Making Everything Easier!™

Édition spéciale EMC Isilon

Le Stockage Big Data POUR LES NULS®

Un ouvrage présenté par

EMC²®

Will Garside
Brian Cox

Le Big Data permet aux organisations d'obtenir des renseignements précieux sur leurs marchés et leurs clients. Il permet à ces organisations de prendre de meilleures décisions en matière de conception, de sécurité et d'assistance, ainsi que d'offrir des produits et des services plus performants et mieux adaptés. Toutefois, pour obtenir ces renseignements, il faut disposer d'une infrastructure solide associée à une solution de stockage hautement évolutive, facilement gérable et bien protégée.

EMC²®

À propos d'EMC Isilon

Isilon, une branche d'EMC, est un leader mondial dans le domaine du stockage scale-out. Isilon offre des solutions puissantes, mais simples aux entreprises qui souhaitent gérer leurs données et non leur stockage. Les produits Isilon sont simples à installer, à gérer, à adapter, quelle que soit leur taille. Contrairement aux systèmes de stockage traditionnels proposé aux entreprises, Isilon ne perd rien de sa simplicité, quels que soient la quantité de stockage ajoutée, le niveau de performances attendu et l'évolution des besoins des sociétés. Isilon encourage les entreprises à penser leur stockage différemment, car lorsqu'elles le font, elles réalisent qu'il existe des moyens bien plus adaptés et plus simples de procéder. Plus d'informations sotn disponibles sur le site www.EMC.com/Isilon.

Le Stockage Big Data POUR LES NULS[MD]

Édition spéciale EMC Isilon

par Will Garside et Brian Cox

FOR DUMMIES®
A Wiley Brand

Le Stockage Big Data pour les Nuls, Édition spéciale EMC Isilon

Publié par :
John Wiley & Sons, Ltd
The Atrium
Southern Gate
Chichester
West Sussex
PO19 8SQ
Angleterre
www.wiley.com

Pour plus d'informations sur la manière de réaliser un livre *Pour les Nuls* destiné à votre entreprise ou organisation, écrivez à l'adresse CorporateDevelopment@wiley.com. Pour plus d'informations sur les licences relatives à la marque *Pour les Nuls* pour les produits et services, écrivez à l'adresse BrandedRights&Licenses@wiley.com.

Visitez notre page Internet à l'adresse www.customdummies.com.

ISBN: 978-1-118-71391-4 (pbk)

Imprimé en Grande-Bretagne par Page Bros

Introduction

Nous avons l'honneur de vous présenter *Le Stockage Big Data pour les Nuls*, un guide qui vous permettra de comprendre les concepts et les technologies nécessaires pour créer une architecture de stockage de données efficaces en vue de gérer des projets sensibles.

Les données sont un ensemble d'informations, telles que des valeurs ou des mesures. Les données peuvent être des nombres, des mots, des observations ou encore des descriptions.

Le stockage et la récupération de grands volumes d'informations, ainsi que la recherche de renseignements au sein d'une masse de données sont au cœur du concept du Big Data, et c'est la raison pour laquelle cette technologie est si importante pour la communauté informatique et la société dans son ensemble.

À propos de ce livre

Concis, ce livre regorge de conseils utiles sur la manière de concevoir, de mettre en œuvre et de gérer des données et des plateformes de stockage.

Partis pris

En écrivant ce livre, nous avons fait quelques suppositions à votre sujet. Nous supposons que :

- Vous faites partie d'une organisation qui prévoit de mettre en œuvre un projet de Big Data.
- Vous êtes responsable d'une équipe ou chef de projet mais pas forcément un expert technique.
- Vous devez être à même de participer à un projet de Big Data et d'y jouer un rôle essentiel. Pour cela, il vous est utile de comprendre les concepts clés du Big Data.

Structure du livre

Le Stockage Big Data pour les Nuls est divisé en sept chapitres concis et riches en informations :

- **Chapitre 1 : Explorer le monde des données.** Ce chapitre vous explique les différents types de données et de structures.
- **Chapitre 2 : Comment le Big Data peut-il aider votre organisation.** Ce chapitre vous aide à comprendre de quelle manière le Big Data peut aider votre organisation à résoudre des problèmes et à obtenir de meilleurs résultats.
- **Chapitre 3 : Créer une infrastructure efficace pour le Big Data.** Découvrez comment des éléments pris individuellement peuvent vous aider à bâtir vos fondamentaux.
- **Chapitre 4 : Améliorer un projet de Big Data à l'aide du stockage scale-out.** Comment une technologie de stockage innovante peut concrétiser vos projets.
- **Chapitre 5 : Bonnes pratiques du stockage scale-out dans le monde du Big Data.** Ces conseils vous permettront de garder votre cap.
- **Chapitre 6 : Autres éléments à prendre en compte pour le stockage de Big Data.** Nous abordons d'autres points complémentaires pour garantir le succès de votre projet de Big Data.
- **Chapitre 7 : Dix astuces permettant de garantir le succès d'un projet de Big Data.** Rendez-vous ici pour la célèbre « Partie des Dix » des livres Pour les Nuls : dix conseils à garder en tête, lorsque vous vous lancez dans un projet de Big Data.

Vous pouvez vous contenter de lire quelques chapitres de ce livre ou le lire du début à la fin, cela ne devrait pas vous prendre trop de temps !

Symboles utilisés dans ce livre

Pour faciliter la lecture et attirer votre attention sur les informations les plus utiles, ces symboles mettent en exergue les parties clés du texte :

La cible attire votre attention sur un conseil avisé.

Ce symbole met en évidence des informations importantes à garder en tête.

Consultez ces exemples de projets de Big Data pour y trouver conseils et inspiration.

Et maintenant ?

Vous pouvez lire ce livre d'un trait ou passer des sections, utiliser les titres des chapitres comme guide pour relever les informations dont vous avez besoin. Peu importe votre choix, vous ne pouvez pas vous tromper. Les deux possibilités mènent au même résultat : les connaissances dont vous avez besoin pour créer une solution de stockage évolutive, facile à gérer et protégée en vue de réaliser un projet de Big Data.

Chapitre 1

Explorer le monde des données

Dans ce chapitre

- Définir les données
- Comprendre les données structurées et non structurées
- Savoir comment exploiter les données
- Stocker et rechercher des données
- Connaître le potentiel et les risques

Le monde évolue au fil de la production d'informations électroniques. Chaque seconde, chaque jour, des ordinateurs et d'autres systèmes électroniques créent, traitent, transmettent et reçoivent d'énormes volumes d'informations. Nous créons environ 2200 pétaoctets de données *chaque jour*. Cet énorme volume représente 2 millions de recherches traitées par Google chaque minute, 4 000 heures de vidéos téléchargées sur YouTube chaque heure et 144 milliards d'e-mails envoyés partout dans le monde chaque jour. Cela équivaut à la totalité du contenu de la Bibliothèque du Congrès américain traversant la toile toutes les 10 secondes !

Dans ce chapitre, nous abordons différents types de données et ce dont nous avons besoin pour les stocker et les rechercher.

Plonger dans l'univers des données

Les données peuvent prendre de nombreuses formes, telles que des sons, des images, des vidéos, des codes-barres, des transactions financières, entre autres, et elles sont réparties en de multiples catégories : données structurées ou non structurées, qualitatives ou quantitatives, discrètes ou continues.

Comprendre les données structurées et non structurées

Indépendamment de leurs sources, les données sont généralement divisées en deux types, à savoir les données structurées ou non structurées :

- **Les données non structurées** sont des informations qui ne présentent souvent pas de modèle de donnée prédéfini ou qui peuvent difficilement être intégrées dans des tableaux ordonnés ou des tableurs. Dans le monde des affaires, les informations non structurées sont souvent lourdes en texte et peuvent contenir des données telles que des dates, des chiffres et des faits. Les fichiers image, vidéo et audio sont souvent qualifiés de non structurés, même s'ils présentent souvent une certaine organisation ; le manque de structure rend la compilation fastidieuse pour faire de l'analyse.
- **Les données structurées** font référence aux informations organisées comme les données commerciales au sein d'une base de données relationnelle. Les ordinateurs peuvent facilement effectuer des recherches et les organiser sur la base de divers critères. Les informations reprises sur un code-barres peuvent sembler méconnaissables pour l'œil humain, mais elles sont hautement structurées et facilement lisibles par les ordinateurs.

Données semi-structurées

Si les données non structurées sont facilement compréhensibles pour les êtres humains et les données structurées destinées aux machines, un grand nombre de données se trouve au milieu de ces deux catégories !

Les e-mails d'un directeur des ventes peuvent être triés par date, heure ou taille, mais s'ils sont totalement structurés, ils peuvent aussi être triés par opportunité de vente ou projet client. Cela reste toutefois difficile, car souvent, les gens n'écrivent pas sur un sujet précisément, même dans un e-mail ciblé. Cependant, le même directeur des ventes peut disposer d'un tableur reprenant les données de ventes actuelles, organisé par client, produit, heure ou date, ou une combinaison de ces points de référence.

Les données peuvent donc présenter diverses caractéristiques :

- Les ***données qualitatives*** sont généralement des informations descriptives et sont souvent subjectives. Par exemple, Bernard Dupont est un homme portant un jean et un T-shirt marron.
- Les ***données quantitatives*** sont des informations numériques et peuvent être soit discrètes, soit continues.
    - Les ***données discrètes*** relatives à Bernard Dupont sont qu'il a deux bras et est le fils de Gilbert Dupont.
    - Les ***données continues*** sont que Bernard Dupont pèse 70 kilos et mesure 1 m 72.

En bref, les données discrètes sont comptées et les données continues sont mesurées.

Si vous voyiez une photo de Bernard Dupont, vous verriez les données structurées sous la forme d'une image, mais c'est votre capacité à estimer son âge, la nature des habits et la perception des couleurs, qui vous permet de réaliser une évaluation *qualitative*. Toutefois, la taille et le poids de Bernard ne peuvent être correctement *quantifiés* que par le biais d'une prise de mesures et ces deux facteurs varieront tout au long de sa vie.

Données audio et vidéo

Un fichier audio ou vidéo présente une structure, mais son contenu comporte également des informations qualitatives, quantitatives et discrètes.

Prenons le fichier de la chanson « Poker Face » de Lady Gaga:

- Les données quantitatives sont que le morceau est une chanson de musique pop chantée par une chanteuse.
- Les données quantitatives continues sont que le morceau dure 3 minutes et 43 secondes et que la chanson est en anglais.
- Les données quantitatives discrètes sont que la chanson a été vendue en 13,46 millions d'exemplaires depuis le 1er janvier 2009. Cependant, ces données ne sont découvertes que via des analyses des données de vente compilées à partir de sources extérieures et pourraient augmenter avec le temps.

Données brutes

Dans le cas de Bernard Dupont ou de la chanson « Poker Face », divers éléments de données ont été traités à partir d'un fichier image ou audio. Toutefois, bon nombre de données sont brutes ou non traitées et consistent surtout en une série de chiffres ou de caractères.

Un météorologiste peut recueillir des données relatives à la température, à l'humidité, à la direction du vent et aux précipitations, mais ce n'est qu'une fois ces données brutes traitées et replacées dans leur contexte qu'elles peuvent être transformées en informations, telles que les probabilités qu'il pleuve ou qu'il neige.

Créer, exploiter et stocker des données

Les informations générées par les systèmes informatiques sont généralement créées à la suite de la réalisation

d'une tâche. La création de données requiert souvent des intrants, un traitement, puis un extrant. Par exemple, dans votre supermarché local, le caissier, qui scanne les codes-barres de chaque article, collecte des données sur ces codes-barres, qui sont lues par le scanner laser placé sur la caisse enregistreuse. Ce processus est relié à un système informatique à distance récupérant les prix et les descriptions, qui sont renvoyés vers la caisse enregistreuse pour impression sur le ticket de caisse. Enfin, le total est calculé et d'autres données, comme celles relatives aux cartes de fidélité, peuvent à leur tour être traitées par la caisse enregistreuse pour calculer les réductions éventuelles. Cet ensemble de tâches est commun dans les systèmes informatiques suivant une méthodologie d'intrant, de traitement et d'extrant.

Tirer parti des données

Ce supermarché dispose peut-être de 10 caisses enregistreuses et la société compte peut-être 10 magasins dans la même ville et des centaines d'autres ailleurs dans le pays. Toutes les données issues de chaque caisse et magasin sont envoyées au siège social, où d'autres systèmes informatiques traitent ces données de ventes pour calculer les stocks et effectuer les commandes.

Les informations financières émanant de tous ces magasins peuvent être envoyées à d'autres systèmes pour calculer les bénéfices et les pertes ou pour aider le département des achats à déterminer quels articles se vendent bien et ceux peu populaires. Le flux de données peut ensuite être transmis soit aux départements marketing, qui prévoient des offres spéciales pour les produits rencontrant peu de succès, soit aux fabricants qui peuvent alors décider de modifier le produit.

Dans l'exemple de la chaine de supermarchés, les données requièrent quatre activités principales :

- La collecte
- La transmission
- Le stockage
- L'analyse

Le stockage des données

Seule la moitié des 7 milliards d'habitants de la planète surfe sur Internet, donc le volume déjà colossal de données numériques augmentera rapidement à l'avenir. Les informations traditionnelles stockées sur des médias physiques, comme les pellicules de film, les livres et les radiographies peuvent être rapidement transformées en un équivalent totalement numérique pouvant être utilisé par les appareils informatiques via les réseaux de communication.

Des données sont créées, traitées et stockées sans arrêt :

- Passer un coup de téléphone, utiliser un distributeur de billets et même faire le plein d'une voiture à la station service sont des activités qui génèrent des kilo-octets d'informations.
- Regarder un film sur Internet demande 1 000 mégaoctets de données.
- Facebook ingère plus de 500 téraoctets de nouvelles données chaque jour.

Des volumes massifs de données doivent être stockés pour des recherches ultérieures. Il peut s'agir de réseaux de télévision souhaitant diffuser un film en particulier, d'agences de presse souhaitant retrouver d'anciens articles sur un penseur français ou encore d'institutions de recherches scientifiques ayant besoin d'examiner d'anciennes cartes aériennes de forêt pour mesurer le taux de déforestation. D'autres organisations peuvent avoir besoin de conserver des dossiers de patients ou des dossiers financiers conformément aux réglementations gouvernementales. Ces données ne requièrent souvent pas d'outils analytiques ou d'autres instruments spécifiques pour découvrir la valeur de l'information. La valeur d'un film, d'une photographie ou d'une carte aérienne se voit tout de suite.

D'autres supports demandent une analyse plus approfondie, afin de déterminer leur valeur : Des pétaoctets d'informations critiques telles que des études géologiques, des images satellites et des résultats d'essais cliniques déferlent sur les réseaux. Ces grands ensembles de données contiennent des informations qui peuvent aider les entreprises à trouver de

nouvelles réserves de ressources naturelles, prédire des tempêtes et mettre au point des traitements de pointe contre le cancer.

Il s'agit du Big Data et la frénésie qui l'entoure porte à la fois sur le stockage et le traitement des ensembles de données brutes nécessaires pour en tirer des avantages tangibles, mais nous couvrirons ce sujet plus en détail au chapitre 4.

Connaître le potentiel et les risques

La croissance massive de la quantité de données électroniques ouvre un grand champ des possibles, comme notamment des avancées majeures dans le domaine scientifique, l'environnement, l'industrie et même de nouveaux modèles commerciaux.

Les données sont précieuses lorsqu'elles sont en de bonnes mains, mais elles intéressent également les criminels, les concurrents industriels, les terroristes ou les états. Que ces données consistent en des appels téléphoniques passant par des réseaux de communication internationaux, en des informations de profil ou des mots de passe sur les médias sociaux, sites de vente en ligne ou même des informations sensibles relatives à des découvertes scientifiques, les données font constamment l'objet d'attaques. Des citoyens, des organisations et même des pays entiers définissent des réglementations et de bonnes pratiques sur la manière de protéger les données et ainsi la vie privée et la confidentialité. Presque tous les grands secteurs disposent de plusieurs réglementations pour gérer la sécurité et la confidentialité des données. Ces lois couvrent généralement :

- La collecte
- Le traitement
- La transmission
- Le stockage
- Le partage
- La destruction

Sécurité et conformité des données

L'une des lois sur la sécurité des données les plus souvent rencontrées concerne les données des cartes de crédit. Ces lois sont définies selon les dispositions du Payment Card Industry (PCI) utilisées par les principaux émetteurs de cartes de crédit pour protéger les informations personnelles de leurs clients et assurer la sécurité des transactions traitées lors de l'utilisation de cartes de paiement. La majorité des institutions financières mondiales doivent se conformer à ces normes si elles souhaitent traiter des paiements par carte de crédit. Sinon elles risquent des amendes et de perdre leur autorisation.

Les principes du PCI ont pour objet de :

- Maintenir une politique de sécurité de l'information
- Protéger les données sensibles par le biais du cryptage
- Mettre en œuvre des mesures strictes de contrôle de l'accès
- Contrôler et tester de manière régulière les réseaux et les systèmes

Chapitre 2

Comment le Big Data peut-il aider votre organisation

Dans ce chapitre

- Respecter les 3 V : volume, vélocité et variété
- Surmonter divers problèmes liés au Big Data
- L'analyse Big Data
- Diviser de grands projets en tâches plus restreintes à l'aide de Hadoop

Le monde est submergé de données numériques et une fois transformées en informations, elles peuvent nous aider dans presque toutes les facettes de notre vie quotidienne. Pour faire simple, on parle de Big Data lorsque les machines et les logiciels informatiques traditionnels ne sont plus à même de contenir, gérer la croissance rapide des données et de protéger de grands volumes, ou lorsqu'ils ne sont plus capables de fournir les renseignements voulus dans des délais raisonnables.

Dans ce chapitre, nous abordons l'aspect « analytique ». Il s'agit d'une méthode d'extraction de nouveaux renseignements et d'informations à partir de la masse de données disponibles. Comme lorsque l'on recherche une aiguille dans une botte de foin, les projets d'analyses du Big Data peuvent consister dans un premier temps à trouver la bonne botte de foin !

Nous parlerons également de Hadoop, un cadre de programmation qui divise les grands projets en tâches plus restreintes.

Identifier les besoins en matière de Big Data

Le terme Big Data est utilisé depuis le début du millénaire et a été proposé pour la première fois par des analystes de Technology Research Gartner avec trois dimensions. Ces paramètres du Big Data sont :

- **Le volume :** Quantité de données très grande ou en augmentation permanente.
- **La vélocité :** La vitesse d'entrée et de sortie des données.
- **La variété :** L'éventail des types et des sources de données.

Ces 3 V (volume, vélocité et variété) caractérisent le Big Data, mais il faut avant tout déterminer si ces données peuvent être traitées pour fournir des renseignements précis et permettre de prendre des décisions avisées dans des délais raisonnables.

Exemples évidents de problèmes liés au Big Data :

- **Un studio de cinéma** produit et stocke un large éventail de productions et réalisations cinématographiques, à partir de séquences brutes non traitées, dans divers formats post-traitement destinés par exemple aux cinémas standard, IMAX, 3 D, à la télévision haute définition, aux Smartphones et aux systèmes de divertissement des compagnies aériennes. Les formats doivent être localisés dans des dizaines de langues, sont de longueurs différentes et doivent être conformes aux normes de censure de chaque pays.
- **Une organisation de soins de santé** inclut dans le dossier d'un patient chaque remarque du médecin, les résultats sanguins, les radiographies, les IRM, les échographies et autres captures d'images médicales tout au long de la vie de ce patient, sachant que des centaines, des milliers, voire des millions de patients sont gérés par cette organisation.
- **Un cabinet d'avocats travaillant sur un important recours collectif** doit non seulement récolter

d'énormes quantités de documents électroniques (e-mails, calendriers électroniques et formulaires), mais également les classer en fonction des éléments du dossier. La capacité à trouver rapidement des modèles, des chaînes de communication et des liens est vitale pour prouver les responsabilités des parties.

- **Pour une entreprise d'ingénierie aérospatiale,** tester les performances, le rendement énergétique et la tolérance d'un nouveau moteur à réaction est un projet de Big Data ambitieux. La construction de prototypes est onéreuse, donc la possibilité de créer une simulation informatique et d'entrer des données pour chaque décollage imaginable, chaque modèle de vol et chaque atterrissage dans des conditions météorologiques différentes permet de réaliser d'importantes économies.
- **Un service de sécurité nationale** utilise un logiciel de reconnaissance faciale pour analyser rapidement les images de plusieurs heures de vidéosurveillance, afin de trouver un fugitif. Il s'agit d'un autre exemple de problème concret pouvant être abordé dans le cadre du Big Data. Assigner cette tâche à des personnes est trop coûteux et l'automatisation exige de résoudre de nombreux problèmes relatifs au Big Data.

Pas vraiment du Big Data ?

Donc, qu'est-ce qui *ne relève pas* du Big Data ? Un directeur des ventes régional tentant de déterminer combien de robes en taille 44 ont été achetées dans un magasin particulier le jour du réveillon de Noël est-il confronté à un problème relevant du Big Data ? Non : cette information est enregistrée par les systèmes de contrôle des stocks du magasin, dans la mesure où chaque article est scanné et payé à la caisse. Bien que la base de données comprenant tous les achats puisse être volumineuse, les informations sont relativement faciles à trouver à partir de la bonne base de données.

Mais . . .

Si l'entreprise souhaite déterminer quel style de robes est le plus populaire auprès des femmes de plus de 30 ans, ou si certaines robes ont fait grimper les ventes d'accessoires, ces

informations peuvent requérir des données supplémentaires en provenance de multiples magasins, de cartes de fidélité ou de sondages, ce qui demande des calculs poussés pour déterminer les bonnes corrélations. Si ces informations doivent être obtenues en urgence pour la campagne de marketing mode de la saison printemps, le problème peut relever du Big Data.

Vous n'êtes pas confronté à un problème de Big Data si :

- Les informations dont vous avez besoin sont déjà rassemblées au sein d'un seul tableur.
- Vous pouvez trouver la réponse à une requête au sein d'une seule base de données et cela vous prend quelques minutes et non plusieurs jours.
- Le stockage et le traitement des informations sont gérés à l'aide d'outils informatiques traditionnels pour un volume modéré de données.

L'analyse de données

L'analyse de données consiste à examiner des données pour y trouver une information ou des renseignements utiles. Le but premier est d'aider les entreprises à prendre de meilleures décisions en permettant à des spécialistes et à d'autres utilisateurs d'analyser d'énormes volumes de données de transaction, ainsi que d'autres sources de données auxquelles des programmes intelligents n'ont pas accès.

Ces autres sources de données peuvent comprendre des journaux de serveur Web et des données de navigation, des rapports d'activité sur les médias sociaux, des enregistrements d'appels passés à partir de téléphones portables et des informations issues de capteurs. Tout comme les données non structurées du même type, les grands systèmes de traitement des transactions et d'autres données hautement structurées sont des types de Big Data pouvant faire l'objet d'analyse de données.

Dans de nombreux cas, le principal critère n'est pas le caractère structuré ou non des données, mais bien de savoir si le problème peut être résolu rapidement et de manière rentable !

Le problème est généralement lié à la capacité de gérer les 3 V (volume, vélocité et variété) des données suffisamment rapidement pour en tirer un avantage. Analysons à présent quelques problèmes analytiques plus en détail.

Un « petit » problème de Big Data

Le gérant d'une cantine scolaire doit accroître ses recettes de 10 %, tout en continuant d'offrir un repas sain aux 1000 étudiants qui prennent leur déjeuner à la cafétéria chaque jour. Les étudiants payent un certain montant pour leur déjeuner, ce montant changeant chaque jour.. Le gérant pourrait tout simplement accroître le coût des plats de 10 %, mais une telle mesure pourrait pousser les étudiants à apporter leur propre casse-croûte. Le gérant décide d'utiliser le broyage de données pour trouver une solution.

- **La première étape** est la création d'un tableau reprenant le nombre de portions de chaque plat qui ont été préparées, quels plats ont été achetés chaque jour et le coût total de chacun de ces plats.
- **La deuxième étape** est une analyse portant sur l'année précédente, où le gérant découvre que les étudiants préfèrent les sandwiches, les paninis et les croque-monsieurs mais qu'ils ne sont pas friands de hotdogs et de lasagnes. En réalité, 30 % des hotdogs étaient jetés à la poubelle !
- **Les résultats** de cette analyse suggèrent qu'en remplaçant simplement les hotdogs par des sandwiches, la cantine pourrait enregistrer une hausse de revenus de 10 %.

Un problème « moyen » de Big Data

Un site de vente en ligne de matériel de bricolage ne sait plus comment faire pour accroître le montant et la fréquence des commandes, surtout depuis que la concurrence s'est faite plus intense dans le secteur. Le directeur des ventes décide qu'une analyse des données est un bon point de départ.

- **La première étape** consiste à créer une base de données reprenant les produits, les clients et les commandes de l'année précédente. 200 000 produits ont été commandés à l'entreprise par environ 20 000 clients. Elle envoie par ailleurs un e-mail publicitaire ciblé chaque mois avec des offres spéciales et dispose d'un programme de fidélisation qui prévoit des ristournes sur la base d'un système de points.
- **La deuxième étape** est de parvenir à mieux cerner la clientèle en ajoutant les profils que les clients ont complétés lors de la procédure d'inscription au programme de fidélité. L'âge, le sexe, le statut marital, le nombre d'enfants et la profession sont les renseignements demandés lors de cette procédure. Le directeur des ventes peut à présent analyser la manière dont certaines données démographiques influencent les ventes à l'aide de références croisées.
- **La troisième étape** consiste à utiliser un logiciel d'analyse de tendance qui détermine que 10 % des clients achètent plus volontiers du papier lorsqu'ils achètent de la peinture. En outre, les détenteurs de cartes de fidélité qui ont des enfants achètent souvent plus d'articles en vrac au début de l'année scolaire.
- **Les résultats** obtenus en effectuant des renvois entre de multiples bases de données et en les comparant à l'efficacité des diverses campagnes permettent au directeur des ventes de créer des rappels « articles proposés » sur son site Internet. De plus, les campagnes marketing ciblant les parents peuvent être plus efficaces.

Un « gros » problème de Big Data

En tant que directrice de la détection des fraudes pour une entreprise de carte de crédit, Sarah tente de repérer d'éventuelles transactions frauduleuses parmi des millions d'activités financières réalisées chaque jour. Sa marge de manœuvre est limitée par plusieurs facteurs, dont la nécessité de discrétion vis-à-vis des clients, la capacité du commerçant à vendre rapidement ses produits et des restrictions légales relatives à l'accès aux données personnelles. La situation se complique encore avec des reglementations spécifiques, les différences culturelles et la distance géographique.

La détection des fraudes à la carte de crédit est un problème de Big Data qui correspond au critère des 3 V : un important volume de données variées qui circulent avec une grande vélocité. Les données parviennent dans le système de détection des fraudes à partir d'un très grand nombre de sources et elles doivent ensuite être analysées en l'espace de microsecondes pour éviter les tentatives de fraude, pour ensuite être de nouveau analysées dans le but de mettre au jour des tendances plus généralisées ou des crimes organisés.

Hadoop : Passer au traitement parallèle des données

Même les ordinateurs les plus évolués peinent à résoudre des problèmes complexes présentant de nombreuses variables et de grands ensembles de données. Imaginez qu'une personne doive trier 26 000 boîtes avec de grandes boîtes contenant chacune 1000 autres boîtes marquées d'une lettre de l'alphabet : cette tâche prendrait des jours entiers. Mais si vous séparez le contenu des 1000 boîtes en 10 boîtes égales plus petites et demandez à 10 personnes d'effectuer ces tâches plus simples, le travail sera réalisé 10 fois plus vite. La notion du *traitement parallèle* est l'une des pierres angulaires de nombreux projets de Big Data.

Apache Hadoop (d'après le nom du jouet en forme d'éléphant du fils du créateur, Doug Cutting) est un cadre de programmation gratuit qui permet de traiter de grands ensembles de données dans un environnement informatique réparti. Hadoop fait partie du projet Apache sponsorisé par la Apache Software Foundation et s'il utilisait Java à l'origine, tout langage de programmation peut être utilisé pour gérer les diverses parties du système.

Hadoop s'inspire du MapReduce de Google, un modèle d'architecture logicielle, où une application est divisée en de nombreuses parties plus petites. Chacune de ces parties (également appelées fragments ou éléments) peut être traitée sur tout ordinateur connecté à un groupe organisé appelé grappe (ou « cluster »). Hadoop permet de lancer des applications sur des milliers d'ordinateurs individuels avec des milliers de téraoctets de données. Son système de fichier réparti facilite des taux de transfert de données rapides entre

les nœuds et permet au système de continuer de fonctionner sans interruption en cas de défaillance du nœud. Cette approche réduit les risques de défaillance système, même si un nombre significatif d'ordinateurs cesse d'être opérationnel.

Premiers secours : le Big Data, une aubaine pour les hôpitaux

Le Boston Children's Hospital a été confronté à des problèmes de stockage avec son réseau de stockage (SAN) traditionnel lorsque les nouvelles technologies ont fait croître rapidement et de manière imprévisible les informations dont dépendaient les chercheurs.

Les chercheurs travaillent sur de nouveaux traitements pour les enfants gravement malades et ils ont besoin de données disponibles immédiatement, à tout moment et en tout endroit.

Pour résoudre les problèmes liés à la croissance rapide des données traitées dans ses opérations de sauvegarde informatique globale, le Boston Children's Hospital a adopté le logiciel SyncIQ de réplication asynchrone des données d'Isilon pour reproduire ses informations de recherches entre deux clusters EMC Isilon.

Cette démarche lui a permis de gagner du temps et de l'argent, de renforcer la fiabilité des données en général et d'éliminer totalement l'impact des données de recherches sur les opérations de sauvegarde informatique globale. Le pool partagé unique de stockage permet aux chercheurs d'accéder immédiatement et à tout moment au volume massif d'archives de données et requiert nettement moins de personnel équivalent temps plein pour le support informatique.

Grâce à EMC Isilon, les chercheurs du Boston Children's Hospital disposent toujours des données dont ils ont besoin, quand ils en ont besoin, ce qui leur permet de faire avancer leur travail sur les maladies infantiles sans interruption.

Chapitre 3

Créer une infrastructure efficace pour le Big Data

Dans ce chapitre

- Comprendre le stockage « scale-up » et « scale-out » de données
- Comprendre le cycle de vie des données peut impacter le stockage
- Utiliser des données actives et inactives

Que les données numériques soient structurées, non structurées, quantitatives ou qualitatives (voir le chapitre 1 pour un rappel de la signification de ces termes si nécessaire), elles doivent toutes être stockées quelque part. Ce stockage peut durer une milliseconde ou toute une vie, en fonction de la valeur des données, de leur utilité, de leur pertinence ou de vos exigences personnelles.

Dans ce chapitre, nous abordons la question du stockage dans le cadre du Big Data. Le stockage du Big Data est composé d'architectures modernes qui ont évolué avec Facebook, les compteurs intelligents et Google Maps. Ces architectures ont été conçues dès le départ pour passer facilement de quantités modérées à massives de données, et ce de manière modulaire.

Les bases du stockage de données

Gardez les points suivants à l'esprit lorsque vous envisagez le stockage de Big Data.

- **Les données sont créées par des actions ou via des procédés.** Généralement, les données proviennent d'une source ou d'une action. Ensuite, elles circulent entre des magasins de données et des clients consommateurs de données. Un magasin de données peut être une grande base de données ou des archives de documents et les clients peuvent être des outils de productivité, des environnements et des cadres de développement, des outils de planification de ressources des entreprises (PRE), des outils de gestion de relation clientèle (CRM) et des systèmes de gestion des contenus web (CMS).
- **Les données sont stockées sous divers formats.** Les bases de données relationnelles font partie des formats les plus souvent rencontrés et sont disponibles dans de nombreuses configurations différentes. Les fichiers texte et numériques, les fichiers XML, les tableurs et les divers types de stockage fermé constituent d'autres types de données, présentant tous leurs propres méthodes d'indexation et d'accès aux données.
- **Les données circulent autour et entre les organisations.** Les données ne se limitent pas à une seule organisation et sont partagées ou regroupées à partir de sources qui se trouvent en dehors du contrôle direct de l'utilisateur. Par exemple :
    - Une compagnie d'assurances automobiles qui calcule une prime d'assurance doit consulter la base de données de la préfecture qui gère les permis de conduire, pour s'assurer que la personne souhaitant une couverture est habilitée à conduire.
    - Ce même assureur effectue une vérification de solvabilité auprès d'autres entités.
    - Les données émanant de ces recherches sont cruciales, mais dans certains cas, l'assureur n'a pas le droit de détenir cette information plus de quelques secondes, afin qu'il ait juste le temps de créer sa police d'assurance. En fait, la rétention prolongée de ces données peut constituer une infraction aux réglementations en vigueur.
- **Les flux de données sont uniques.** La manière dont circulent les données à travers une organisation dépend de l'environnement, des procédures de fonctionnement,

du secteur et même des règlementations en vigueur. Toutefois, quelle que soit l'organisation concernée, la structure de la technologie sous-jacente, des systèmes de stockage, des éléments de traitement et des réseaux qui caractérisent ces flux est souvent très similaire.

Scale-up ou Scale-out ? Examen des options pour le stockage des données

Le stockage d'importants volumes de données numériques est un problème majeur pour les organisations de toutes tailles et de toutes sortes. L'évolution technologique qui s'est opérée depuis les balbutiements du stockage de données sur les premiers disques magnétiques au début des années 1960 est phénoménale. L'unité de disque demeure la technologie de stockage la plus répandue, mais la manière de l'utiliser a radicalement changé pour répondre à de nouvelles exigences. Les deux principales tendances sont le « scale-up », qui consiste à acheter un plus grand système de stockage, et le « scale-out », qui revient à acheter des systèmes multiples pour ensuite les regrouper.

Imaginez que vous lancez l'entreprise Rapido Orange dont le travail est de livrer des palettes d'oranges :

- **Scale-up :** Vous achetez un grand entrepôt pour réceptionner et stocker les oranges livrées par l'agriculteur, ainsi qu'un grand camion. Mais votre entreprise continue de croître. Vos clients existants et nouveaux exigent des temps de livraison plus courts ou un plus grand volume d'oranges livrées chaque jour. L'option du scale-up consiste à acheter un plus grand entrepôt et un plus grand camion, afin de pouvoir gérer plus de livraisons.

  Cette option peut être rentable dans un premier temps, lorsque l'entreprise ne dispose que de quelques grands clients locaux. Toutefois, cette démarche présente plusieurs dangers, comme l'éventualité d'un incendie dans l'entrepôt ou d'une panne du camion. Dans ces cas-là, personne ne reçoit d'oranges. En outre, une fois

que l'entrepôt et le camion ont de nouveau atteint leur capacité maximale, les livraisons à effectuer chez ne fut-ce que quelques clients supplémentaires requièrent un investissement majeur.

- **Scale-out :** Vous achetez quatre petits dépôts régionaux pour réceptionner et stocker les oranges de l'agriculteur. Vous achetez également quatre camionnettes plus rapides capables de transporter de nombreuses petites palettes chez chaque client. Mais votre entreprise continue de croître. L'option du scale-out consiste à acheter plusieurs dépôts régionaux supplémentaires plus proches des clients, ainsi que de nouvelles camionnettes.

  Grâce au scale-out, si l'un des dépôts prend feu ou si l'une des camionnettes tombe en panne, les autres pans de l'activité sont sauvés et le commerçant peut continuer à livrer ses oranges, tout en ayant peut-être même la possibilité d'absorber les pertes sans changer ses volumes de livraison et sans mettre ses clients dans l'embarras. Au fil des nouvelles opportunités commerciales, l'entreprise peut continuer sur cette voie en augmentant le nombre de dépôts et de camionnettes de manière flexible avec moins de dépenses.

Avec ces deux options, l'entreprise Rapido Orange peut accroître ses capacités et la performance de ses opérations. Il n'existe pas de règle absolue sur la manière de choisir l'une ou l'autre option, dans la mesure où tout dépend de la situation.

Les architectures de type scale-up pour les données numériques peuvent mieux convenir à des applications très structurées, volumineuses et prévisibles, telles que des bases de données, tandis que les systèmes de type scale-out peuvent mieux convenir aux flux en pleine croissance, moins prévisibles et non structurés, comme les journaux de requêtes sur Internet ou de grandes quantités de fichiers image. Consultez le tableau 3-1 pour déterminer lequel de ces systèmes vous conviendrait le mieux.

De nombreuses organisations utilisent les deux méthodes pour répondre à des exigences différentes. En bref, pour l'entreprise Rapido Orange, l'adoption d'une double approche pourrait être de disposer d'un grand entrepôt central qui alimente les petits dépôts à l'aide de grands camions, tandis que le réseau de dépôts régionaux continue de croître avec de

nouvelles petites installations et de nouvelles camionnettes pour les livraisons.

Tableau 3-1 Scale-out ou Scale-up ?

Scale-out	*Scale-up*
La quantité de données que nous devons stocker à des fins de traitement croît de plus de 20 % par an	Nos données n'augmentent pas de manière significative
Le système de stockage doit prendre en charge un grand nombre d'appareils qui accèdent simultanément au système	La plupart de nos données sont stockées au sein d'une grande base de données qui est hautement optimisée pour notre charge de travail
Les données peuvent être réparties dans de nombreuses machines et regroupées lorsqu'une recherche est nécessaire.	Toutes les données sont synchronisées vers un répertoire central
Nous préférerions un accès plus lent que pas d'accès du tout en cas de problème mineur	Les exigences d'accès à nos magasins de données sont très prévisibles
Nos données sont principalement non structurées, présentes en grands volumes avec des taux d'accès très imprévisibles	Les ensembles de données sont très structurés et relativement petits

Comprendre le cycle de vie des données pour aboutir à un meilleur stockage

Peu importe d'où proviennent les données, où elles sont traitées et où elles sont stockées, elles présentent toujours une durée de vie utile. Une vidéo numérique du mariage d'un proche doit être conservée pour toujours. Toutefois, le code à trois chiffres indiqué à l'arrière d'une carte de crédit utilisé à des fins de vérification ne doit jamais être stocké dans les registres de ventes d'un commerçant après le traitement.

Les données en temps réel doivent être disponibles rapidement

Certaines données sont essentielles pour des analyses en temps réel et doivent donc être presque instantanément rendues disponibles pour d'autres systèmes ou utilisateurs. Par exemple, un officier de police sur le point de donner une contravention doit savoir rapidement si la plaque d'immatriculation de la voiture interceptée est connectée à une éventuelle infraction.

L'accessibilité et le stockage à long terme de données sont très importants en termes de coûts. En général, les données qui sont consultées fréquemment ou de manière continue dans le cadre du fonctionnement d'une entreprise ou d'autres opérations requièrent un matériel plus performant et des spécifications plus poussées que le stockage de données inactives consultées moins souvent.

Voir l'encadré « Stockage de données en temps réel : Jaguar Land Rover » pour un exemple de stockage de données en temps réel.

Gérer des données utilisées moins fréquemment

L'archivage des données consiste à déplacer des données qui ne sont plus activement utilisées dans un dispositif de stockage séparé pour une conservation à long terme. Les archives de données sont composées de données plus anciennes qui restent toujours importantes et nécessaires pour l'avenir, ainsi que des données qui doivent être conservées à des fins règlementaires. Les archives de données sont indexées, dotées de capacités de recherches, afin que les fichiers et les éléments de fichiers puissent être facilement localisés et retrouvés. Voir l'encadré « Stockage de données d'archive : HathiTrust » pour un exemple d'archivage de données.

Stockage de données en temps réel : Jaguar Land Rover

Jaguar Land Rover conçoit, crée et produit certains des véhicules les plus prisés au monde et son succès dépend de sa politique d'innovation.

Dans le cadre de ses activités de conception et de production, les ingénieurs de Jaguar Land Rover utilisent des procédés d'ingénierie assistée par ordinateur (IAO) notamment à des fins de simulation. Mais la réalisation de modèles coûte de l'argent et prend du temps. Jaguar Land Rover souhaitait disposer d'un procédé innovant qui lui permettrait de renforcer son efficacité, sa flexibilité et sa rentabilité, tout en réduisant les délais de mise sur le marché.

Pour relever ce défi, l'entreprise devait repenser son infrastructure informatique à l'aide d'un environnement informatique à haute performance (HPC) qui gérerait les simulations virtuelles pour tous ses ingénieurs.

Les simulations virtuelles de Jaguar Land Rover génèrent plus de 10 To de données par jour et l'entreprise utilise les capacités de stockage scale-out d'EMC Isilon X-Series pour augmenter la capacité de sa configuration de stockage originale de 500 To. En six mois, grâce à EMC Isilon, l'environnement HPC a augmenté de plus de 250 %. La capacité de stockage a crû de plus de 500 % et l'architecture de gestion du réseau a été multipliée par dix.

Les programmes de simulation virtuelle, gérés par les technologies d'EMC Isilon, permettent aux équipes d'examiner les problèmes plus en détail, d'aisément tester de nouvelles idées et de réaliser des changements bien plus rapidement que jamais auparavant. Les ingénieurs peuvent à présent créer des images en 3D et relever les défis survenant avant la réalisation des prototypes, ce qui permet de réduire les coûts de manière significative.

Dans la mesure où les équipes ont rapidement accès à des centaines de TB d'itérations de conception sur le système d'EMC Isilon, ils peuvent revoir de nouvelles idées en quelques jours et envisager de nouveaux modèles avant la réalisation des prototypes. À présent, Jaguar Land Rover réalise des simulations dès les premières phases de son travail, avant même que les données géométriques et de conception aient été créées. L'équipe peut consulter les informations en temps réel pour comprendre le processus de la simulation et décider d'apporter des modifications si nécessaire.

Stockage de données d'archive : HathiTrust

En 2008, l'université du Michigan (U-M) en collaboration avec le Committee on Institutional Cooperation (CIC) s'est lancée dans un grand projet visant à collecter et à conserver un répertoire numérique partagé de connaissances intitulé HathiTrust.

L'objet premier de ce partenariat était de préserver et de donner accès au contenu des collections de la bibliothèque des partenaires, composées de livres numériques et de revues. Le principal défi consistait à créer une infrastructure de stockage des données suffisamment solide pour prendre en charge plus de 10 millions d'éléments numériques et de gérer l'évolution rapide de ce projet ambitieux.

Le système NAS scale-out d'EMC Isilon constitue le principal répertoire de la bibliothèque numérique HathiTrust. En partenariat avec Google, entre autres, HathiTrust est parvenu à numériser plus de 10,5 millions d'ouvrages — 3,6 milliards de pages — à partir des bibliothèques du partenariat, en vue de créer un immense répertoire numérique de matériel représentant plus de 470 téraoctets.

Les données actives et d'archive sont tout aussi importantes

De nombreux projets de Big Data utilisent à la fois des données actives et d'archive pour fournir des renseignements. Par exemple, les données actives ou en temps réel émanant de la bourse peuvent aider un trader à acheter ou à vendre des titres, tandis que les données d'archive relatives à la stratégie à long terme d'une entreprise, la croissance du marché et ses produits sont utiles pour mieux gérer un portefeuille de manière globale. Les informations en temps réel relatives aux indices boursiers doivent être disponibles dès que possible, mais les rapports plus anciens ou les tendances de marché peuvent être récupérés dans un fichier d'archive et être analysés par la suite.

Un accès plus rapide aux données coûte généralement plus cher

Pour faire simple, les données en temps réel, actives ou continues qui permettent des prises de décision rapides sont généralement conservées sur les supports de stockage les plus rapides. Souvent, plus le support est rapide, plus il est cher en comparaison avec la capacité disponible. C'est ce que l'on appelle le « Coût par gigaoctet (Go) ou téraoctet (To) ». Ces différents types de performance de stockage et coûts sont souvent considérés comme des classes de stockage distinctes, comme le montre l'encadré 3-1.

Plus de rapidité = Coûts plus élevés par Go/To
RAM : mémoire de l'ordinateur qui requiert une alimentation constante pour conserver les informations.
SSD : un disque à circuits intégrés qui utilise un type de RAM pour stocker les données.
Disques SATA/SAS : deux types de support magnétique, SAS étant la technologie la plus rapide.
Banque magnétique : support magnétique qui ne requiert aucune alimentation pour conserver les données indéfiniment, mais la vérification de l'intégrité des données est bien plus difficile.
Plus lent = Coûts moins élevés par Go/To

Encadré 3-1: Plus de rapidité équivaut à des coûts plus élevés par Go ou To.

Chapitre 4

Améliorer un projet de Big Data à l'aide du stockage scale-out

Dans ce chapitre

- Comprendre le concept des nœuds de stockage
- Construire un cluster de stockage
- Éviter les problèmes liés à l'infrastructure de stockage du Big Data

Les données existent sous diverses formes et présentent des exigences différentes en matière de volume, de vélocité et de variété. Les projets de Big Data peuvent nécessiter des éléments de données structurées, déstructurées, en temps réel et d'archive pour aboutir à des résultats, et toutes ces données doivent être stockées dans un endroit accessible pour les applications d'analyse.

Notre exemple de l'entreprise Rapido Orange au chapitre 3 explique les deux philosophies de base pour la construction d'une architecture de stockage, le scale-up et le scale-out, et vous pouvez utiliser l'une d'entre elles ou les deux ensemble pour parvenir au résultat voulu. Toutefois, les exigences des projets de Big Data qui gèrent à la fois des ensembles de données structurées et non structurées alliés à une vélocité, un volume et une variété de données importants demandent souvent des technologies de stockage de type scale-out.

Explorer l'architecture commune de type scale-out

De nombreuses technologies de stockage scale-out suivent un modèle architectural similaire éprouvé comme l'approche à adopter. La plupart des systèmes reposent sur une structure simple qui les rend reconnaissables comme penchant vers le scale-out ou une approche mixte scale-out/scale-up.

Le nœud de stockage ou le premier élément constituant

L'un des aspects fondamentaux de l'architecture scale-out est, comme son nom l'indique, sa capacité à agrandir l'échelle (« scale » en anglais) du stockage. Pour atteindre cet objectif, les spécialistes utilisent le principe du nœud. Un nœud est un dispositif de stockage autonome qui fonctionne avec d'autres nœuds pour stocker et véhiculer des données entre les producteurs et les utilisateurs des données.

Voici une comparaison simple pour vous aider. Un nœud est comme un seau et l'eau qu'il contient représente les données. Si vous souhaitez stocker plus d'eau, vous avez besoin de plus de seaux. À présent, vous avez plus d'eau et plus de personnes à même de puiser simultanément de l'eau dans les seaux. Toutefois, si un seau se vide, les personnes doivent patienter dans une file d'attente jusqu'à ce qu'un autre seau rempli d'eau soit disponible.

Pour résoudre ce problème, chaque seau dispose de tuyaux qui le relient aux autres, afin que plusieurs personnes puissent accéder à la source d'eau commune en même temps, à partir de différents seaux. Si vous ajoutez plus d'eau dans un seau, elle coule via les tuyaux de connexion et est distribuée de manière équitable à tous les seaux connectés, comme le montre l'image 4-1. Si un seau comporte une fuite, vous pouvez tout simplement le déconnecter des autres seaux et déplacer son tuyau vers un autre seau pendant que vous le réparez.

De nombreuses architectures scale-out découpent les fichiers et répartissent les éléments à travers les nœuds, ce qui leur permet de circuler comme de l'eau.

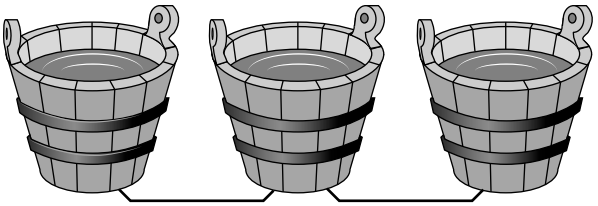

Image 4-1: Les seaux (nœuds) distribuent l'eau (données).

Attention : au sein des architectures scale-up tentant d'agir comme de réels systèmes scale-out, les fichiers sur chaque nœud ne sont pas divisés en divers éléments et répartis, donc, contrairement à l'eau, les données demeurent souvent dans le nœud avec parfois une copie envoyée à un autre client.

Fouiller à l'intérieur du nœud

Un nœud de stockage scale-out est une sorte de serveur hautement optimisé avec une application logicielle spécialement conçue pour gérer le stockage et les flux de données entre sa propre structure et les autres nœuds du groupe. Chaque nœud peut également communiquer avec les clients externes via le réseau pour stocker ou envoyer des données, comme le montre l'image 4-2.

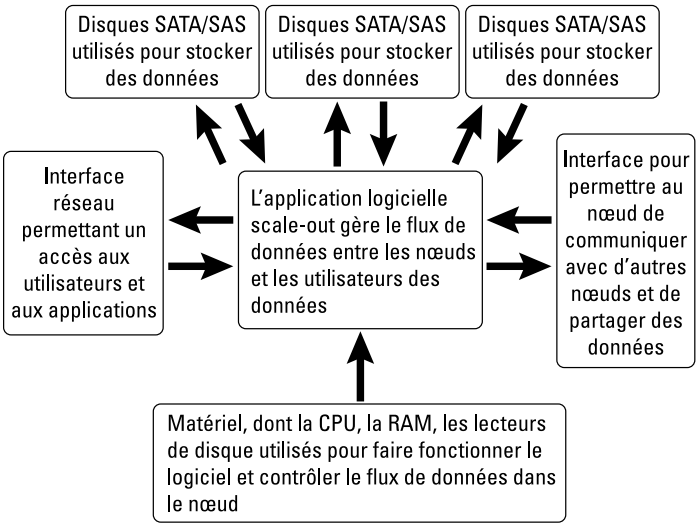

Image 4-2: Architecture scale-out.

Au sein d'une réelle architecture scale-out, chaque nœud contient une ou plusieurs unités centrales de traitement (UC), un certain volume de mémoire à accès aléatoire (RAM) et un nombre déterminé de lecteurs de disque dur. L'unité dispose en outre d'une connexion au réseau et souvent d'une méthode d'interconnexion des nœuds.

Connecter les nœuds pour former un cluster

Chaque nœud de stockage scale-out est généralement connecté de deux manières. Il présente d'abord une connexion qui le relie aux autres nœuds par le biais d'une carte réseau pour former une grappe de stockage (cluster), comme le montre l'image 4-3.

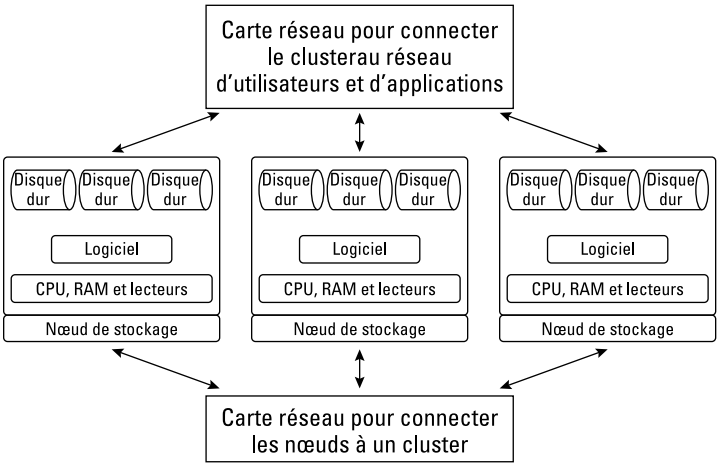

Image 4-3: Les nœuds sont reliés pour former un cluster.

Cette interconnexion permet au cluster de partager des données entre les nœuds, ce qui offre une certaine résilience aux données et permet d'accroître les performances, car chaque nœud peut fournir les données stockées dans les autres nœuds du cluster. Ces interconnexions utilisent généralement une interface de réseau rapide de type Ethernet 1 Go (giga-octet) ou 10 Go, voire de type Infiniband avec 40 Go.

Connecter le cluster au réseau dans son ensemble

Le cluster scale-out est également connecté à l'ensemble du réseau pour permettre aux applications et aux utilisateurs d'y accéder, tant localement qu'à partir du réseau plus large. La plupart des clusters scale-out (surtout dans le cas des projets de Big Data) disposent souvent de liens les plus rapides vers les applications d'analyse fonctionnant sur les serveurs d'applications. Il s'agit généralement de connexions Ethernet 1 Go et 10 Go.

Communiquer en toute confiance

Le cluster scale-out peut communiquer avec les applications de diverses manières. Ces *protocoles de communication* permettent d'accéder aux différentes normes de communication adoptées par les différents fournisseurs de systèmes d'exploitation. Par exemple :

- Common Internet File System (CIFS), également connu sous le nom de Server Message Block (SMB) est communément utilisé par les applications Windows.
- Network File System (NFS) est un système de fichiers réparti souvent utilisé avec les applications libres d'UNIX et Linux.
- File Transfer Protocol (FTP) est un protocole de réseau plus ancien souvent utilisé pour un transfert basique de fichiers.
- Hypertext Transfer Protocol (HTTP) est un protocole d'application surtout utilisé pour les applications web.

Le cluster de stockage scale-out peut souvent utiliser plusieurs systèmes de fichiers et protocoles de communication, mais certains offrent une meilleure compatibilité entre les applications, tandis que d'autres présentent un net avantage en termes de performance. Il faut trouver le juste équilibre entre la compatibilité et les performances et il est donc crucial, lors de la conception de l'architecture de stockage pour les projets de Big Data, de bien comprendre la manière dont les applications devront communiquer avec les clients, le stockage et les utilisateurs.

Comprendre les avantages et les limites potentiels

En général, les projets de Big Data présentent à la fois un grand volume, une vélocité importante et une grande variété de données. Un système de détection des fraudes peut présenter ces trois éléments, tandis qu'un système de simulation de moteur à réaction dans une soufflerie peut n'en présenter que deux. Les deux projets représentent un défi pour le système de stockage qui doit rassembler des données ou alimenter les systèmes d'analyse avec ces données.

Grâce aux architectures de stockage scale-out, il est relativement facile d'accroître les capacités et les performances à l'aide d'une architecture à nœuds basique. Chaque nœud comporte une capacité de stockage, une connectivité réseau et des unités de traitement puissantes (UC) lui permettant de recevoir, stocker et transmettre des données. Chaque nœud permet d'accroître la capacité globale du cluster.

- **Avantages :** Plus de nœuds signifie plus de capacité. Par exemple, un seul nœud dispose de 100 téraoctets de capacité de stockage, 2 Go de bande passante réseau et 256 Go de RAM pour la mise en cache des données, afin d'améliorer les taux de transfert. En ajoutant un autre nœud, ces divers éléments sont multipliés par deux. Si vous utilisez quatre nœuds, vous disposez de quatre fois plus de capacité de stockage, de bande passante et de cache.
- **Limites :** Le fait d'ajouter des nœuds supplémentaires ne constitue pas toujours un vecteur d'accroissement. Comme lorsque l'on installe un moteur plus rapide dans une voiture, à un moment donné, d'autres facteurs limitent la vitesse maximale, comme la résistance du vent, les pneus ou la durée du parcours.

  Dans le cas d'un cluster scale-out, la connectivité du réseau constitue souvent une limite lorsqu'il s'agit d'augmenter les performances. En outre, les disques durs de chaque nœud disposent d'un taux de transfert théorique maximum. Même en utilisant des disques à circuits intégrés plus rapides, vous pouvez être confronté à des limites.

> La taille physique peut également constituer un frein. Bien que chaque nœud soit relativement petit, construire un cluster scale-out pouvant stocker toutes les photos de Facebook demanderait l'espace d'un terrain de football et suffisamment d'électricité pour faire fonctionner une petite ville.

Anticiper les problèmes

Malheureusement, tout système mécanique finit par rencontrer des problèmes. Même les machines les plus simples finissent par s'enrayer, tomber en panne ou s'user. La technologie numérique a remplacé la plupart des éléments mécaniques tels que les valves ou les relais par des circuits intégrés et des puces en silicone qui accroissent de manière significative la fiabilité du système. Toutefois, même cette amélioration ne permet pas à un ordinateur ou à un nœud scale-out de durer pour toujours.

Une panne surviendra, c'est certain

Certains éléments tomberont en panne à coup sûr et vous pouvez prédire cette probabilité en fonction de l'utilisation du système. Par exemple, la technologie des disques durs utilisée pour stocker la plus grande partie des informations numériques repose sur des plaques tournantes qui présentent des zones codées magnétisées. Ces plateaux s'usent de manière naturelle, à mesure que le dispositif de lecture flotte à quelques nanomètres au-dessus des plaques tournantes. À terme, cette magnétisation constante de la surface entraîne une usure naturelle qui a pour effet de dégrader soit l'inscription, soit la lecture des données. Au pire des cas, le moteur constamment en action qui fait tourner ces plateaux finit par lâcher.

Pour éviter cette usure naturelle, les ingénieurs ont mis au point divers systèmes visant à protéger les données numériques. La méthode la plus communément utilisée est de simplement faire une copie de toutes les données ou d'une partie de celles-ci d'un disque à un autre. C'est ce qu'on

appelle le Redundant Array of Independent Disks (ou RAID). Le RAID présente toutefois trois problèmes majeurs :

- Le RAID n'est pas très pratique. Au vu du nombre de données produites chaque jour dans le monde, conserver une copie complète, ou même partielle, de ces données n'est pas toujours faisable.
- Récupérer les données à partir de la copie pour les envoyer sur une nouvelle source prend du temps et a un impact sur l'efficacité globale de la plateforme de stockage.
- À mesure que le volume des données originales et des copies augmente, les deux ensembles de données peuvent subir des pertes critiques simultanément.

C'est pour ces raisons que le RAID ne convient généralement pas aux projets de Big Data.

Au lieu de cela, de nombreuses architectures scale-out divisent les données et les répartissent sur de multiples disques et nœuds. Nombre de ces petits ensembles de données sont également copiés et les systèmes créent des sommes de contrôle pour permettre de récupérer les données qui seraient perdues. Un peu comme avec un Sudoku, si vous disposez de suffisamment de cases et êtes doué en mathématiques, vous pouvez reconstituer l'ensemble des données.

La répartition des données peut également évoluer avec le nombre de nœuds et de disques pour diminuer la probabilité que des pannes simultanées n'effacent des ensembles clés de données.

Même si un disque et un nœud entier sont perdus, les données restent disponibles, ce qui constitue un autre avantage. Ce type de protection, grâce à laquelle les données sont réparties à divers endroits ou envoyés sur de multiples canaux, est connue sous le nom de répartition des données ou de correction d'erreurs sans circuit de retour.

Chapitre 5

Bonnes pratiques du stockage scale-out dans le monde du Big Data

Dans ce chapitre

- Comprendre les niveaux de données, les quotas et l'allocation granulaire
- Utiliser un disque à circuits intégrés
- Problèmes liés à la sécurité et au respect des lois

Nous avons abordé les principes fondamentaux du stockage des données et les éléments dont vous avez besoin pour créer une architecture de stockage adaptée à un projet de Big Data. Toutefois, un certain nombre de technologies, de procédés de gestion et de bonnes pratiques peuvent vous faire gagner du temps, économiser de l'argent et vous permettre de rendre votre système plus sûr. Ce chapitre a pour objet d'aborder ces points.

Niveaux de données : Examiner de plus près l'architecture scale-out

Bien que chaque organisation présente des exigences et une approche légèrement différentes, des éléments communs interagissent avec les architectures de stockage scale-out au sein d'une entreprise classique, comme le montre l'image 5.1.

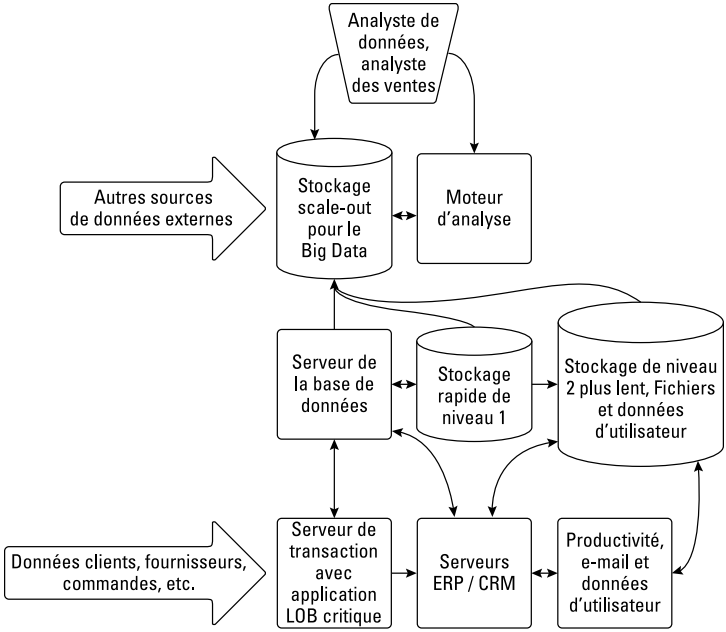

Image 5-1: Des éléments communs interagissent avec les architectures de stockage scale-out.

Comprendre les niveaux des données et leur impact financier

Les projets de Big Data ont souvent pour objectif d'extraire des informations utiles à partir de données brutes. Les données brutes et les informations obtenues présentent une valeur intrinsèque. Cette valeur indique la manière dont les données sont stockées, qui y a accès et où elles se trouvent sur les dispositifs de stockage physiques. Les données doivent aussi être disposées de manière à être accessibles pour les moteurs d'analyse ou correspondre à des exigences particulières de performance ou de capacité.

Effectuons une comparaison pour expliquer le concept : les éléments importants d'une voiture (le volant, les pédales, le boîtier de vitesses et le siège) sont disposés au sein de l'habitacle selon des critères d'accessibilité et de sécurité. Les éléments moins importants, comme l'interrupteur du toit

ouvrant ou celui du lecteur de CD sont éloignés des éléments critiques comme l'accélérateur, dans la mesure où ils sont utilisés moins souvent.

Les niveaux de données indiquent de manière logique les caractéristiques de grands ensembles de données. Ces niveaux peuvent à la fois décrire le type d'informations et leurs exigences en matière d'accès. C'est particulièrement le cas pour les projets de Big Data, pour lesquels l'analyse effectuée peut porter sur des données plus anciennes ou de moindre valeur pouvant être situées à un niveau de stockage plus lent. La manière la plus simple d'aborder les différents niveaux est de les envisager comme une hiérarchie :

Niveau 1	Données cruciales et récemment utilisées	Requiert le degré de performance, de fiabilité et d'accessibilité le plus élevé
Niveau 2	Données rarement utilisées, comme les informations sur les ventes de l'année précédente	Requiert des performances moindres
Niveau 3	Données d'archive qui peuvent être conservées pour des raisons de conformité avec la législation	Requiert de faibles performances, mais beaucoup de capacité

Améliorer l'efficacité à l'aide de la répartition par niveau

De nombreuses plateformes de stockage scale-out disposent d'un logiciel qui permet de gérer cette approche par niveaux. Le logiciel utilise un ensemble de règles paramétrables pour automatiser le processus de déplacement des données au sein du cluster et vers des dispositifs de stockage externes disponibles. Par exemple, les règles peuvent être configurées pour déplacer des données qui n'ont pas été consultées au cours de l'année vers un disque à performances inférieures ou, à l'inverse, déplacer un élément souvent demandé, comme une vidéo populaire, vers une zone plus performante.

Gérer les exigences croissantes des utilisateurs à l'aide de quotas

Les projets informatiques ont tendance à occuper tout l'espace disponible s'ils ne sont pas contrôlés. À partir d'une simple boîte de réception d'e-mail en passant par des zones de transition pour des projets, et même des fichiers archivés sans date de suppression, les données continuent de s'accumuler. Pour les administrateurs aux ressources limitées, les *quotas* permettent de définir l'espace dont dispose chaque utilisateur ou projet, et de déterminer une politique ou de confier la responsabilité de la gestion de cette capacité à une agence responsable de ces données.

Les quotas deviennent des outils de plus en plus flexibles et peuvent être reliés à des niveaux pour créer des zones utilisées à des fins spécifiques. Ces quotas peuvent également se voir associer un coût monétaire pour permettre une facturation interne ou une refacturation.

Lorsque les équipes techniques se voient allouer des ressources limitées en matière de gestion, elles s'attachent davantage à nettoyer elles-mêmes les données inutiles !

Réduire la perte d'espace à l'aide de l'allocation granulaire

Les technologies traditionnelles de gestion du stockage produisent beaucoup de déchets. Les fournisseurs parlent de capacité brute (la taille de chaque disque multipliée par le nombre de disques) et non de la capacité utile et ces chiffres peuvent beaucoup varier. Cet écart découle de plusieurs facteurs, dont :

- Le différentiel entre la capacité brute et utile
- Méthode de protection des données

- Métadonnées utilisées pour garder une trace des fichiers
- Capacités inutilisées dans de nombreux conteneurs de stockage

Un grand nombre de données est reproduite, ce qui occupe également une capacité de stockage précieuse. Toutefois, les principales coupables sont les technologies d'allocation inefficaces en raison desquelles la capacité de la plateforme de stockage peut être allouée, sans être pour autant utilisée.

Le stockage scale-out combiné à la technologie de l'allocation granulaire constitue une approche qui contribue à la résolution de ce problème, sans assigner de capacité à des groupes par avance. Ce système alloue des capacités de stockage uniquement sur demande. Cette méthode répond aux applications, qui vérifient l'espace de stockage disponible avant l'inscription de données, en informant le système qu'il y a suffisamment de capacité pour que l'opération puisse être effectuée. Dans certains cas, le système prévu au sein du moteur d'allocation granulaire peut travailler en tandem avec un stockage par niveau automatisé. Ceci permet le déplacement des données présentes dans les zones de stockage principales et coûteuses, mais qui ne sont jamais utilisées, vers un autre niveau de stockage (voir l'image 5-1) qui coûte moins cher ou plus adapté à un archivage à long terme.

Boostez votre projet Big Data à vitesse grand V à l'aide d'un disque à circuits intégrés

Votre projet d'analyse du Big Data est à présent lancé. Les données circulent, les applications vous fournissent de nouvelles informations, mais vous avez besoin de performances encore plus importantes. Que faire ?

Une solution simple consiste à accélérer la performance des données alors qu'elles circulent à travers le cluster de stockage. Les disques physiques rotatifs présentent un débit de données maximum. Ce débit est limité par la vitesse à laquelle le disque peut tourner et à laquelle les données

peuvent être lues à partir de ceux-ci sous la forme d'un signal magnétique. L'utilisation d'un support sans disque, comme les puces de mémoire à accès aléatoire peut constituer une méthode plus rapide. Un disque à circuits intégrés (SSD) consiste, comme son nom l'indique, en un disque utilisant des puces de mémoire au lieu de disques rotatifs. Cette technologie se présente sous deux axes :

- Les SSD Flash conviennent aux applications en lecture seule et aux applications de mobilité.
- Les SSD DRAM présentent de bien meilleures performances d'écriture et de lecture avec un meilleur rendement par unité de performance que la version Flash, mais ils coûtent plus cher au début par Go de stockage.

Toutefois, le simple fait de changer tous les disques au sein d'une plateforme de stockage scale-out pour passer de disques rotatifs à un système SSD coûte extrêmement cher. Par ailleurs, les SSD n'ont pas une durée de vie infinie, tout comme les disques. En outre, au fil de l'augmentation de la taille des mémoires Flash, la fiabilité est remise en question en comparaison avec les anciens disques rotatifs.

Les architectures scale-out utilisent souvent les SSD de diverses manières. L'une des façons de le faire est d'utiliser un SSD pour accélérer la recherche des informations demandées par le client. Donc, avec un cluster doté de 20 nœuds et de plusieurs milliers d'éléments de données, le processus de recherche de données spécifiques au sein du cluster peut prendre une seconde. Déplacer cette carte (parfois appelée *métadonnée*) où se trouve physiquement chaque élément de données sur un SSD (et non sur des disques rotatifs plus lents) peut réduire ce délai. Utilisé intelligemment, le SSD augmente les performances globales du système sans devoir remplacer chaque disque physique par un équivalent SSD.

Garantir la sécurité

Les données numériques ont de la valeur. Un projet de Big Data qui vise à générer de nouvelles informations ou à donner lieu à des découvertes scientifiques est comme une pierre précieuse pour un voleur, qui peut tenter d'en voler les

résultats voire le matériel source. La sécurité des informations fait l'objet d'inquiétudes constantes, donc ne prenez pas cela à la légère lorsque vous travaillez sur un projet. Un projet de Big Data peut requérir une protection renforcée, au vu des dégâts potentiels pouvant résulter de l'enregistrement de tant d'informations sensibles en un seul endroit.

Les nombreuses préoccupations entourant la sécurité du stockage sont les suivantes :

- S'assurer que le réseau est facilement accessible pour les personnes, les entreprises et les agences autorisées à y accéder.
- Il doit être extrêmement difficile pour un pirate informatique potentiel de compromettre le système.
- Le réseau doit être fiable et stable dans un large éventail de conditions et peu importe les volumes traités.
- Fournir une protection contre les menaces en ligne, comme les virus.
- Segmenter les accès selon chaque département.
- Assigner certaines actions ou privilèges à un individu en fonction de ses responsabilités.
- Crypter les données sensibles.
- Désactiver les services inutiles, afin de minimiser les failles de sécurité.
- Installer régulièrement les mises à jour du système d'exploitation et des appareils.
- Informer tous les utilisateurs des principes et politiques adoptés pour gérer l'utilisation du réseau.

C'est génial, mais est-ce légal ?

Parfois, un projet de Big Data peut pousser une organisation à collecter et à stocker des types d'informations qui n'étaient précédemment pas conservées. Dans certains cas, l'entreprise peut avoir besoin de données provenant d'une source extérieure pour les comparer avec ses propres données et, ce faisant, l'entreprise entre dans un nouveau domaine du droit. Par exemple, si une compagnie d'assurances allemande

souhaite analyser les résultats cliniques de différentes procédures chirurgicales par rapport aux types de politiques utilisées et aux rendements, le projet peut requérir d'énormes volumes de données issues du monde entier.

Si les données proviennent des États-Unis, leur stockage devra se conformer au Healthcare Insurance Portability and Accountability Act (HIPAA).

Lorsque les données nécessaires à l'alimentation des projets de Big Data traversent les frontières, il faut souvent prendre en compte les réglementations locales. Par exemple, la directive relative à la protection des données de l'Union européenne stipule que les organisations qui ne parviennent pas à sécuriser les données ou qui présentent des failles risquent des amendes et, dans certains cas plus graves, leurs directeurs risquent l'emprisonnement.

Les principaux cadres de conformité à prendre en compte aux États-Unis sont :

- Le **Healthcare Insurance Portability and Accountability Act (HIPAA)**, qui vise à protéger le caractère privé des informations relatives à la santé.
- Le **Sarbanes Oxley Act**, qui vise le secteur de la comptabilité.
- Le **Gramm-Leach-Bliley Act (GLB)**, qui oblige les institutions financières à garantir la sécurité et la confidentialité des informations de leurs clients.
- Le **Bank Secrecy Act**, utilisé par le gouvernement américain pour poursuivre les fraudeurs.

Chapitre 6

Autres éléments à prendre en compte pour le stockage de Big Data

Dans ce chapitre

- Optimiser le centre de données
- Planification à plus long terme pour réduire les coûts
- Envisager la virtualisation et le cloud computing

Dans ce chapitre, nous abordons les autres aspects qui peuvent être concernés par les projets de Big Data. Nous envisageons également plusieurs objectifs et stratégies à long terme, qui peuvent fournir une alternative à la gestion des projets de Big Data en interne.

N'oubliez pas le datacenter !

Diverses estimations montrent que le stockage représente 35 % de l'énergie utilisée dans les centres de données (ou datacenters). Le fardeau sur les centrales électriques devrait s'alourdir à mesure que les internautes génèrent et consomment des contenus numériques. Alors que les coûts de l'énergie grimpent en flèche et que l'on craint des surcharges électriques, la consommation d'énergie devient une préoccupation majeure. Lorsque vos projets de Big Data se développent, et avec eux, de nouveaux espaces de stockage et des centres de serveur, suivez ces conseils :

- **Réduisez les points chauds des datacenters pour réduire les coûts du refroidissement.** Lorsque les datacenters grandissent sans que les exigences en matière d'alimentation et de refroidissement aient été prises en compte, un point chaud peut poser des problèmes dans le fonctionnement de l'équipement informatique. Les dispositifs de stockage prennent beaucoup de place, et une fois installés sur le sol, il est difficile de les déplacer sans causer d'interruption du fonctionnement des applications. Choisissez plutôt de répartir les unités de manière stratégique sur le site.
- **Configurer l'espace alloué à l'équipement avec des couloirs chauds et froids.** La plupart des appareils informatiques rejettent de l'air chaud de leur face arrière. Si la rangée arrière reçoit l'air chaud provenant de la rangée juste devant, le flux d'air froid est interrompu, ce qui oblige les dispositifs d'air conditionné à générer de l'air froid plus coûteux. Assurez-vous que l'équipement est installé avec des dispositifs d'échappement pour expulser l'air chaud vers des zones non utilisées ou vers l'extérieur.
- **Déplacez les charges de travail pour économiser de l'énergie.** Un logiciel de virtualisation et de gestion du stockage peut aider les datacenters à réorganiser les tâches informatiques et de stockage réalisées physiquement au sein du centre. Ces logiciels peuvent permettre de répartir de manière égale (du moins en théorie) ou de déplacer les charges de travail vers des serveurs moins utilisés et d'éteindre les nœuds de stockage « vides » ou les serveurs inutilisés sans avoir à déplacer le matériel.
- **Une plus grande densité peut accroître l'espace au sol.** Envisagez d'accroître la densité des disques durs utilisés pour le stockage des données. Bien qu'un disque de 4 To jouisse de quatre fois plus de capacité qu'un disque de 1 To, il n'utilise pas quatre fois la même quantité d'énergie. Dans certaines architectures de stockage scale-out, il est relativement facile de changer de disque sans interruption du système. Si ces augmentations de densité sont réalisées sur un seul nœud à la fois, un cluster de 100 To peut passer à 400 To en prenant la même place qu'avant pour seulement quelques points de pourcentage de consommation en plus.

Planification à plus long terme pour réduire les coûts

Que votre projet de Big Data soit petit, moyen ou grand, votre infrastructure informatique devient probablement plus importante. Même avec l'arrivée de la virtualisation, qui permet aux ordinateurs de fonctionner plus efficacement, l'importance critique des systèmes informatiques entraîne une dépendance vis-à-vis de systèmes plus grands et plus complexes.

Le stockage est aujourd'hui plus puissant et prend moins de place physiquement. Le coût par giga-octet de capacité de stockage a chuté, alors que la densité, la vitesse et les performances de stockage ont augmenté massivement.

Les technologies fondées sur les disques pour le stockage des données sont les solutions les plus souvent utilisées pour passer au niveau supérieur. Alors que les disques standard dépasseront les 4 To de capacité pour atteindre parfois 16 To par unité au cours des cinq prochaines années, l'aptitude des organisations à accroître leur capacité in situ au sein du même pool de stockage présente un avantage majeur.

Une autre stratégie à plus long terme consiste à déplacer automatiquement les disques durs dotés de la technologie Serial Attached SCSI (SAS) à haute performance et les SSD vers un dispositif de stockage plus lent et moins cher, tel que des disques Serial AT Attachment (SATA). Les doubles entrées de données sont supprimées et les informations statistiquement moins importantes sont retirées. Ces projets de gestion du cycle de vie de l'information (ILM) peuvent contribuer à augmenter la viabilité de l'architecture de stockage.

S'atteler à la virtualisation

La virtualisation est l'une des tendances technologiques les plus significatives de la dernière décennie. Il s'agit toutefois d'un terme qui regroupe de nombreux concepts :

- **Virtualisation de serveurs :** Elle permet à un serveur de faire fonctionner plusieurs systèmes d'exploitation en même temps, ce qui diminue le nombre de serveurs physiques nécessaires au fonctionnement de multiples applications serveur. Un serveur virtualisé peut ne pas offrir un élément visuel à l'utilisateur et peut simplement fonctionner via un processus non interactif comme un proxy serveur ou une tâche de traitement des données.
- **Virtualisation du poste de travail :** Souvent connu sous le nom d'infrastructure de bureau virtuel (Virtual Desktop Infrastructure ou VDI), le concept de la virtualisation du poste de travail permet aux préférences de chaque ordinateur, au système d'exploitation, aux applications et aux fichiers d'être hébergés sur un serveur à distance. Les utilisateurs peuvent alors utiliser un client d'accès, comme un PC, ou un client léger pour visualiser et interagir avec ce poste de travail à distance sur un réseau. La virtualisation du poste de travail présente plusieurs avantages, tant pour les utilisateurs finaux que la DSI, dans la mesure où un appareil à faible consommation, tel qu'une tablette, peut faire fonctionner des applications complexes et où la gestion des données est simplifiée, étant donné qu'elle ne quitte jamais le serveur central.
- **Virtualisation du stockage :** Il s'agit de la consolidation du stockage physique à partir de dispositifs multiples de stockage, au sein de ce qui apparaît comme un dispositif de stockage unique géré à partir d'un endroit central. La virtualisation du stockage est le concept fondamental du stockage scale-out : un ensemble de nœuds de stockage peut être ajouté sur demande pour accroître la capacité et les performances d'un seul pool de stockage, sans interruption pour les utilisateurs ou les applications. Cet outil présente de nombreux avantages en termes de gestion réduite, de besoins d'espace physique limités et de capacité à réduire la redondance des données. La virtualisation du stockage simplifie et réduit souvent le nombre de dispositifs de stockage physiques nécessaires pour tout volume de données grâce à des gains d'efficacité.

La technologie du cloud computing pour les projets de Big Data

Au vu des exigences rigoureuses des projets de Big Data en matière de réseaux, de stockage et de serveurs, il n'est pas surprenant que certains clients externalisent les tracasseries et les coûts auprès d'une autre entité. Il s'agit d'un domaine dans lequel le cloud computing peut être utile.

Le cloud computing public ou privé consiste à fournir des ressources matérielles et logicielles en tant que service sur un réseau, notamment Internet.

Les nuages (« cloud » en anglais) peuvent avoir plusieurs finalités (comme le montre l'image 6-1) et comprennent :

- **Infrastructure en tant que service (IaaS) :** Un ou plusieurs ordinateurs avec un stockage et une connectivité réseau auxquels vous pouvez accéder via une connexion réseau.
- **Logiciel en tant que service (SaaS) :** Accès à une application logicielle spécifique dotée de vos propres données par le biais d'une connexion réseau.
- **Plateforme en tant que service (PaaS) :** Fournit les éléments fondamentaux, tels que les outils de développement logiciel, nécessaires pour mettre au point votre propre environnement informatique à distance, auquel les utilisateurs peuvent accéder, notamment via des postes de travail virtuels ou via un réseau.
- **Stockage en tant que service (STaaS) :** Une plateforme de stockage à distance qui présente un coût spécifique par Go pour le stockage et le transfert des données.

Certains projets de Big Data peuvent être adaptés à une utilisation dans le cloud public, dans la mesure où sa flexibilité lui permet de croître rapidement. En outre, de nombreux clouds publics permettent de louer des ressources à court terme en évitant les coûts initiaux souvent très élevés.

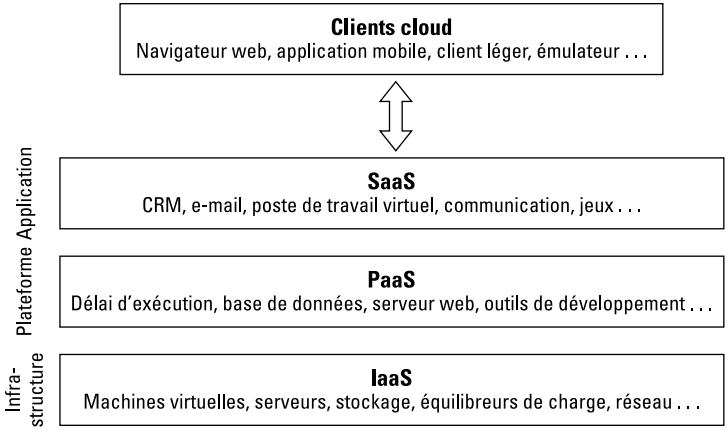

Image 6-1: Différents aspects du cloud computing.

Toutefois, la sécurité, la fiabilité, les performances et le transfert de données à l'aide des technologies liées au cloud public font l'objet de certaines préoccupations :

- Pour les projets qui nécessitent de déplacer de grandes quantités de données sur Internet, les limites et le coût de la bande passante réseau peuvent en réalité rendre cette solution plus coûteuse pour un projet de Big Data qu'un équivalent sur site ou via un cloud privé.
- Pour les organisations qui disposent d'informations assorties de droits de la propriété intellectuelle ou d'informations personnelles très sensibles, comme des dossiers médicaux ou des dossiers d'étudiants, conserver des données dans un endroit inconnu géré par des inconnus peut poser problème. En réalité, de nombreuses entités publiques disposent de lois relatives à la résidence ou la souveraineté des données, imposant la localisation des données dans leur juridiction de création, et interdisant leur stockage à l'étranger. En outre, les politiques de protection de données et les procédures d'information au sein du cloud public ne peuvent que difficilement être contrôlées.
- La performance des données inscrites et lues à partir d'un cloud public peut être lente et coûteuse selon la distance et le type de réseau utilisé, et les tarifs que le fournisseur de cloud publics impose pour l'écriture et la récupération de ces données.

- Une fois que de grands volumes de données sont stockés au sein d'un cloud public, il peut s'avérer difficile et coûteux de déplacer ces données vers un autre fournisseur. Le changement d'un fournisseur de cloud public peut vous coûter cher !

De nombreuses organisations adoptent une stratégie prévoyant le recours à un cloud privé, dans le cadre de laquelle la liberté d'accès au cloud est permise via Internet, mais le contrôle de l'accès reste aux mains de l'organisation.

En outre, la sécurité physique, la sauvegarde, la réparation en cas de problème et les performances des données sont contrôlées par l'organisation.

Chapitre 7

Dix astuces permettant de garantir le succès d'un projet de Big Data

Dans ce chapitre

- Identifier les types et les flux de données
- Se préparer à l'accroissement des données
- Éviter les erreurs relevant de la gestion des données sensibles
- Se préparer aux pires scénarios

Si vous lisez ce chapitre en premier, nous supposons que c'est parce que vous souhaitez éviter de commettre des erreurs qui pourraient faire échouer votre projet de Big Data. Voici quelques points à prendre en compte.

- **Commencez tout projet de Big Data avec un examen des données et un processus de classification.** Déterminer si les données sont structurées, non structurées, qualitatives ou quantitatives peut être utile pour concevoir les architectures de stockage (voir le chapitre 1 pour un petit rappel). Il est également judicieux d'estimer la croissance des données en fonction des tendances passées et des stratégies futures.
- **Créez un aperçu simple de la manière dont les données circulent au sein de votre organisation.** Disposer d'un diagramme simple montrant où les données sont créées, stockées et circulent est utile lorsque vous travaillez avec un groupe de travail. Mettre tout le monde sur la même longueur d'onde peut vous permettre d'éviter des malentendus qui coûtent cher.

- **Envisagez vos futures exigences en matière de stockage de données sur la base du succès du projet de Big Data.** Les projets de Big Data peuvent faire apparaître de nouvelles informations ou vous obliger à modifier les processus opérationnels. Les informations émanant du projet peuvent à leur tour nécessiter des capacités de stockage supplémentaire, ce qui entraîne une croissance exponentielle des besoins de capacité. Pensez toujours à plus long terme.
- **Soyez flexible.** De nombreux projets reposent à la fois sur les technologies de stockage scale-up et scale-out (voir le chapitre 3). Chaque organisation et chaque projet sont uniques. Le choix d'une technologie de stockage doit être axé sur l'objectif à atteindre et non sur une architecture technique particulière. De nombreux fournisseurs proposent des produits scale-up et scale-out qui peuvent fonctionner ensemble.
- **Les exigences en matière de stockage de données peuvent augmenter, mais envisagez de déplacer automatiquement les données peu consultées vers un dispositif de stockage moins coûteux et plus lent.** La suppression est également une option viable à plus long terme. Peu importe d'où proviennent les données, où elles sont traitées et où elles sont stockées, elles présentent toujours une durée de vie utile. Décider de supprimer des données est une tâche complexe, mais cela peut vous permettre de réaliser d'énormes économies à plus long terme. Déplacer automatiquement des données vers un dispositif de stockage plus lent est une option plus facile qui comporte toujours de grands avantages.
- **Demandez aux fournisseurs ce qu'il se passera lorsque vous aurez atteint les limites de capacité ou de performances théoriques.** Même si vous commencez par un petit projet de Big Data, celui-ci prendra sûrement plus d'ampleur avec le temps. Comprendre de quelle manière la technologie choisie peut évoluer, vous permettra d'éviter les mauvaises surprises dans les années à venir.
- **Préparez-vous au pire.** Même les machines les plus simples finissent par tomber en panne ou s'user. Demandez à votre fournisseur ce qu'il se passerait si divers éléments de la plateforme de stockage venaient à tomber. Un système bien conçu ne devrait jamais présenter de point de défaillance.

- **Créez un système de quota dès le début du projet pour prévenir les futurs problèmes de gestion.** Les projets informatiques ont tendance à occuper tout l'espace disponible s'ils ne sont pas contrôlés. Les quotas permettent de définir la quantité d'espace dont jouit chaque utilisateur ou projet. Confiez la responsabilité de la gestion de cette capacité à une entité responsable de ces données ou définissez une politique.
- **Impliquez toujours des experts de la sécurité informatique dans vos projets de Big Data.** Les données numériques ont de la valeur. Même si le projet de Big Data ne concerne qu'un seul groupe de recherche, l'équipe de sécurité informatique doit être impliquée dès le début, afin que la sécurité demeure au cœur du projet.
- **Rappelez-vous de prendre en compte le temps de gestion lors du calcul des coûts du stockage.** Les coûts totaux du stockage doivent inclure le temps nécessaire à l'approvisionnement et à la gestion de la plateforme. Un système résilient et hautement automatisé, qui ne nécessite pas un administrateur à temps plein, permet d'économiser bien plus d'argent à long terme qu'un matériel moins cher, qui requiert beaucoup de travail manuel.

CAPACITE PERFORMANCE EVOLUTIVITE SIMPLICITE

BIG DATA

IT'S ON ISILON

UNE NOUVELLE APPROCHE DU STOCKAGE DE DONNÉES AU SERVICE DE VOTRE MÉTIER

Faîtes face aux défis techniques et économiques posés par la croissance exponentielle des volumes de données.

Quel que soit votre secteur d'activité, les solutions de stockage EMC Isilon vous permettent de valoriser votre capital numérique et d'améliorer l'agilité et la compétitivité de votre organisation.

EMC.COM/ISILON

EMC²

Un guide pratique pour la conception, la mise en œuvre et la gestion de votre architecture de stockage dans le cadre de projets de Big Data.

La croissance massive de la quantité de données élargit le champ des possibles et apporte des avancées majeures notamment dans le domaine scientifique, l'environnement et l'industrie. Obtenir des informations clé en tirant parti du Big Data impose des architectures de stockage, des procédés de gestion et des pratiques innovantes qui permettront de gagner du temps, d'économiser de l'argent et de renforcer votre sécurité.

- ***Comprendre les éléments fondamentaux qui composent les données – se familiariser avec la structure et l'utilisation des données***
- ***Construire une architecture évolutive – accroître vos connaissances des éléments liés au stockage Big Data***
- ***Apprendre les bonnes pratiques – découvrir les compétences à acquérir pour réussir vos projets Big Data***

Ouvrez le livre et découvrez:

- **Ce que vous devez savoir pour construire une architecture de stockage complète**
- **Les techniques permettant de garantir le succès d'un projet Big Data**
- **Les meilleures pratiques en matière de sécurité et conformité des données**
- **Des exemples de Big Data**

Visitez le site Dummies.com® pour visualiser des vidéos, des tutoriels, des articles ou pour faire vos achats!

Will Garside est un éditeur indépendant, un auteur et un journaliste informatique de plus de 17 ans d'expérience dans les domaines de l'Internet, de l'impression et de la radio. Il est fier d'être un féru de technologies et adore en parler avec les passionnés. **Brian Cox** est le directeur senior produits et solutions marketing d'EMC Isilon. Il a vécu l'émergence des premiers ordinateurs personnels et a eu un TI-99/4A et le modèle original de deux lecteurs de disquette IBM PC.

FOR DUMMIES® A Wiley Brand

FSC www.fsc.org MIX Paper from responsible sources FSC® C013056

ISBN: 978-1-118-71391-4