GSA | Remarques en situation

Google Search Appliance

Architectures de déploiement

Août 2014

Architectures de déploiement

Google Search Appliance (GSA) propose plusieurs fonctionnalités de déploiement dans différentes architectures pour répondre aux divers besoins de disponibilité, de vitesse et d'évolutivité. Ce document présente les consignes d'utilisation de ces fonctionnalités d'architecture, ainsi que d'autres composants de déploiement d'assistance pour la conception d'une architecture de solution GSA appropriée.

À propos de ce document

Les recommandations et les informations contenues dans ce document sont issues de notre travail sur le terrain avec plusieurs clients et dans différents environnements. Nous remercions nos clients et partenaires d'avoir partagé leurs expériences et impressions.

Sujets abordés	Ce document donne quelques exemples d'architectures de déploiement classiques et indique comment déployer le système Google Search Appliance pour répondre aux besoins d'une grande communauté d'utilisateurs caractérisée par des exigences de performance très élevées et une très faible tolérance aux pannes.
Rôles et responsabilités	● Administrateurs GSA : déployer et configurer les systèmes Google Search Appliance de la manière la plus adaptée aux besoins généraux de recherche des utilisateurs de l'organisation. ● Administrateurs réseau : vérifier que le réseau de l'entreprise dispose de la capacité et la disponibilité nécessaires pour faire face au trafic que pourront engendrer les différentes configurations. ● Propriétaires de contenu : distribuer l'accès aux diverses sources de contenus auxquelles le système Google Search Appliance devra accéder pour permettre l'exploration, l'indexation et la recherche.
Environnement informatique	Système Google Search Appliance, SNMP, équilibreur de charge, commutateur de réseau, serveur proxy de recherche
Phases de déploiement	Planification du déploiement
Autres sources d'information	● Learngsa.com fournit des ressources pédagogiques pour le système Google Search Appliance. ● La documentation du produit GSA fournit des informations complètes sur le système Google Search Appliance. ● Le portail d'assistance Google for Work permet d'accéder à l'assistance Google.

Sommaire

Chapitre 1 Composants de l'architecture

Lorsque vous concevez l'architecture de mise en œuvre du système Google Search Appliance, vous devez tenir compte des diverses fonctions du système, ainsi que des points d'intégration avec les composants d'assistance à son déploiement. Ce chapitre offre une vue d'ensemble du rôle que chaque composant, comme les connecteurs GSA et les mécanismes de sécurité, joue au sein de l'architecture d'un déploiement du système Google Search Appliance.

Fonctions de base de l'architecture du système Google Search Appliance

Le système GSA offre la possibilité de déployer trois fonctions principales pour répondre à vos besoins de haute disponibilité, de performances et d'évolutivité :

- Mise en miroir GSA
- Fédération de GSA
- Exploration et traitement distribués de GSA

Mise en miroir GSA

La mise en miroir GSA réalise la réplication d'un index de recherche et de la plupart des paramètres de configuration sur un ou plusieurs systèmes de recherche "réplica".

La mise en miroir des architectures peut être conçue pour une configuration active-passive ou active-active, et peut permettre de bénéficier d'une haute disponibilité et d'un débit élevé lors d'un déploiement d'un système GSA.

Pour plus d'informations sur la mise en miroir GSA, consultez les articles du centre d'aide GSA (en anglais) Configuring GSA Mirroring (Configuration de la mise en miroir GSA) et Specifications and Limits: GSA Mirroring (Spécifications et limites d'utilisation : Mise en miroir GSA).

Fédération de GSA

La fédération de GSA permet de configurer un groupe de systèmes de recherche de sorte à pouvoir rechercher les documents indexés séparément sur plusieurs systèmes de recherche à l'aide d'une seule requête. Les systèmes de recherche dans la configuration indexent chacun différents ensembles de documents et sont installés avec leurs propres collections et frontaux, ainsi que d'autres fonctions configurables par l'administrateur. Lorsqu'un utilisateur effectue une recherche, les systèmes de recherche communiquent entre eux afin de fusionner les résultats à partir de leur index respectif.

Un environnement GSA fédéré est généralement utilisé lorsque la taille du corpus de documents dépasse les capacités de recherche et d'indexation d'un système Google Search Appliance unique ou pour rassembler des corpus répartis géographiquement.

Dans cette configuration, l'un des systèmes de recherche est désigné comme le système, ou nœud, principal, tandis que les autres systèmes sont désignés comme étant les systèmes, ou nœuds, secondaires.

Pour plus d'informations sur la mise en miroir GSA, consultez les articles du centre d'aide GSA (en anglais) Configuring GSA Mirroring (Configuration de la mise en miroir GSA) et Specifications and Limits: GSA Mirroring (Spécifications et limites d'utilisation : Mise en miroir GSA).

Exploration et traitement distribués de GSA

La fonctionnalité d'exploration et de traitement distribués du système Google Search Appliance améliore la capacité d'indexation de documents du système de recherche. Grâce à cette fonctionnalité, plusieurs systèmes de recherche gèrent les requêtes comme s'ils ne constituaient qu'un seul système de recherche. Lorsqu'elle est activée, toutes les tâches d'exploration, d'indexation et de traitement sont configurées sur
un système de recherche unique, appelé le "maître administrateur", tandis que les autres sont des systèmes "non maîtres".

Par exemple, si deux systèmes de recherche possèdent chacun une licence pour explorer 100 millions de documents, l'activation de la fonctionnalité d'exploration et de traitement distribués permet d'explorer un total de 200 millions de documents.

Pour plus d'informations sur l'exploration et le traitement distribués de GSA, consultez les articles du centre d'aide GSA (en anglais) Configuring Distributed Crawling and Serving (Configuration de l'exploration et du traitement distribués) et Specifications and Limits: GSA Distributed Crawling and Serving (Spécifications
et limites d'utilisation : Exploration et traitement distribués de GSA).

Combinaison de la mise en miroir GSA, de la fédération de GSA et de l'exploration et du traitement distribués

Depuis GSA version 7.2, les combinaisons d'architectures suivantes sont possibles :

La mise en miroir GSA peut être utilisée dans les environnements suivants :

- Réseau indépendant de mise en miroir **GSA** avec un système maître lié à un ou plusieurs réplicas.
- Réseau de fédération de **GSA**.
- Réseau d'exploration et de traitement distribués **GSA**, dans lequel il est possible de créer des réplicas de chaque système non-maître.

La fédération de GSA peut être utilisée dans les environnements suivants :

- Réseau indépendant de fédération de **GSA** avec un système principal lié à un ou plusieurs nœuds.
- Réseau de fédération de **GSA avec mise en miroir GSA** de n'importe quel nœud.

La fonctionnalité d'exploration et de traitement distribués GSA peut être utilisée dans les environnements suivants :

- Réseau indépendant d'exploration et de traitement distribués **GSA** avec un système maître lié à un ou plusieurs systèmes non-maîtres.
- Réseau d'exploration et de traitement distribués **GSA avec mise en miroir GSA** de tous les nœuds.

Chapitre 2 Architecture pour une haute disponibilité

Haute disponibilité du système GSA

Pour bénéficier d'une haute disponibilité lors du traitement des requêtes par le système GSA, déployez un équilibreur de charge sur plusieurs systèmes qui contiennent le même index de recherche. Grâce aux fonctionnalités de contrôle et de détection des pannes, en cas de basculement, l'équilibreur de charge va permettre de rediriger le trafic vers un réplica GSA. Cette procédure peut être effectuée manuellement ou automatiquement (en fonction de l'équilibrage de charge ou de la solution de contrôle que vous utilisez). Le système Google Search Appliance ne propose pas de fonctionnalités d'équilibrage de charge et de basculement automatiques. Ces procédures doivent donc être gérées par des composants externes.

La mise en miroir GSA permet d'obtenir un index de recherche cohérent entre différents systèmes. Toutefois, lorsqu'il est impossible d'utiliser cette fonctionnalité (par exemple, en raison d'une connexion réseau insuffisante entre les systèmes), vous devez configurer chaque système de recherche pour explorer, indexer et traiter le contenu identique afin de garantir la cohérence des résultats en cas de basculement.

Remarque sur les licences : Si l'ensemble du trafic est entièrement dirigé vers le système GSA principal, cette architecture est appelée configuration *active-passive*. Dans ce cas, le système principal ne requiert qu'une licence GSA de "production", tandis que les systèmes réplicas peuvent utiliser une licence de "sauvegarde".

Pour en savoir plus sur l'équilibrage de charge GSA, consultez l'article du centre d'aide GSA (en anglais) Configuring Search Appliances for Load Balancing or Failover (Configuration des systèmes de recherche pour l'équilibrage de charge ou le basculement).

Surveillance et détection des défaillances

Utilisation du SNMP pour surveiller le système GSA

Le système Google Search Appliance accepte l'intégration du protocole SNMP (Simple Network Management Protocol), ce qui vous permet de recevoir des messages en cas de modification de son état de fonctionnement. Un signal est émis pour indiquer les requêtes SNMP au système de recherche sur le port UDP 161,, compatible avec SNMP v1, v2 et v3.

Le serveur SNMP du système Google Search Appliance fournit un sous-ensemble des informations d'état du système de recherche disponibles dans la console d'administration. Le système de recherche accepte les commandes SNMP `Get` et `GetNext`, mais pas la commande `Trap`, ni la définition de valeurs via la commande SNMP `Set`.

Remarque : Cette fonctionnalité n'est généralement configurée que si vous utilisez déjà le protocole SNMP pour gérer d'autres appareils sur votre réseau, tels que des routeurs, des commutateurs ou encore des serveurs d'application ou de stockage. Dans le cas contraire, nous vous conseillons de mettre en œuvre un système de surveillance de serveur personnalisé (voir ci-après).

Pour plus d'informations, consultez la page d'aide de la console d'administration GSA : **Administration > Configuration SNMP**.

Utilisation d'un système de surveillance personnalisé
En l'absence de système de surveillance pour vérifier l'état de fonctionnement du système GSA, vous pouvez configurer un système personnalisé à l'aide d'une simple application Web. Cette application Web peut utiliser l'API d'administration pour accéder aux informations d'état du système GSA, pour ensuite les transmettre dans une page d'état consultable par l'administrateur.

Vous pouvez également utiliser l'API d'administration pour surveiller le système Google Search Appliance et stocker les informations dans une base de données ou des fichiers journaux. Vous pouvez ensuite les consulter ultérieurement à des fins d'analyse.

Vous trouverez les bonnes pratiques de mise en œuvre d'un système de surveillance pour les requêtes de recherche dans l'article Monitoring GSA Serving (Surveillance du fonctionnement de GSA).

Haute disponibilité avec les connecteurs

Deux fonctions principales des connecteurs GSA peuvent avoir une incidence sur la disponibilité : 1) le balayage et la recherche de contenu ; et 2) les services de sécurité gérant l'authentification et l'autorisation.

Haute disponibilité avec les connecteurs gérant le balayage de contenu

Le balayage de contenu n'est pas le service le plus utilisé pour bénéficier d'un système à haute disponibilité, toutefois il se révèle parfois indispensable dans les environnements dont la priorité est l'actualisation du contenu. En cas d'utilisation d'un connecteur, l'approche à privilégier pour obtenir une haute disponibilité lors des tâches d'exploration dépend du type de connecteur utilisé et de la version de la structure (par exemple, v3.x ou v4.x).

Pour la plupart des connecteurs présents dans la structure du gestionnaire de connecteurs 3.x (à l'exception du connecteur du système de fichiers), les interruptions (liées au connecteur lui-même) qui se produisent lors de l'exploration requièrent une intervention manuelle : pour redémarrer le connecteur ou activer un autre connecteur pour l'exploration. Il s'agit de connecteurs à états, par conséquent vous devez réinitialiser le balayage ou redémarrer les processus de récupération faisant suite à des interruptions d'exploration gérée par les connecteurs. Dans certains cas, vous pouvez sauvegarder les informations d'état du balayage relatives au connecteur (comme le fichier d'état XML du connecteur SharePoint) afin de recommencer le balayage à partir de l'état actuel, au lieu d'avoir à effectuer un balayage complet.

Pour garantir une haute disponibilité pour l'exploration avec les connecteurs de la structure 4.x (ainsi que le connecteur du système de fichiers 3.x), déployez des instances de "sauvegarde" du connecteur et liez-les à un équilibreur de charge. Vous devez désactiver les tâches de recherche et de balayage complet pour ces connecteurs secondaires afin qu'ils réalisent uniquement la récupération de documents spécifiques à la

demande du système GSA. De cette façon, si un connecteur est en panne, l'équilibreur de charge est en mesure de diriger les requêtes d'exploration vers un autre connecteur et de maintenir la disponibilité des tâches d'exploration. Dans ce cas, une intervention manuelle est tout de même indispensable pour afficher la liste des résultats et procéder à la recherche de nouveaux contenus à partir du connecteur principal configuré pour le balayage.

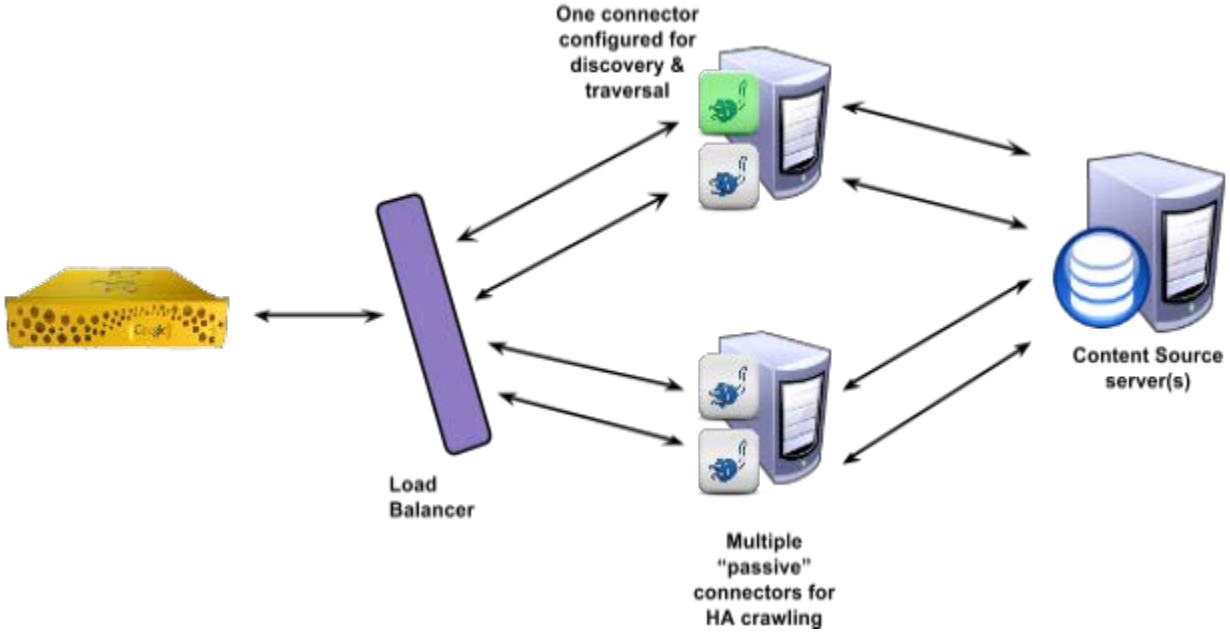

Diagramme 1 : Exemple d'architecture pour obtenir une haute disponibilité pour le balayage de contenu via des connecteurs

Haute disponibilité avec les connecteurs gérant les mécanismes de sécurité

Pour les services d'authentification et d'autorisation, vous pouvez déployer plusieurs connecteurs et mettre en place un équilibreur de charge afin de répartir le trafic entre les connecteurs. Pour plus d'informations sur cette configuration, reportez-vous au document (en anglais) GSA Notes from the Field: Introduction to Content Integration.

Haute disponibilité des mécanismes de sécurité

Dans les environnements de recherche sécurisée, la haute disponibilité des mécanismes de sécurité est primordiale, car en cas d'interruption des mécanismes d'authentification et d'autorisation, la recherche sécurisée n'est plus assurée. Pour garantir la haute disponibilité de ces mécanismes, la stratégie la plus courante consiste à déployer un équilibreur de charge pour plusieurs instances du mécanisme de sécurité. Par exemple, vous pouvez déployer un équilibreur de charge pour plusieurs connecteurs traitant les requêtes d'authentification ou d'autorisation ou configurer une URL d'équilibrage de charge pour les systèmes d'authentification unique.

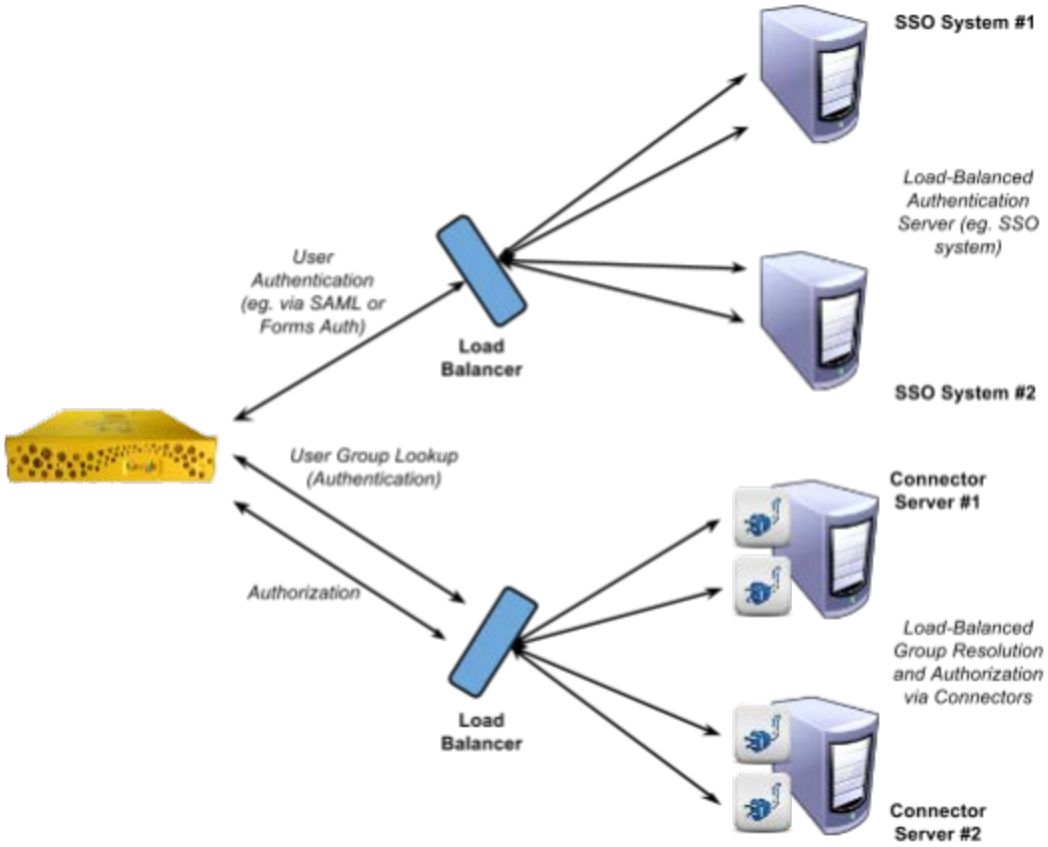

Diagramme 2 : Exemple d'architecture pour optimiser la disponibilité et les performances des mécanismes de sécurité

Chapitre 3 Architecture pour des performances élevées

Équilibrage de charge du système GSA pour optimiser les performances des requêtes

Le nombre de requêtes par seconde qu'un seul système de recherche peut traiter dépend du temps nécessaire pour effectuer la requête, du nombre de requêtes simultanées et du type de fonctionnalités du frontal activées. Un seul système de recherche peut traiter jusqu'à 50 requêtes simultanées, mais le nombre de requêtes par seconde qu'un système GSA peut gérer est inférieur, notamment lors de recherches sécurisées. Pour plus d'informations sur les requêtes simultanées, consultez les articles du centre d'aide : Designing a search solution (Création d'une application de recherche) et How many concurrent users can a GSA handle? (Combien d'utilisateurs un système GSA peut-il gérer simultanément ?).

Pour atteindre les niveaux de débit de requêtes souhaités, déployez plusieurs systèmes associés à un équilibreur de charge qui répartit équitablement le trafic de recherche entre les systèmes GSA. Chaque système doit contenir le même index, via la mise en miroir GSA ou l'exploration indépendante du même contenu. Dans ce cas, les systèmes GSA se trouvent dans une configuration *active-active* (présentation active des résultats) et doivent tous disposer d'une licence en tant que système de "production".

Dans les configurations d'équilibrage de charge pour la recherche sécurisée, il est recommandé d'utiliser les sessions permanentes, car chaque session de recherche sécurisée est associée à un système spécifique. Par conséquent, si un utilisateur déjà authentifié est redirigé vers un système différent, il est de nouveau invité à saisir ses identifiants.

Performances du mécanisme de sécurité

Un goulot d'étranglement nuisant aux performances se produit régulièrement lors du processus d'autorisation d'accès aux résultats de recherche, notamment en cas d'utilisation de liaisons tardives (contrôles de sécurité en temps réel au niveau du système de contenu) pour déterminer les droits d'accès à un document.

Il est recommandé d'utiliser les liaisons précoces pour attribuer les autorisations en raison des gains de performances réalisés lors des contrôles de sécurité sur le système GSA. Toutefois, dans certains cas, vous devez utiliser les liaisons tardives en raison de l'impossibilité d'extraire les listes de contrôle d'accès (LCA) ou lorsque les autorisations de sécurité pour l'accès aux documents changent fréquemment, ce qui nécessite des contrôles en temps réel.

Pour améliorer les performances du processus d'autorisation, différentes stratégies sont possibles :

- Si une surcharge dégrade les performances du processus d'autorisation, déployez plusieurs instances du mécanisme d'autorisation bénéficiant de l'équilibrage de charge (par exemple, plusieurs fournisseurs SAML).
- Si un connecteur gère le processus d'autorisation, ainsi que l'exploration et le balayage de contenu, déployez d'autres connecteurs dédiés au processus d'autorisation et employez

un proxy ou un équilibreur de charge pour séparer le trafic d'exploration de celui des requêtes d'autorisation depuis le système GSA.

Les problèmes de performances du processus d'authentification sont moins souvent évoqués (du fait que les utilisateurs n'y ont recours qu'une fois par session). Néanmoins, si vous souhaitez optimiser les performances de ces mécanismes, vous pouvez employer les stratégies précédentes. Voir le diagramme 2 pour un exemple d'optimisation des performances des mécanismes de sécurité.

Considérations sur les performances des fonctionnalités du système GSA

Bien que le système GSA offre de nombreuses fonctionnalités pour améliorer l'expérience utilisateur, lorsque vous concevez un déploiement GSA dans le but d'optimiser les performances, plusieurs éléments sont à prendre en compte concernant l'impact de l'activation des fonctions de recherche sur les performances :

- **Si vous utilisez la navigation dynamique** avec la recherche sécurisée, le système GSA doit traiter un plus grand nombre de requêtes d'autorisation. Sans la navigation dynamique, le système GSA traite les requêtes d'autorisation jusqu'à récupérer 10 résultats que l'utilisateur est autorisé
  à consulter, dans le cas de liaisons tardives (en supposant que 10 résultats soient demandés), et 1 000 résultats autorisés dans le cas de liaisons précoces. Avec la navigation dynamique, le système GSA traite les requêtes d'autorisation jusqu'à récupérer *10 000 résultats autorisés*. Ce chiffre a un impact considérable sur les performances dans une configuration avec des liaisons tardives et peut également avoir des conséquences sur une configuration avec des liaisons précoces si l'utilisateur a un accès limité au contenu. Pour plus d'informations, consultez l'article du centre d'aide : Search results are slow when dynamic navigation is enabled (L'affichage des résultats de recherche est lent lorsque la navigation dynamique est activée)

- **La mise en clusters dynamique des résultats** effectue de nombreuses requêtes d'autorisation afin de déterminer les clusters de résultat et envoie également une requête HTTP supplémentaire au système GSA.

- **Les requêtes OneBox** sont gérées de manière synchrone et donc augmente le temps de réponse global aux requêtes, notamment lorsque les requêtes sont envoyées à un système tiers.

- **Les suggestions de requête** envoient des requêtes HTTP supplémentaires au système GSA pour afficher des suggestions à l'utilisateur pendant la saisie. Lorsque le système déployé est utilisé par de nombreux utilisateurs, cette fonctionnalité peut ajouter une charge considérable au système GSA. Pour en réduire l'impact sur les performances, vous pouvez ajuster le délai d'attente après lequel l'appui sur une touche envoie la requête à la fonctionnalité de suggestions de requêtes du système GSA. Consultez l'article du centre d'aide : Impact of query suggest on the general search performance (Incidence des suggestions de requêtes sur les performances de recherche générales).

- **Les filtres des résultats de recherche** (par exemple, les filtres de dossiers, de titres ou d'aperçus en double) peuvent nuire aux performances en cas d'utilisation de la recherche sécurisée en raison du nombre croissant de requêtes d'autorisations nécessaires pour renvoyer un ensemble de résultats. Par exemple, si un utilisateur demande 100 résultats dans une configuration à liaisons tardives, et que les 100 premiers résultats rencontrés par le système

GSA, et dont il dispose des droits d'accès, se trouvent dans le même répertoire, via le filtrage de répertoires en double, le système GSA les regroupe dans un seul résultat et doit encore trouver 99 résultats autorisés avant d'afficher les résultats de recherche.

Chapitre 4 Architecture pour les index à grande échelle

Pour les déploiements concernant un volume important de documents, le temps nécessaire à la réindexation complète du contenu peut durer plusieurs jours voire plusieurs semaines. Dans cette situation, la mise en place de stratégies d'architecture pour optimiser la méthode d'indexation des documents représente un atout majeur.

Stratégies d'exploration du système Google Search Appliance

Pour l'exploration du contenu Web avec le système GSA, plusieurs stratégies d'optimisation de la vitesse d'indexation sont possibles :

- **Vous pouvez définir les paramètres de la charge d'hôte** pour accroître le nombre de fils utilisé pour explorer les serveurs Web, ainsi que les périodes de temps de contrôle pendant lesquelles l'exploration peut être réalisée plus activement.
- Utilisez plusieurs **chemins de démarrage** si possible, car cela génère des fils d'exploration distincts. Si vous ne disposez que d'un seul chemin de démarrage pour un volume important de contenu, la vitesse d'exploration et de recherche est plus lente que dans une configuration avec plusieurs chemins de démarrage.
- Veillez à placer **le réseau du système GSA** au plus près possible du serveur Web ou de manière la plus optimale afin de limiter la latence lors de l'exploration et du téléchargement des fichiers.
- Si le débit d'exploration est d'importance majeure, il est recommandé d'utiliser **l'exploration et le traitement distribués GSA**, afin de répartir la charge d'exploration entre les différents systèmes de recherche dans le réseau d'exploration et de traitement distribués.

Stratégies de balayage géré par les connecteurs

Pour des déploiements à grande échelle impliquant l'utilisation de connecteurs, il est recommandé d'employer une stratégie de balayage géré par des connecteurs afin de tirer parti de l'indexation parallèle, et de tenir compte de l'importance et des priorités des différents contenus indexés.

Voici quelques exemples de stratégies courantes pour l'indexation de référentiels contenant plusieurs millions de documents :

- Déployez un connecteur par gestionnaire de connecteurs, et maximisez le nombre de fils utilisés à l'intérieur d'un connecteur (le cas échéant, comme avec le connecteur du système de fichiers, vous pouvez le spécifier dans les fichiers de propriétés du connecteur).
- Configurez chaque connecteur pour qu'il indexe une partie du contenu (par exemple, 3 millions de documents par connecteur). En cas de difficulté pour fractionner le contenu en catégories, vous pouvez utiliser les "Formats à suivre" dans le connecteur pour diviser le contenu en format (par exemple, pour répartir les répertoires entre les connecteurs par ordre alphabétique : [a-g], [h-m], etc.).

- Augmentez la charge d'hôte du système GSA ou déployez plusieurs connecteurs pour accroître la vitesse de balayage (à l'instar de l'approche d'optimisation de la disponibilité décrite ci-dessus).
- Si certains documents ont une priorité plus élevée et sont plus importants que le reste du contenu, spécifiez cela dans la stratégie de balayage. Vous pouvez définir une vitesse de balayage plus élevée sur ces connecteurs ou des planifications de balayage afin que le contenu prioritaire soit indexé en premier.

1