour entretenir ces ressources.

Stockage et traitement en colonnes

Comme mentionné précédemment, le stockage en colonnes améliore considérablement l'efficacité et la performance du stockage, de la récupération et de l'analyse des données, permettant aux utilisateurs du système d'accéder plus rapidement aux résultats.

Disques SSD

Contrairement aux lecteurs de disques durs (HDD), les SSD stockent les données sur des puces de mémoire flash, ce qui accélère le stockage, la récupération et l'analyse des données. Ces améliorations augmentent la puissance de calcul des datawarehouses conçus pour utiliser efficacement les SSD.

NoSQL

Le *NoSQL*, abréviation de « *not only structured query language* » (SQL ou language d'interrogation structuré), décrit une technologie qui permet le stockage et l'analyse de nouvelles formes de données, comme celles générées par des machines et des médias sociaux, afin d'enrichir et d'étendre l'analyse des données d'une organisation. Les datawarehouses traditionnels sont mal adaptés pour ces types de données. Par conséquent, de nouvelles approches, comme JSON, Avro et XML, sont apparues ces dernières années pour traiter ces formes de données semi-structurées.

Certains de ces systèmes NoSQL ont été conçus dans l'intention de remplacer les datawarehouses traditionnels mais ont fini par les seulement les compléter. Pour tirer profit de données semi-structurées, les organisations doivent souvent extraire et transformer les données d'un système NoSQL, puis les charger dans un datawarehouse traditionnel pour que les utilisateurs professionnels puissent y accéder facilement. De ce fait, cela ajoute une couche supplémentaire de complexité et de coût pour les entreprises (comme Jana ; voir l'étude de cas précédente) qui tentent de capitaliser sur les avantages des deux types de systèmes.

Par conséquent, le datawarehouse moderne établi dans le cloud doit intégrer et optimiser l'ingestion et l'interrogation de formats de données structurées (traditionnelles) et semi-structurées afin que les organisations évitent de payer et de gérer deux systèmes.

DANS CE CHAPITRE

- Choisir la bonne solution de datawarehouse
- Obtenir un rapport performance/prix élevé
- Donner la priorité à la sécurité, la protection et la gouvernance des données

Chapitre 3
Les critères de sélection d'un datawarehouse moderne

Les tendances évoquées au chapitre 2 ont conduit à la nécessité et à l'opportunité d'un nouveau type de datawarehouse : conçu pour le volume, la variété et la rapidité des données actuelles, et pour les nouvelles façons dont les organisations utilisent leurs données. Une telle solution doit tirer parti des principales innovations technologiques, y compris le cloud.

Si vous êtes à la recherche d'un datawarehouse, une liste de critères vous aidera à déterminer quelle alternative répond le mieux à vos besoins. Considérez ce chapitre comme votre checklist pour trouver la meilleure solution de datawarehouse pour votre organisation.

Répondre aux besoins actuels et futurs

La véritable élasticité a ses avantages commerciaux, mais ce n'est pas tout. Vous devez être en mesure d'adapter les ressources de calcul et de stockage de manière indépendante, pour ne pas devoir ajouter plus de stockage lorsque vous avez simplement besoin de plus de calcul, et vice versa. Ce sont des capacités clés d'un datawarehouse élastique.

Stockage et intégration de toutes les données en un seul endroit

Les données non traditionnelles, ou semi-structurées, comme nous l'avons vu dans les chapitres précédents, peuvent enrichir l'analyse des données au-delà des limites des données traditionnelles. Mais cela nécessite une nouvelle approche pour charger et transformer ces nouveaux types de données avant qu'une organisation puisse en faire l'analyse. La plupart des datawarehouses traditionnels sacrifient la performance ou la flexibilité pour traiter ces types de données. Un datawarehouse moderne doit éliminer la nécessité de concevoir et de modéliser d'emblée des structures traditionnelles rigides qui nécessiteraient la transformation de données semi-structurées avant leur chargement. Il doit également optimiser les performances des requêtes par rapport à ces types de données, tout en gardant leur forme native. Dans l'ensemble, le datawarehouse devrait supporter des données diverses avec souplesse et éviter les problèmes de performance.

Il est essentiel de charger efficacement toutes vos données en un seul endroit. Mais l'intégration de tous ces divers types de données pour une analyse plus précise est autre chose. Un datawarehouse moderne doit automatiquement intégrer vos données semi-structurées, autrefois confinées aux systèmes NoSQL, aux données structurées inhérentes à une base de données relationnelle d'entreprise traditionnelle. Il ne doit rien y avoir à installer et à configurer, et le réglage et les performances doivent être intégrés. Plus important encore, vous n'avez pas à entretenir et à payer deux systèmes distincts pour gérer toutes vos données.

Soutien des compétences, des outils et de l'expertise existants

Les datawarehouses traditionnels sont dépassés uniquement parce que la technologie s'étend sur quatre décennies et n'est pas facilement reconfigurée pour le cloud. Cela signifie également que le langage sur lequel ils s'appuient, le SQL, reste un pilier de l'industrie. C'est pourquoi il existe un large éventail d'outils établis et émergents de gestion, de transformation, d'intégration, de visualisation, de veille économique et d'analyse de données qui communiquent avec un datawarehouse SQL. Le rôle bien établi du SQL standard signifie également qu'un grand nombre de personnes ont des compétences en SQL.

Les datawarehouses traditionnels supportent le langage SQL, mais ne disposent pas des capacités nécessaires pour stocker et traiter efficacement les données semi-structurées. De nombreuses organisations se sont donc tournées vers d'autres approches, comme les solutions NoSQL.

ÉTUDE DE CAS

ANALYSE DE DONNÉES DISPARATES

Chime est une banque plus intelligente pour la génération mobile. Chime recueille et analyse des données sur les plates-formes mobiles, le web et les serveurs back-end afin d'améliorer l'expérience de ses membres tout en apportant de la valeur à son entreprise.

L'analyse des principaux paramètres commerciaux chez Chime a été laborieuse et a nécessité la collecte et l'analyse de données provenant d'un grand nombre de services, notamment les services publicitaires de Facebook et Google. Chime a également tiré des évènements d'autres outils d'analyse tiers, dont la plupart fournissaient des données semi-structurées comme de type JSON.

Chime a satisfait aux exigences suivantes avec son nouveau datawarehouse dans le cloud :

- Fournir efficacement des données structurées et semi-structurées et les rendre disponibles pour des recherches en temps quasi réel, en utilisant des tables de base de données SQL standard.
- Simplifier son pipeline de données sans avoir à concevoir un nouveau modèle pour chaque nouveau type de données chargé dans son datawarehouse.
- Redimensionner à la volée sa capacité pour répondre aux exigences de la charge de travail et contrôler les coûts.
- Intégrer rapidement et facilement des outils d'analyse de données tiers.
- Activer le SQL au lieu d'autres options qui nécessitent des langages de programmation compliqués pour extraire et analyser les données.
- Les analystes de Chime modélisent désormais plus de scénarios pour améliorer les services aux membres, passent moins de temps à attendre les résultats des requêtes et plus de temps à analyser les données.

Les limites de ces systèmes posent un autre problème. Ils demandent des connaissances et des compétences spécialisées qui ne sont pas largement disponibles et pourraient ne pas supporter le SQL. Un datawarehouse moderne doit être conçu avec une technologie de pointe mais s'appuyer sur des standards inclusifs et établis (comme le SQL), et il doit être compatible avec d'autres compétences et outils couramment disponibles dans l'industrie, comme les langages informatiques Spark, Python et R.

Économies d'argent pour votre organisation

Un datawarehouse classique peut coûter des millions de dollars : en frais de licence, matériel et services ; en temps et pour l'expertise nécessaires à la mise en place, à la gestion, au déploiement et au réglage du

datawarehouse, ainsi qu'en coûts de sécurisation et de sauvegarde des données. En outre, l'établissement d'un datawarehouse répondant aux besoins des entreprises et tirant pleinement parti du volume et de la variété des données actuelles est souvent d'un coût prohibitif pour toute organisation.

Un datawarehouse moderne devrait permettre de relever ces défis à un prix beaucoup plus bas. Par exemple, le stockage et le calcul sont-ils dimensionnés séparément pour que vous ne payiez que les ressources dont vous avez besoin ? Est-ce qu'il permet également d'adapter les charges de travail et la simultanéité ? Va-t-il supporter diverses structures de données et intégrer diverses données en un seul endroit ? Connaîtra-t-il des temps d'arrêt minimes ou inexistants et offrira-t-il le choix de fournir des mises à jour automatiques ou par étapes ? Et enfin, peut-il faire tout cela automatiquement sans la complexité, les dépenses et le casse-tête que représentent les réglages manuels du système pour obtenir les meilleures performances ? (Voir le chapitre 5 pour comparer les datawarehouses dans le cloud).

RAPPEL

Avec le stockage de données dans le cloud, vos frais de service devraient tout couvrir pour une petite fraction du coût d'une solution conventionnelle on premise. Mais toutes les solutions basées sur le cloud ne sont pas les mêmes. Leurs différences déterminent également le montant qu'un client doit payer, d'une manière ou d'une autre, pour obtenir des données précieuses.

Résilience et récupération des données

Les datawarehouses peuvent avoir de nombreux types de défaillances qui peuvent entraîner des pertes de données ou des incohérences. Par conséquent, votre datawarehouse doit assurer la sécurité, la mise à jour et la disponibilité de vos données. Les datawarehouses traditionnels protègent généralement les données en effectuant des sauvegardes périodiques, qui consomment de précieuses ressources informatiques et interfèrent avec les charges de travail en cours. Les sauvegardes périodiques nécessitent également un stockage supplémentaire et n'incluent souvent pas les données les plus récentes, ce qui entraîne des incohérences dans les données.

Un datawarehouse moderne doit s'autogérer lorsqu'il s'agit de garantir la durabilité, la résilience et la disponibilité du système. Il ne doit pas interférer avec les charges de travail en cours, ni dégrader les performances, ni entraîner une indisponibilité du service en raison de processus de sauvegarde exécutés en arrière-plan. Et il devrait être bon marché, doté de moyens astucieux de préserver vos données sans avoir à les copier et les transférer ailleurs. Enfin, une architecture multi-cloud vous permet de transférer les données et les charges de travail à mesure que votre

entreprise se développe, à la fois entre les régions géographiques et entre les principaux fournisseurs cloud, tels qu'Amazon, Microsoft et Google.

Sécurité des données au repos et en transit

La sécurité des données couvre les deux principaux domaines suivants :

- **Confidentialité :** empêcher l'accès non autorisé aux données
- **Intégrité :** s'assurer que les données ne sont pas modifiées ou corrompues, qu'elles sont correctement gérées et que leur qualité est préservée.

Un datawarehouse moderne doit également supporter le *contrôle d'accès basé sur les rôles (role-based access control – RBAC)* à plusieurs niveaux. Cela garantit que les utilisateurs n'ont accès qu'aux données qu'ils sont autorisés à voir. Pour une meilleure sécurité, exiger une authentification multifactorielle *(multi-factor authentication ou MF)*. Avec l'AMF, lorsqu'un utilisateur se connecte, le système envoie une demande de vérification secondaire, souvent à un téléphone portable. Le code envoyé au téléphone doit ensuite être saisi. Ainsi, l'accès au système est impossible à une personne non autorisée qui a volé un nom d'utilisateur et un mot de passe.

La gouvernance des données garantit que les données d'une entreprise sont correctement accessibles et utilisées, et que toutes les données sont gérées et sauvegardées afin de contrecarrer les infractions et de se conformer à des règlements détaillés. Une surveillance rigoureuse est également impérative pour maintenir la qualité des données que votre entreprise partage avec ses constituants. De mauvaises données peuvent conduire à des décisions d'affaires manquées ou inadéquates, à une perte de revenus et à une augmentation des coûts. Les gestionnaires de données – chargés de superviser la qualité des données – peuvent identifier quand les données sont corrompues ou inexactes, quand elles ne sont pas rafraîchies assez souvent pour être pertinentes, ou quand elles sont analysées hors contexte.

Le cryptage des données, qui consiste à appliquer un algorithme d'encodage pour traduire le texte en clair en texte chiffré, est une autre caractéristique de sécurité requise. Une grande partie de la solution est la gestion des clés. Une fois que vous aurez crypté vos données, vous utiliserez une clé de cryptage pour les décrypter. En plus de protéger les données, vous devez protéger la clé qui décode les données. Combien de temps utilisez-vous la même clé ? Que se passe-t-il si la clé est compromise ? Tout cela doit être géré. Le datawarehouse devrait utiliser une approche hiérarchique d'encapsulation des clés, qui permet de crypter les clés de cryptage, ainsi qu'un processus robuste de rotation des clés, qui limite le nombre de fois où une seule clé est utilisée.

En outre, avec un datawarehouse dans le cloud moderne, le prestataire de solutions doit effectuer des tests de sécurité périodiques, appelés *tests de pénétration*, pour vérifier de manière proactive les vulnérabilités. Le fournisseur doit administrer ces mesures de manière cohérente et automatique sans que cela n'ait d'incidence sur les performances.

Pour une discussion complète sur la sécurité et la gouvernance des datawarehouses dans le cloud, voir le chapitre 8.

RAPPEL

Choisissez un datawarehouse avec une sécurité de bout en bout conforme aux normes de l'industrie. Trouvez une solution qui a passé avec succès les audits de sécurité tels que SOC 1/SOC 2 Type II et ISO/IEC 27001.

Rationalisation du pipeline de données

Le *pipeline de données* se réfère principalement aux processus *d'extraction, de transformation et de chargement (extract, transform, and load – ETL)* qui importent des données dans le datawarehouse et sous un format qui supporte les requêtes. Un pipeline de données lent oblige les utilisateurs, tels que les analystes, à à attendre trop longtemps pour accéder aux données. La croissance rapide de la diversité, du nombre et de la taille des données non relationnelles provenant de sources multiples aggrave le problème.

Un datawarehouse moderne devrait réduire la complexité globale du processus afin de faire circuler les données plus rapidement dans le pipeline de données. Les solutions modernes devraient être capables de charger efficacement des données semi-structurées sous leur format natif et de les rendre immédiatement disponibles pour une interrogation sans avoir besoin de systèmes supplémentaires et complexes, tels que le NoSQL, pour transformer les données. Cela permet aux utilisateurs d'accéder immédiatement aux données de la même manière qu'ils interrogent une base de données SQL. Ces solutions peuvent permettre d'accéder aux nouvelles données de manière exponentiellement plus rapide, en réduisant le processus d'ingestion et de transformation d'une journée à moins d'une heure.

Optimisation de votre valeur temps

Le déploiement d'une solution ne devrait pas être une entreprise majeure, et les aspects cruciaux qui étaient autrefois manuels devraient être automatisés. Surtout, tous les utilisateurs doivent pouvoir accéder à tout moment à la solution de votre choix, qui doit englober tous les types de données à une fraction du coût des systèmes traditionnels. Un tel système devrait fournir un aperçu immédiat des données pour aider à rationaliser une organisation et accroître sa capacité à servir ses clients et à diriger son industrie.

DANS CE CHAPITRE

- Compression de l'écart entre le temps et la valeur
- Réduire les coûts de stockage et de calcul
- Profiter de l'élasticité dynamique
- Externalisation de l'administration et de la sécurité

Chapitre 4
Le Datawarehouse on-premise ou dans le cloud

Si vous êtes à la recherche d'un nouveau datawarehouse, le premier choix à faire est celui du lieu d'implantation de votre datawarehouse : dans le datacenter de votre organisation ou dans le cloud et fourni sous forme de logiciel en tant que service (SaaS). Le stockage traditionnel de données on premise est une technologie éprouvée et bien établie, conçue bien avant que le cloud ne devienne une plate-forme viable. L'adoption rapide du cloud a créé un besoin de solutions de datawarehouses qui peuvent tirer pleinement profit de ce qu'il apporte. Dans ce chapitre, nous présentons les éléments clés relatifs au stockage de données dans le cloud en le comparant aux systèmes traditionnels on premise.

Évaluation de la valeur temps

Le déploiement d'un datawarehouse traditionnel (voir chapitre 3)) peut prendre au moins un an et s'étendre à un projet sur plusieurs années avant que vous ne puissiez extraire des informations de vos données. L'agilité des entreprises aujourd'hui signifie que les principaux acteurs qui soutiennent le projet, et les principaux facilitateurs commerciaux et techniques responsables de la réussite du projet, peuvent quitter l'équipe ou l'entreprise avant que le projet ne soit opérationnel. Un cycle aussi long expose également le projet à des ralentissements économiques, à des manques à gagner pour l'entreprise et au risque de ne jamais mettre en œuvre le projet en raison d'une dérive des objectifs.

De plus, les solutions on premise ne sont pas adaptées au traitement des données semi-structurées d'aujourd'hui. Il faut une plate-forme NoSQL open source en plus, qui ajoute une autre couche de complexité et allonge la phase de mise en œuvre d'un nouveau datawarehouse.

Bien fait, un datawarehouse dans le cloud peut être opérationnel en quelques semaines ou quelques mois seulement. Par conséquent, la majeure partie du temps nécessaire à la mise en route devrait être consacrée à l'extraction des données de vos autres sources de données et à la configuration d'un outil d'analyse frontal pour extraire des informations du datawarehouse.

Prise en compte des coûts de stockage et de calcul

Les datawarehouses on premise sont coûteux en termes de matériel, de logiciels et d'administration. Les coûts du matériel peuvent comprendre les coûts des serveurs, des périphériques de stockage supplémentaires, de l'espace dans le datacenter pour abriter le matériel, d'un réseau à haut débit pour accéder aux données, et de l'alimentation électrique et des blocs d'alimentation redondants nécessaires pour maintenir le système en état de marche. Si votre datawarehouse est essentiel à votre mission, ajoutez les coûts de configuration d'un site de reprise après sinistre. Les organisations paient aussi fréquemment des centaines de milliers de dollars en frais de licence pour les logiciels de datawarehouse et les progiciels complémentaires. Les utilisateurs finaux supplémentaires, y compris les clients et les fournisseurs qui ont accès au datawarehouse, peuvent augmenter considérablement ces coûts. Ajoutez ensuite le coût permanent des contrats de support annuels, qui représentent souvent 20 % du coût initial de la licence. De plus, un datawarehouse on premise a besoin de personnel spécialisé dans les *technologies de l'information* (IT) pour déployer et maintenir le système. Cela peut créer un goulot d'étranglement lorsque des problèmes surviennent et fait assumer la responsabilité du système au client, et non au fournisseur.

Un datawarehouse dans le cloud remplace le CapEx initial et le coût permanent d'un système on premise par une simple tarification basée sur l'utilisation de l'OpEx. Vous payez des frais mensuels basés sur la quantité de stockage et de ressources informatiques que vous utilisez réellement. D'un point de vue conservateur, le coût annualisé d'une solution de datawarehouse dans le cloud peut être dix fois moins élevé que celui d'un système similaire on premise.

Dimensionnement, équilibrage et réglage

Pour obtenir des performances optimales, un datawarehouse on premise doit être modélisé, dimensionné, équilibré et réglé, ce qui nécessite un investissement initial important ainsi que des coûts de surveillance et d'administration permanents. Dans une telle configuration, entrent souvent en jeu :

- le nombre et la vitesse des unités centrales de traitement (CPU)
- la quantité de mémoire
- le nombre et la taille des disques pour la capacité de stockage requise
- la *bande passante d'entrée/sortie (I/O)* (une mesure de la quantité de données pouvant être transférées à un moment donné)
- un modèle de données personnalisé définissant la structure du datawarehouse, les types de données inclus et la fréquence de mise à jour

Avec un datawarehouse on premise, les organisations dimensionnent souvent leur système en fonction des pics d'utilisation, qui peuvent ne représenter qu'une petite période de l'année. Par exemple, une entreprise peut n'avoir besoin de la pleine puissance du datawarehouse qu'à la fin de chaque trimestre ou exercice comptable. Mais elle doit payer pour cette capacité de pointe 24 heures sur 24, 7 jours sur 7, pour contrecarrer le manque d'évolutivité du système.

Le stockage élastique de données dans le cloud offre deux avantages clés :

- Les complexités et le coût de la planification et de l'administration des capacités – dimensionnement, équilibrage et réglage du système – devraient être intégrés dans le système, automatisés et couverts par vos frais d'abonnement.
- Il en va de même pour l'approvisionnement dynamique des ressources de stockage et de calcul à la volée afin de répondre aux exigences fluctuantes de vos charges de travail en période de pointe et d'utilisation régulière. La capacité, c'est avoir tout ce dont on a besoin quand on en a besoin. Mais toutes les charges de travail ne sont pas égales. Un datawarehouse élastique dans le cloud vous donne des informations très précises pour savoir quelles ressources sont affectées à quel utilisateur et à quelles charges de travail.

Prise en compte des coûts de préparation des données et d'ETL (extraction, transformation, chargement)

Un datawarehouse on premise doit extraire des données de toutes vos sources. Il doit ensuite transformer ces données pour qu'elles adhèrent à la structure souvent rigide des données à l'intérieur du système *avant* de les charger dans le datawarehouse. L'un des principaux défis consiste à respecter une quantité limitée et coûteuse de capacité de traitement et de stockage. Par conséquent, la transformation des données doit se faire en dehors des heures de travail normales pour éviter de concurrencer d'autres travaux de traitement des données. Cela coûte cher. En outre, les données semi-structurées ne sont pas livrées dans les rangées et les colonnes cohérentes inhérentes aux structures de données traditionnelles. Il s'agit également de données de grand volume et de grande vitesse.

Les meilleures solutions établies dans le cloud peuvent charger directement des données semi-structurées sans les transformer. Ces solutions peuvent permettre d'accéder à des données fraîches jusqu'à 50 fois plus rapidement qu'un datawarehouse traditionnel. En outre, le coût moins élevé du stockage illimité dans le cloud permet aux analystes d'accéder à toutes les données au lieu d'être limités à des agrégats périodiques de ces données.

ÉTUDE DE CAS

L'OPTIMISATION D'UN PIPELINE DE DONNÉES

DoubleDown, un studio de jeux en ligne, a ajouté un système NoSQL à son pipeline de données pour préparer les données à charger dans son datawarehouse. Mais cette approche signifiait que le traitement du fichier journal quotidien des évènements de DoubleDown (clics des utilisateurs et autres données générées par les activités des joueurs) prenait beaucoup de temps. L'entreprise ne pouvait pas accéder aux données d'une journée avant 15 heures le lendemain. Pire encore, si l'un de ses clusters informatiques tombait en panne, l'entreprise perdait des données.

DoubleDown a choisi un système qui pouvait charger directement ses données semi-structurées sans les transformer au préalable, rendant ces données immédiatement disponibles pour des requêtes. Cela a permis d'améliorer la qualité et les performances de son pipeline de données en transmettant les données aux analystes près de 100 fois plus rapidement : en 15 minutes contre 24 heures, en éliminant presque toutes les défaillances fréquentes du précédent pipeline de la société, en fournissant aux analystes une granularité complète des données au lieu d'agrégats périodiques, et en réduisant le coût du pipeline de données de DoubleDown de 80 %.

Les analystes de DoubleDown ont désormais un accès immédiat aux données des nouvelles versions de produits pour prendre des décisions plus rapides et fondées sur les données.

Coût supplémentaire d'outils d'analyse commerciale spécialisés

Comme mentionné au chapitre 3, les datawarehouses traditionnels on premise ne sont pas conçus pour traiter le volume, la variété et la vitesse des données actuelles. En conséquence, les organisations exploitent deux plates-formes de données : un datawarehouse SQL d'entreprise on premise pour le stockage des données relationnelles traditionnelles, et une plate-forme Big Data pour les données NoSQL, qui peut s'exécuter on premise ou dans le cloud, pour le stockage des données non relationnelles.

Malheureusement, ces nouveaux systèmes sont très complexes à gérer et nécessitent des outils et une expertise spécialisés qui ne sont pas aussi répandus que les outils et l'expertise SQL. Après tout, le SQL existe depuis des décennies, alors que les systèmes NoSQL sont relativement nouveaux.

La solution idéale de stockage de données dans le cloud offre le meilleur des deux mondes : la flexibilité d'intégrer des données relationnelles et non relationnelles, ainsi que le support pour les outils et les compétences SQL facilement disponibles pour interroger ces données.

CONSEIL

Si vous êtes à la recherche d'un nouveau datawarehouse, prenez en compte le coût et la disponibilité des compétences et de l'expertise requises pour gérer le datawarehouse, ainsi que les nombreux outils d'analyse ou autres utilisés conjointement avec un datawarehouse.

Prise en compte de l'évolutivité et de l'élasticité

Les datawarehouses traditionnels sont sujets à des ralentissements et à des plantages du système, car les utilisateurs et les processus sont en concurrence pour des ressources limitées. Ces systèmes relient étroitement le stockage et le calcul sur une seul *cluster* d'ordinateurs (un groupe d'ordinateurs), ce qui fait qu'il est coûteux d'augmenter l'un sans augmenter l'autre.

Les solutions de datawarehouse plus récentes, établies dans le cloud, offrent un stockage et un calcul virtuellement illimités ; toutefois, envisagez un datawarehouse où le stockage est dimensionné séparément du calcul (voir la figure 4-1). Idéalement, l'évolutivité du datawarehouse dans le cloud devrait être accomplie en trois volets :

- **Stockage :** le stockage dans le cloud est par nature évolutif, il s'ajuste facilement en fonction de la quantité de stockage et des besoins.
- **Calcul :** les ressources utilisées pour le traitement des charges de données et des requêtes doivent pouvoir être facilement augmentées ou réduites, à tout moment, en fonction de l'évolution du nombre et de l'intensité des charges de travail.
- **Utilisateurs et charges de travail (simultanéité) :** les solutions avec des ressources informatiques fixes ralentissent à mesure que les utilisateurs et les charges de travail augmentent. Les organisations sont souvent obligées de répliquer les données dans des datamarts séparés, de transférer certaines charges de travail en dehors des heures de travail normales et de mettre les utilisateurs en file d'attente pour préserver les performances. Seul le cloud peut permettre à un datawarehouse d'évoluer en ajoutant des clusters de calcul dédiés de n'importe quelle taille à un nombre presque infini d'utilisateurs ou de charges de travail qui accèdent tous à une seule copie des données, mais sans impact sur les performances des autres.

Cherchez une solution de cloud qui découple le stockage du calcul, afin que les deux puissent s'adapter facilement et indépendamment l'un de

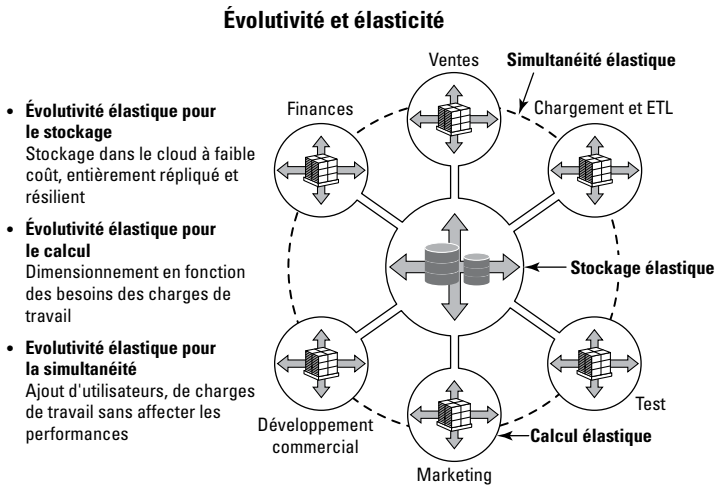

FIGURE 4-1 : Le datawarehouse idéal devrait évoluer de trois façons.

l'autre pour maintenir les coûts à un faible niveau. La solution devrait également évoluer, ou s'étendre horizontalement, pour supporter davantage d'utilisateurs et de charges de travail sans nuire aux performances.

Diminution des retards et des temps d'arrêt

Les entreprises qui disposent de solutions on premise sont nombreuses à se plaindre de deux choses. Elles doivent attendre des heures ou plus d'une journée avant que les données collectées la veille soient dans le datawarehouse et disponibles. Elles doivent attendre le même temps pour qu'une requête complexe soit exécutée sur un vaste ensemble de données. Dans certains cas, de multiples processus simultanés peuvent geler ou faire crasher le système, ce qui prolonge les délais et les temps d'arrêt.

Avec des ressources de stockage et de calcul virtuellement illimitées, les solutions de datawarehouse dans le cloud, conçues pour être dynamiquement élastiques, sont mieux équipées pour s'adapter aux demandes croissantes. Toutefois, pour réduire les retards et éliminer les temps d'arrêt imprévus, il ne suffit pas d'augmenter les ressources du système. De meilleures solutions permettent de rationaliser le pipeline de données et de stocker les données pour que les requêtes s'exécutent plus efficacement sans réglage manuel.

Cherchez des solutions qui répondent à tous ces types de problèmes de performance et qui minimiseront les temps d'arrêt. La rapidité avec laquelle vous pouvez accéder à vos données et à vos analyses peut affecter considérablement vos opérations et votre capacité à maintenir un avantage concurrentiel.

Prise en compte des coûts liés aux questions de sécurité

Une seule violation peut rapidement se transformer en un cauchemar de relations publiques et entraîner des pertes commerciales et des amendes élevées de la part des organismes de réglementation. Bien que le cloud suscite la crainte de risques de sécurité, il peut être plus sûr que votre datacenter.

Si vous optez pour un datawarehouse on premise, vous êtes seul responsable de la sécurisation des données sensibles, ce qui implique une attention minutieuse et constante à la protection par pare-feu ; aux protocoles de sécurité ; au cryptage des données, au repos et en transit ; aux rôles et privilèges des utilisateurs, ainsi qu'à la surveillance et à l'adaptation aux nouvelles menaces de sécurité.

Une sécurité des données efficace est complexe et coûteuse à mettre en œuvre, notamment en termes de ressources humaines. Des mesures de sécurité mal appliquées vous exposent à des coûts encore plus élevés en cas de violation.

Comme les fournisseurs de datawarehouses dans le cloud desservent un certain nombre de clients, ils sont en mesure d'offrir l'expertise et les ressources nécessaires pour assurer une sécurité de bout en bout des datawarehouses de qualité industrielle. Cherchez un fournisseur qui assure un cryptage de bout en bout conforme aux normes du secteur pour sécuriser les données au repos et en transit.

Le prix de la protection et de la récupération des données

Les datawarehouses on premise sont vulnérables aux pertes de données dues aux pannes d'équipement, aux coupures ou surtensions électriques, au vol ou au vandalisme, et aux catastrophes (incendie, inondation, tremblement de terre, etc.). Pour protéger vos données, vous devez les sauvegarder régulièrement et stocker les sauvegardes dans un endroit distant. Une alimentation électrique de secours est également nécessaire pour prévenir la perte de données et garantir que votre datawarehouse est toujours disponible pour traiter les données et les requêtes entrantes. En cas de catastrophe, vous aurez besoin de personnel qualifié pour récupérer les données, en utilisant les sauvegardes les plus récentes. Si votre datawarehouse est essentiel à votre mission, vous pouvez également avoir besoin d'un site de reprise après sinistre géographiquement séparé (un datacenter supplémentaire) ainsi que des logiciels, licences et processus pour assurer un basculement automatique afin qu'il n'y ait aucune interruption de service.

Le cloud offre une solution idéale pour la protection et la récupération des données. De par sa nature, il stocke les données hors des locaux. Certaines solutions basées sur le cloud sauvegardent automatiquement les données vers deux ou plusieurs lieux physiques distincts. Si les datacenters sont géographiquement isolés, ils offrent également une reprise après sinistre intégrée. Les datacenters sur le cloud disposent d'une alimentation électrique redondante, de sorte qu'ils restent opérationnels même en cas de longues coupures de courant. Les fournisseurs de services sur le cloud peuvent offrir ces protections à un coût beaucoup plus faible que vous, en répartissant le coût sur des milliers de clients.

CONSEIL

Si vous ne souhaitez pas gérer vos propres sauvegardes de données, demandez bien à votre fournisseur potentiel de datawarehouse sur le cloud comment il configure son service. De même, si vous avez besoin d'une protection de reprise après sinistre, confirmez que l'architecture du fournisseur utilise des centres géographiquement séparés. Demandez également si votre fournisseur offre sa solution à travers plusieurs fournisseurs de cloud au cas où une catastrophe vous obligerait à passer à une instance de votre datawarehouse dans un autre cloud.

DANS CE CHAPITRE

- Considération des facteurs qui affectent les performances
- Choix d'une solution qui assure la protection et la sécurité des données
- Évaluation des économies réalisées sur les frais administratifs

Chapitre 5
Comparaison des solutions de datawarehouses dans le cloud

L'adoption croissante du cloud a incité les fournisseurs on premise et les nouveaux venus sur le marché à proposer des versions cloud de leurs produits de stockage de données. Bien sûr, il n'y a pas deux solutions identiques. Dans ce chapitre, nous expliquons certaines des différences et ce qu'il faut rechercher parmi les datawarehouses dans le cloud.

Explication des approches du datawarehouse dans le cloud

Les approches du cloud suivantes offrent des capacités de stockage de données très différentes :

- **Infrastructure en tant que service (Infrastructure-as-a-service – IaaS) :** le client est obligé d'installer un logiciel de datawarehouse traditionnel sur les ordinateurs fournis par le prestataire de la plate-forme dans le cloud. Le client gère tous les aspects du matériel du cloud et du logiciel de datawarehouse. Les capacités du datawarehouse sont identiques à celles du logiciel utilisé pour le matériel on premise.
- **Plate-forme en tant que service (Platform-as-a-service – PaaS) :** Avec cette approche hybride, le fournisseur du datawarehouse fournit

le matériel et les logiciels en tant que service dans le cloud, et il gère le déploiement du matériel, l'installation et la configuration des logiciels. Le client gère, règle et optimise les logiciels.

- **Logiciel en tant que service (Software-as-a-service – SaaS) :** Le fournisseur du datawarehouse fournit tout le matériel et les logiciels, y compris tous les aspects de la gestion du matériel et des logiciels. Généralement inclus dans le service : mises à jour des logiciels et du matériel, sécurité, disponibilité, protection des données et optimisation.

Dans tous ces scénarios, la tâche d'acheter, de déployer et de configurer l'espace du datacenter et le matériel pour supporter le datawarehouse passe du client au fournisseur. Au-delà de cet avantage, les avantages et les inconvénients des différentes offres varient depuis la facilité d'utilisation jusqu'à la sécurité et à la disponibilité.

RAPPEL

Si un fournisseur de datawarehouse se contente de fournir un accès à son datawarehouse traditionnel via le cloud, la solution ressemblera probablement à son architecture et à ses fonctionnalités d'origine, on premise.

Comparaison des architectures

De nombreux fournisseurs proposent un datawarehouse dans le cloud conçu et déployé à l'origine pour les environnements on premise. Ces architectures traditionnelles ont été créées bien avant que le cloud et ses avantages n'apparaissent comme une option viable. Par ailleurs, toute solution de datawarehouse établie pour le cloud devrait capitaliser sur les avantages du cloud (voir figure 5-1).. Pour identifier une solution bâtie sur une architecture optimisée pour le cloud, recherchez les caractéristiques suivantes :

- Stockage centralisé de toutes les données
- Mise à l'échelle indépendante des ressources de calcul et de stockage
- Simultanéité quasi illimitée sans concurrence pour les ressources
- Chargement et interrogation simultanés des données sans dégradation des performances
- Reproduction des données à travers plusieurs régions et clouds pour améliorer la continuité des activités et simplifier l'expansion
- Partage de données sans mise en place d'API ou de procédures lourdes d'ETL
- Un service de *métadonnées* robuste qui s'applique à l'ensemble du système. (Les métadonnées sont des données concernant d'autres données, telles que la taille du fichier, l'auteur et la date de création). Une

architecture optimisée pour le cloud tire également profit du stockage en tant que service, où la taille du stockage des données s'adapte automatiquement et de manière transparente pour l'utilisateur. Le stockage de données conçu pour des architectures plus anciennes est coûteux et sa capacité d'évolution est limitée.

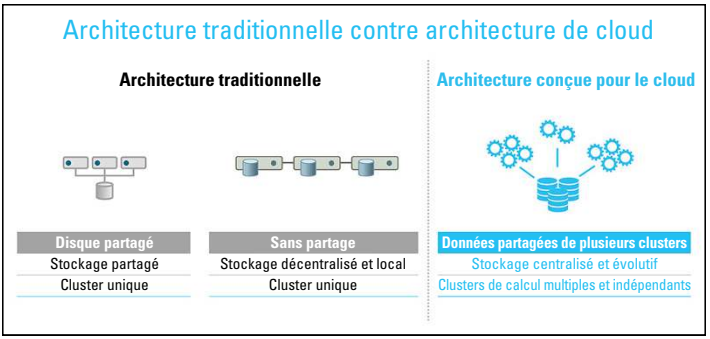

FIGURE 5-1: Comment une architecture optimisée pour le cloud rationalise les performances.

Évaluation de la gestion de la diversité des données

L'un des facteurs clés de l'adoption du stockage de données dans le cloud découle du volume croissant de données qui proviennent du cloud – en dehors du datacenter d'une entreprise. Dans la plupart des cas, ces données non relationnelles doivent être transformées avant d'être chargées dans un datawarehouse traditionnel sur place ou dans le cloud. Cette approche ajoute une complexité et des délais importants à l'accès aux nouvelles données.

Avec ce plus grand volume et cette plus grande variété de données, le cloud est devenu un point d'intégration naturel. Un moyen idéal de résoudre ce problème est de disposer d'un datawarehouse dans le cloud qui peut traiter à la fois les données relationnelles et non relationnelles, et ce, sans avoir à transformer les données non relationnelles ou à compromettre les performances pendant le chargement des données ou le traitement des requêtes.


RAPPEL

Les données doivent être transformées avant d'être chargées dans un datawarehouse traditionnel basé dans le cloud. Sinon, l'organisation doit acheter et maintenir un système supplémentaire pour traiter les données non relationnelles.

Estimation de l'évolutivité et de l'élasticité

Tous les datawarehouses dans le cloud ne présentent pas le même type d'élasticité. Les solutions avancées peuvent s'adapter, à la volée, et sans mettre le système hors ligne ou en mode lecture seule.

CONSEIL

Réfléchissez aux inconvénients des solutions qui ne permettent pas une évolution facile :

- Un datawarehouse dans le cloud qui nécessite une reconfiguration manuelle implique une planification et une coordination minutieuses avec le fournisseur afin de dimensionner les ressources.
- Un redimensionnement peut nécessiter un temps d'arrêt ou un passage en mode lecture seule pour redistribuer les données et reconfigurer le système.
- La plupart des offres de datawarehouses dans le cloud regroupent le calcul et le stockage sur le même nœud, ce qui oblige les clients à adapter les deux lorsqu'ils ont besoin d'augmenter seulement l'un ou l'autre.
- La plupart sont des versions adaptées au cloud de solutions on premise, vous devrez donc acheter une configuration surdimensionnée mais sous-utilisée pour les pics d'utilisation. Vous finirez par dépasser les ressources disponibles et vous serez confronté à des mises à niveau coûteuses.

Comparaison des capacités de simultanéité

La *simultanéité* est la capacité d'effectuer deux ou plusieurs tâches simultanément ou de permettre à deux utilisateurs ou plus d'accéder à une solution informatique. Dans un datawarehouse traditionnel, les ressources fixes de calcul et de stockage limitent la simultanéité. Avec le cloud, cependant, le calcul et le stockage ne sont pas fixes. Les architectures optimisées pour le cloud favorisent la simultanéité de deux façons :

- Plusieurs utilisateurs peuvent interroger les mêmes données simultanément sans compromettre la performance.
- Le chargement et l'interrogation peuvent se faire simultanément, ce qui permet des charges de travail multiples et simultanées sans conflit de ressources.

Garantie d'un support de SQL et d'autres outils

Presque tous les outils d'informations commerciales (business intelligence - BI), d'extraction, de transformation et de chargement (ETL) et d'analyse des données peuvent communiquer avec un datawarehouse

qui utilise le langage SQL standard. Cependant, les solutions de stockage de données dans le cloud ne supportent pas toutes le SQL standard. Par exemple, les grandes solutions de données positionnées comme datawarehouses dans le cloud sont souvent des solutions NoSQL et n'ont qu'un support de SQL incomplet ou non standard. Bien qu'il soit important de supporter ces nouveaux outils analytiques, le SQL reste la norme de l'industrie pour l'interrogation des données. Votre datawarehouse doit prendre en charge les outils SQL pour la gestion des données, leur transformation, leur intégration, la visualisation, les informations commerciales et d'autres types d'analyse.

Vérification des dispositifs de sauvegarde/récupération

Avec beaucoup de solutions de stockage des données on premise et dans le cloud, les clients doivent protéger leurs propres données à l'aide d'outils de sauvegarde et de réplication des données. Toutefois, certaines solutions de datawarehouse dans le cloud incluent la protection des données dans le cadre du service.

RAPPEL

Pour une protection optimale, recherchez une solution qui sauvegarde automatiquement les versions antérieures des données ou qui les duplique automatiquement en tant que moyen de sauvegarde en ligne. La solution devrait également permettre la récupération en libre-service des données perdues ou corrompues par le biais d'une réplication entre régions au sein du même fournisseur de cloud ou entre plusieurs fournisseurs de cloud pour une parfaite continuité des activités.

Confirmation de la résilience et de la disponibilité

La *résilience* est la capacité du datawarehouse à continuer à fonctionner automatiquement en cas de défaillance d'un composant, d'un réseau ou même d'un datacenter. La *disponibilité* (appelée temps de service) est la capacité qu'ont les utilisateurs à accéder au système à tout moment. Les services de datawarehouse dans le cloud varient en fonction de la responsabilité du client en matière de disponibilité et de résilience. Au niveau le plus élémentaire, un service de datawarehouse dans le cloud peut exiger du client qu'il assure la surveillance du système pour détecter et éventuellement prévenir une défaillance. Le client peut également avoir à gérer la réplication des données, de sorte qu'un double du datawarehouse soit disponible en cas de défaillance. À l'autre extrémité du spectre, le fournisseur assure la surveillance, la réplication et le basculement automatique dans le cadre du service.

La disponibilité est également un facteur pour les mises à jour des logiciels. Différents fournisseurs adoptent des approches différentes pendant la mise à niveau :

- **basique :** Les clients gèrent les mises à niveau et les temps d'arrêt qui en découlent.
- **améliorée :** Le fournisseur gère les mises à niveau et informe les utilisateurs des mises à niveau à venir, afin qu'ils puissent planifier les temps d'arrêt.
- **optimale :** Le fournisseur fournit des mises à niveau transparentes sans impliquer les utilisateurs ni les soumettre à des temps d'arrêt. Le fournisseur permet également aux clients d'opter pour ou contre les mises à niveau automatiques, afin qu'ils puissent les recevoir quand ils le souhaitent.

CONSEIL

Vérifiez le nombre de 9 pour décrire la disponibilité que supporte la solution de datawarehouse dans le cloud (99,9XX % de temps de service).

Optimisation des performances

L'une des grandes promesses du cloud est la possibilité de disposer d'énormes quantités de ressources à ne payer que lorsque vous en avez besoin. Cherchez une solution de datawarehouse dans le cloud qui puisse optimiser les performances à la demande et qui élimine les efforts administratifs pour intégrer de nouvelles ressources.

RAPPEL

Évitez les datawarehouses qui perturbent ou retardent l'activité pour ajouter ou soustraire des ressources. Certaines solutions nécessitent également un travail administratif, notamment la redistribution des données et le recalcul des métadonnées.

Évaluation de la sécurité des données dans le cloud

Le cloud est souvent perçu comme moins sûr que le stockage des données on premise, mais les solutions de cloud sont de plus en plus acceptées en raison des effractions dans les datacenters sécurisés on premise. Ces incidents révèlent que les entreprises sont limitées dans leur capacité à sécuriser leurs propres données. Les offres de stockage de données dans le cloud confient la responsabilité de la sécurité physique du datacenter au fournisseur de la solution, mais attention : les dispositifs de sécurité varient selon les fournisseurs.

- L'offre de base des datawarehouses dans le cloud ne fournit que certaines capacités de sécurité, laissant au client le soin de s'occuper de choses comme le cryptage, le contrôle d'accès et la surveillance de la sécurité.
- D'autres solutions offrent des fonctionnalités comme le cryptage et les contrôles d'accès, que les clients peuvent choisir d'activer, mais elles laissent le système vulnérable s'il n'est pas activé.

- Les datawarehouses dans le cloud qui sont davantage axés sur les services intègrent des fonctions de sécurité et fournissent dans le cadre du service le cryptage, la gestion des clés de cryptage, la rotation des clés, la détection des intrusions, et plus encore.

Prise en compte de l'administration

Les datawarehouses traditionnels nécessitent une quantité importante de temps, d'efforts et d'expertise de la part du client. Un ou plusieurs administrateurs de base de données (database administrators – DBA) doivent effectuer des correctifs et des mises à jour de logiciels, le partitionnement et la répartition des données, la gestion des index, la gestion des charges de travail, la mise à jour des statistiques, la gestion et la surveillance de la sécurité, les sauvegardes et la réplication, le réglage et la réécriture des requêtes, etc.

Au niveau basique, une solution de datawarehouse dans le cloud qui repose sur une technologie plus ancienne, on premise, exige encore du client qu'il gère tous ces aspects. Les nouvelles offres de stockage de données réduisent ou éliminent une grande partie de ces frais de gestion grâce à de nouvelles conceptions et à l'automatisation.

Possibilité d'un partage sécurisé des données

De nombreuses entreprises peuvent améliorer leurs opérations en exploitant des référentiels, des services et des flux de données tiers. Les méthodes traditionnelles de partage des données, comme le FTP, les API et le courrier électronique, exigent que vous copiez les données et les envoyiez aux consommateurs. Ces méthodes lourdes, coûteuses et risquées sont basées sur le partage de données statiques, qui deviennent rapidement obsolètes et doivent être continuellement mises à jour avec des versions plus récentes. Le chapitre 6 explique en détail comment un datawarehouse établi dans le cloud permet un partage de données en direct, régulé et sécurisé.

CONSEIL

Les méthodes de partage de données robustes d'aujourd'hui vous permettent d'échanger des données en direct sans les transférer d'un endroit à l'autre.

Possibilité de réplication des données mondiales

La *réplication des données* crée de multiples copies de vos données dans le cloud. Ce type d'empreinte mondiale n'est pas seulement essentiel pour la reprise après sinistre et la continuité des activités : il est également utile si vous souhaitez partager des données avec une clientèle mondiale

sans avoir à mettre en place des pipelines ETL entre les régions. Les principaux fournisseurs de datawarehouses vous permettent de partager facilement des données entre régions géographiques et dans plusieurs clouds , notamment Amazon Web Services (AWS), Microsoft Azure et Google Cloud Platform (GCP). Ces capacités de réplication mondiale élargissent vos marchés, facilitent l'engagement de partenaires et permettent un écosystème plus complet pour l'analyse et le partage des données.

Garantie d'isolement des charges de travail

Un facteur clé de la rapidité et de la performance d'un datawarehouse est sa capacité à isoler les charges de travail. Pour être efficace, le datawarehouse dans le cloud doit pouvoir configurer facilement plusieurs pools de ressources de calcul (de tailles variables) afin de séparer les charges de travail des utilisateurs et les processus qui doivent s'exécuter simultanément. Cela permet d'éliminer les conflits et de disposer de ressources adaptées à chaque charge de travail. Idéalement, ces charges de travail distinctes devraient accéder simultanément aux mêmes données et s'activer et se désactiver facilement, en fonction des besoins.

Permettre tous les cas d'utilisation

Dans les environnements traditionnels, des systèmes de données différents traitent des cas d'utilisation différents : un datawarehouse pour les rapports opérationnels, des datamarts de données pour les rapports et les analyses des services, des data lakes pour l'exploration des données et des outils spécialisés pour des activités comme l'analyse prédictive. Chacun d'entre eux nécessite du matériel, une copie des données, une gestion individuelle, etc.

Pour rassembler ces divers cas d'utilisation dans le cloud, un datawarehouse doit permettre de cloner rapidement et efficacement des copies multiples de tables, de schémas et de bases de données, mais sans le casse-tête ni les coûts de stockage qu'impliquent les formes traditionnelles de duplication des données. Un datawarehouse dans le cloud devrait également permettre une reprise facile à la suite d'erreurs ou de problèmes créés par des travaux de transformation des données, grâce à des fonctionnalités de retour en arrière, qui permettent un accès simple et un rollback aux versions précédentes des données.

DANS CE CHAPITRE

- Reconnaître l'importance du partage des données
- Mettre en place une architecture efficace de partage des données
- Profiter des possibilités de partage des données

Chapitre 6
Possibilité de partage des données

Le partage des données est l'acte de donner accès aux données : à la fois au sein d'une entreprise et entre les entreprises qui ont déterminé qu'elles avaient de précieux atouts à partager. L'organisation qui met ses données à disposition, ou qui partage ses données, est un *fournisseur de données*. L'organisation qui veut utiliser les données partagées est un *consommateur de données*. Toute organisation peut être fournisseur de données, consommateur de données ou les deux.

En plus de toutes les données que les organisations génèrent et partagent en interne, beaucoup d'entre elles améliorent leurs opérations en exploitant des référentiels, des services et des flux de données de tiers. Par exemple, une organisation de services financiers peut exploiter divers indicateurs de marché, financiers et économiques pour créer de meilleurs modèles de données, qui l'aideront à créer de nouvelles offres de produits pour ses clients.

Il est possible de dégager un niveau élevé de valeur des sources croissantes de données mondiales, en interne et sur les marchés et échanges extérieurs. Jusqu'à récemment, il n'existait cependant aucune technologie permettant de partager des données sans risque, coût, difficultés ou délais importants. Bien que l'utilisation commerciale du partage de données existe depuis près d'un siècle, toutes les méthodes utilisées jusqu'à présent ont été limitées. Imaginez les possibilités si toutes les organisations pouvaient avoir accès à la demande à des données en direct et prêtes à l'emploi et pouvaient les utiliser immédiatement. Les données n'auraient plus à être déconstruites par le fournisseur de données, transmises au consommateur de données et reconstruites par ce dernier. Elles seraient instantanément accessibles et prêtes à être utilisées dans un environnement sécurisé et régulé.

Relever les défis techniques

Les méthodes traditionnelles de partage des données, telles que le protocole de transfert de fichiers (File Transfer Protocol ou FTP), le stockage dans le cloud (Amazon S3, Box, Dropbox, etc.), les interfaces de programmation d'applications (application programming interfaces ou API) et le courrier électronique, vous obligent à faire une copie des données partagées et à l'envoyer à vos consommateurs de données. Ces méthodes lourdes, coûteuses et risquées produisent des données statiques, qui deviennent rapidement obsolètes et doivent être actualisées avec des versions plus récentes, ce qui nécessite un mouvement et une gestion constants des données.

Les nouvelles technologies de partage des données permettent aux organisations de partager facilement des tranches de leurs données, et de recevoir des données partagées, de manière sécurisée et régulée. Elles ne nécessitent pas de transfert de données, de technologie d'extraction, de transformation et de chargement (ETL), ni de mises à jour constantes pour actualiser les données. Il n'est pas nécessaire de transférer des données via FTP ou de configurer des API pour relier des applications. Comme les données sont partagées plutôt que copiées, aucun stockage supplémentaire dans le cloud n'est nécessaire. Grâce à cette nouvelle architecture, les fournisseurs de données peuvent facilement et en toute sécurité publier des données pour que les consommateurs de données puissent les découvrir, les interroger et les enrichir instantanément, comme le montre la figure 6-1.

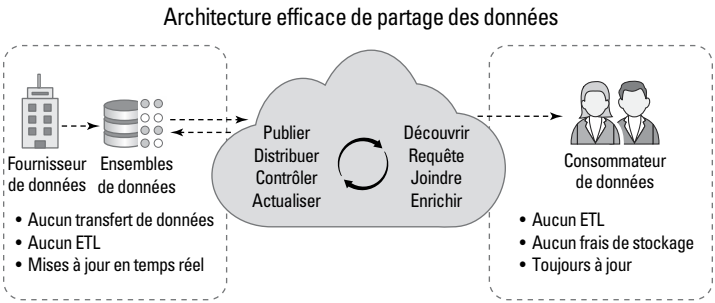

FIGURE 6-1 : Une architecture efficace pour le partage des données en temps réel.

Un datawarehouse multi-tenant établi dans le cloud fournit la plate-forme idéale pour un service de partage de données car il permet aux membres autorisés d'un écosystème de cloud d'accéder à des versions en direct et en lecture seule des données. Les fournisseurs de données peuvent partager des données avec les fournisseurs, les partenaires de la chaîne d'approvisionnement, les partenaires logistiques, les clients et de nombreux autres bénéficiaires. Ces solutions cloud tirent parti des dernières avancées en matière de cloud et de stockage de données. Plutôt que de transférer physiquement les données à des consommateurs internes ou externes, le datawarehouse

permet un accès en lecture seule à une partie régulée de l'ensemble des données en direct via SQL.

Réussir le partage des données

La plupart des organisations qui s'engagent dans le partage de données suivent une progression familière.

1. **Collaboration interne :** les données sont partagées au sein de l'entreprise entre ses unités commerciales et ses filiales, ce qui permet d'améliorer la collaboration et de briser les silos de données.
2. **Perspectives commerciales :** le fait de disposer de données plus complètes améliore la collaboration et permet d'avoir une meilleure vue d'ensemble des activités, le partage des données devenant la norme.
3. **Analyse de la clientèle :** l'entreprise développe des analyses axées sur le client pour améliorer la valeur d'un produit ou d'un service – la première étape vers la monétisation des données.
4. **Analyse avancée :** comme les clients demandent plus de données, l'entreprise développe des services d'analyse personnalisés pour fournir aux clients des informations riches à partir de ses données.
5. **Services de données :** l'entreprise s'appuie sur des ensembles de données internes pour fournir également à ses clients des services d'enrichissement des données, comme la modélisation des données, l'enrichissement des données et l'analyse des données.
6. **Échange de données :** l'entreprise cherche des moyens d'améliorer ses produits de données en se procurant des données externes et en offrant ses produits de données à un public plus large, généralement par le biais d'un marché ou d'un échange de données.

Monétiser vos données

La plupart des organisations partagent déjà des données ou prévoient de le faire, mais elles risquent de négliger la manière de les monétiser. Il existe un marché immense et en pleine expansion de monétisation des données. Dans les « Prédictions 2019 pour la transformation numérique » d'IDC, le cabinet de recherche a prédit que 80 % des entreprises créeront des capacités de gestion et de monétisation des données d'ici 2020, et que d'ici 2023, 95 % des établissements auront intégré de nouveaux ensembles d'indicateurs clés de performance (KPI) numériques.

CONSEIL

Avec la bonne architecture de partage des données, vous pouvez facilement analyser davantage de vos données pour découvrir de nouveaux produits, services et opportunités de marché.

ÉTUDE DE CAS

MAXIMISER LES POSSIBILITÉS DE REVENUS

Environics Analytics est l'une des principales entreprises d'analyse de données en Amérique du Nord. Pour fournir des informations fondées sur des données à plus de 3 000 clients, Environics ingère et analyse de grandes quantités de données démographiques, de lieu et de consommation.

Environics a récemment transféré ces activités analytiques dans un datawarehouse dans le cloud qui peut gérer n'importe quelle quantité de données et n'importe quel nombre de charges de travail. Un service d'échange de données intégré permet aux clients de découvrir et d'obtenir instantanément de nouvelles données. Selon Sean Howard, vice-président senior du développement des produits chez Environics, un service de partage de données sécurisé offre un mécanisme pratique de fourniture de données et présente d'immenses possibilités d'accroissement des revenus. La plate-forme de cloud s'adapte rapidement aux besoins analytiques de chaque utilisateur, sans l'aide de l'équipe informatique.

Auparavant, les scientifiques d'Environics stockaient des ensembles de données sur leurs ordinateurs et partageaient les produits finis avec leurs clients via FTP, ce qui créait la confusion en interne et freinait la croissance. L'exploration d'énormes ensembles de données contenant des milliards de rangées d'évènements a nécessité un soutien constant de l'équipe informatique pour installer le matériel, établir des environnements SQL Server, optimiser les performances des requêtes et surveiller l'utilisation des ressources de stockage et de calcul.

Aujourd'hui, le fait de disposer d'un environnement analytique qui s'adapte à la demande permet aux scientifiques de prototyper en toute confiance de grands ensembles de données provenant de n'importe quelle industrie, source ou type de fichier. Ils peuvent convertir des milliards de points de données brutes en produits de données viables. Le service sécurisé de partage de données renforce la fidélité des clients, réduit les coûts d'exécution et élimine les transferts de fichiers inutiles, tout en simplifiant considérablement la gestion des versions.

Les détaillants, les banques, les coopératives de crédit, les sociétés immobilières, les organisations à but non lucratif et les agences gouvernementales utilisent l'échange de données pour les aider à prendre des décisions éclairées sur les consommateurs et les marchés. Environics expérimente actuellement les données de l'Internet des objets (Internet of Things ou IoT) et d'autres grandes sources de données, grâce à un service d'ingestion continue de données qui accélère le chargement des données et permet une analyse en temps quasi réel. Selon Howard, « La participation à l'échange de données favorisera une véritable croissance des affaires et nous aidera à faire parvenir nos données à davantage de clients potentiels ».

DANS CE CHAPITRE

- Renforcer la reprise après sinistre et la continuité des activités
- Permettre la portabilité entre les clouds, sans verrouillage des fournisseurs
- Utiliser les initiatives d'expansion mondiale
- Simplifier la sécurité et l'administration dans les environnements multi-cloud

Chapitre 7
Maximiser les options grâce à une stratégie multi-cloud

Le fait de disposer d'un datawarehouse pouvant couvrir plusieurs régions et plusieurs clouds offre d'énormes avantages pour le partage des données, la continuité des activités et la pénétration géographique. Selon le rapport « L'état du cloud en 2019 » de Flexera, 84 % des organisations ont une stratégie multi-cloud, ce qui reflète les réalités du marché. Qu'il s'agisse d'Amazon Web Services, de Microsoft Azure ou de Google Cloud Platform, chaque service de cloud répond à des besoins légèrement différents.

Pour les organisations qui veulent avoir une portée mondiale grâce à leur datawarehouse, une stratégie cross-cloud a du sens : elle favorise la circulation libre et sécurisée des données partout dans le monde tout en vous permettant de sélectionner les fournisseurs de stockage dans le cloud qui répondent le mieux à vos besoins. Par exemple, chacun des services de votre organisation peut avoir des besoins uniques en matière de cloud. Plutôt que d'exiger que toutes les unités commerciales utilisent le même fournisseur, une stratégie multi-cloud permet à chaque unité d'utiliser le cloud qui fonctionne le mieux pour elle. Si cette flexibilité est importante pour vous, recherchez un fournisseur qui prend en charge plusieurs environnements du cloud et qui offre un support cross-cloud.

Comprendre le cross-cloud

Le *multi-cloud* signifie que vous pouvez stocker vos données dans plusieurs clouds différents. Le *cross-cloud* signifie que vous pouvez accéder aux données de tous ces clouds simultanément, migrer de manière transparente les opérations analytiques d'un cloud vers un autre et partager les données entre les clouds. C'est le Saint Graal du stockage de données dans le cloud, car vous n'êtes pas lié à un seul fournisseur de données dans le cloud. Pourquoi est-ce si important ?

- Il s'agit d'un avantage stratégique pour les entreprises mondiales, car tous les fournisseurs de services de cloud n'opèrent pas dans toutes les régions.
- C'est utile si vous acquérez une entreprise qui s'est uniformisée dans un cloud différent au vôtre.
- Si vous prévoyez de partager ou de monétiser vos données, vous élargirez votre marché potentiel si vous disposez d'une plate-forme de gestion de données unifiée qui couvre les régions et les clouds.

Dans les sections suivantes, nous passons en revue les technologies qui rendent possible un datawarehouse cross-cloud.

CONSEIL

Collaborez avec un fournisseur de datawarehouse qui a fourni les efforts voulus pour résoudre les différences entre les configurations de cloud et construit sa solution sur un codebase commun qui couvre tous les clouds.

Le levier de la réplication mondiale

La *réplication des données* est le processus de stockage des données dans plus d'un endroit pour assurer la disponibilité des données pendant une panne régionale. C'est aussi la technologie fondamentale qui permet de partager des données entre les régions et les clouds. Les datawarehouses ont besoin d'une technologie avancée de réplication des données pour maximiser les options de déploiement régional, permettre la continuité des activités et étendre les opérations dans le monde entier.

Votre plate-forme de datawarehouse doit permettre la réplication inter-régionale et cross-cloud sans réduire les performances opérationnelles par rapport à vos données primaires.

Minimiser les interruptions de service

La réplication des datawarehouses cross-cloud est importante pour les scénarios de reprise après sinistre critiques pour les entreprises. En cas

de panne, elle vous permet de reprendre instantanément les activités de traitement des données sans aucun temps d'arrêt (voir figure 7-1). Cependant, sans la bonne technologie de réplication des données, la restauration des géo-sauvegardes pour les grands datawarehouses peut prendre des heures, voire des jours. Atteindrez-vous ainsi vos objectifs en matière de temps de récupération ?

Demandez à votre fournisseur de datawarehouse s'il prend en charge l'accès et la récupération instantanés pour les bases de données de n'importe quelle taille, dans n'importe quel cloud et dans toutes les régions. Si une catastrophe se produit dans une région particulière du monde, vous devriez pouvoir accéder immédiatement aux données reproduites dans une autre région ou un autre service de cloud. Vérifiez que votre fournisseur de datawarehouse réplique les bases de données et les synchronise en permanence entre les plate-formes cloud et les régions.

Réplication cross-region et cross-cloud

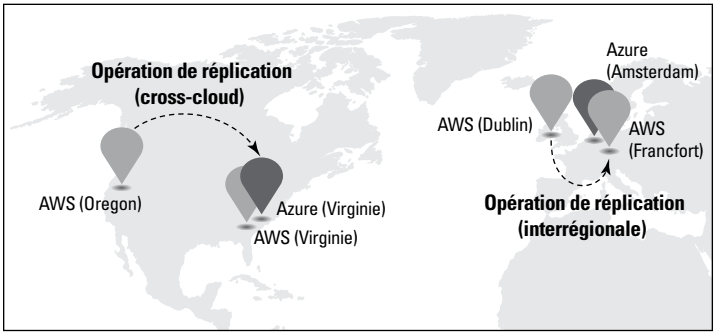

FIGURE 7-1 : La réplication des données à l'échelle mondiale assure la continuité des activités pendant les pannes.

Un support multi-sites

La portabilité des données est un défi de taille pour toutes les organisations qui disposent de grandes quantités de données. Chaque fournisseur de cloud public a différents niveaux de pénétration régionale. Le transfert des données et des charges de travail entre les régions géographiques et les clouds est plus facile avec une architecture cross-cloud.

La portabilité des données simplifie le respect de la réglementation si votre secteur d'activité exige que vos données restent dans un certain pays ou une certaine région. La fusion ou l'acquisition d'une autre entreprise dont le fournisseur de cloud est différent est également plus facile dans ce cas .

Respecter la souveraineté des données

Au fur et à mesure que votre entreprise se développe, vous voudrez peut-être implanter vos activités informatiques dans les régions que vous desservez. Une stratégie multi-cloud vous donne la possibilité de sélectionner le cloud le plus fort dans chaque région. Vous pouvez ainsi mettre en place une architecture qui minimise la latence, respecte les exigences de géo-résidence et se conforme aux mandats de souveraineté des données. Vous pourrez étendre vos activités dans des régions éloignées sans sacrifier l'accès aux données et vous découvrirez la valeur d'une source de vérité unique pour toute votre organisation.

La réplication des données permet également de faciliter le partage et la monétisation des données et de faire participer les partenaires à un échange, tout en respectant le principe fondamental du partage des données : les données existent localement dans une source unique, à partir de laquelle on peut y accéder plutôt que de les transférer.

Simplifier la sécurité

Lorsque vous travaillez avec plusieurs clouds, comment vous assurer que les mêmes configurations et techniques de sécurité s'appliquent à tous vos fournisseurs de cloud ? Devrez-vous résoudre les différences dans les pistes d'audit et les journaux d'évènements ? Vos experts en cybersécurité devront-ils composer avec différents ensembles de règles, ou bricoler plusieurs systèmes de gestion de clés pour crypter les données ? Un codebase unifié couvrant toutes les plates-formes de cloud simplifie toutes ces opérations. Vous n'aurez pas besoin d'embaucher des personnes dotées de compétences spécifiques ou de vous familiariser avec les nuances de plusieurs clouds.

RAPPEL

La technologie de réplication avancée vous permet de partager facilement des données entre de nombreuses régions et entre différents clouds de fournisseurs, sans avoir à mettre en place des pipelines de données, à copier des données ou à résoudre des différences de sécurité. Cela élargit vos marchés, facilite l'engagement de partenaires et vous donne un écosystème solide pour l'analyse et le partage des données.

DANS CE CHAPITRE

- Établissement d'une sécurité complète des données
- Respect de la réglementation en matière de protection de la vie privée
- Vérification des attestations et des certifications
- Amélioration de la conservation, de la protection et de la disponibilité des données

Chapitre 8
Sécurisation de vos données

Les faits sur la sécurité du cloud : Dans la plupart des cas, vos données sont plus sûres dans le cloud que dans votre propre datacenter. Une enquête menée en 2019 par Deloitte auprès des responsables informatiques, rédigée par une équipe comprenant Tom Davenport, Ashish Verma et David Linthicum, a révélé que plus de 90 % des entreprises conservent principalement leurs données sur des plates-formes cloud. Selon l'enquête, la sécurité des données et la gouvernance sont les principaux facteurs qui incitent les organisations à migrer leurs données vers le cloud.

Les fournisseurs de cloud SaaS desservent des milliers, voire des millions de clients. Ils peuvent se permettre les ressources nécessaires pour assurer une sécurité des données de niveau industriel de bout en bout. Cependant, tous les fournisseurs de cloud ne font pas l'effort de sécuriser vos données. Regardez bien, et vous verrez que les capacités de sécurité sont très variables.

Étude des fondamentaux

La protection de vos données et le respect des réglementations pertinentes doivent être fondamentaux pour l'architecture, la mise en œuvre et le fonctionnement d'un service de datawarehouse dans le cloud. Tous les aspects du service doivent être centrés sur la protection de vos

données dans le cadre d'une stratégie de sécurité à plusieurs niveaux, qui tient compte des menaces de sécurité actuelles et en évolution. Cette stratégie devrait porter sur les interfaces externes, le contrôle d'accès, le stockage des données et l'infrastructure physique, en conjonction avec une surveillance complète, des alertes et des pratiques de cybersécurité vérifiables.

Cryptage des données par défaut

Crypter des données signifie appliquer un algorithme de cryptage pour traduire le texte clair en texte crypté. C'est fondamental pour la sécurité. Cryptez les données dès qu'elles quittent vos locaux, via Internet, et qu'elles entrent dans le datawarehouse : lorsqu'elles sont stockées sur disque, lorsqu'elles sont déplacées dans un lieu de transit, lorsqu'elles sont placées dans un objet de base de données et lorsqu'elles sont mises en cache dans un datawarehouse virtuel. Les résultats des requêtes doivent également être cryptés. Tout cela doit impérativement être intégré.

Le fournisseur doit également protéger les clés de décryptage qui décodent vos données. Les meilleurs fournisseurs de services utilisent le cryptage AES 256 bits avec un modèle de clé hiérarchique. Cette méthode permet de crypter les clés de cryptage et de déclencher une rotation des clés qui limite la durée d'utilisation d'une seule clé.

RAPPEL

Vos données vivent probablement dans de nombreux endroits. Vous devez protéger et contrôler le flux de données à chaque point. Toutes les données doivent être cryptées de bout en bout et automatiquement, en transit et au repos.

Application du contrôle d'accès

La sécurisation des données n'est qu'un aspect de la sécurité globale. Les violations de données résultent souvent du fait que les utilisateurs choisissent des mots de passe faibles associés à des procédures d'authentification rudimentaires. Un service de datawarehouse dans le cloud devrait toujours autoriser les utilisateurs, authentifier les identifiants et accorder aux utilisateurs l'accès uniquement aux données auxquelles ils ont droit.

Le point de départ est le *contrôle d'accès basé sur les rôles*, qui garantit que les utilisateurs ne peuvent accéder qu'aux données qu'ils sont autorisés à voir. Le contrôle d'accès doit être appliqué à tous les objets de la base de données, y compris les tableaux, les schémas et toute extension virtuelle du datawarehouse. Pour un maximum de commodité et de sécurité, votre datawarehouse dans le cloud devrait également fournir une

authentification multifactorielle, qui nécessite une vérification secondaire telle qu'un code de sécurité unique envoyé au téléphone portable d'un utilisateur.

Les procédures de signature unique et l'*authentification fédérée* permettent aux personnes de se connecter plus facilement au service de datawarehouse directement à partir d'autres applications sanctionnées. L'authentification fédérée centralise la gestion des identités et les procédures de contrôle d'accès, ce qui permet à votre équipe de gérer plus facilement les privilèges d'accès des utilisateurs.

CONSEIL

Votre fournisseur de datawarehouse dans le cloud ne devrait pas avoir accès aux données non cryptées des clients à moins que vous ne lui accordiez explicitement cet accès.

Correctifs, mises à jour et surveillance du réseau

Les correctifs logiciels et les mises à jour de sécurité doivent être installés sur tous les composants logiciels pertinents dès qu'ils sont disponibles. Le fournisseur doit également faire procéder à des tests de sécurité périodiques (également appelés tests de pénétration) par une société de sécurité indépendante afin de vérifier de manière proactive les vulnérabilités.

Les mesures de sécurité physique dans le datacenter doivent inclure des contrôles d'accès biométriques, des gardes armés et une surveillance vidéo pour s'assurer que personne n'obtienne un accès non autorisé. Toutes les machines physiques et virtuelles doivent être contrôlées davantage grâce à des procédures logicielles rigoureuses d'audit, de surveillance et d'alerte. En guise de protection supplémentaire, les outils de surveillance de l'intégrité des fichiers (file integrity monitoring ou FIM) garantissent que les fichiers système critiques ne sont pas altérés, et les listes blanches d'adresses IP vous permettent de limiter l'accès au datawarehouse exclusivement à des réseaux de confiance. (Une liste blanche est une liste d'adresses électroniques ou de noms de domaine sur laquelle s'appuiera un programme de blocage du courrier électronique pour permettre la réception de messages).

Les évènements de sécurité, générés par les systèmes de contrôle de la cybersécurité qui surveillent le réseau, doivent être automatiquement consignés dans un système inviolable de gestion des informations et des évènements de sécurité (security information and event management ou SIEM). Des alertes automatiques doivent être envoyées au personnel de sécurité lorsqu'une activité suspecte est détectée.

Garantir la protection, la conservation et la redondance des données

En cas d'incident, vous devez être en mesure de restaurer instantanément ou d'interroger les versions précédentes de vos données dans un tableau ou une base de données pendant une période de conservation spécifiée, telle que régie par votre accord de niveau de service (SLA) avec le fournisseur du datawarehouse dans le cloud. Une stratégie complète de conservation des données doit aller au-delà de la duplication des données au sein d'une même région ou zone du cloud : elle devrait reproduire ces données dans plusieurs zones de disponibilité pour assurer une redondance géographique. Éventuellement, le basculement automatique vers ces autres zones peut assurer la continuité des opérations commerciales.

Exiger l'isolement des environnements

Si votre fournisseur de solution de datawarehouses utilise un environnement de cloud multi-tenant, dans lequel de nombreux clients partagent la même infrastructure physique, assurez-vous que chaque client dispose d'un datawarehouse virtuel isolé de tous les autres datawarehouses. Pour le stockage, cette isolation devrait s'étendre jusqu'à la couche de la machine virtuelle : l'environnement de stockage des données de chaque client doit être isolé de l'environnement de tous les autres clients, régulé par des répertoires indépendants et des clés de cryptage uniques. Certains fournisseurs proposent également des réseaux privés virtuels (VPN) dédiés et des passerelles entre les systèmes d'un client et le datawarehouse dans le cloud. Ces services dédiés garantissent que les éléments les plus sensibles de votre datawarehouse sont complètement séparés de ceux des autres clients.

Maintenir la gouvernance et la conformité

La gouvernance des données garantit que les données de l'entreprise sont correctement accessibles et utilisées, et que les pratiques quotidiennes de gestion des données sont conformes à toutes les exigences réglementaires pertinentes. Les politiques de gouvernance établissent des règles et des procédures pour contrôler la propriété et l'accessibilité de vos données. Les types d'informations qui relèvent généralement de ces directives comprennent les informations relatives aux cartes de crédit, les numéros de sécurité sociale, les dates de naissance, les informations relatives au réseau IP et les coordonnées de géolocalisation.

Vérification des attestations et des certifications

La conformité n'est pas seulement une question de pratiques de cybersécurité solides. Il s'agit également de s'assurer que votre fournisseur de datawarehouse peut prouver qu'il a mis en place les procédures de sécurité requises. Réparer une violation de données peut coûter des millions de dollars et peut endommager de façon permanente les relations avec vos clients.

Des rapports d'attestation conformes aux normes du secteur vérifient que les fournisseurs de services de cloud utilisent des contrôles de sécurité appropriés. Par exemple, un fournisseur de datawarehouse dans le cloud doit démontrer qu'il surveille et répond de manière adéquate aux menaces et aux incidents de sécurité et qu'il a mis en place des procédures de réponse aux incidents suffisantes (voir figure 8-1).

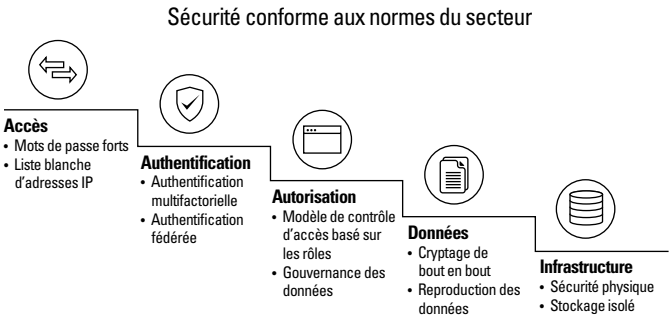

FIGURE 8-1 : Vérifiez que tout le trafic de données est crypté et sécurisé, et que vos fournisseurs de cloud détiennent toutes les certifications nécessaires.

En plus des certifications technologiques standard du secteur telles que ISO/IEC 27001 et SOC 1/SOC 2 Type II, vérifiez que votre fournisseur de cloud respecte également toutes les réglementations gouvernementales et du secteur applicables. Selon votre entreprise, cela peut inclure les certifications PCI, HIPAA/Health Information Trust Alliance (HITRUST) et FedRAMP.

Demandez des preuves à vos fournisseurs et qu'ils vous présentent une copie intégrale du rapport pour chaque norme pertinente, et pas seulement les lettres d'accompagnement. Par exemple, le rapport SOC 2 de type II vérifie que les contrôles techniques et administratifs appropriés ont été mis en place de manière cohérente au cours des 12 derniers mois.

L'attestation de conformité PCI-DSS révèle si votre fournisseur stocke et traite correctement les informations relatives aux cartes de crédit. Si vous manipulez des informations médicales protégées, exigez de vos fournisseurs qu'ils se conforment aux directives de l'HIPAA.

CONSEIL

La conformité et les attestations prouvent que votre fournisseur de datawarehouse est sérieux et transparent en matière de sécurité.

Les fournisseurs de cloud doivent également donner la preuve que les fournisseurs de logiciels tiers avec lesquels ils travaillent sont conformes et qu'ils effectuent régulièrement des audits de sécurité. La sécurité de vos données dépend du maillon le plus faible de la chaîne technologique. Veillez donc à ce que tous les acteurs disposent de contrôles de sécurité solides et se conforment aux pratiques de sécurité standard du secteur. Si une preuve de conformité fait défaut, procurez-vous des documents justificatifs.

RAPPEL

Ne travaillez qu'avec des fournisseurs de services de cloud qui démontrent qu'ils respectent de bout en bout les pratiques de sécurité sanctionnées par l'industrie et que des auditeurs indépendants le confirment. Ces considérations de conformité doivent comprendre vos exigences minimales pour cet important référentiel de données.

Insister sur un système de sécurité complet

Assurer une bonne sécurité est coûteux et nécessite des connaissances spécialisées. Les pannes d'équipement, les brèches dans le réseau et les incidents de maintenance peuvent entraîner des pertes de données et introduire des incohérences dans vos données. Une pratique de sécurité complète englobe de nombreux aspects. Votre fournisseur de datawarehouse dans le cloud doit disposer de procédures de protection contre une destruction accidentelle ou intentionnelle. Certains fournisseurs offrent des capacités de sécurité rudimentaires, laissant le cryptage, le contrôle d'accès et la surveillance de la sécurité à vous, le client. La sécurité devrait être un des fondements du service de datawarehouse ; vous ne devriez pas avoir à faire quoi que ce soit de plus pour sécuriser vos données.

CONSEIL

Un fournisseur transparent sur ses certifications de sécurité est beaucoup plus susceptible d'avoir un programme de sécurité solide.

DANS CE CHAPITRE

- Créer un environnement de stockage rentable
- Obtenir la meilleure valeur et la meilleure performance grâce à l'architecture et à la tarification

Chapitre 9
Minimiser vos coûts de datawarehouse

Dans ce chapitre, nous examinons comment gérer un datawarehouse dans le cloud et comment votre fournisseur de datawarehouse peut vous aider à minimiser les coûts sur le long terme.

Minimiser vos coûts de datawarehouse

Plus vous pouvez stocker de données, plus vous pouvez en tirer des informations. Heureusement, le stockage dans le cloud d'Amazon, de Microsoft et de Google est devenu relativement peu coûteux. Vous ne devriez donc pas vous sentir limité dans la quantité et les types de données que vous stockez. Examinez les conditions pour vous assurer que votre fournisseur de datawarehouse dans le cloud ne majore pas ces coûts de stockage bruts. Le fournisseur doit vous transmettre directement la grille tarifaire. Votre fournisseur de datawarehouse peut apporter une valeur ajoutée en *comprimant* vos données de trois à cinq fois. Une compression triple signifie que vous avez un tiers de la quantité de données à stocker, pour un tiers du coût.

Examiner les termes de l'accord d'utilisation : vous ne devriez avoir à payer que pour le stockage que vous utilisez, et non pour la capacité de stockage excédentaire ou réservée. Vous ne devez pas non plus payer pour cloner des bases de données dans votre datawarehouse pour des activités de développement et d'essai. Vous devriez pouvoir référencer, et non copier, vos données plusieurs fois et donc ne pas avoir à payer de supplément pour le stockage.

Votre datawarehouse dans le cloud devrait également vous permettre de stocker et d'interroger des données structurées et semi-structurées telles que JSON. Enfin, recherchez un fournisseur qui offre des *capacités multi-cloud*, car cela peut vous faire économiser des coûts futurs si vous migrez votre datawarehouse vers un autre environnement de stockage dans le cloud.

Maximiser la productivité de calcul

Les ressources de calcul étant plus coûteuses que les ressources de stockage, votre service de datawarehouse doit vous permettre de dimensionner chaque ressource de manière indépendante et faciliter l'affectation des ressources de calcul dont vous avez besoin selon un modèle de tarification basé sur l'utilisation. Le fournisseur doit vous facturer uniquement les ressources que vous utilisez – à la seconde près – et suspendre automatiquement le calcul des ressources lorsque vous cessez de les utiliser, afin d'éviter des coûts excessifs. La tarification basée sur l'utilisation ou l'abonnement vous permet de choisir comment vous consommez les ressources.

Des conditions flexibles devraient également vous permettre de dimensionner correctement vos clusters de calcul en fonction de chaque charge de travail. Si vous exécutez un travail d'extraction, de transfert, de chargement (ETL) avec de faibles exigences de calcul, vous pouvez faire correspondre un petit cluster à cette charge de travail plutôt que d'encourir le coût d'un cluster surprovisionné. Si vous avez besoin de tester de nouveaux modules d'apprentissage machine, vous pouvez utiliser un grand cluster. Cela vous donne une évolutivité fine pour chaque charge de travail tout en minimisant les coûts d'utilisation. Votre datawarehouse coûtera moins cher à gérer que les datawarehouses on-premise et leurs cousins en version cloud qui rampent, utilisent d'énormes ressources et produisent des résultats limités. Les charges de travail ne ralentiront pas et ne se bloqueront même pas non plus, grâce aux clusters de calcul dédiés à chaque charge de travail.

DANS CE CHAPITRE

- Lister vos besoins et vos critères de réussite en matière de datawarehouse
- Tenir compte de tous les facteurs pour le coût total de la propriété
- Faire un essai de votre datawarehouse avant d'acheter

Chapitre 10
Six étapes pour démarrer datawarehouse dans le cloud

Dans ce chapitre, nous vous guidons à travers six étapes clés pour choisir un datawarehouse dans le cloud pour votre organisation. Le processus commence par l'évaluation de vos besoins de datawarehouse et se termine par l'essai de votre premier choix. À la fin, vous disposerez d'un plan pour vous aider à choisir votre solution en toute confiance.

Étape 1 : évaluez vos besoins

Le datawarehouse qui vous convient doit répondre à vos besoins actuels et pouvoir s'adapter à vos besoins futurs. Par conséquent, tenez compte de la nature de vos données, des compétences et des outils déjà en place, de vos besoins d'utilisation, des projets futurs de votre entreprise et de la manière dont un datawarehouse peut amener votre entreprise plus loin que vous ne l'aviez imaginé.

- **Les données :** Quels types de données le datawarehouse doit-il contenir ? A quel rythme de nouvelles données sont-elles créées ? À quelle fréquence les données seront-elles transférées dans le datawarehouse ? Quelles sont les données essentielles auxquelles vous ne pouvez pas accéder aujourd'hui ?
- **S'adapter aux compétences, outils et processus existants :** Quels sont les outils et les compétences de votre équipe qui s'appliqueront aux différentes options du datawarehouse dans le

cloud ? Quels sont les processus sur lesquels un datawarehouse dans le cloud aura un impact ?

- **Utilisation :** Quels utilisateurs et applications auront accès au datawarehouse ? Quels types de requêtes allez-vous exécuter ? À quelle quantité de données les utilisateurs doivent-ils accéder, et à quelle vitesse ? Comment les charges de travail varieront-elles dans le temps ? Quelles sont les performances requises par vos utilisateurs et vos applications ? Combien d'utilisateurs devraient déjà accéder au datawarehouse, mais ne le font pas aujourd'hui faute de ressources ?
- **Partage des données :** Prévoyez-vous de partager les données de manière sécurisée au sein de votre organisation et avec vos clients et/ou partenaires ? Si oui, quels types de données allez-vous partager, et allez-vous créer un marché ou un échange de données pour les monétiser également ? Allez-vous permettre à ces consommateurs d'accéder aux données brutes ou allez-vous enrichir ces données en offrant également des services de données tels que des analyses ?
- **Accès mondial :** Envisagez-vous de stocker des données dans espace de stockage d'objets public, comme Amazon S3, Microsoft Azure ou Google Cloud Platform ? Avez-vous des exigences spécifiques en matière de souveraineté fonctionnelle, régionale ou en matière de données qui nécessitent le maintien de ces relations ? Avez-vous besoin d'une architecture cross-cloud pour maximiser les options de déploiement régional, pour renforcer la reprise après sinistre ou pour assurer la continuité des activités à l'échelle mondiale ?
- **Ressources :** Quelles sont les ressources humaines disponibles pour gérer le datawarehouse ? Quel investissement souhaitez-vous faire pour contrôler et gérer la disponibilité, les performances et la sécurité ? Disposez-vous d'une expertise ciblée en matière de développement et de test de datawarehouses, ou d'une équipe DevOps pour les rationaliser ?

Étape 2 : migrez ou partez de zéro

Chaque projet de datawarehouse dans le cloud devrait commencer par l'évaluation de la proportion de votre environnement existant qui devrait migrer vers le nouveau système et de ce qui devrait être établi en partant de zéro pour un datawarehouse dans le cloud. Ces décisions peuvent porter sur tous les aspects, de la conception des processus d'extraction, de transformation et de chargement (ETL) aux modèles de données et aux méthodes du cycle de développement de logiciels. Réfléchissez aux points suivants :

- **S'agit-il d'un tout nouveau projet ?** Dans ce cas, il est souvent judicieux de concevoir le projet de manière à tirer pleinement parti des capacités d'un datawarehouse dans le cloud plutôt que de poursuivre une mise en œuvre existante avec des contraintes.
- **Quelles sont les parties de votre système actuel qui constituent le plus gros casse-tête ?** Une migration bien planifiée pourrait se concentrer sur le transfert en premier lieu des charges de travail les plus problématiques vers le datawarehouse dans le cloud. Ou alors, vous pouvez migrer les charges de travail les plus simples pour obtenir des gains rapides.
- **Quels sont les aspects de votre système actuel qui tiennent compte de contraintes inexistantes dans un datawarehouse dans le cloud ?** Les outils et les processus conçus pour contourner les contraintes de ressources, pour éviter les efforts perturbateurs nécessaires pour ajouter de la capacité ou pour optimiser les coûts peuvent être inutiles avec la bonne solution de cloud.
- **Comment les utilisateurs et les applications actuels accèdent-ils au datawarehouse ?** Les utilisateurs et les applications qui reposent sur des interfaces standard, comme le SQL, et qui utilisent des outils ETL et d'information de gestion standard, subiront moins de changements en s'adaptant à une nouvelle approche.
- **Comment vos besoins en matière de données et d'analyses sont-ils susceptibles d'évoluer à l'avenir ?** Une solution conçue pour évoluer est susceptible de durer plus longtemps que prévu et de révéler de nouvelles possibilités qui tirent parti de capacités avancées telles que le partage sécurisé des données et leur accès à l'échelle mondiale.

RAPPEL

Si vous disposez d'un datawarehouse traditionnel, vaste et complexe, faites migrer une petite partie du système pour vous familiariser avec l'utilisation d'un datawarehouse dans le cloud. Vous pouvez ensuite étendre itérativement votre empreinte dans le cloud.

Étape 3 : établir des critères de réussite

Comment allez-vous mesurer le succès du passage à un nouveau datawarehouse dans le cloud ? Choisissez des exigences commerciales et techniques importantes. Les critères doivent être axés sur la performance, la simultanéité, la simplicité et le coût total de propriété (TCO).

RAPPEL

Si votre nouveau datawarehouse dans le cloud possède des capacités qui n'étaient pas disponibles dans votre ancien système, et que ces capacités sont pertinentes pour évaluer le succès commercial et technique de votre nouvelle solution, assurez-vous de les inclure.

Lorsque vous établirez les critères de réussite de votre nouvelle solution, déterminez comment vous mesurerez cette réussite en décidant quels critères sont quantifiables et lesquels sont qualitatifs, comment vous mesurerez les critères quantifiables et comment vous évaluerez les critères qualitatifs.

ÉTUDE DE CAS

RÉSOUDRE LES PROBLÈMES DE LATENCE

White Ops est l'un des principaux fournisseurs de services de cybersécurité. Contrairement aux approches traditionnelles qui font appel à l'analyse statistique, White Ops combat l'activité criminelle en faisant la distinction entre l'interaction robotique et humaine, en s'efforçant de découvrir et de caractériser de nouveaux modèles de fraude. Ce processus constant nécessite le stockage et le traitement de quantités massives de données.

White Ops s'appuyait auparavant sur les systèmes NoSQL pour stocker et traiter ces données. Cependant, le temps de latence pour les résultats était d'au moins 24 heures, en fonction de la charge de travail. Plus il y avait de requêtes, plus les délais étaient longs.

Pour accroître la productivité et les performances, White Ops a mis en place un datawarehouse dans le cloud avec SQL comme langage de base et fourni sous forme de service. Le e datawarehouse permet à White Ops d'avoir toutes les données en un seul endroit, d'évoluer de manière élastique, d'interroger diverses données avec le SQL standard et d'accélérer l'évolution de ses offres de prévention de la fraude.

White Ops peut désormais consolider et mettre à l'échelle d'énormes quantités de données, en permettre l'accès sans avoir recours à des spécialistes ayant des compétences poussées en programmation, et aider ses clients à éviter les effets dévastateurs de la fraude en ligne.

Étape 4 : évaluer les solutions

Une fois que vous aurez déterminé les besoins de votre e datawarehouse et les critères de réussite, vous serez prêt à commencer à évaluer les solutions. Tout au long de ce livre, nous détaillons les différences entre les options disponibles (voir les chapitres 3, 4 et 5). En les comparant, assurez-vous que les critères suivants sont satisfaits :

- Satisfaction des besoins actuels et futurs
- Intégration de données structurées et semi-structurées, stockage du tout en un seul endroit sans créer de silos de données
- Soutien des compétences, des outils et de l'expertise existants

- Protection contre la perte de données et possibilité de les récupérer facilement
- Sécurisation de vos données grâce à une protection par mot de passe et un cryptage conformes aux normes du secteur
- Disponibilité permanente des données et des analyses
- Rationalisation du pipeline de données afin que les nouvelles données soient disponibles pour analyse dans les plus brefs délais
- Optimisation de la la valeur temps, afin que vous puissiez profiter au plus vite des avantages de votre nouveau e datawarehouse
- Ressources dédiées à des charges de travail isolées
- Partage des données sans avoir à copier ou à transférer les données en direct et mise en relation facile des fournisseurs de données et des consommateurs
- Reproduction des bases de données et synchronisation entre les comptes, les plateformes de cloud et les régions afin d'améliorer la continuité des activités et de rationaliser l'expansion
- Clonage zero-copy des bases de données pour le développement et les essais, et pourdes cas d'utilisation multiples, tels que la communication et l'exploration de données et l'analyse prédictive
- Possibilité de récupération facile des données perdues à cause d'erreurs ou d'attaques en revenant aux versions précédentes des données
- Mise à l'échelle indépendante et automatique du calcul, du stockage et de la simultanéité sans ralentir les performances

Étape 5 : calcul du TCO

Si vous choisissez un e datawarehouse dans le cloud en fonction du prix, considérez le coût total de propriété (TCO) d'un e datawarehouse classique, qui comprend le coût des licences, généralement basé sur le nombre d'utilisateurs ; le matériel (serveurs, périphériques de stockage, mise en réseau) ; le datacenter (espace de bureau, électricité, administration, maintenance et gestion continue) ; la sécurité des données (protection par mot de passe et cryptage) ; les solutions pour assurer la disponibilité et la résilience ; le support en matière de dimensionnement et de simultanéité ; et la création d'environnements de développement et de simulation.

Pour certaines solutions, vous devrez peut-être envisager des coûts supplémentaires, comme pour la création et la gestion de plusieurs datamarts, pour avoir plusieurs copies de données dans différents datamarts, pour la formation, pour avoir plusieurs systèmes (par exemple, SQL et NoSQL) pour traiter des données diverses, etc.

Le calcul des coûts des options de datawarehouse dans le cloud est généralement plus facile, mais il varie selon les services du fournisseur. En supposant que vous sous-traitez tout au fournisseur en choisissant un e datawarehouse comme service (Data-Warehouse-as-a-Service ou DWaaS), vous pouvez calculer le TCO sur la base des frais d'abonnement mensuels. Si vous optez pour une solution de type infrastructure en tant que service (infrastructure-as-a-service ou IaaS) ou plate-forme en tant que service (platform-as-a-service ou PaaS) (voir chapitre 5), vous devez ajouter les coûts des logiciels, d'administration et des services que la solution n'inclut pas.


CONSEIL

Les organisations calculent généralement le TCO sur la durée de vie prévue du datawarehouse, qui est généralement de un à trois ans. Mise en garde essentielle : on suppose souvent qu'un système de cloud fonctionne à haute capacité 24 heures sur 24 et 7 jours sur 7, en négligeant les économies possibles lorsqu'une solution de cloud est mise à l'échelle de manière dynamique selon l'évolution de la demande, et ne facture qu'à la seconde près.

Étape 6 : démonstration de faisabilité (Proof of Concept - POC)

Après avoir étudié les différentes options de datawarehouse dans le cloud, visionné les démonstrations, posé des questions et rencontré l'équipe de chaque fournisseur, recourez à une POC avant de choisir. La POC teste une solution pour déterminer dans quelle mesure elle répond à vos besoins et à vos critères de réussite. Considérez-la comme un essai sur route. Elle dure généralement un jour ou deux, mais elle peut s'étaler sur plusieurs semaines. Vous demandez une POC à un fournisseur potentiel en sachant que si la solution fonctionne de manière satisfaisante, vous achèterez le produit. Ou, dans le cas du stockage de données dans le cloud, vous vous abonnerez au service.


CONSEIL

Lorsque vous établissez votre POC, énumérez toutes les exigences et tous les critères de réussite – pas seulement les problèmes que vous essayez de résoudre, mais tout ce qui est possible avec une solution cloud.

Élaborez une liste de contrôle complète des besoins en matière de stockage de données et de vos critères de réussite comme point de départ. Assurez-vous que le nouveau e datawarehouse fait tout ce que votre e datawarehouse actuel fait mais en mieux, et qu'il surmonte les inconvénients du système actuel. Si vous demandez une POC à plusieurs fournisseurs, utilisez la même liste de contrôle pour chacun d'entre eux.

Obtenir un avantage concurrentiel grâce à la puissance du datawarehouse dans le cloud

Les organisations modernes ont désormais accès à des quantités exponentielles de données qu'elles peuvent analyser pour obtenir les informations les plus approfondies possibles. En outre, les organisations veulent partager des données en toute sécurité – et acquérir des données partagées – entre leurs unités commerciales, au sein de leurs écosystèmes commerciaux et au-delà, en utilisant les échanges de données pour les monétiser. Mais l'accès à ces données pose des problèmes encore plus importants qui continuent de nuire aux plate-formes d'analyse de données traditionnelles. Les entreprises modernes réalisent aujourd'hui que le stockage des données dans le cloud est le moyen le plus efficace et le plus rentable de stocker et d'analyser toutes leurs données pour tous leurs utilisateurs professionnels. Ce livre révèle ce qui est disponible et comment votre organisation peut bénéficier de cette nouvelle et passionnante technologie.

À l'intérieur…

- Pourquoi le datawarehouse dans le cloud a vu le jour
- Comment le datawarehouse dans le cloud soutient la comparaison
- Comment évaluer les différents datawarehouses
- Pourquoi la sécurité et la gouvernance sont importantes
- Les avantages d'une solution cross-cloud
- Comment le partage moderne des données permet d'obtenir des informations encore plus précises
- Études de cas réels

snowflake®

Joe Kraynak est un auteur chevronné d'ouvrages pour les Nuls qui a écrit ou co-écrit des dizaines de livres sur des sujets variés. **David Baum** est un chroniqueur économique indépendant spécialisé dans les sciences et la technologie.

Allez sur le site Dummies.com®
pour voir des vidéos, des photos étape par étape, des articles pratiques ou pour vous procurer des ouvrages !

ISBN: 978-1-119-71448-4

pour les nuls®

Également disponible en e-book

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.