LIVRE BLANC

Les Big Data prennent tout leur sens sur les serveurs nus

Les performances des big data : une priorité

SYNTHÈSE

De nos jours, les entreprises créent et collectent des volumes de données toujours plus importants, provenant de différentes sources, sous des formats structurés ou non structurés. Stocker, traiter et dégager de la valeur ajoutée de ces « big data » n'est pas chose facile. Les professionnels de l'informatique provisionnent souvent des serveurs de cloud public pour pouvoir faire évoluer le stockage et la puissance de traitement en fonction de ce flux permanent de données, mais ces ressources virtualisées ne fournissent pas les performances et la stabilité des serveurs nus.

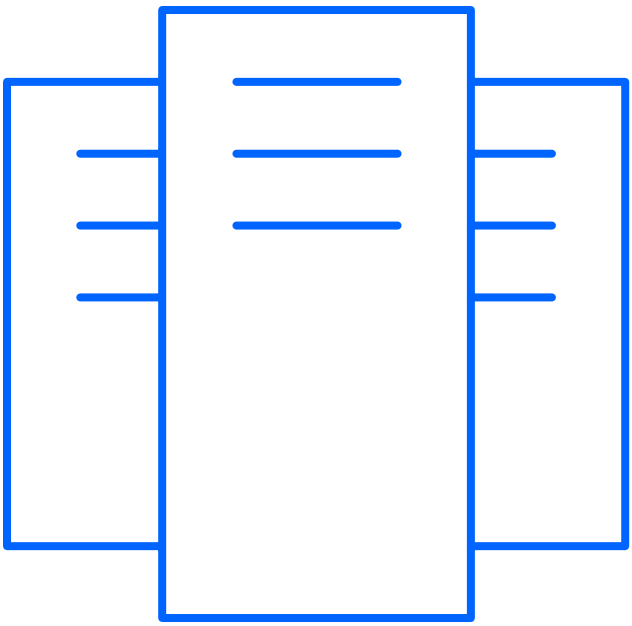
IBM Cloud a testé les performances et la stabilité des charges de travail de big data sur des serveurs virtuels et des serveurs nus pour comparer l'adéquation de ces plateformes aux applications qui stockent et traitent des volumes importants de données. Grâce à ces résultats, les professionnels de l'informatique peuvent prendre de meilleures décisions lorsqu'ils sélectionnent leurs ressources cloud pour les charges de travail intensives en stockage et en puissance de traitement.

Les « Big Data », qu'est-ce que c'est ?

Avec l'évolution des technologies de stockage et la baisse des coûts des capacités, les entreprises disposent de nouvelles possibilités pour collecter et traiter davantage d'informations. En traitant ces informations, les entreprises peuvent dégager des perspectives susceptibles d'apporter de la valeur ajoutée. Le défi est d'organiser et d'analyser les données afin d'élaborer de nouvelles stratégies et prendre des décisions.

Jusqu'à récemment, les outils de prédilection pour organiser et analyser les données étaient les systèmes de gestion des bases de données relationnelles (RDBMS) tirant parti du Structured Query Language (SQL). Les solutions SQL s'appuient sur des ensembles de données structurés, généralement stockés et administrés sur un serveur unique. Lorsque la taille de l'ensemble de données atteint le plafond de capacité du serveur existant, la solution évolue en passant à un serveur plus important, avec davantage de puissance de traitement, de stockage et de RAM. Une telle évolution peut s'avérer chronophage et augmenter considérablement les coûts.

Les données affluant désormais plus rapidement à partir de nombreuses sources et sous des formats multiples, les administrateurs de bases de données doivent maximiser l'efficacité et l'évolutivité de leurs solutions. Ainsi, nombreux sont ceux qui ont commencé à tirer parti de bases de données NoSQL (Not Only SQL), qui exploitent des ensembles de données non relationnels et non structurés. Cette architecture de « big data » assure le stockage des données sur plusieurs systèmes, permettant ainsi aux applications NoSQL d'évoluer par ajout incrémental afin d'offrir une augmentation des capacités à la demande et une meilleure rentabilité.



Ces architectures de big data peuvent donner du sens à des volumes colossaux de données mais, pour y parvenir, l'infrastructure doit répondre à certaines exigences :

- Un stockage adapté au volume de données
- Une RAM permettant de déplacer et de charger les données en fonction des besoins
- Une puissance de traitement correspondant au niveau de performances nécessaire dans la solution
- Un réseau capable de connecter les magasins de données distribués avec de faibles temps de latence pour optimiser les performances

Conscientes de ces exigences, de nombreuses entreprises tirent parti du cloud computing en tant qu'infrastructure sous-jacente pour faire évoluer horizontalement leurs environnements de big data. Dans ces environnements, les composants les plus courants sont les serveurs de cloud public et les serveurs nus.

Les **quatre V** des Big Data

Volume : pensez en pétaoctets. Historiques Internet, archives publiques, documents confidentiels internes : les entreprises stockent tout.

Variété : Volumes importants de données structurées et non structurées, provenant notamment des e-mails, des réseaux sociaux, des vidéos, des images, des données météorologiques, des blogs, etc.

Vélocité : Des données sont générées à chaque seconde et les requêtes sur des informations pertinentes doivent être satisfaites en temps réel.

Valeur ajoutée : Des perspectives pertinentes provenant des big data et allant au-delà des résultats de l'intelligence traditionnelle des requêtes et rapports. Ces perspectives peuvent être transformées en analyse prévisionnelle pour repérer les tendances et les schémas.

Les serveurs nus face aux serveurs virtuels

Envisagez les serveurs nus et les serveurs virtuels comme deux outils dans la même boîte à outils. Il est impossible de dire que l'un est meilleur que l'autre ; chacun présente ses points forts et ses points faibles.

Les serveurs nus offrent aux clients un accès direct et exclusif aux ressources matérielles brutes. Les serveurs virtuels sont des instances cloud indépendantes provisionnées par un hyperviseur sur un nœud matériel public (partagé) ou privé.

Serveurs nus : une puissance de traitement brute

Les serveurs nus (parfois appelés serveurs dédiés) représentent la solution idéale pour les charges de travail intensives en puissance de traitement et en E/S. Ces serveurs sont intégralement dédiés à un seul utilisateur. Cela signifie que vos performances ne seront pas affectées par des voisins encombrants.

De même, étant donné que les serveurs nus ne s'exécutent pas sur un hyperviseur, les charges de travail ne paient pas la « taxe hyperviseur » : une légère réduction des performances due au fait que l'hyperviseur sert d'intermédiaire entre le système d'exploitation et le matériel.

Sans hyperviseur pour isoler le matériel, les serveurs nus sont généralement bien plus longs à provisionner et à configurer que les serveurs virtuels. Lorsque l'infrastructure doit pouvoir évoluer rapidement, les serveurs nus sont généralement à éviter. Pour corriger ce problème, IBM Cloud a automatisé le déploiement et le contrôle des serveurs nus, mettant ainsi en service les configurations sélectionnées en 20 à 30 minutes et les serveurs entièrement personnalisés (votre choix de processeur, de cœurs, de RAM, de stockage, de ports, etc.) en deux à quatre heures.

Serveurs virtuels : flexibilité et évolutivité

Les applications et charges de travail dont la taille peut faire le grand écart ou qui doivent rester réactives sur un marché en évolution constante conviennent parfaitement aux serveurs virtuels. Les serveurs virtuels sont provisionnés sur un hyperviseur, dans un environnement de

cloud public à un seul ou plusieurs utilisateurs. Les ressources de serveurs virtuels peuvent être déployées en seulement cinq minutes, avec une tarification mensuelle ou horaire ; cela vous permet ainsi d'évoluer de manière horizontale en ajoutant des serveurs très rapidement.

Serveurs nus et serveurs virtuels associés

IBM Cloud provisionne simultanément des serveurs nus et des serveurs virtuels dans un seul environnement cloud unifié afin de proposer aux clients un choix et un contrôle sur les ressources qui alimenteront leurs différentes charges de travail.

Les besoins stratégiques évoluent. Notre solution est conçue de manière à ce que vous puissiez vous concentrer sur les besoins actuels sans vous soucier de leur évolution dans quelques jours, semaines ou mois.

La plateforme et l'infrastructure IBM Cloud sont entièrement évolutives :

- Ajouter des serveurs nus et des serveurs virtuels à la demande
- Revoir les ressources à la baisse lorsque cela est nécessaire pour réduire les coûts
- Payer sur une base horaire ou mensuelle afin de répondre aux différents calendriers de projets
- Pas de contrats à long terme

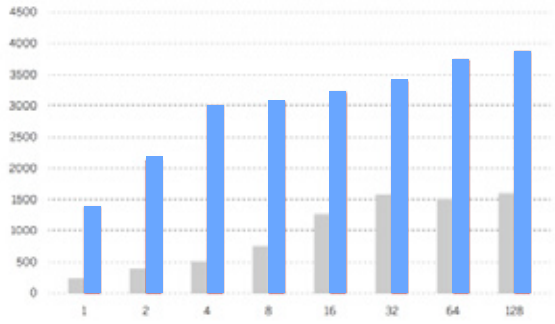
Big Data : performances

Pour déterminer si les applications de big data sont mieux adaptées à des serveurs nus ou à des serveurs virtuels, nous avons réalisé des tests d'évaluation qui mesurent les performances et la stabilité des deux plateformes. Afin de mesurer les performances de manière précise, un ingénieur IBM Cloud a configuré de manière équivalente un environnement de serveurs nus et un environnement de serveurs virtuels pour qu'ils effectuent des recherches et mettent à jour un ensemble de données MongoDB en utilisant l'outil d'évaluation à disposition (détails dans l'Annexe A).

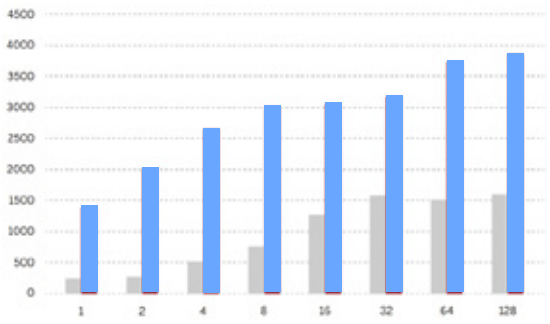
L'outil d'évaluation a enregistré le nombre d'opérations de lecture et d'écriture par seconde de chaque grappe, en fonction du nombre de clients concurrents impliqués. Les résultats des tests ont été sans appel. Tous les environnements de serveurs nus ont été plus performants que les environnements équivalents avec serveurs virtuels en nombre moyen de lectures/écritures.

Les serveurs virtuels face aux serveurs nus

Nombre moyen d'opérations de lecture par seconde par client concurrent



Nombre moyen d'opérations d'écriture par seconde par client concurrent



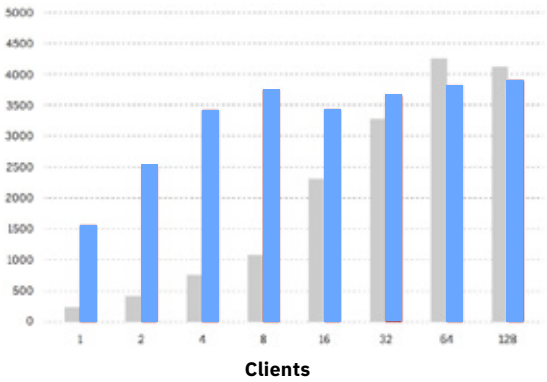
Serveurs virtuels **Serveurs nus**

Etant donné que l'environnement de serveurs nus pouvait tirer directement parti des ressources matérielles du serveur sans avoir à les partager avec d'autres utilisateurs, il a affiché des performances jusqu'à six fois supérieures à celles de l'environnement équivalent doté de serveurs virtuels.

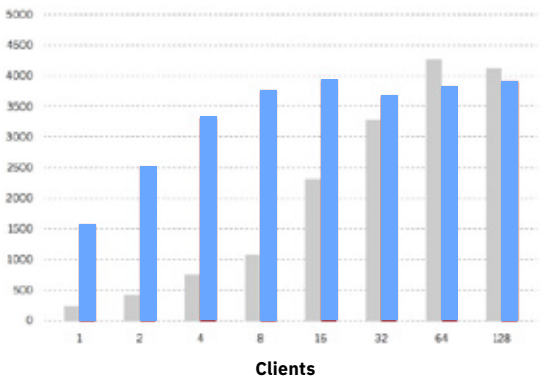
Outre le recensement des opérations de lecture et d'écriture par seconde, l'outil d'évaluation a également enregistré les pics de performances de chaque environnement, et ces résultats ne sont pas non plus négligeables (pour une autre raison) :

Les serveurs virtuels face aux serveurs nus

Nombre moyen d'opérations de lecture par seconde par client concurrent



Nombre moyen d'opérations d'écriture par seconde par client concurrent



Serveurs virtuels **Serveurs nus**

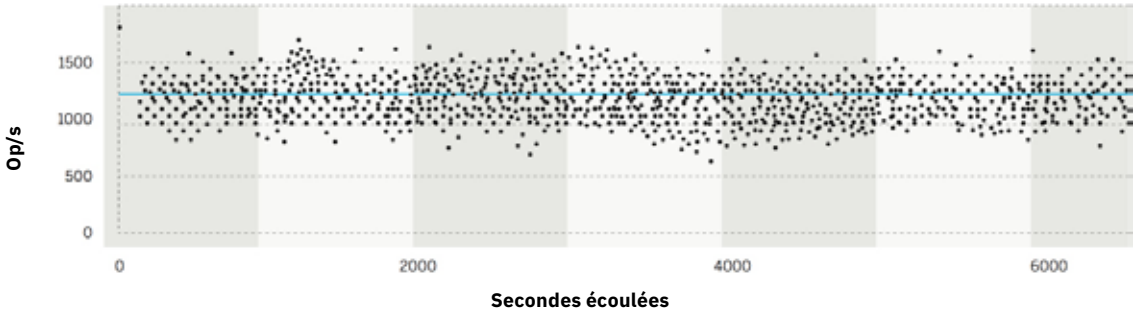
Les pics de lectures et d'écritures par seconde dans les environnements de serveurs nus étaient très proches du nombre moyen de lectures et d'écritures enregistré pour cet environnement. En revanche, les pics s'écartaient nettement des moyennes dans l'environnement de serveurs virtuels. Dans deux des scénarios, les serveurs virtuels n'ont pas atteint de pic supérieur au pic des serveurs nus. Mis en contexte, avec le nombre moyen d'opérations par seconde enregistré par l'environnement de serveurs virtuels, ces résultats mettent en évidence l'autre indicateur de performances clé pour les charges de travail de big data : **la stabilité.**

Big Data : stabilité

Les performances présentent un intérêt uniquement si elles sont stables. Dans notre test de performances, l'environnement de serveurs virtuels a pu atteindre 4 500 opérations de lecture par seconde à son pic mais, en moyenne, son résultat est de 1 500 opérations de lecture par seconde. Si les performances d'un environnement varient autant d'une seconde à l'autre, il est extrêmement compliqué de mettre en œuvre une solution capable de gérer une charge de travail en pleine évolution. Pour comparer la stabilité des résultats sur les serveurs nus par rapport aux serveurs virtuels, un ingénieur IBM Cloud a configuré deux grappes Riak à cinq nœuds et a simulé des déploiements de charges en utilisant Basho Bench (détails dans l'Annexe B). Ce test a observé et consigné les opérations par seconde sur une période de deux heures :

Serveurs nus :
opérations par seconde en charge (2 heures)

Serveurs virtuels :
opérations par seconde en charge (2 heures)



L'environnement de serveurs nus affiche un nombre supérieur d'opérations par seconde tout au long du test mais, surtout, les résultats sont bien plus concentrés autour de la moyenne. Si les performances fluctuent considérablement d'une seconde à l'autre dans l'environnement de serveurs virtuels, la planification des capacités devient compliquée. Quelles statistiques sont à prendre en compte au moment de décider si l'environnement doit être augmenté ou réduit ? Si vous adaptez votre environnement aux pires résultats enregistrés et que les performances sont satisfaisantes, vous parviendrez à la conclusion que vous avez sur-provisionné les ressources. Opter pour des capacités basiques en fonction des meilleurs résultats entraînera probablement un manque de performances de l'environnement. Et calquer les capacités sur les résultats moyens est un jeu de pile ou face entre ces deux alternatives.

Les entreprises s'appuient sur **des résultats stables** pour prévoir les tendances, établir des budgets et prendre des décisions importantes. La planification des environnements d'infrastructure cloud doit suivre la même règle.

Les big data doivent s'appuyer sur des serveurs nus

Les promesses de déploiement simple et rapide sont extrêmement alléchantes. Même si certaines applications s'exécutent mieux sur des serveurs virtuels dans un environnement de cloud public, ce n'est pas le cas des big data.

Il est important de noter que :

- Les deux caractéristiques les plus importantes pour un environnement cloud exécutant des charges de travail intensives en E/S telles que les big data sont **les performances** et **la stabilité**.
- Les serveurs nus peuvent être configurés et optimisés de manière à **fournir des résultats inégalés** lors de la distribution et **du traitement de volumes importants de données**.
- Les serveurs virtuels traitant des charges de travail intensives en E/S peuvent **être entravés par l'utilisation de ressources des autres clients** lorsque plusieurs utilisateurs partagent le même nœud hôte de serveur virtuel.
- Les ressources de serveurs nus **sont locales et non partagées.** Ainsi, les charges de travail s'exécutent de manière beaucoup plus stable que dans les environnements de serveurs virtuels partagés ou en réseau.
- Les serveurs virtuels peuvent être provisionnés et évoluer horizontalement bien plus vite que les serveurs nus mais les charges de travail n'ayant pas besoin de ce niveau de réactivité tirent un meilleur partir des performances et de la stabilité des serveurs nus.

Pourquoi IBM Cloud est-il un fournisseur idéal de charges de travail de big data ?

Une technologie inégalée : IBM Cloud vous propose l'infrastructure cloud la plus performante qui existe. Que vos big data s'étendent au niveau mondial ou local, nos centres informatiques internationaux, ainsi que nos serveurs virtuels et nos serveurs nus de pointe, peuvent relever le défi.

Un réseau transparent : Notre réseau de pointe intègre des réseaux publics, privés et de gestion interne afin de proposer un débit élevé ; ce qui constitue un élément essentiel dans le transfert et l'analyse des big data.

Gestion et automatisation complètes : Nous avons développé une nouvelle sorte de solution cloud : une plateforme automatisée tout-en-un. Chaque serveur, périphérique de stockage, service de gestion et élément de sécurité peut être contrôlé par le biais d'un seul système de gestion, entièrement accessible par notre API, notre portail client et même les applications mobiles.

Découvrez les serveurs nus et les serveurs virtuels IBM Cloud à l'adresse http://ibm.co/bare-metal et apprenez en plus sur les solutions de big data à la demande et les meilleures pratiques spécifiques aux applications pour Riak, Hadoop et MongoDB à l'adresse http://ibm.co/big-data.

D'autres questions ? Interrogez un expert : http://ibm.co/contact-us ou appelez-nous au : 214-442-0600.

Exécutez vos big data sur des serveurs nus. Nos experts IBM Cloud vous aideront à créer une infrastructure cloud hautes performances afin de répondre au mieux aux besoins de vos big data.

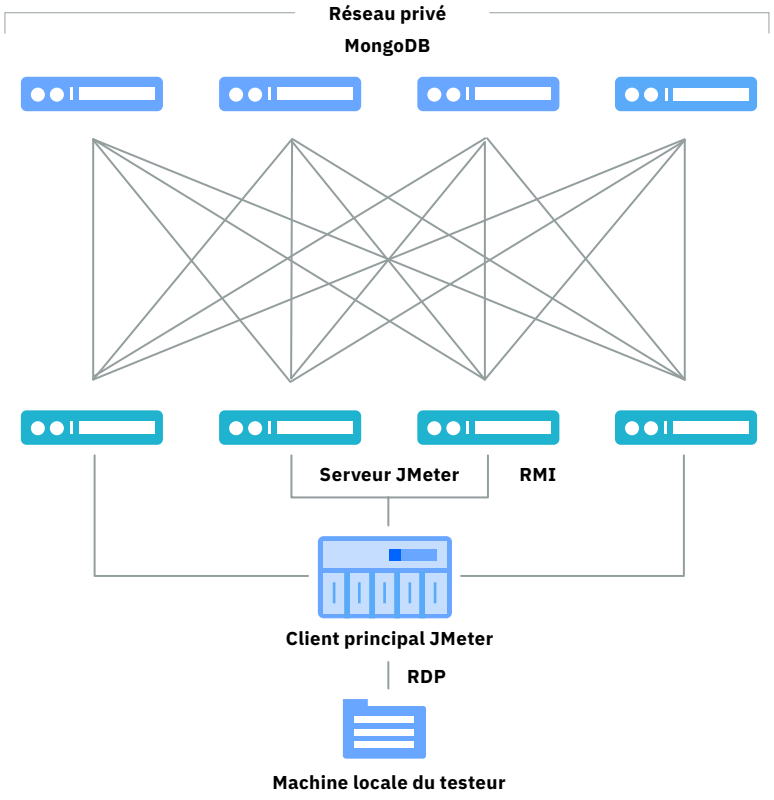
Annexe A

Méthode de test des performances des big data : MongoDB

Des ensembles de données constitués de documents de 500 Ko ont été préchargés dans des instances MongoDB uniques sur chaque serveur. Les ensembles de données ont été créés avec différentes tailles par rapport à la mémoire disponible afin de pouvoir tester des ensembles plus importants (2x) ou plus petits que cette mémoire disponible. Le test a également veillé à ce que l'ensemble de données soit modifié assez fréquemment lors de l'exécution pour éviter la mise en cache de toutes les données dans la mémoire.

One fois les ensembles de données créés, des instances de serveur JMeter à 4 cœurs et 16 Go de RAM ont été utilisées pour réaliser l'évaluation avec les outils MongoDB. Le diagramme ci-dessous illustre notre configuration de l'environnement de test.

Ces serveurs Jmeter jouent le rôle de clients générant du trafic sur les instances MongoDB. Chaque client a généré des requêtes et des demandes de mises à jour aléatoires avec un ratio de six requêtes par mise à jour (Les demandes de mise à jour du test servaient à s'assurer que les données n'étaient pas autorisées à se mettre entièrement en cache dans la mémoire et à ne jamais subir de lectures de la part du disque). Ces tests ont été conçus de manière à créer une charge extrême sur les serveurs, provenant d'un nombre exponentiel de clients, jusqu'à ce que le ressources du système soient saturées. Nous avons enregistré les performances de l'application MongoDB dans ce contexte.



Configuration de test

- Ensemble de données (32 Go de documents de 0,5 Mo)
- 200 itérations d'opérations 6:1 requête/mise à jour
- Les connexions client concurrentes ont augmenté de manière exponentielle de 1 à 128
- Le test s'est déroulé sur 48 heures

Annexe A (suite)

Méthode de test des performances des big data : MongoDB

Les serveurs nus face aux serveurs virtuels

	Nœud de serveur nu	Nœud de serveur virtuel
Cœur	Doubles processeurs Intel 5670 à 6 cœurs	26 unités de calcul virtuelles
Système d'exploitation	CENTOS 64 bits	CENTOS 64 bits
RAM	36 Go de RAM	30 Go de RAM
RAID	SSD 2 x 64 Go RAID1 (journal)	2 x 64 Go de stockage réseau RAID1 (journal)
SAS	SSD 4 x 400 Go RAID10 (données)	SSD 4 x 300 Go RAID10 (données)
Réseau	Réseau 1 Go | Connecté	Réseau 1 Go

Annexe B

Méthode de test des performances des big data : Riak

Des grappes de 5 nœuds dotées de Riak 1.3.1 ont été créées sur des serveurs nus et sur des serveurs virtuels de cloud public. Des ajustements ont été réalisés au niveau du système d'exploitation de chaque serveur (exécutant CentOS 64 bits) pour optimiser les performances Riak :

Noatime

Nodiratime

barrier=0

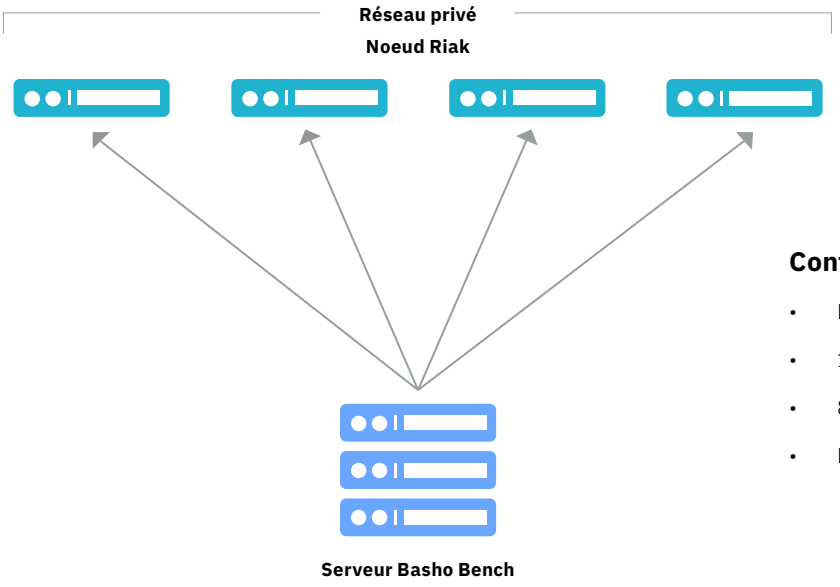
data=writeback

ulimit -n 65536

Les paramètres Noatime et Nodiratime communs éliminent le besoin d'écriture au cours des lectures afin d'augmenter les performances et de réduire l'usure des disques. Les paramètres barrier et writeback sont un peu moins communs et peuvent ne pas correspondre à ce que vous définiriez en temps normal. Bien que ces paramètres présentent un risque très mince de perte des données en cas de panne de disque, il est à noter que la solution Riak est déployée dans des anneaux de cinq nœuds avec des données disponibles de manière redondante dans les différents nœuds de l'anneau.

Sur la base de ces éléments et sachant que chaque nœud est également déployé avec un système de stockage RAID10, le risque mineur de perte de données en cas de panne d'un disque dans toute la solution n'impacterait aucunement l'intégrité de l'ensemble de données (car il existe de nombreuses copies redondantes de ces données). Etant donné le peu de risques induits, les augmentations de performances de ces deux paramètres justifient leur utilisation.

Outre l'ajustement et la configuration de tous les nœuds dans les grappes, nous avons mis en place l'outil de test Basho (Basho Bench) afin de simuler à distance la charge sur les déploiements. Basho Bench vous permet de créer un projet de test configurable pour les grappes Riak en paramétrant plusieurs collaborateurs susceptibles d'utiliser un type de disque et de générer de la charge. Il est associé à une application Erlang avec un exemple de fichier de configuration que vous pouvez modifier afin d'indiquer les spécifications de la concurrence, la taille de l'ensemble de données et la durée de vos tests. Les résultats peuvent être affichés en tant que données CSV et il existe une solution facultative qui vous permet de générer des graphiques. Exemple de graphique simplifié de notre environnement de test :



Configuration de test

- Ensemble de données : 400 Go
- 10:1 opérations requête/mise à jour
- 8 connexions de clients concurrentes
- Durée du test : 2 heures

Annexe B (suite)

Méthode de test des performances des big data : Riak

Riak - Test de stabilité

Grappe à 5 nœuds sur serveur nu par rapport à une grappe à 5 nœuds sur serveur virtuel

	Nœud de serveur nu	Nœud de serveur virtuel
Cœur	Doubles processeurs Intel 5670 à 6 cœurs	26 unités de calcul virtuelles
Système d'exploitation	CENTOS 64 bits	CentOS 64 bits
RAM	36 Go de RAM	30 Go de RAM
RAID	SAS 15K 4 x 300 Go | RAID10	4 x 300 Go de stockage réseau
SAS	Réseau 1 Go - Connecté	Réseau 1 Go

Compagnie IBM France
17, avenue de l'Europe,
92275 BOIS COLOMBES
CEDEX

IBM Ireland Limited enregistré en Irlande en tant que société numéro 16226.

Please Recycle

KUW12525-FRFR-02