Répondre au besoin de sécurisation de Hadoop®

Une approche holistique pour un Data Lake totalement sécurisé

Livre blanc Hortonworks
JULLET 2015

©2015 Hortonworks

Sommaire

©2015 Hortonworks

Introduction

Les entreprises souhaitent aujourd'hui tirer rapidement parti des big data dans le cadre de leurs activités, mais de nouveaux modèles de fonctionnement empêchent parfois les services informatiques de bien comprendre l'ensemble des répercussions associées. Il est indéniable que la création d'un Data Lake basé sur Hadoop offre des bases robustes pour les outils d'analyse et d'information de nouvelle génération. Il est toutefois important d'étudier la sécurité avant de lancer une initiative Hadoop ou d'en étendre la portée. En vous assurant que votre environnement dédié aux big data intègre protection des données et gouvernance, vous pouvez tirer pleinement profit d'analyses avancées sans exposer votre entreprise à de nouveaux risques.

Hortonworks comprend l'importance de la sécurité et de la gouvernance pour les entreprises, quelle que soit leur forme. Pour garantir une protection efficace à nos clients, nous nous appuyons sur une approche holistique qui repose sur cinq piliers :

- Administration
- Authentification et sécurité du périmètre
- Autorisation
- Audit
- Protection des données

Dans chacun de ces domaines, Hortonworks dispose de capacités différenciées qui vont au-delà de celles que proposent les autres fournisseurs afin d'aider ses clients à mettre en œuvre le meilleur niveau de protection possible. Résultat : big data ne rime pas systématiquement avec risques. Les entreprises peuvent ainsi en tirer parti en toute tranquillité.

Comprendre les défis relatifs à la sécurité du Data Lake

Large constat partagé par les plus grandes entreprises, tous secteurs confondus : les données représentent un nouveau levier générateur d'avantage concurrentiel. Hadoop joue un rôle essentiel au sein de l'architecture de données moderne en offrant un stockage de données économique en mode « scale-out » et un traitement créateur de valeur. Une mise en œuvre Hadoop réussie débute typiquement avec l'optimisation de l'architecture de données ou le déploiement de nouvelles applications analytiques avancées qui donnent naissance à un Data Lake. Les types de données, nouveaux et existants, issus de capteurs et de machines, de journaux de serveurs, de parcours de navigation et d'autres sources se déversent dans le Data Lake, qui sert de référentiel central basé sur des services Hadoop partagés. Ces derniers permettent de tirer parti d'informations organisationnelles détaillées recueillies dans un éventail de données vaste et varié.

Il est indispensable de protéger le Data Lake à l'aide de mesures de sécurité complètes. De grands volumes en perpétuelle croissance de données disparates sont stockés dans le Data Lake, qui devient ainsi le coffre-fort de votre entreprise, abritant les données vitales et souvent extrêmement sensibles qui soutiennent vos activités depuis la création de votre organisation. Néanmoins, l'écosystème externe de données et de systèmes opérationnels qui alimentent le Data Lake est exceptionnellement dynamique et peut régulièrement vous exposer à de nouvelles menaces de sécurité. Les utilisateurs de différentes entités opérationnelles peuvent librement accéder au Data Lake et affiner, explorer et enrichir les données qui s'y trouvent à leur guise, à l'aide des méthodes de leur choix, augmentant ainsi les risques d'exposition à des utilisateurs non autorisés. Tout vol interne ou externe de ces données d'entreprise peut être catastrophique, du point de vue de la confidentialité et du respect des réglementations en passant par la mise en péril de la réputation de l'entreprise et de sa valeur actionnariale à long terme. Pour éviter toute atteinte aux activités de l'entreprise, à ses clients, à ses finances et à sa réputation, les responsables informatiques doivent veiller à ce que leur Data Lake réponde aux mêmes normes de sécurité que tout environnement de données déjà en place.

©2015 Hortonworks

La sécurité, un facteur à ne pas sous-estimer

Une protection fragmentée n'est pas plus efficace pour un Data Lake qu'elle le serait pour un référentiel traditionnel. Il est inutile de sécuriser le chemin d'accès principal au Data Lake si un utilisateur peut facilement accéder aux mêmes données d'une autre façon.

Une sécurité Hadoop efficace doit reposer sur une approche holistique, les équipes d'Hortonworks en sont convaincues. Notre framework dédié à une sécurité de bout en bout s'articule autour de cinq piliers : administration, authentification/sécurité du périmètre, autorisation, audit et protection des données.

Les cinq piliers de la sécurité en entreprise

Administration Central management and consistent security	How do I set policy across the entire cluster?
Authentication Authenticate users and systems	Who am I/ prove it?
Authorization Provision access to data	What can I do?
Audit Maintain a record of data access	What did I do?
Data Protection Protect data at rest and in motion	How can I encrypt data at rest and over the wire?

Figure 1 : Impératifs en matière de sécurité d'entreprise

Les administrateurs de la sécurité doivent répondre à ces questions et mettre en œuvre une couverture métier pour chacun de ces piliers à mesure qu'ils conçoivent l'infrastructure qui permettra de sécuriser les données sous Hadoop. Si l'un de ces piliers est vulnérable, il introduira des vecteurs de menaces dans l'ensemble du Data Lake. Dans cette optique, votre stratégie en matière de sécurité Hadoop doit englober l'ensemble de ces cinq piliers et reposer sur une approche cohérente de mise en œuvre afin de garantir son efficacité.

©2015 Hortonworks

Inutile de préciser qu'une protection complète pour l'ensemble de la pile Hadoop ne peut être assurée via une approche ponctuelle. La sécurité doit faire partie intégrante de la plateforme sur laquelle votre Data Lake s'appuie, avec une approche mixte, ascendante et descendante. Seule cette stratégie permet d'appliquer et de gérer la sécurité dans l'ensemble de la pile via un point centralisé d'administration et d'éviter tout écart ou toute incohérence. Cette approche est particulièrement importante pour les déploiements Hadoop qui intègrent en permanence de nouvelles applications ou de nouveaux moteurs de données sous la forme de nouveaux projets Open Source, un scénario dynamique qui peut rapidement exacerber les vulnérabilités du système.

Hortonworks aide ses clients à maintenir le niveau élevé de protection qu'exigent leurs données d'entreprise grâce à l'intégration de fonctions centralisées d'administration et de gestion de la sécurité au cœur de sa solution Hortonworks Data Platform (HDP). HDP est une plateforme de données dédiée aux entreprises dotées de capacités avancées qui couvre sécurité, gouvernance et exploitation. En appliquant des mesures de sécurité au niveau de la plateforme, Hortonworks garantit la gestion homogène de la sécurité pour chaque application basée sur la plateforme de données et facilite le développement ou le retrait d'une application de données, sans impact sur la sécurité.

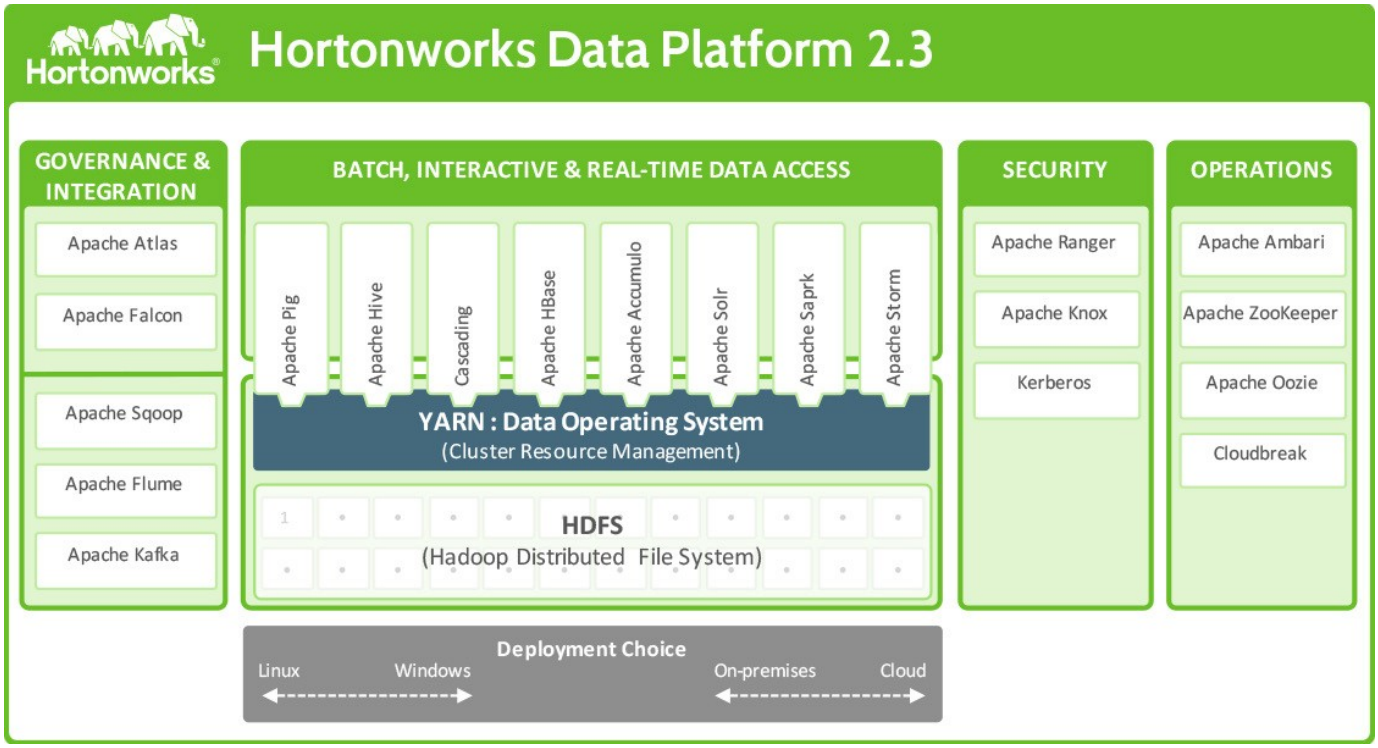

Figure 2 : Hortonworks Data Platform 2.3

©2015 Hortonworks

Notre objectif : accompagner les entreprises

Créée dans le but d'adapter le système Hadoop aux modèles de fonctionnement des entreprises, Hortonworks dispose d'un historique bien établi de réalisations significatives en la matière. Cet objectif, rendre Hadoop accessible aux entreprises, a encouragé l'équipe Hadoop initiale chez Yahoo! à développer Kerberos, le système d'authentification robuste qui servira de base à Hadoop. Depuis lors, Hortonworks poursuit ses investissements majeurs dans le domaine de la sécurité. En mai 2014, Hortonworks acquiert XA Secure, une entreprise leader du secteur de la sécurité des données, pour accélérer son offre d'une approche complète dédiée à la sécurité sous Hadoop. Pour honorer sa mission de développer et distribuer une plateforme de données Apache Hadoop 100 % open source tout en fournissant une assistance technique, Hortonworks a immédiatement intégré la technologie de XA Secure à sa solution Hortonworks Data Platform (HDP), tout en transposant la solution commerciale en projet Apache communautaire et open source intitulé Apache Ranger.

Grâce à HDP, Hortonworks offre une sécurité complète qui couvre les cinq piliers de la sécurité. Avec cette approche basée sur une plateforme, HDP permet aux services informatiques de répondre aux impératifs de sécurité sous Hadoop, et ce mieux que n'importe quelle autre solution.

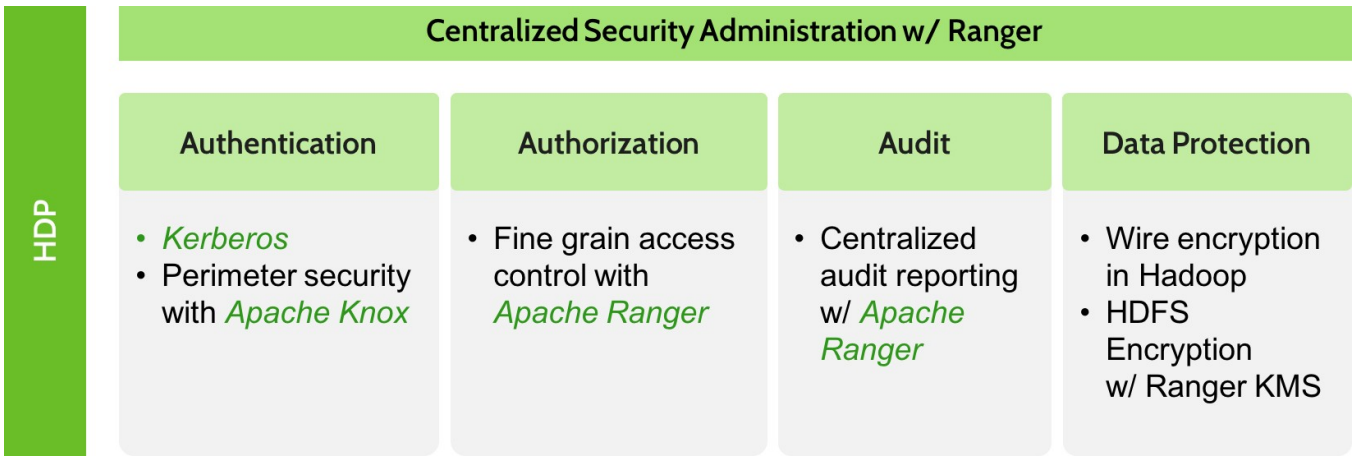

Figure 3 : Sécurité intégrale dans HDP

Administration

Afin d'harmoniser l'administration et la gestion de la sécurité, les administrateurs Hadoop ont besoin d'une interface utilisateur centralisée, un seul écran permettant de définir, d'administrer et de gérer les politiques de sécurité de façon homogène pour tous les composants de la pile Hadoop. Hortonworks répond à cet impératif avec Apache Ranger, un système complètement intégré à HDP qui offre un point central d'administration consacré aux quatre autres piliers fonctionnels de la sécurité sous Hadoop.

©2015 Hortonworks

Ranger améliore la productivité des administrateurs de la sécurité et réduit les erreurs potentielles en leur permettant de définir des politiques de sécurité uniques pouvant être appliquées à d'autres composants de la pile Hadoop à partir d'une interface centrale.

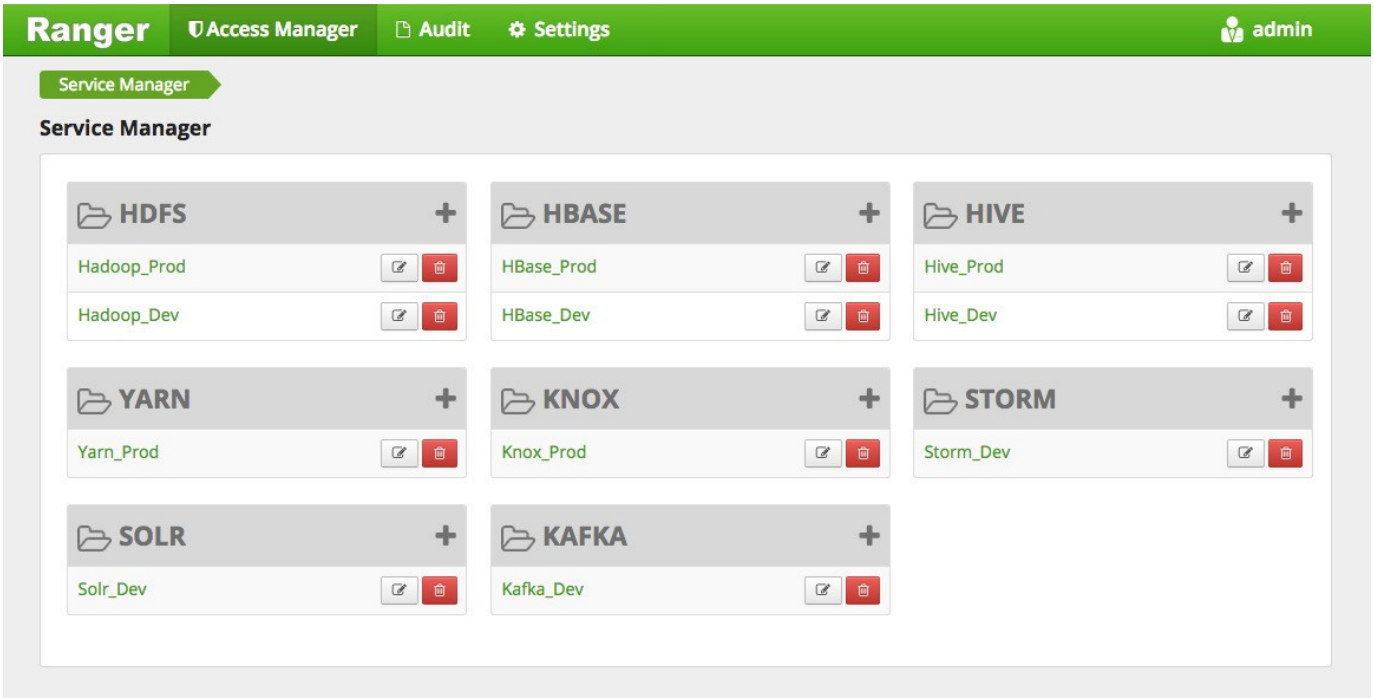

Figure 4 : Apache Ranger, un seul et unique écran pour l'administrateur de la sécurité

D'autres solutions consacrées à la sécurité sous Hadoop en entreprise offrent uniquement une administration partielle avec authentification, autorisation, audit et protection/chiffrement des données. L'administration et la gestion centralisées nécessaires à une sécurité efficace et complète ne sont pas au rendez-vous.

		APACHE RANGER
Centralized security administration	●	Apache Ranger provides a centralized platform for security policy administration

Authentification et sécurité du périmètre

L'établissement de l'identité des utilisateurs par le biais d'une authentification robuste est le principe fondamental à respecter pour sécuriser l'accès sous Hadoop. Les utilisateurs doivent pouvoir s'identifier en toute fiabilité et leur identité doit par la suite pouvoir être propagée dans le cluster Hadoop afin qu'ils accèdent aux ressources (fichiers, répertoires, etc.) et réalisent des tâches, sous MapReduce par exemple. Hortonworks utilise Kerberos, une référence dans le secteur, pour authentifier les utilisateurs et les ressources au sein du cluster Hadoop. Hortonworks a également simplifié l'installation, la configuration et la maintenance de Kerberos avec Ambari 2.0.

©2015 Hortonworks

La passerelle Apache Knox garantit une sécurité du périmètre aux clients Hortonworks. Grâce à Knox, les entreprises peuvent, en toute confiance, étendre leur API REST Hadoop à de nouveaux utilisateurs sans la complexité de Kerberos, tout en maintenant la conformité aux politiques de l'entreprise en matière de sécurité. Point d'accès unique à Hadoop, Knox fournit une passerelle centralisée pour les API REST Hadoop disposant de différents degrés d'autorisation, d'authentification et de capacités SSL et SSO.

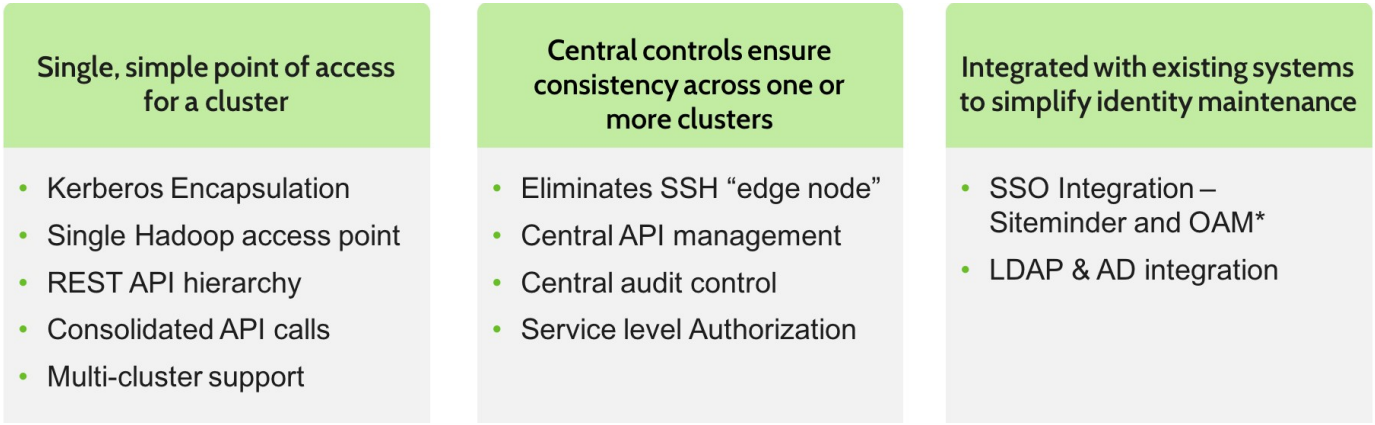

Figure 5 : Sécurité du périmètre avec Apache Knox

Les autres acteurs du marché ne parviennent pas à proposer une solution complète dans ce domaine et se contentent de s'appuyer sur Kerberos pour la sécurité du périmètre. Kerberos est une étape essentielle à l'authentification des utilisateurs, mais le système à lui seul ne suffit pas : il ne permet pas de masquer des points d'entrée au cluster et de bloquer l'accès au périmètre. Comparativement, Apache Knox a été conçue pour servir de plateforme API sécurisée pour Hadoop et est en mesure de bloquer les services aux frontières du cluster. Lorsqu'Apache Knox est utilisée pour les API REST, plusieurs points d'accès au cluster peuvent être rendus invisibles aux utilisateurs finaux, pour une couche supplémentaire de protection au niveau du périmètre.

Apache Knox est un framework universel, mais aussi un nouveau service d'API REST qui peut être facilement ajouté en définissant des services configurables (Knox Stacks).

		KERBEROS
Kerberos-based authentication	●	Ambari simplifies the setup, configuration and maintenance of Kerberos
	●	Ambari includes support for Apache Ranger installation and configuration
		APACHE KNOX
Perimeter security	●	Provide security to all of Hadoop's REST and HTTP services

©2015 Hortonworks

Autorisation

Ranger permet de gérer un contrôle d'accès granulaire via une riche interface utilisateur qui garantit l'harmonisation de l'administration des politiques pour l'ensemble des composants d'accès aux données Hadoop. Les administrateurs de la sécurité ont ainsi la flexibilité de définir des politiques de sécurité pour une base de données, une colonne et une table ou un fichier et d'administrer les permissions pour des protocoles LDAP spécifiques basés sur des groupes ou des utilisateurs individuels. Les conditions dynamiques basées sur des règles, comme l'heure/la date ou le lieu, peuvent également être ajoutées à une politique existante. Ranger est un modèle d'autorisation facile à intégrer et extensible à toutes les sources de données dont la définition est basée sur des services.

Les administrateurs peuvent utiliser Ranger pour définir des politiques de sécurité centralisées pour les composants suivants :

- Apache Hadoop HDFS
- Apache Hadoop YARN
- Apache Hive
- Apache HBase
- Apache Storm
- Apache Knox
- Apache Solr
- Apache Kafka

Ranger s'appuie sur des API d'autorisation standard pour chacun des composants Hadoop et est en mesure d'appliquer de façon centralisée les politiques administrées pour toutes les méthodes d'accès au Data Lake.

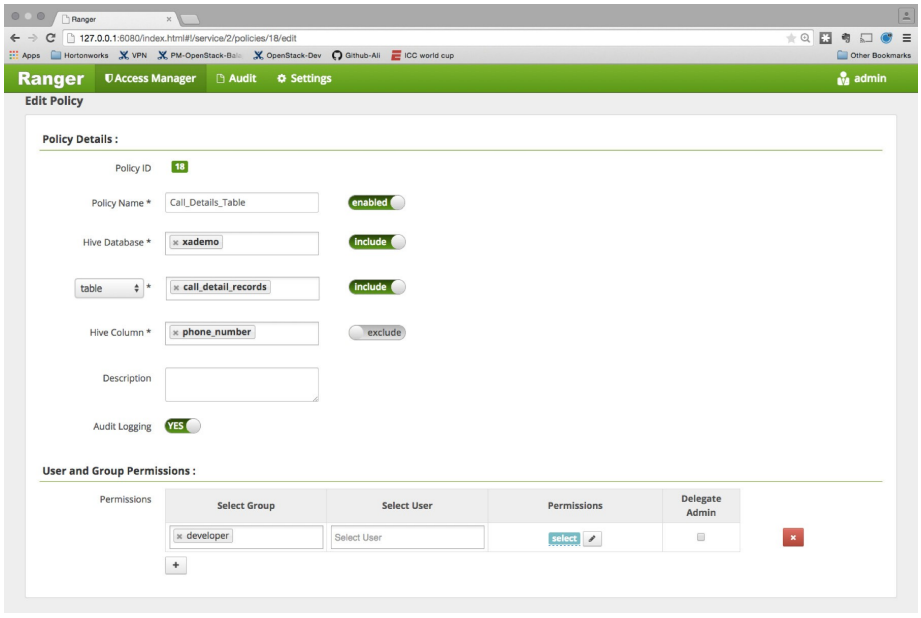

Figure 6 : Définition granulaire des politiques de sécurité avec Apache Ranger

©2015 Hortonworks

Les solutions concurrentes manquent de flexibilité et ne proposent pas d'interface utilisateur assez riche pour permettre aux administrateurs de configurer des politiques de sécurité pour des groupes et des utilisateurs individuels.

Au contraire, Ranger offre aux administrateurs une visibilité détaillée sur le processus d'administration de la sécurité requis dans le cadre d'audits. L'association de la riche interface utilisateur de Ranger et de la visibilité élevée qu'il offre pour les audits rend ce système extrêmement intuitif, pour une productivité accrue des administrateurs de la sécurité.

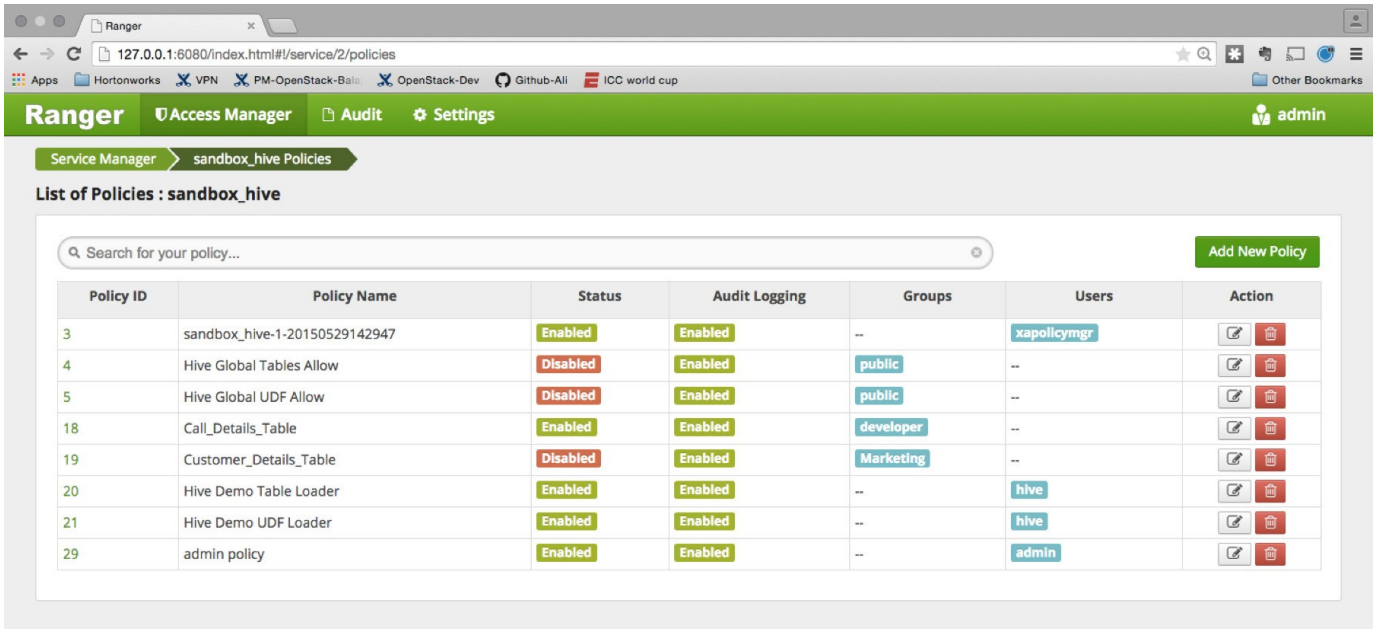

Figure 7 : Avec Apache Ranger, les administrateurs disposent d'une visibilité complète sur le processus d'administration de la sécurité

		APACHE RANGER
Platform-wide coverage across Hadoop stack	●	Coverage across HDFS, YARN, Hive, HBase, Storm, Knox, Solr and Kafka
Fine grain authorization	●	Authorize security policies for a database, table and column or a file as well as LDAP based groups or individual user
Provide hooks for dynamic policy-based authorization	●	Specify dynamic conditions in service definitions Flexibility to define unique conditions by service (HDFS, Hive etc.)
Built on pluggable service-based model	●	Custom plugins can be created for any data store

©2015 Hortonworks

Audit

À mesure que les clients déploient Hadoop au sein d'environnements d'entreprise consacrés aux données et à leur traitement, la gouvernance des données et des métadonnées fait partie des aspects essentiels à tout Data Lake d'entreprise. C'est pourquoi Hortonworks a créé l'initiative Data Governance Initiative (DGI) aux côtés d'Aetna, Merck, Target et SAS. Cette dernière vise à introduire une approche commune en matière de gouvernance des données Hadoop au sein de la communauté open source. Cette initiative a depuis évolué en projet open source intitulé Apache Atlas. Apache Atlas est un ensemble fondamental de services de gouvernance, ces services indispensables permettent aux entreprises de satisfaire, avec la plus grande efficacité et effectivité, aux exigences de conformité et d'intégrer la totalité de l'écosystème des données de l'entreprise sous Hadoop. Ces services incluent :

Recherche et lignage des ensembles de données
Contrôle d'accès aux données basé sur des métadonnées
Événements opérationnels d'audit centralisés indexés et recherches
Gestion du cycle de vie des données de l'ingestion à l'utilisation
Échange de métadonnées avec d'autres outils

Ranger fournit également un framework centralisé permettant la collecte de l'historique d'audit d'accès et la génération facile de rapports sur ces données, avec notamment la possibilité de filtrer les données en fonction de différents paramètres. Associé à Apache Atlas, il permet aux utilisateurs de bénéficier d'une vision complète sur le lignage des données et les audits d'accès, tout en leur offrant la possibilité de filtrer les audits et d'effectuer des recherches en fonction de la classification des données, des utilisateurs ou des groupes, et d'autres filtres.

		APACHE ATLAS AND APACHE RANGER
Data lineage	●	Reporting by entity type or instance
Consolidated audit	●	Ranger provides security audit which can be combined with data lineage in Atlas to provide a comprehensive view
Metadata services	●	Open extensible system with a policy rules engine
Third party support	●	HDP fosters a rich ecosystem of 3rd party vendors

©2015 Hortonworks

Protection des données

La protection des données ajoute une couche robuste de sécurité en rendant illisibles les données en transit sur le réseau ou au repos, stockées sur un disque. HDP chiffre le trafic réseau sous la forme de données entrant et circulant dans le cluster Hadoop via RPC, HTTP, Data Transfer Protocol (DTP) et JDBC. Le trafic réseau sur chacun de ces protocoles peut être chiffré afin de garantir la confidentialité des mouvements de données.

HDP répond également aux contraintes des entreprises en matière de sécurité et de conformité grâce au chiffrement des données au repos. HDP prend en charge le chiffrement des fichiers stockés sous Hadoop, notamment grâce à un système open source de gestion des clés intégré dans Ranger. Ranger offre aux administrateurs de la sécurité la possibilité de gérer les clés et les politiques d'autorisation pour le système de gestion de clés. Hortonworks collabore étroitement avec son partenaire spécialiste du chiffrement pour intégrer le chiffrement HDFS aux frameworks d'entreprise dédiés à la gestion de clés. Avec HDP, nos clients ont la flexibilité de tirer parti du système de gestion de clés open source ou d'utiliser une solution de gestion de clés d'entreprise offerte par leurs partenaires.

Le chiffrement dans HDFS, associé aux politiques d'accès du système de gestion de clés gérées dans Ranger, empêche les administrateurs Linux ou Hadoop malintentionnés d'accéder aux données et prend en charge la séparation des tâches relatives à l'accès aux données et au chiffrement.

	HDP TDE (HDFS) + APACHE RANGER (KMS)
Open source	● 100% open source and open community
Partner ecosystem	● Flexibility to use open sources KMS or existing enterprise solutions ● Hortonworks is currently working with select encryption partners to integrate HDFS encryption with enterprise grade KMS solutions

©2015 Hortonworks

Framework universel et évolutivité

Apache Ranger et Knox offrent une architecture basée sur des services pour l'administration et l'application des politiques qui permet aux clients d'intégrer facilement de nouveaux composants ou moteurs de données. Les applications intégrées à Ranger peuvent tout simplement utiliser l'architecture universelle pour s'appuyer sur les politiques de sécurité existantes dans Ranger sans entraîner la modification intégrale de la définition de ces dernières. Les utilisateurs peuvent également concevoir des services personnalisés sous la forme de plug-ins pour n'importe quel système de gestion de données et créer et gérer des services de façon centralisée pour leurs applications de business intelligence reposant sur des big data. De la même façon, les API REST peuvent être facilement intégrées à Apache Knox au moyen d'une configuration basée sur des services. Le framework configurable permet en outre aux solutions partenaires de fonctionner de façon fluide avec Ranger et Knox afin d'étendre un nouveau service et de prendre en charge un nouveau composant ou moteur de données.

Comparatif des piliers de la sécurité

Point par point, pilier par pilier, HDP offre de riches fonctionnalités pour protéger les données sensibles qui se trouvent dans votre Data Lake Hadoop. Pour chaque aspect, l'offre d'Hortonworks atteint ou dépasse les possibilités offertes par les solutions concurrentes. La conception ascendante de la plateforme de l'architecture de sécurité d'Hortonworks garantit que tous les piliers se complètent, pour une protection intégrale. La centralisation de l'administration et de la gestion permet aux services informatiques d'adopter une approche holistique en matière de sécurité, tout en augmentant la productivité et l'efficacité des systèmes. En appliquant une approche basée sur une conception universelle et complètement open source, Hortonworks permet aux partenaires de l'écosystème des entreprises d'introduire facilement de nouvelles fonctionnalités de sécurité. Ainsi, les entreprises peuvent tirer parti d'un système de sécurité complet et facile à utiliser pour leurs données et appliquer, comme il se doit, la même rigueur pour Hadoop et leurs environnements de données traditionnels.

	HDP	OTHER DISTRIBUTIONS
Administration	●	◑
Authentication	●	◑
Perimeter Security	●	○
Authorization	●	◑
Audit	◕	◑
Data Protection	◕	◑

Figure 8 : Analyse concurrentielle des piliers de la sécurité

©2015 Hortonworks

Synthèse

Aucune entreprise ne peut se permettre de tirer parti des big data au détriment de sa sécurité. Lors de la planification de votre stratégie Hadoop, veillez à ce que la plateforme que vous sélectionnez offre une approche complète et holistique pour la protection de votre Data Lake et des précieuses informations qu'il contient. Grâce à Hortonworks, les entreprises peuvent mettre en œuvre une plateforme qui couvre les cinq piliers de la sécurité Hadoop au sein de leur architecture : administration, authentification/sécurité du périmètre, autorisation, audit et protection des données. Contrairement aux autres plateformes Hadoop qui offrent uniquement des mesures de sécurité partielles, HDP inclut tout ce dont les services informatiques ont besoin pour encourager une stratégie complète de sécurisation des big data, pour des avantages concurrentiels au service des utilisateurs dans l'ensemble de l'organisation, sans exposer l'entreprise à de nouveaux risques.

À propos d'Hortonworks

Hortonworks développe, distribue et soutient la seule plateforme de données Apache Hadoop 100 % open source. Notre équipe compte le plus grand nombre de concepteurs et d'architectes de l'écosystème Hadoop, qui représentent et dirigent les critères généraux pour les entreprises au sein de ces communautés. Hortonworks Data Platform est une plateforme ouverte qui s'intègre parfaitement aux investissements déjà existants en informatique, à partir de laquelle les entreprises peuvent élaborer et déployer des applications basées sur Hadoop. Hortonworks entretient des relations étroites avec des partenaires stratégiques de datacentres, qui permettent à nos clients de profiter des meilleures opportunités d'Hadoop. Pour plus d'informations, rendez-vous à l'adresse : www.hortonworks.com.

©2015 Hortonworks