GSA | Remarques en situation

Google Search Appliance

Manuel des scénarios de déploiement

Mai 2014

Manuel des scénarios de déploiement

Ce document décrit les scénarios de déploiement d'un système GSA (Google Search Appliance).

À propos de ce document

Les recommandations et informations rassemblées dans ce document sont le résultat de notre travail sur le terrain avec de nombreux clients et partenaires dans des environnements variés. Nous tenons à les remercier chaleureusement d'avoir partagé avec nous leurs expériences et leurs observations.

Thèmes abordés	Ce guide décrit des configurations avancées du système GSA susceptibles d'être requises dans une architecture lorsque de nombreuses sources de contenu sont intégrées dans le système GSA.
Lecteurs cibles	Les administrateurs Google Search Appliance, novices ou expérimentés, et les analystes fonctionnels.
Environnement informatique	Système GSA configuré pour la recherche en mode public sur Internet, les sites Intranet et les systèmes de partage de fichiers.
Phases de déploiement	Configuration initiale du système GSA et intégration de sources de contenu supplémentaires dans ce dernier.
Autres ressources	● Le site Web Learngsa.com fournit des ressources pédagogiques sur le système GSA. ● La documentation produit de GSA fournit des informations complètes sur le système. ● Le Portail d'assistance Google for Work permet d'accéder à l'assistance Google. ● Le guide GSA en pratique permet de concevoir et de déployer une solution de recherche en entreprise basée sur le système Google Search Appliance (GSA).

Sommaire

Chapitre 1 : Recherche basique sur un site Web public

Présentation du scénario

Acme Inc. est une importante multinationale dont l'activité est la fabrication de produits électroniques grand public. Elle bénéficie d'une large présence sur le Web. Les contenus Web d'Acme Inc. incluent des informations générales sur l'entreprise, ainsi que des supports marketing spécifiques à chaque unité commerciale. Des forums d'assistance sur les produits sont également disponibles. Dans le scénario à l'étude, l'entreprise souhaite utiliser Google Search Appliance pour mettre en place un champ de recherche global permettant d'effectuer des recherches sur l'ensemble des contenus, ainsi qu'un champ de recherche spécifique pour chaque unité commerciale. Tous les contenus Web externes appartenant à Acme Inc.sont publics, sans restriction d'accès à certains utilisateurs et/ou groupes.

Exigences requises

- Indexer l'ensemble des contenus publics disponibles sur le Web
- Mettre à disposition un champ de recherche global renvoyant les résultats portant sur l'ensemble du contenu indexé, y compris celui des unités commerciales.
- Proposer des champs de recherche spécifiques renvoyant les résultats portant sur une unité commerciale donnée.
- Personnaliser le style du champ de recherche et de la page des résultats aux couleurs de l'entreprise en fonction des normes relatives à la marque établies par Acme Inc.
- Indexer le forum d'assistance d'Acme Inc. toutes les heures, son contenu évoluant rapidement.
- Gérer 20 requêtes par seconde pendant les périodes de forte affluence et maintenir un haut niveau de disponibilité en cas de problème ou de panne du système GSA.
- Ne pas mélanger les contenus écrits dans différentes langues dans les résultats de recherche.

Hypothèses

- Le contenu Web d'Acme Inc. est publié pour chaque langue sur des pages distinctes.
- L'entreprise dispose de sites Web pouvant accueillir des champs de recherche.

Points clés à prendre en considération

- Veillez à avoir suffisamment de bande passante pour gérer 20 requêtes par seconde lors des périodes de forte affluence.
- Choisissez la manière de présenter les résultats : directement à partir du système de recherche ou par l'intermédiaire de la couche de présentation d'une application Web.
- Utilisez la création de rapports et les statistiques pour évaluer la manière dont l'utilisateur interagit avec les outils de recherche.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre d'un système de recherche basique sur un site Web public couvre les domaines suivants :

- Architecture de déploiement
- Configuration de l'exploration et de l'indexation
- Configuration des frontaux
- Éléments d'administration

Architecture de déploiement

En raison des fonctionnalités de charge et de basculement, Acme Inc. utilisera un total de trois systèmes GSA dans une configuration de production. Deux de ces trois systèmes de recherche seront configurés en mode actif/actif pour une planification appropriée de la capacité. Le troisième système sera utilisé pour la sauvegarde à chaud en cas de basculement.

Acme Inc. configurera les trois GSA en miroir. L'un des systèmes sera considéré comme le système de recherche principal et c'est dans celui-là que toutes les modifications relatives à la configuration devront être effectuées. Pour que la configuration actif/actif fonctionne et assure un haut niveau de disponibilité, un répartiteur de charge sera déployé en amont des systèmes GSA. Le répartiteur de charge assurera les deux fonctions suivantes :

- Répartir activement et équitablement le trafic généré par les requêtes de recherche entre les deux systèmes GSA actif/actif.
- Effectuer un test ping sur les deux systèmes GSA actif/actif et basculer sur l'unité de sauvegarde à chaud en cas d'échec de réponse d'une des unités actives.

GSA est déployé dans une application Web existante. Nous vous recommandons donc l'approche qui consiste à traiter les requêtes et les réponses provenant du système GSA par l'intermédiaire de la couche de présentation de l'application Web. Dans ce cas, le système GSA sera utilisé en tant que service et la couche de l'application Web enverra les requêtes de recherche et analysera la réponse XML en fonction des consignes marketing et des exigences relatives à la marque pour la mise en forme de la page. La page de recherche ne sera pas directement en contact avec le système GSA. Ce dernier ne doit donc pas être en contact avec le public et doit être protégé par un pare-feu sur le réseau d'Acme Inc. en utilisant un périmètre de sécurité via les pare-feu du réseau.

Configuration de l'exploration et de l'indexation

Acme Inc. va configurer des collections pour chacune des langues des sites Web de l'entreprise. Ainsi, le paramètre `site` peut être utilisé pour distinguer les requêtes associées à une langue donnée en fonction de la page à partir de laquelle l'utilisateur a effectué la recherche. Chaque unité commerciale souhaite pouvoir effectuer une recherche uniquement sur ses propres documents. Acme Inc. va donc également configurer une collection pour chaque unité commerciale.

Acme Inc. configurera des URL de début pour les pages racines. En ce qui concerne le contenu faisant l'objet de fréquentes modifications, la fréquence du robot d'exploration permet de s'assurer que le contenu est exploré au moins une fois par jour. Si un contrôle approfondi des heures d'exploration est nécessaire, l'API Admin ou un flux Web peuvent être utilisés pour veiller à ce que des pages spécifiques reviennent dans la file d'exploration plusieurs fois par jour.

Configuration des frontaux

Chaque champ de recherche déployé sur les sites Web d'Acme Inc. sera associé à des paramètres de recherche. Ces paramètres seront transmis au système GSA en même temps que la requête afin de définir le type de résultats qui devra apparaître sur la page des résultats de recherche.

Par exemple, un champ de recherche déployé sur une page produit rédigée en anglais doit transmettre les paramètres de la collection associés à la langue anglaise, ainsi que l'unité commerciale concernée. La DTD des résultats doit être consultée pour voir dans quels éléments XML le système GSA renvoie les informations. Ces éléments doivent être analysés par le frontal et affichés en conséquence sur la page.

Éléments d'administration

Acme Inc. utilisera la fonctionnalité de rapports détaillés sur les recherches pour savoir quels sont les éléments recherchés par les utilisateurs et ceux faisant l'objet de clic sur la page des résultats de recherche. Ces rapports doivent être générés et analysés fréquemment, car ce sont de bons indicateurs du niveau de satisfaction général des utilisateurs du système de recherche.

Approche alternative

Au lieu d'utiliser une couche d'application Web en amont du système GSA, Acme Inc. peut présenter la recherche directement sur le système GSA en personnalisant la feuille de style d'un frontal. Quoiqu'il soit plus difficile de la personnaliser totalement, cette approche est plus simple à développer et facilite la mise en oeuvre des nouvelles fonctionnalités de frontaux qui viendraient à être créées sur le système GSA.

Si vous choisissez cette option, veillez à ce qu'aucun contenu sécurisé ne soit marqué comme "public" dans l'index du système GSA, car les utilisateurs auront un accès direct pour exécuter des requêtes sur le système de recherche. Vous pouvez créer un proxy inverse pour restreindre l'accès au système GSA en mettant en liste blanche des formats d'URL pouvant être utilisés.

Afin que le nombre de collections définies sur le système GSA reste raisonnable (inférieur à 200), vous pouvez utiliser un paramètre de métadonnées spécifique pour chaque unité commerciale au lieu d'utiliser des collections distinctes. Il sera indexé avec le contenu. Ce paramètre de métadonnées permettra de filtrer les requêtes associées à une unité commerciale donnée de façon que le système GSA ne récupère que les contenus lui appartenant.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et les activités d'un projet de mise en oeuvre d'une recherche basique sur un site Web public.

Tâche	Activités
Planifier l'architecture de déploiement	• Installer et câbler les systèmes de recherche • Configurer les systèmes et le mécanisme de mise en miroir • Configurer le répartiteur de charge en amont du système GSA • Configurer le périmètre de sécurité autour du système GSA
Configurer l'exploration et l'indexation	• Définir des URL de début pour explorer le contenu • Configurer des collections identifiées pour les langues et les unités commerciales • Identifier le contenu fréquemment renouvelé et s'assurer qu'il soit indexé au moins une fois par jour
Configurer les frontaux	• Activer la couche d'application Web existante pour l'ajout de champs de recherche • Analyser le code XML de réponse du système GSA et afficher les résultats conformément aux consignes de l'entreprise relatives à l'interface utilisateur

Améliorations sur le long terme

- Améliorer la recherche et les fonctionnalités en fonction des schémas de recherche des utilisateurs émanant des rapports
- Identifier les correspondances dans le contenu
- Activer des listes de synonymes plus complexes

Activer la navigation dynamique pour une navigation par attributs basées sur des métadonnées

Chapitre 2 : Recherche interne basique

Présentation du scénario

Acme Inc. bénéficie d'une large présence Web en interne, et ce dans le monde entier. Dans le scénario à l'étude, l'entreprise souhaite regrouper en un seul et même endroit la recherche portant sur l'ensemble de ses sites Web internes pour que leurs employés n'aient pas à accéder à différents sites Web pour trouver les informations qu'ils recherchent. Bien que l'ensemble des utilisateurs puissent accéder à l'Intranet d'Acme Inc., tout le monde n'a pas accès à l'ensemble des informations des différents sites de leur domaine d'entreprise. Par exemple, l'accès aux informations relatives aux ressources humaines est souhaitable, mais il est important que les informations personnelles soient sécurisées.

Exigences requises

- Indexer les contenus suivants :
    - Systèmes de fichiers d'entreprise partagés
    - Pages Web internes
    - Informations relatives aux ressources humaines
- Mettre à disposition un champ de recherche global renvoyant les résultats portant sur l'ensemble du contenu indexé, y compris celui des unités commerciales.
- Mettre à disposition des champs de recherche spécifiques renvoyant les résultats portant sur une unité commerciale donnée.
- Personnaliser le style du champ de recherche et de la page des résultats aux couleurs de l'entreprise en fonction des normes relatives à la marque établies par Acme Inc.
- Afficher le contenu sécurisé dans les résultats de recherche uniquement pour les utilisateurs autorisés à le consulter.
- Fournir une fonctionnalité de basculement en cas de panne ou de problème avec un système GSA.

Hypothèse

Un mécanisme permet d'authentifier les utilisateurs.

Points clés à prendre en considération

- Choisissez la manière de présenter les résultats : directement à partir du système de recherche ou par l'intermédiaire de la couche de présentation d'une application Web.
- Choisissez la manière de gérer la sécurité : via le système de recherche ou via l'application en amont du système GSA.
- Choisissez comment configurer le connecteur Google Search Appliance Connector pour les systèmes de fichiers afin d'indexer le contenu des fichiers partagés.
- Utilisez la création de rapports et les statistiques pour évaluer la manière dont l'utilisateur interagit avec les outils de recherche.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre de la recherche interne basique couvre les domaines suivants :

- Architecture de déploiement
- Configuration de l'exploration et de l'indexation
- Configuration de la recherche sécurisée
- Configuration des frontaux
- Éléments d'administration

Architecture de déploiement

En raison des fonctionnalités de basculement, Acme Inc. utilisera au total deux systèmes GSA dans une configuration de production. Les deux GSA seront utilisés en tant que systèmes de recherche actif/passif (un système principal et une sauvegarde à chaud pour le basculement). Acme Inc. configurera les deux GSA en miroir. Un des systèmes sera considéré comme le système de recherche principal. Toutes les modifications relatives à la configuration devront être effectuées sur ce système. Pour que la configuration actif/passif fonctionne, un répartiteur de charge sera déployé en amont des systèmes GSA. Le rôle du répartiteur de charge est d'effectuer un ping du système GSA actif et de basculer sur la sauvegarde à chaud en cas de réponse non concluante de l'unité active.

Le système GSA est déployé en interne. Il est donc recommandé d'afficher les résultats d'emblée en les personnalisant à l'aide de la feuille de style du système GSA. Dans ce cas, Acme Inc. peut modifier la feuille de style à l'aide de l'application d'assistance à la mise en page, un assistant XSLT sur le système GSA permettant d'ajouter rapidement des fonctionnalités à l'écran. Si des modifications supplémentaires sont nécessaires, la feuille de style peut être modifiée manuellement. Veuillez noter que l'ESO n'est pas compatible avec les modifications XSLT personnalisées.

Si la fidélité de la requête est requise, l'architecture nécessite un proxy inverse pour s'assurer que les paramètres des requêtes de recherche ne soient pas altérés ou ne puissent pas être soumis tels quels au système GSA. Si le contenu sécurisé est marqué dans l'index en tant que "public" et que la sécurité est appliquée par une couche applicative basée sur les métadonnées, un proxy inverse doit être utilisé en amont des systèmes GSA pour filtrer les requêtes de recherche de sorte que personne ne puisse y accéder directement pour soumettre ses propres requêtes. Ceci permet de s'assurer que l'URL n'est pas manipulée par un utilisateur pour afficher des éléments issus de métadonnées ou de collections qu'il n'est pas autorisé à voir.

Google Search Appliance Connector pour les systèmes de fichiers permet d'indexer le contenu des fichiers partagés et doit être hébergé sur un serveur externe dans un environnement de production. Le connecteur s'exécute dans un environnement JVM et est intégré à Tomcat.

Configuration de l'exploration et de l'indexation

Acme Inc. configurera des URL de début pour toutes les pages racines. Pour distinguer le contenu basé sur les départements d'Acme Inc., des collections peuvent être établies pour chaque département.

Google Search Appliance Connector pour les systèmes de fichiers doit être utilisé pour indexer les fichiers partagés. Le connecteur prend en charge les éléments suivants :

- l'autorisation par liaison précoce (LCA) ;
- la nécessité de conserver des dates de dernier accès aux fichiers et aux dossiers balayés ;
- le partage lorsque celui-ci est un partage racine au niveau du domaine DFS Windows non HTTP.

Configuration de la recherche sécurisée

Acme peut utiliser les stratégies suivantes pour sécuriser le contenu, selon que l'autorisation est requise ou non :

- Seule l'authentification est requise et non l'autorisation
- L'authentification et l'autorisation sont requises

Seule l'authentification est requise

Si l'authentification est requise, mais pas l'autorisation :

- Explorez le contenu avec un compte d'administrateur et marquez-le comme "public".
- Placez ce contenu exploré dans une collection.
- Le niveau applicatif au-dessus du système GSA gère l'authentification. Une fois qu'un utilisateur a été authentifié sur une page comportant un champ de recherche, une recherche est exécutée sur la collection dans laquelle le contenu a été placé.
- Cette stratégie mélange les contenus sécurisés et publics. Si vous souhaitez empêcher certains utilisateurs d'afficher le contenu sécurisé, vous devez utiliser un proxy inverse en amont du système GSA afin que les bonnes requêtes soient envoyées au système GSA. Le proxy inverse permet de s'assurer que le système GSA n'est pas directement exposé aux utilisateurs non authentifiés si les utilisateurs peuvent définir leurs propres paramètres de recherche dans les requêtes.

Veuillez noter qu'un proxy inverse ajoute un composant supplémentaire à l'architecture. Pour en savoir plus, consultez le Chapitre 7, Mise en oeuvre d'un proxy inverse pour le périmètre de sécurité, entre autres.

L'authentification et l'autorisation sont requises

Si à la fois l'authentification et l'autorisation sont requises :

- Explorez le contenu avec le compte d'un utilisateur qui peut y accéder et ne marquez pas le contenu comme "public".
- Pour exécuter une recherche, les utilisateurs doivent soumettre leurs identifiants et les résultats sont autorisés en fonction des points terminaux du service à l'aide de vérifications par requête HEAD.

- Déterminez comment effectuer l'intégration avec le mécanisme d'authentification disponible. Les possibilités sont notamment les suivantes :
    - Kerberos (pour en savoir plus, consultez le scénario Kerberos décrit au Chapitre 6)
    - Authentification Windows NTML intégrée avec SAML Bridge
    - Invite Basic ou LDAP de saisie des nom d'utilisateur et mot de passe par le système GSA
    - Transposition de cookie à intégrer avec une authentification par formulaire pour fournir un nom d'utilisateur validé au système GSA.

Configuration des frontaux

Chaque champ de recherche déployé sur les sites Web d'Acme Inc. est associé à un ensemble de paramètres de recherche. Ces paramètres seront transmis au système GSA en même temps que la requête afin de définir le type de résultats qui devra apparaître sur la page des résultats de recherche.

Par exemple, un champ de recherche déployé sur la page du département des ressources humaines doit transmettre les paramètres de la collection de ce département. Google recommande la personnalisation des résultats par Acme Inc. en utilisant l'Application d'assistance à la mise en page. Ainsi, certaines fonctionnalités peuvent être activées ou désactivées. L'utilisation de cet assistant XSLT a également l'avantage d'offrir une meilleure compatibilité avec les futures versions de XSLT.

Éléments d'administration

Acme Inc. utilisera la fonctionnalité de rapports détaillés sur les recherches pour savoir quels sont les éléments recherchés par les utilisateurs et ceux faisant l'objet de clic sur la page des résultats de recherche. Ces rapports doivent être générés et analysés fréquemment, car ce sont de bons indicateurs du niveau de satisfaction général des utilisateurs du système de recherche.

Approches alternatives

Pour la configuration de recherche sécurisée "Seule l'authentification est requise", effectuez l'authentification à l'aide de la fonctionnalité de périmètre de sécurité sur le système GSA au lieu d'une application en amont du système GSA. Cela garantit que le système de recherche n'affiche aucun résultat si l'utilisateur n'est pas authentifié. Lorsque le périmètre de sécurité est activé, le système de recherche doit authentifier un utilisateur avec l'un des mécanismes d'authentification configurés avant d'afficher un quelconque résultat. Si l'authentification échoue, le système GSA n'affiche aucun résultat, même si ces derniers sont publics.

Utiliser les règles LCA pour le contenu auquel seuls les utilisateurs authentifiés peuvent accéder. Avec cette approche, un groupe "tous" peut être utilisé pour gérer l'accès à ce contenu. Cette approche requiert la résolution du groupe "tous" au moment de l'authentification.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et les activités du projet pour la mise en oeuvre d'un système de recherche interne basique.

Tâche	Activités
Planifier l'architecture de déploiement	• Installer et câbler les systèmes de recherche • Configurer les systèmes de recherche et le mécanisme de mise en miroir • Configurer le répartiteur de charge en amont du système GSA. • Configurer le périmètre de sécurité autour du système GSA • Fournir le serveur au connecteur du système de fichiers hébergeur
Configurer l'exploration et l'indexation	• Définir des URL de début pour explorer le contenu • Configurer des collections identifiées pour les départements • Installer et configurer le connecteur File System • Identifier les mécanismes de sécurité et configurer l'accès du robot d'exploration
Configurer les frontaux	• Activer la couche d'application Web existante pour l'ajout de champs de recherche • Configurer les modifications XSLT sur chaque frontal à l'aide de l'assistant embarqué

Améliorations sur le long terme

- Améliorer la recherche et les fonctionnalités en fonction des schémas de recherche des utilisateurs émanant des rapports.
- Identifier les correspondances dans le contenu.
- Activer des listes de synonymes plus complexes.
- Activer la reconnaissance d'entité pour enrichir automatiquement les documents avec des métadonnées en se servant de dictionnaires au format texte, de termes ou d'expressions régulières.
- Activer la navigation dynamique pour une navigation par attributs basée sur des métadonnées.
- Activer la recherche de spécialistes pour les listes par bureau et/ou par département.
- Identifier les domaines pour lesquels les modules OneBox peuvent être utiles.

Chapitre 3 : Recherche interne sur un Intranet, dans un système de fichiers et dans SharePoint

Présentation du scénario

Acme Inc. dispose de différents corpus traités sur différents serveurs du réseau d'entreprise. L'accès à ces recueils de données s'effectue par le biais de différentes applications de gestion des données (SharePoint, par exemple), ainsi que des systèmes de partages de fichiers sécurisés.

Le fait de devoir passer par de multiples applications pour trouver les informations s'avère fastidieux et très chronophage pour les employés. En outre, le fait d'avoir à localiser l'information recherchée affecte la productivité et commence à avoir des répercussions sur le chiffre d'affaires, du fait des recherches répétées dans des systèmes distincts et de l'inefficacité des outils de recherche existants.

Exigences requises

- Indexer le contenu suivant, tout en veillant à ce qu'il reste sécurisé :
    - Systèmes de partage de fichiers sécurisés
    - Données du portail SharePoint hébergeant des sites internes
- Afficher le contenu sécurisé dans les résultats de recherche uniquement pour les utilisateurs autorisés à le consulter.
- Créer une interface utilisateur standard pour l'accès aux données.
- Créer des interfaces personnalisées pour les utilisateurs internes et externes.
- Les bénéfices du déploiement pour l'entreprise doivent être mesurables.

Hypothèses

- Plus de 500 000 documents sont stockés dans SharePoint.
- Une solution de statistiques automatisée est souhaitable.

Points clés à prendre en considération

- Choisir d'utiliser ou non Google Search Appliance Connector pour SharePoint et décider d'embarquer ou non le connecteur.

- Opter pour l'exploration de systèmes Web de partage de fichiers SMB ou pour le connecteur de système de fichiers.

- Choisissez la manière de présenter les résultats : directement à partir du système GSA ou par l'intermédiaire de la couche de présentation d'une application Web.

- Choisir la manière de gérer la sécurité : par le système de recherche ou par l'application en amont du système GSA.

- Déterminer si l'authentification silencieuse est nécessaire lorsque le système GSA ne redemande pas aux utilisateurs leurs identifiants.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre de la recherche sur l'Intranet, le système de fichiers et SharePoint couvre les domaines suivants :

- Analyse des avantages
- Architecture de déploiement
- Configuration de l'exploration et de l'indexation
- Configuration de l'autorisation et de l'authentification lors de la présentation des résultats

Analyse des avantages

Pour évaluer les avantages pour l'entreprise de la solution de recherche, Acme Inc. conduit une étude rapide afin de déterminer le temps passé sur les plates-formes existantes. Des outils automatisés doivent permettre de recueillir les informations lorsque cela est possible. Si des outils d'analyse existent déjà, ils doivent être utilisés pour recueillir les informations sur l'utilisation du système de recherche ou le temps requis pour trouver les informations sur les systèmes actuels. Si de tels outils n'existent pas, Acme Inc. doit envisager de mettre en oeuvre une solution d'analyse pour l'évaluation automatique de l'efficacité.

Une fois le déploiement effectué, Acme Inc. mène une évaluation de la nouvelle solution et estime son efficacité. Pour déterminer quels sont les bons critères d'évaluation, des cas d'utilisation similaires avant le déploiement doivent être comparés.

Architecture de déploiement

Acme Inc. déploie le connecteur SharePoint non embarqué, car le nombre total de documents SharePoint dépasse les 500 000. Si les fichiers sont partagés sur le Web, ils peuvent être directement explorés par le système GSA.

Les résultats peuvent être affichés directement à partir du système GSA en utilisant des frontaux personnalisés pour les différents répertoires de données. Dans le cas de la recherche avec SharePoint, le champ de recherche pour SharePoint sera déployé et utilisé.

Envisagez d'utiliser Google Search Appliance Connector pour les systèmes de fichiers afin d'indexer les fichiers partagés. Voici des scénarios qui exigent l'utilisation du connecteur :

- Autorisation par liaison précoce (LCA).
- Nécessité de stocker les dates de dernier accès sur les fichiers et les dossiers qui sont balayés.
- Le partage est un partage racine de domaine DFS Windows exposé non HTTP.

Configuration de l'exploration et de l'indexation

Acme Inc. configurera l'indexation et l'exploration pour les types de sources de contenu suivants :

- SharePoint : pour indexer le contenu sur SharePoint, Acme installera le connecteur SharePoint et le configurera sur un serveur distinct. Les services Web Google pour SharePoint seront également installés sur chaque frontal Web SharePoint existant dans la ferme. Les LCA seront introduites dans le système GSA en tant qu'option de configuration de connecteur. Le connecteur SharePoint est utilisé pour indexer le contenu SharePoint. Active Directory Groups Connector est donc requis pour résoudre les adhésions au groupe Active Directory au moment de la présentation des résultats. Ces adhésions sont nécessaires pour l'autorisation du contenu et pour le mappage des utilisateurs/groupes Active Directory sur les groupes locaux SharePoint.
- Systèmes de partages de fichiers : pour indexer des fichiers partagés, Acme configurera le partage de fichiers Web sur la page **Sources de contenu > Exploration Web > URL de début et URL bloquées** (avant la version 7.2 : **Explorer et indexer > URL d'exploration)** dans la console d'administration du système GSA.

Configuration de l'autorisation et de l'authentification lors de la présentation des résultats

Acme Inc. utilisera Kerberos comme mécanisme d'authentification privilégié entre le système GSA et le serveur de contenu. Pour ce faire, les tâches suivantes seront réalisées :

- Création d'un compte de service Active Directory pour le système GSA.
- Configuration de Kerberos sur le système GSA.
- Configuration du connecteur SharePoint pour soumettre les flux de contenu au système GSA avec l'actualisation des LCA activée.
- Configuration du connecteur ADGroups pour mettre en cache les objets Active Directory dans une base de données pour une résolution rapide au moment de la présentation des résultats.
- Configuration de l'instance de connecteur SharePoint en tant que mécanisme d'authentification afin de résoudre des groupes pour une session d'utilisateur. Ces groupes permettent de supprimer les LCA de l'index.

Les flux de contenu comportant des LCA doivent être envoyés par SharePoint au système GSA. Le contenu doit donc être autorisé dans l'index du système GSA à l'aide des LCA qui ont été intégrées lors de l'exploration.

Approches alternatives

- Indexer le contenu des systèmes de partage de fichiers avec Google Search Appliance Connector pour les systèmes de fichiers.
    - Avantage de cette approche : les LCA sont intégrées avec le contenu, ce qui permet d'effectuer une autorisation par liaison précoce, plus performante.
    - Pour que cette approche fonctionne, les groupes appropriés pour l'utilisateur doivent être résolus au moment de l'authentification, car ils sont nécessaires pour l'autorisation LCA par liaison précoce. Le connecteur ADGroups, qui est également requis pour SharePoint, peut être utilisé dans ce but.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et les activités du projet pour la mise en oeuvre d'un système de recherche interne sur Intranet, dans un système de fichier et dans SharePoint.

Tâche	Activités
Planifier l'architecture de déploiement	• Configurer le serveur du connecteur SharePoint/ADGroups
Configurer l'exploration et l'indexation	• Configurer le connecteur SharePoint • Configurer le connecteur ADGroups • Configurer les emplacements des systèmes de partage de fichiers sous "Explorer et indexer" dans la Console d'administration du système GSA
Configurer les frontaux	• Personnaliser les frontaux pour les répertoires de données
Configurer l'autorisation/authentification lors de la présentation des résultats	• Activer le système GSA pour Kerberos • Configurer l'autorisation basée sur Connector • Configurer le mécanisme de résolution de groupes ○ Si le connecteur SharePoint est utilisé, l'authentification se fera certainement par un connecteur SharePoint et sera configurée uniquement pour la résolution de groupes.

Amélioration sur le long terme

Déployer le champ de recherche Google Search Box pour SharePoint afin de présenter les résultats provenant de SharePoint.

Chapitre 4 : Indexation par l'intermédiaire des flux

Présentation du scénario

Acme Inc. dispose de ses propres boutiques gérées sous deux noms de marque différents. Dans le présent scénario, l'entreprise souhaite que les employés des boutiques puissent effectuer des recherches dans la base de données des produits. Les informations sur les produits sont actuellement stockées dans une base de données et certaines données sont contenues dans des applications de l'entreprise. Le robot d'exploration ne peut pas indexer les pages de produits directement. Un frontal Web affiche les informations sur les produits lorsque ces dernières sont accompagnées du numéro du produit dans l'URL.

Exigences requises

- Indexer le contenu relatif aux produits.
- Fournir un système de recherche au sein de marques de distribution spécifiques.
- Activer la navigation par paramètre dans un volet gauche par :
    - Prix
    - Catégorie
    - Heure d'affichage

Hypothèses

- Un frontal Web existe et permet d'afficher les pages des produits. Ce frontal Web ne contient pas toutes les métadonnées requises pour l'indexation des produits.
- Les informations sur les produits ne sont pas sécurisées et tout le monde peut y avoir accès.

Points clés à prendre en considération

- Choisir d'utiliser ou non Google Search Appliance Connector pour bases de données afin d'intégrer au système GSA les enregistrements relatifs aux produits.
- Choisir d'utiliser un flux de contenus ou un flux Web pour intégrer au système GSA les enregistrements relatifs aux produits.
- Définir des métadonnées à indexer avec le contenu pour gérer la navigation dynamique et les fonctionnalités de recherche avancée.

Approche recommandée

L'approche recommandée par Google pour l'indexation via les flux couvre les domaines suivants :

- Architecture de déploiement
- Configuration de l'exploration et de l'indexation
- Importance des métadonnées
- Configuration des frontaux

Architecture de déploiement

Les enregistrements associés aux métadonnées requises ne peuvent pas être construits exclusivement à l'aide des requêtes sur la base de données. C'est la raison pour laquelle le connecteur de base de données n'est pas utilisé pour indexer le contenu. Acme Inc. utilisera à la place un flux personnalisé. Un flux personnalisé est une application construisant un code XML contenant les enregistrements à indexer sur le système GSA. L'étape principale dans l'intégration de code XML avec des enregistrements dans le système GSA est une action POST sur l'interface de protocole de flux sur le système GSA.

Outre le système GSA, un autre serveur (Windows ou Linux) est requis pour héberger l'application de flux. L'application exécutera des logiques pour construire un enregistrement et envoyer ce dernier GSA.

Configuration de l'exploration et de l'indexation

L'approche recommandée consiste à utiliser les flux de contenu pour intégrer au système GSA les enregistrements relatifs aux produits. Le contenu peut ainsi être personnalisé pour l'indexation. Cette méthode s'appuie également sur la fonctionnalité de mise en cache du système GSA permettant l'enregistrement d'une page de produits personnalisée dans son index. Ainsi, les gérants des boutiques peuvent décider d'afficher la version en cache qui peut s'avérer plus facile à afficher que le frontal Web existant.

L'application de flux doit être conçue de façon que, pour chaque enregistrement de produit dans la base de données, le code HTML requis et les métadonnées associées puissent être construits et envoyés au système GSA. Un mécanisme est nécessaire pour suivre tous les enregistrements supprimés, modifiés et ajoutés.

Importance des métadonnées

Il est extrêmement important de se concentrer sur les métadonnées en ce qui concerne le contenu des produits. Elles permettent aux utilisateurs finaux d'effectuer des recherches avancées plus puissantes et d'affiner ces dernières à l'aide de différentes catégories définies. Acme Inc. peut identifier certaines métadonnées pour créer des en-têtes de navigation dynamique afin que les utilisateurs puissent, d'un simple clic, examiner en détail différentes catégories de valeurs de métadonnées. Les autres valeurs de métadonnées peuvent servir pour limiter les requêtes à un ensemble de contenus spécifique.

Configuration des frontaux

L'avantage d'utiliser les flux de contenu pour envoyer le contenu dans l'index du système GSA est que la version mise en cache du contenu personnalisé créé pour le flux sera enregistrée sur le système GSA et pourra être sélectionnée pour être affichée sur le frontal.

L'indexation de pages destinées à l'impression et affichées sous forme de feuilles de données est un exemple d'utilisation possible. Lorsqu'un utilisateur souhaite afficher une page de données, il peut référencer le contenu mis en cache sur le système GSA. S'il souhaite une vision plus approfondie, il peut cliquer sur le lien et est redirigé vers le frontal Web qui affiche la page d'informations détaillées sur les produits pour l'élément donné. Acme Inc. doit modifier le XSLT du système GSA pour afficher les produits et les métadonnées associées.

Acme Inc. modifiera également le frontal afin d'activer les fonctionnalités de recherche avancées pour les utilisateurs en fonction de collections et de métadonnées définies. Le fait de sélectionner un élément dans une liste déroulante ou de cliquer sur un bouton radio peut permettre d'associer les métadonnées à la requête en tant que termes de recherche et de limiter cette dernière au corpus de produits souhaité.

Approches alternatives

- Si l'ensemble du contenu et des métadonnées peut être dérivé de requêtes sur la base de données, utilisez le connecteur de base de données pour intégrer tous les produits.
- Si l'application du frontal dédiée à l'affichage des produits peut afficher toutes les informations nécessaires à l'indexation, utilisez un flux Web pour indexer tous les produits.
- Au lieu de définir le processus d'alimentation en métadonnées au moment de l'indexation, vous pouvez configurer des règles de reconnaissance d'entités par l'intermédiaire de dictionnaires ou en définissant des expressions régulières XML pour associer automatiquement des entités aux documents au moment de l'indexation.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et les activités du projet relatif à la mise en oeuvre d'un système d'indexation par l'intermédiaire de flux.

Tâche	Activités
Planifier l'architecture de déploiement.	• Installer et câbler les systèmes de recherche. • Mettre un serveur à disposition pour héberger l'application de flux.
Configurer l'exploration et l'indexation.	• Configurer des URL à suivre pour l'alimentation en contenu. • Configurer des collections identifiées pour les marques. • Concevoir la logique de construction du contenu des flux. • Concevoir l'application de flux pour l'écriture des enregistrements XML et l'envoi de ces derniers au système GSA.
Configurer les frontaux.	• Activer la navigation dynamique. • Modifier le code XSLT pour afficher les enregistrements avec les métadonnées souhaitées. • Créer une page ou une partie de page de recherche avancée limitée aux requêtes contenant les métadonnées souhaitées.

Chapitre 5 : Transposition de cookie avec authentification silencieuse

Présentation du scénario

Dans le présent scénario à l'étude, Acme Inc. souhaite intégrer au système GSA son SSO SiteMinder et trois sources de contenu différentes, à l'aide des mécanismes suivants :

- LCA par URL
- Autorisation par connecteur
- Contenu public

L'entreprise préfère que ses utilisateurs bénéficient d'un système de recherche d'authentification silencieuse qui soit transparent, une fois connectés au portail principal de l'entreprise.

Exigences requises

- Indexer les contenus suivants :
    - Livelink
    - Application d'annuaire sur le Web
    - Lotus Connections
- Fournir un champ de recherche général. La page de résultats renvoyée doit comporter les liens les plus pertinents parmi l'ensemble du contenu indexé.
- Afficher le contenu sécurisé dans les résultats de recherche uniquement pour les utilisateurs autorisés à le consulter.
- Fournir une authentification silencieuse transparente pour la recherche une fois que l'utilisateur s'est connecté au portail principal.
- Le contenu de Lotus Connections est sécurisé en fonction des groupes natifs Lotus Connections, ainsi que des groupes LDAP.

Hypothèses

- SSO SiteMinder permet d'authentifier les utilisateurs auprès des sources de contenu sécurisées : Livelink et Connections.
- Toutes les sources de contenu utilisent la même identité.
- Le connecteur existant Google Search Appliance Connector pour Livelink sera utilisé pour intégrer le système GSA dans Livelink.
- Le contenu Connections sera intégré au système GSA.
- Les LCA peuvent être intégrées dans le contenu Connections pour tirer parti de la fonctionnalité LCA par URL du système GSA.

Points clés à prendre en considération

- Confirmer l'hypothèse que toutes les sources de contenu utilisent la même identité commune.
- Déterminer si les sources de contenu utilisent des groupes natifs de source de contenu ou des groupes synchronisés avec Active Directory/LDAP.
- Confirmer l'hypothèse selon laquelle Lotus Connections peut se servir de la fonctionnalité LCA par URL du système GSA pour intégrer les informations des LCA avec le flux de contenu.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre de la transposition de cookie fournissant l'authentification silencieuse couvre les domaines suivants :

- Présentation de l'architecture
- Authentification
- Autorisation

Présentation de l'architecture

Acme Inc. intégrera dans le système GSA les sources de contenu répertoriées dans le tableau suivant.

Source de contenu	Méthode d'intégration
Lotus Connections	• Flux contenant le contenu Connections et les LCA de chaque document. • Les groupes natifs Connections et les groupes LDAP sont intégrés avec le contenu. • Le contenu est synchronisé à l'aide de la fonctionnalité SeedList de Lotus Connections.
Open Text Livelink	• Le contenu est balayé et intégré dans le système GSA avec le connecteur Livelink. • Le connecteur synchronise le contenu. • Les LCA ne sont pas intégrées avec le contenu. • L'autorisation par connecteur est utilisée pour autoriser plusieurs documents en même temps.
Contenu de l'annuaire Web	• Disponible sur le Web et directement exploré par le système GSA.

Authentification

En fonction des mécanismes d'autorisation d'Acme Inc., les sources de contenu sécurisées requièrent une identité validée par un nom d'utilisateur afin d'autoriser le contenu pour les utilisateurs lors de la présentation des résultats. Le connecteur Livelink requiert un nom d'utilisateur validé pour effectuer l'autorisation par connecteur. Un nom d'utilisateur valide, ainsi que les groupes associés doivent être fournis au système GSA afin d'autoriser le contenu avec des LCA dans l'index.

Le mécanisme SSO SiteMinder d'Acme Inc. est appliqué aux deux sources de contenu. Les utilisateurs disposeront donc d'un cookie, dans leur session, leur permettant d'accéder aux sources après s'être connectés au portail de recherche principal pour effectuer une recherche. Une règle d'authentification par formulaire de connexion universelle sera configurée pour récupérer un modèle d'URL, protégé par SiteMinder, dans le cadre du processus d'authentification. Si un utilisateur est déjà connecté au portail et dispose d'un cookie SiteMinder, il dispose de l'autorisation nécessaire et ne doit pas fournir d'identifiants.

Autorisation

Les mécanismes d'autorisation des deux sources de contenu requièrent des identifiants :

- Livelink requiert une identité validée par un nom d'utilisateur.
- Connections requiert un nom d'utilisateur et des groupes.

Par conséquent, la page du modèle d'URL protégée par l'authentification unique SiteMinder doit renvoyer les informations suivantes sur l'utilisateur fournissant le cookie SSO pour l'authentification :

- Nom d'utilisateur
- Liste des groupes associés au compte utilisateur

Cette procédure est appelée "déchiffrement de cookie".

Il s'agit de créer un code JSP ou ASP.NET qui, après validation de l'authenticité du cookie SSO, renvoie une réponse HTTP au système GSA. La réponse comprend un code d'état 200 OK et un nom d'utilisateur validé, ainsi qu'une liste de groupes associés dans l'en-tête, notamment "X-Username" et "X-Groupes". Veuillez noter que les groupes natifs Connections et les groupes LDAP de l'utilisateur sont renvoyés pour assurer la compatibilité avec les LCA Connections dans l'index.

Les étapes suivantes décrivent un flux d'autorisation et d'authentification complet et abouti.

1. L'utilisateur se connecte sur le portail Intranet de l'entreprise via une page de connexion SiteMinder. Un cookie SiteMinder est créé pour cette session.
2. L'utilisateur effectue une recherche sur le système GSA en transmettant les cookies applicables.
3. GSA récupère l'URL d'exemple en transmettant les cookies dédiés à la récupération. Le cookie sert à valider l'authentification auprès de la page protégée de SiteMinder.
4. La page renvoie la valeur "200", ainsi que le nom d'utilisateur validé et les groupes associés au compte utilisateur dont le cookie a été transmis à la page.
5. La récupération est considérée comme aboutie par le système GSA et ce dernier associe le nom d'utilisateur et le groupe au groupe d'identification de l'identité validée.
6. Le nom d'utilisateur et les groupes sont utilisés pour autoriser le contenu dans le système GSA. Le nom d'utilisateur est transmis afin d'effectuer une authentification par connecteur pour Livelink. Le nom d'utilisateur et les groupes sont transmis afin d'effectuer les vérifications d'autorisation pour les LCA associées au contenu dans l'index.

Approche alternative

Pour l'ensemble du contenu, effectuez des vérifications d'autorisation par requête HEAD. Ces dernières ne requièrent pas de groupes ni de nom d'utilisateur avec l'identité validée.

Présentation des tâches du projet

Le tableau suivant répertorie les activités et tâches du projet pour la mise en oeuvre du déchiffrement de cookie qui permet l'authentification silencieuse.

Tâche	Activités
Planifier l'architecture de déploiement	• Concevoir une application qui, en fonction d'un cookie, renvoie le nom d'utilisateur et les groupes associés au compte utilisateur. • Déployer l'application sur un serveur d'application ou un serveur Web protégé par SiteMinder. Pour Apache, un plug-in SiteMinder peut être utilisé pour effectuer l'intégration à SSO.
Configurer l'exploration et l'indexation.	• Configurer Connectors pour l'indexation du contenu. • Configurer le robot d'indexation pour l'indexation du contenu public.
Configurer l'authentification par cookie à l'aide d'une URL d'exemple sous "Mécanismes d'authentification de connexion universelle".	• Une fois la configuration effectuée, GSA récupère une URL d'exemple permettant d'accéder à la page. La page de déchiffrement de cookie renvoie alors le nom d'utilisateur, ainsi que les groupes associés à l'identité validée effectuant la recherche.

Améliorations sur le long terme

- Vérifier l'architecture existante lorsque de nouvelles sources de contenu sont créées, afin de voir si elle doit être modifiée pour intégrer le nouveau contenu.
- Si les LCA pour Livelink viennent à être disponibles, ajustez le déchiffrement de cookie afin de renvoyer également les groupes LiveLink. Envisagez d'utiliser un outil de déchiffrement de cookie dans un groupe d'identification (espace de noms) distinct en cas de conflit entre les noms des groupes.

Chapitre 6 : Authentification silencieuse - Intégration avec NTLM et SAML Bridge

Présentation du scénario

Acme Inc. utilise NTLM et Integrated Windows Authentication (IWA) avec un serveur Active Directory. Dans le présent scénario à l'étude, il est préférable que les utilisateurs disposent d'un système transparent d'authentification silencieuse pour effectuer la recherche après s'être connectés au domaine Windows et avoir utilisé le navigateur Internet Explorer.

Exigences requises

- Indexer et afficher le contenu protégé NTLM de manière sécurisée.
- Fournir un champ de recherche général. La page de résultats renvoyée doit comporter les liens les plus pertinents de l'ensemble du contenu indexé.
- Afficher le contenu sécurisé dans les résultats de recherche uniquement pour les utilisateurs autorisés à le consulter.
- Fournir une authentification silencieuse transparente pour la recherche une fois que l'utilisateur s'est connecté au domaine Windows.

Hypothèses

- Le serveur IIS hébergeant le contenu accepte les requêtes HEAD.
- L'ensemble du contenu exploré est associé au même domaine Windows que celui auquel les utilisateurs se sont connectés.

Points clés à prendre en considération

- Vérifiez l'hypothèse selon laquelle toutes les sources de contenu utilisent le domaine Windows auquel se connectent les utilisateurs.
- Assurez-vous que tous les serveurs figurant dans l'architecture de déploiement sont synchronisés avec le même serveur de temps.
- Veillez à ce qu'un certificat soit disponible dans IIS et puisse être utilisé pour envoyer des requêtes de liaison Post SAML.
- Pour que la communication avec SAML Bridge puisse se faire avec HTTPS, configurez le système GSA avec le certificat racine de SAML Bridge.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre de l'authentification silencieuse par l'intermédiaire de l'intégration NTLM et SAML Bridge couvre les domaines suivants :

- Authentication
- Autorisation
- Flux de processus pour l'authentification et l'autorisation avec SAML Bridge

Authentication

Acme Inc. utilisera SAML Bridge pour authentifier les utilisateurs avec Active Directory, en s'appuyant sur l'authentification silencieuse fournie par IWA. Pour que cela fonctionne, le contrôleur de domaine exécutant Active Directory doit répondre aux exigences suivantes :

- L'extension Kerberos Windows 2003 doit être disponible, car Kerberos est utilisé pour l'authentification entre SAML Bridge et le serveur de contenu.
- Le niveau de fonctionnement du domaine doit être défini sur Windows Server 2003.
- Active Directory doit être configuré afin d'autoriser SAML Bridge à utiliser les identifiants de délégation de l'utilisateur pour accéder au contenu du serveur de contenu.

Pour que SAML Bridge puisse être utilisé par le système GSA afin d'effectuer l'authentification, configurez-le sur la page **Rechercher > Recherche sécurisée > Mécanismes d'authentification de connexion universelle** (avant la version 7.2 : **Traitement > Mécanismes d'autorisation de connexion universelle > SAML**), dans la Console d'administration.

L'option de liaison Post SAML fait son apparition dans la version 2.8 de SAML Bridge. Il est donc recommandé de l'utiliser pour l'authentification avec SAML Bridge. Pour connaître la procédure de configuration de SAML Bridge pour la liaison Post, accédez au wiki suivant : http://code.google.com/p/google-saml-bridge-for-windows/wiki/SAMLBridge28features

Autorisation

L'autorisation avec NTLM est requise. SAML Bridge est donc aussi utilisé pour l'autorisation. Dans ce cas, SAML Bridge doit être configuré en tant que fournisseur d'autorisation sur la page **Rechercher > Recherche sécurisée > Contrôle d'accès** (avant la version 7.2 : **Traitement > Contrôle d'accès**) de la Console d'administration. GSA délègue alors les vérifications d'autorisation pour les documents individuels à SAML Bridge. SAML Bridge répond avec la valeur "PERMIT" ou "DENY", selon le cas.

Flux de processus pour l'authentification et l'autorisation avec SAML Bridge

1. Un utilisateur crée une requête de recherche sur du contenu sécurisé.
2. La SPI d'authentification du système GSA est utilisée pour déléguer l'authentification à SAML Bridge. NTLM, configuré sur le navigateur de l'utilisateur, permet d'authentifier l'utilisateur.
3. Après avoir authentifié l'utilisateur, le système GSA détermine les résultats les plus pertinents pour lui. Si ces résultats incluent des documents sécurisés, le système GSA utilise la SPI d'autorisation pour déléguer les vérifications d'autorisation de ces documents à SAML Bridge.
4. SAML Bridge obtient un ticket Kerberos pour le compte de l'utilisateur et agit en son nom auprès du serveur de contenu.
5. SAML Bridge renvoie un message PERMIT ou DENY au système GSA et ce dernier affiche les résultats que l'utilisateur est autorisé à voir sur une page de résultats de recherche.

Approche alternative

Choisir Kerberos en tant que mécanisme d'authentification pour les serveurs de contenu. En choisissant Kerberos, il est possible de contourner le déploiement de SAML Bridge en configurant l'authentification silencieuse.

Présentation des tâches du projet

Le tableau suivant répertorie les activités et tâches du projet pour la mise en oeuvre de l'authentification silencieuse en intégrant NTLM et SAML Bridge.

Tâche	Activités
Planifier l'architecture de déploiement.	• Préparer la configuration d'Active Directory pour l'installation de SAML Bridge. • Déployer SAML Bridge sur le contrôleur de domaine et apporter les modifications nécessaires à la configuration. • Configurer les certificats pour la liaison POST et la communication par HTTPS.
Configurer SAML Bridge sur le système GSA.	• Configurer la SPI d'authentification de sorte qu'elle utilise SAML Bridge. • Configurer la SPI d'autorisation de sorte qu'elle utilise SAML Bridge.

Amélioration sur le long terme

Vérifier l'architecture existante à chaque fois que de nouvelles sources de contenu sont créées, afin de voir si elle doit être modifiée pour intégrer le nouveau contenu.

Chapitre 7 : Mise en oeuvre d'un proxy inverse pour le périmètre de sécurité, entre autres

Présentation du scénario

Acme Inc. dispose de documents de recherche et développement très sensibles. Dans ce scénario, l'entreprise souhaite restreindre l'accès à ces documents en imposant le passage de toutes les recherches par un proxy. Le proxy met en place l'authentification via le système d'authentification unique (SSO) de l'entreprise avant d'autoriser l'accès au système GSA. Il limite également les requêtes pouvant être soumises au système GSA.

Exigences requises

- Mettre en place la connexion SSO avant d'accéder au système GSA.
- Limiter les recherches exécutées sur le système GSA à une collection spécifique en restreignant les paramètres de requête par URL.

Hypothèses

- Pour cet exemple, nous partons du principe qu'un serveur Web Apache est utilisé. Veuillez noter que d'autres serveurs Web peuvent être utilisés pour des proxys inverses sur le système GSA.
- Un serveur Apache est disponible.
- Un plug-in Apache est disponible pour le système SSO d'Acme.

Points clés à prendre en considération

- Si GSA est utilisé pour des recherches sécurisées :
  - la transmission du trafic par un proxy HTTPS est requise ;
  - les appels au gestionnaire de sécurité du système GSA doivent également passer par le proxy.
- En cas d'accès au système GSA par HTTPS, le trafic SSL doit également passer par le proxy.
- Le système GSA est protégé par un pare-feu et l'accès est limité au serveur de proxy.

Approche recommandée

L'approche recommandée par Google pour la mise en oeuvre d'un périmètre de sécurité par l'intermédiaire d'un proxy inverse couvre les domaines suivants :

- Intégration d'Apache dans un système SSO
- Transmission des requêtes au système GSA via un proxy
- Limitation de l'ensemble du trafic par le proxy inverse

Intégration d'Apache dans un système SSO

Pour protéger l'instance d'Apache avec le protocole SSO, Acme Inc. installera le plug-in SSO Apache correspondant au système d'authentification unique utilisé. Selon que le plug-in contient ou non une interface de configuration, des options de protection d'application peuvent être affichées sous forme d'un ensemble d'assistants. Il se peut également que la configuration exige de définir des filtres de ressources appropriés pour le trafic d'Apache.

Lorsque le plug-in SSO est configuré, à chaque fois qu'un accès à l'hôte Apache est détecté avec le champ d'application approprié du domaine du cookie, un utilisateur est authentifié avec le système SSO. Si l'utilisateur n'a pas de cookie pour sa session, il doit être redirigé vers la page de connexion SSO pour en obtenir un. Il sera alors autorisé à accéder au système GSA.

Transmission des requêtes envoyées au système GSA via un proxy

Cette opération est généralement réalisée grâce à un bloc d'hôte virtuel, mais vous pouvez également utiliser la configuration du serveur principal. Pour configurer un hôte virtuel de sorte qu'il oriente le trafic vers le proxy :

```
<VirtualHost *:80>
  ProxyRequests Off
    <Proxy *>
      Order Deny,Allow
      Deny from all
      Allow from [gsa_ip]
  </Proxy>

  ProxyPass / http://gsa32.example.com
  ProxyPassReverse / http://gsa32.example.com
</VirtualHost>
```

Si la configuration est telle que la recherche sécurisée est activée, le plug-in Apache mod_ssl est nécessaire pour faire passer le trafic par le proxy. L'émission d'un certificat pour le serveur Apache est également nécessaire. Ce certificat devra être installé sur le système GSA, afin que les requêtes orientées vers le proxy puissent être reconnues comme signées.

Limitation de l'ensemble du trafic par le proxy inverse

Une fois le proxy inverse mis en oeuvre, Acme Inc. configurera une règle de pare-feu pour autoriser l'accès au système GSA uniquement pour le trafic provenant de l'hôte Apache. Cette procédure oblige le passage de toutes les requêtes souhaitant accéder au système GSA par le proxy inverse Apache.

Approche alternative

Utiliser un autre serveur Web pour mettre en oeuvre le proxy inverse. Vous pouvez, par exemple, utiliser IIS pour filtrer le trafic.

La fonctionnalité de périmètre de sécurité de la version 6.14 de GSA permet de mettre en oeuvre un tel mécanisme. Vous devez configurer un mécanisme de sécurité sur le système GSA, pour l'authentification uniquement. Quand c'est le cas, les résultats publics ne sont pas affichés aux utilisateurs qui ne sont pas authentifiés auprès du système GSA.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et activités du projet pour la mise en oeuvre d'un périmètre de sécurité avec un proxy inverse.

Tâche	Activités
Planifier l'intégration d'Apache avec SSO.	• Protéger l'URL Apache. • Configurer Apache de sorte que le serveur utilise un plug-in SSO et définisse des filtres de ressources appropriés pour filtrer le trafic vers les ressources protégées par SSO.
Configurer les requêtes envoyées au système GSA par Apache et transitant par le proxy.	• Configurer l'hôte virtuel pour faire passer le trafic allant au système GSA par le proxy et inversement. • Si la recherche sécurisée ou l'accès au système GSA par HTTPS est requis, mod_ssl est nécessaire pour faire transiter le trafic HTTPS par le proxy.
Configurer le pare-feu de sorte qu'il bloque l'accès au système GSA pour tous à l'exception de l'hôte Apache.	• Configurer une règle de pare-feu pour établir un périmètre de sécurité autour du système GSA afin que le seul moyen d'y accéder soit via le proxy Apache.

Améliorations sur le long terme

- Envisager d'autres utilisations du proxy inverse : URL de nettoyage, configuration d'un tunnel pour traverser le pare-feu ou mise en cache pour améliorer les performances.

- La mise en cache avec Apache peut grandement améliorer les temps de réponse et les capacités de traitement du système GSA. Par exemple, une configuration memcache peut être ajoutée à la section d'hôte virtuel :

  ```
  CacheEnable mem /
  MCacheSize 4096
  MCacheMaxObjectCount 1000
  MCacheMinObjectSize 1
  MCacheMaxObjectSize 4096
  ```

  Ce code permet de mettre en cache les 1000 réponses les plus récentes renvoyées par le système GSA, dans la limite de 4 Ko de mémoire.

Chapitre 8 : Test de pertinence

Présentation du scénario

Acme Inc. a déployé son système GSA et y a intégré les sources de contenu suivantes :

- Contenu Livelink
- Site Intranet exploré
- Application d'annuaire

Dans le présent scénario à l'étude, l'entreprise souhaite conduire un test de pertinence pour s'assurer que les utilisateurs sont satisfaits des résultats de recherche renvoyés par le système GSA avant de mettre
en production de la solution de recherche.

Exigence requise

Veillez à ce que les résultats de recherche renvoyés aux utilisateurs soient pertinents par rapport aux termes de recherche employés.

Hypothèses

- Le contenu a déjà été intégré au système GSA. Il est disponible dans l'index.
- Les responsables du test connaissent le contexte du contenu d'entreprise figurant dans l'index du système GSA.

Points clés à prendre en considération

- Le niveau de pertinence n'est pas une mesure scientifique absolue. C'est un concept difficile à appréhender, différent pour chaque personne.

- Les résultats renvoyés par les algorithmes de pertinence prêts à l'emploi du système GSA s'avèrent très pertinents.

Approche recommandée

L'approche recommandée par Google pour le test de pertinence couvre les domaines suivants :

- Préparation de l'analyse et du test
- Exécution d'un exemple de test
- Fonctionnalités à prendre en compte pour le réglage de la pertinence

Préparation du test

- Identifier différents groupes d'utilisateurs appartenant à l'organisation et souhaitant utiliser la fonction de recherche.
- En fonction du contenu indexé dans le système GSA, déterminer des contextes d'entreprise sur le type de recherches pouvant être effectuées par différents utilisateurs et les documents attendus dans les résultats de recherche.
- Développer une liste prédéterminée de requêtes que vos utilisateurs devront exécuter afin de juger de la pertinence des résultats. Outre le jeu de requêtes imposé, demander aux utilisateurs de créer environ trois requêtes personnelles au cours du processus de test afin de pallier les éventuels oublis.
- Identifier un jeu de documents que vous jugez les plus pertinents pour une requête et un utilisateur donnés. Ce jeu sera utilisé comme référence pour la notation.
- Développer une échelle de notation à utiliser pour estimer le niveau de pertinence des résultats renvoyés et la transmettre aux utilisateurs effectuant les tests. Par exemple, établissez une échelle allant de 1 à 5 où :
    - 1 : excellent niveau de pertinence. Tous les résultats renvoyés apparaissant sur la première page sont tous extrêmement pertinents. Le document identifié pour cette requête spécifique est renvoyé sur la première page des résultats.
    - 5 : niveau de pertinence très faible. Les résultats attendus ne sont pas renvoyés dans les deux premières pages des résultats de la recherche. Le document identifié comme pertinent pour cette requête n'apparaît qu'après le 60e résultat. Une source de contenu empêche l'apparition de tous les autres résultats.

Exécution d'un exemple de test

- Avant d'effectuer une quelconque modification, développez des statistiques d'analyse en faisant exécuter le jeu de requêtes prédéfinies par un ensemble d'utilisateurs identifiés et appartenant à différents services/départements au sein de l'entreprise. Choisissez des utilisateurs destinés à utiliser la solution de recherche en production.
- Dans une feuille de calcul de pointage, demandez aux utilisateurs de noter tous les résultats des requêtes exécutées selon l'échelle prédéfinie. Demandez également aux utilisateurs de commenter chacune des recherches.
- Après avoir analysé les statistiques de la configuration de pertinence par défaut du système GSA (pas de synonymes pour l'extension de requêtes, de règle de pondération, etc.), modifiez la configuration de pertinence en fonction des retours/commentaires. Demandez ensuite aux utilisateurs d'effectuer un nouveau test et une nouvelle notation après chaque cycle de modification afin de connaître l'impact des modifications précédentes sur la perception de la pertinence.

Fonctionnalités à envisager pour améliorer la pertinence

Le tableau suivant répertorie les fonctionnalités GSA à considérer pour améliorer la pertinence.

Fonctionnalité	Commentaire
Pondération de la source	À l'aide de la correspondance de format, pondérez une source par rapport à une autre.
Pondération de la date, pondération des métadonnées	Pondérez les documents associés à des métadonnées spécifiques.
Correspondances	Utilisez les correspondances pour privilégier des documents pour certaines requêtes.
Extension des requêtes	Utilisez une règle d'extension des requêtes pour élargir les termes de recherche à d'autres termes (synonymes).
Marqueur intelligent	Lorsque la génération de rapports de recherche avancée est activée, le système GSA utilise la fonctionnalité "Marqueur intelligent" pour analyser les données de flux de clics et mettre en avant certains résultats de recherche au fil du temps. Par exemple, pour une requête de recherche donnée, si les utilisateurs cliquent toujours sur le second résultat de la page au lieu du premier, c'est ce second résultat qui finira par prendre la première position sur la page.
Regroupement et filtrage d'hôtes	le système GSA filtre toutes les combinaisons de : • résultats provenant du même chemin ; • résultats comportant le même titre et les mêmes extraits.
Structure de classement	Spécifier une pondération par URL Veuillez noter qu'il s'agit d'une solution très complexe à gérer. Elle doit toujours être tentée en dernier recours.
Mots vides (fonctionnalité introduite dans la version 6.10 de GSA)	Utilisez les mots vides pour empêcher certains termes de la requête d'être utilisés lors d'une recherche. Veuillez utiliser cette fonctionnalité avec précaution, car elle peut avoir d'importantes répercussions sur le classement si elle permet de résoudre un problème spécifique.
Collections	Répartissez le contenu dans plusieurs collections afin de limiter le corpus de documents pour une requête de recherche.
Affichage des métadonnées et/ou des entités dans une navigation dynamique pour optimiser l'expérience utilisateur.	Au lieu d'afficher la pertinence du système GSA, envisagez d'enrichir l'interface en ajoutant des catégories supplémentaires de navigation dynamique pour les sources de métadonnées ou les entités qui ont été définies dans la reconnaissance d'entités. Il ne s'agit pas vraiment d'une option de réglage. La navigation dynamique peut toutefois permettre d'enrichir l'interface utilisateur afin que ce dernier puisse parcourir l'ensemble des résultats et trouver ceux qu'il recherche.

Approche alternative

Pondération de l'adresse au moment de l'indexation. Utilisez un flux de contenu et spécifiez un classement pour des documents spécifiques. L'attribut "Pagerank" vous permet de spécifier manuellement le classement d'un document. La valeur de l'attribut peut atteindre "99" pour un classement très élevé. La valeur par défaut pour tous les documents dont on récupère le contenu était précédemment de 96.

Présentation des tâches du projet

Le tableau suivant répertorie les tâches et les activités du projet pour le test de pertinence.

Tâche	Activités
Planifier le test.	• Identifier les groupes d'utilisateurs responsables de l'exécution du test et des commentaires relatifs à la pertinence. • Développer une liste de requêtes devant être exécutées par chaque utilisateur lors du test. • Identifier un jeu de documents "pertinents" par requête et par utilisateur. • Développer une échelle de pertinence pour estimer la qualité des résultats de recherche.
Exécuter le test.	• Développer des statistiques d'analyse en demandant aux utilisateurs d'effectuer les tests avant toute modification du système GSA. • Demander aux utilisateurs d'effectuer les tests et de pointer les résultats/commentaires.
Répéter l'opération et effectuer un nouveau test.	• En fonction des commentaires reçus, modifiez la pondération du système GSA et demandez aux utilisateurs d'effectuer un nouveau test. • Affiner et réitérez l'opération jusqu'à ce que les résultats soient satisfaisants.

Amélioration sur le long terme

Développer un processus et un mécanisme pour la collecte des commentaires des utilisateurs et poursuivre l'amélioration de la pertinence des recherches une fois le système mis en production.

Récapitulatif

Chaque déploiement GSA apporte son lot de défis en fonction de votre environnement informatique. Les hypothèses, études, approches, tâches du projet et améliorations sont indiquées à titre d'exemple et ne doivent pas être considérées comme des planifications de référence à appliquer à la lettre. Votre propre environnement et vos délais peuvent refléter une plus grande complexité. Lorsque vous planifiez un projet de déploiement, prenez en compte les exigences commerciales et techniques spécifiques à votre entreprise. N'oubliez pas de tenir compte des aléas dans vos planifications.