LIVRE BLANC

Relever les défis du stockage à l'échelle des pétaoctets pour applications **Big Data et Analytiques**

SCALE TO WIN

INFINIDAT

Résumé

Le Big Data et les charges analytiques constituent une nouvelle frontière pour les entreprises. Les données collectées émanent de sources inexistantes il y a 10 ans. Les données des smartphones, celles générées par des machines et par les interactions avec les sites web sont collectées et analysées. De plus, dans un contexte où les budgets IT sont déjà sous pression, les empreintes Big Data s'étendent et posent d'énormes problèmes de stockage.

Ce livre blanc délivre des informations sur les problématiques liées aux applications Big Data pour les systèmes de stockage et explique en quoi le choix de la bonne infrastructure de stockage peut rationaliser et consolider les applications d'analyse du Big Data sans faire sauter la banque.

Introduction: le Big Data “stresse” l’infrastructure de stockage

Les applications Big Data font exploser les volumes de données qu’une entreprise doit maintenir en ligne pour pouvoir les analyser. Ceci a pour effet de faire exploser la part du coût du stockage dans le budget IT global. Une étude récente auprès d’entreprises du Big Data adeptes de l’analytique a révélé cinq grands vecteurs de cette explosion des données :

1. Meilleur service client et support

2. Sécurité numérique, détection d’intrusion, détection de la fraude et prévention

3. Analyse opérationnelle

4. Explosion du Big Data

5. Augmentation de la taille des entrepôts de données

L’amélioration du service client est une question centrale dans les entreprises de toute taille, grandes et petites. « Comment puis-je améliorer la relation avec mon client ? Je sais que mon client préférera acheter mes produits et ce plus souvent si je cultive et soigne la relation. » Généralement, on s’appuie sur les données de préférence émanant de multiples sources, comme la navigation en ligne et les critères de recherche. Collectées sur tout le cycle de vie du client, ces données sont extraites et consignées dans les dossiers clients.

La sécurité numérique et la détection d’intrusion sont très importantes pour les clients. Ces données sont collectées et analysées en temps réel et généralement générées par des machines. Les résultats d’analyse doivent être rendus immédiatement pour que cette activité soit pertinente. Ceci suppose un stockage rapide avec de grandes capacités, dans la mesure où les données générées par les machines et capteurs sont consommateurs de larges volumes.

L’analyse opérationnelle suppose de collecter les données (souvent issues de capteurs d’autres machines) à utiliser pour identifier les potentiels d’amélioration des opérations, isoler les pannes et analyser la résolution possible. Les sociétés de fabrication industrielle collectent des données à la seconde sur les activités robotiques dans leurs usines pour en connaître le statut, mais aussi améliorer le processus. Comme pour la détection d’intrusion, ces données sont générées et analysées en temps réel et les résultats doivent être stockés et renvoyés dans la chaîne pour pouvoir être exploités. Mais à l’inverse de la détection d’intrusion, toutes les données sont intéressantes et révèlent des tendances relatives aux machines et processus, utilisables ultérieurement.

Exploration du Big Data : Comment savoir ce qu’est le Big Data tant que vous n’avez pas découvert ce que vous collectez ou non et identifié ce qui manque ? Normalement, ceci passe par la collecte de toujours plus de données.

Augmentation de la taille des entrepôts de données : « Comment faire pour prendre les données analytiques existantes, généralement dans un entrepôt ou un magasin, et augmenter les flux de données de sources extérieures pour plus de précision, pour réduire les délais d’exécution et obtenir les réponses attendus sans réinventer la roue ? » L’adoption des entrepôts de données se généralise, même dans les plus petites entreprises, maintenant que l’analyse des données transactionnelles s’impose à tous les niveaux des entreprises.

Ces scénarios d’utilisation requièrent plus de stockage et de puissance de calcul. On considère désormais le Big Data comme des données de production, si bien que leur disponibilité, leur capacité de restauration et leur performance sont aussi importantes que pour les systèmes transactionnels. Et comme indiqué précédemment, la tendance va plutôt au resserrement des budgets IT que l’inverse. Ces forces diamétralement opposées induisent un changement dans l’industrie du stockage. Comment faire plus avec moins, sachant que vous ne voulez faire aucun compromis de fiabilité, d’efficacité ni de performance des systèmes et applications ?

Voici un résumé des exigences que les charges Big Data et analytiques font peser sur le stockage d’entreprise :

- Obligation d’excellents niveaux de performance
- Densité extrêmement élevée
- Disponibilité excellente, haute fiabilité
- Facile à utiliser, à gérer et à provisionner
- Coût total de possession bas et attractif

C’est là qu’une toute nouvelle architecture de stockage comme Infinidat InfiniBox® peut aider.

Haute performance

Les gros fichiers de données, les applications analytiques lourdes et les demandes urgentes de résultats font que le stockage sous-jacent doit absolument délivrer de hauts niveaux de performance. Avec InfiniBox, on atteint un niveau de performance maximum sans réglage ni optimisation. InfiniBox utilise des composants standard vendus dans le commerce (CPU/mémoire/disques durs/SSD) réunis dans une solution de stockage sophistiquée pour extraire des performances maximales des 480 disques SAS Near-Line de l'architecture InfiniBox. L'un des éléments clés développés dans le code du système réside dans la capacité d'analyser de vrais profils d'application et de définir précisément des algorithmes pre-fetch et destage du cache. Le système cible spécifiquement ces profils et délivre des performances optimales dans ces conditions. Cette fonctionnalité est au centre de l'architecture InfiniBox.

DE GROS FICHIERS DE DONNEES

Les grands fichiers de données posent un défi unique aux baies de stockage de données d'entreprise en proposant un profil E/S imprévisible qui submerge souvent les systèmes de stockage matériels traditionnels. Ceci se traduit par de fortes latences, qui freinent l'exécution des charges analytiques. Certaines activités analytiques sont très sensibles à la latence avec souvent des répercussions sur la population des utilisateurs de l'application. Nombre de ces charges submergent les plateformes de stockage ayant une taille de cache limitée, mais pas InfiniBox. InfiniBox utilise un algorithme avancé de gestion du cache (Neural Cache) avec DRAM et SSD pour améliorer les réponses du cache et réduire la latence.

Scénarios Big Data et analytiques

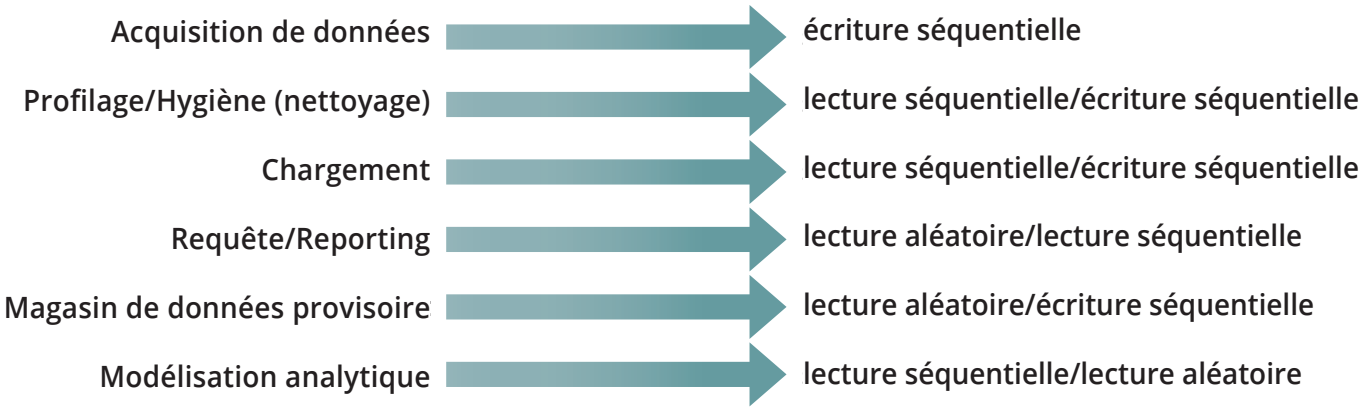
- Exploration
- Service client amélioré
- Sécurité et protection contre la fraude
- Analyse opérationnelle
- Simulation DWH

Exigences de stockage Big Data et analytique

- Performance stable
- Haute capacité
- Disponibilité et fiabilité optimales
- Facilité de gestion et de provisioning
- Coût de possession attractif

PROFILS D'E/S ET PATTERNS SPECIFIQUES AU BIG DATA

De nombreux environnements analytiques ont des profils E/S avec les caractéristiques suivantes :



Plusieurs de ces caractéristiques peuvent se produire simultanément et d'autres sont déclenchées par des activités spécifiques, comme les sauvegardes ou le chargement des données/ETL. InfiniBox soutient parfaitement un large éventail d'E/S en même temps. L'architecture de données visualise le stockage de chaque volume et alimente chacune des 480 broches de l'InfiniBox, en parallèle, en suivant un agencement distribué sophistiqué avec parité des données.

De plus, InfiniBox utilise des fonctionnalités très avancées d'amélioration des opérations d'écriture. Grâce à un mécanisme unique et breveté d'écriture multimodale de journal, Infinidat améliore nettement l'efficacité de déplacement des E/S d'écriture du cache. C'est très important pour les phases d'acquisition de données et ETL de cet exemple.

LA GESTION DE LA TAILLE DES BLOCS COMPTE

De nombreuses charges analytiques peuvent changer le profil d'E/S à la volée. Mais en général, la grande majorité des applications Big Data et analytiques utilisent les E/S de grands blocs, chargeant les données depuis le stockage, les réduisant, les triant et les comparant pour ensuite écrire les données agrégées. Or les grands blocs sont connus pour poser des problèmes aux plateformes de stockage traditionnelles car la plupart des environnements de stockage ne sont pas pensés pour supporter les grands blocs.

Haute densité

Infinidat sait configurer un système avec plusieurs pétaoctets de capacité effective en un seul rack 19 pouces. Le système de stockage InfiniBox est un système de contrôleur (nœud) all-active moderne, avec grille entièrement symétrique et une architecture de cache avancée sur plusieurs couches. L'architecture de données englobe un modèle de distribution de données à double parité (wide-stripe). Ce modèle utilise une combinaison unique de distribution aléatoire des données et de protection de la parité. C'est la clé d'une disponibilité maximale des données tout en réduisant l'empreinte des données. Chaque volume créé sur un système InfiniBox stocke de petites quantités de données sur chacun des 480 disques. La capacité de stockage utilisable par système InfiniBox est la plus importante de toute l'industrie du stockage.

Hauts niveaux de disponibilité et de fiabilité

La disponibilité analytique est critique pour un système de stockage. L'architecture InfiniBox fournit un environnement de stockage robuste et hautement disponible, à sept 9. Ceci correspond à moins de trois secondes d'indisponibilité par an ! Nos clients ne déplorent aucune perte de données, même en cas de panne de plusieurs disques. InfiniBox offre des fonctions de continuité des opérations, y compris la mise en miroir asynchrone à distance et les instantanés. Avec des instantanés, il est possible de réduire le délai de récupération d'une base de données au temps de mise en correspondance des volumes et des hôtes, en minutes plutôt qu'en heures, selon un processus plus traditionnel de sauvegarde et restauration.

Provisioning et administration simples et automatisés

L'architecture InfiniBox, avec l'élégante simplicité de son interface web GUI et de son interface à ligne de commande intégrée, permet le déploiement et l'administration simples et rapides du système de stockage. Le temps gagné par rapport à l'administration d'un système de stockage traditionnel est considérable. Et grâce à l'architecture ouverte InfiniBox et au support de l'API RESTful, les plateformes comme OpenStack et Docker permettent d'effectuer les tâches d'administration du stockage au niveau applicatif, sans qu'il faille utiliser l'interface GUI d'InfiniBox.

InfiniBox fournit un système d'administration capable d'isoler des pools et volumes de stockage pour des utilisateurs en particulier. Des fonctions de mutualisation permettent aux utilisateurs d'applications, dans un environnement cloud privé par exemple, de voir et gérer le stockage attribué à telle ou telle communauté.

Un coût total de possession très bas

Haut niveau de performance, disponibilité extrême, plus forte densité de données et facilité d'utilisation convergent vers un TCO incomparable. C'est important pour les environnements où il convient de consolider des bases de données stratégiques en empreintes physiques de plus en plus restreintes.

Splunk et InfiniBox

Un des nombreux scénarios Big Data concerne le support des clusters Splunk. Le modèle de déploiement par défaut de clusters Splunk consiste à créer un cluster pour de nombreux nœuds abordables. Chaque nœud a des capacités équivalentes de calcul, de mémoire et de stockage dédié. La phase de conception initiale de la plupart des clients qui lancent leur premier environnement Splunk consiste à trouver la bonne combinaison de ressources de ces nœuds. Ils enchaînent ensuite ces nœuds jusqu'à créer un environnement de calcul suffisamment grand pour leur application analytique préférée. Le problème que rencontrent la plupart des clients est qu'ils arrivent à bout de la capacité de stockage bien avant qu'ils épuisent les cycles de calcul dans le cluster. La seule option pour les clients qui utilisent un stockage dédié consiste à ajouter davantage de nœuds au cluster. C'est correct, sauf qu'avec l'ajout de puissance de calcul (qui peut s'avérer superflue), ceci peut aboutir très souvent à la rupture du modèle de stockage dédié.

En faisant le choix du stockage SAN de bloc InfiniBox au lieu de disques durs dédiés par nœud, vous n'êtes plus limité par la capacité de stockage potentielle de chaque nœud. Vous pouvez ajouter des LUN de façon dynamique et plus d'espace par LUN pour chaque nœud de cluster. Plus besoin d'ajouter de la puissance de calcul tant que ce n'est pas nécessaire. Le niveau de redondance des données Splunk peut être réduit, étant donné que les données sont parfaitement protégées avec une disponibilité de sept 9.

Conclusion

InfiniBox comble l'écart entre haute performance et haute capacité des applications Big Data. InfiniBox permet à toute entreprise qui déploie des projets Big Data et analytique d'atteindre ses objectifs stratégiques : réduction des coûts, montée en charge continue de la capacité et gestion simple et efficace, sans compromis de performance ni de fiabilité. Tout ceci vient soutenir avec efficacité les applications Big Data pour un prix exceptionnellement attractif.