Hortonworks

Une architecture de données moderne Avec Apache™ Hadoop®

Le chemin vers le data lake

Un document de
présentation technique de
Hortonworks – mars 2014

Résumé

Apache Hadoop n'a pas bouleversé les centres de données, ce sont les données qui ont engendré ce bouleversement.

Peu après que les services informatiques centraux des entreprises ont adopté des systèmes à grande échelle pour la gestion des données, les entrepôts de données de l'entreprise (EDW) se sont avérés être l'emplacement de stockage logique pour toutes les données d'entreprise. Aujourd'hui, chaque entreprise dispose d'un entrepôt de données qui sert à modéliser et concentrer le cœur de l'activité à partir des systèmes de l'entreprise.

L'explosion de nouveaux types de données ces dernières années – qu'il s'agisse d' entrées à partir du Web, d'appareils connectés ou encore de volumes considérables d'enregistrements – a exercé une pression phénoménale sur les EDW.

En réponse à ces bouleversements, un nombre toujours plus important d'entreprises ont choisi Apache Hadoop pour les aider à gérer l'extraordinaire augmentation de leurs données tout en maintenant la cohérence de leur entrepôt de données.

Ce document présente Apache Hadoop, ses capacités en tant que plate-forme de données et la manière dont le cœur de Hadoop et son écosystème environnant de fournisseurs de solutions permettent d'intégrer les exigences de l'entreprise dans l'entrepôt de données et les autres systèmes de données de l'entreprise, afin de constituer une architecture de données moderne, première étape sur le chemin qui mène au data lake d'une entreprise.

Le data lake d'une entreprise lui permet d'obtenir les avantages substantiels suivants

Des gains d'efficacité pour l'architecture des données grâce à un coût de stockage nettement plus faible et à l'optimisation des charges de traitement de données, telles que les données de transformation et d'intégration.

De nouvelles opportunités commerciales grâce à un accès flexible du type « schéma à la lecture » (schema-on-read) sur toutes les données d'entreprise et grâce à un traitement de données multiutilisation, multicharge sur les mêmes ensembles de données, qu'il s'agisse de traitement par batch ou en temps réel.

Apache Hadoop offre tous ces avantages grâce à une technologie fondamentale comprenant :

Hadoop Distributed Filesystem. HDFS est un système de fichiers basé sur Java qui autorise un stockage de données évolutif et fiable permettant de couvrir des clusters importants de serveurs standard.

Apache Hadoop YARN. YARN fournit une architecture insérée et une gestion des ressources pour les moteurs de traitement de données, afin de pouvoir échanger avec les données stockées dans HDFS.

Pour obtenir une analyse indépendante de la plate-forme de données Hortonworks, téléchargez Forrester Wave™ : Big Data Hadoop Solutions, Q1 2014 *de Forrester Research.*

Bouleversement au niveau des données

Depuis de nombreuses années, les services informatiques centraux des entreprises ont dû faire face à des problèmes de données à grande échelle. La très grande majorité des données produites au sein de l'entreprise provient des systèmes très étendus de planification des ressources de l'entreprise (ERP), des systèmes de gestion de la relation client (CRM) et des autres systèmes qui supportent une fonction particulière de l'entreprise. Peu de temps après que ces « systèmes d'enregistrement » sont devenus la norme, l'entrepôt de données s'est imposé comme l'emplacement logique de stockage des données extraites de ces systèmes, afin de pouvoir débloquer les applications de « business intelligence » : c'est ainsi qu'un nouveau secteur d'activité est né. Aujourd'hui, chaque entreprise dispose d'entrepôts de données qui servent à modéliser et concentrer le cœur de l'activité à partir des systèmes de l'entreprise.

Le défi posé par les nouveaux types de données

L'émergence et l'explosion de nouveaux types de données au cours des dernières années ont mis une pression énorme sur tous les systèmes de données au sein de l'entreprise. Ces nouveaux types de données proviennent de « systèmes d'engagement », comme les sites Web, ou de l'utilisation plus importante des appareils connectés.

Les données provenant de ces sources ont plusieurs caractéristiques qui constituent un important défi pour l'entrepôt de données :

Croissance exponentielle. Pour une estimation de 2,8 Zo de données en 2012, on pense parvenir à 40 Zo en 2020. 85 % de cette croissance devrait provenir de nouveaux types de données. On estime que les données générées par les machines vont être multipliées par 15 en 2020. (Source IDC)

Grande diversité. Les données entrantes peuvent être faiblement ou pas structurées du tout, ou bien leur structure peut changer trop fréquemment pour permettre la création d'un schéma fiable au moment de leur acquisition.

Valeur des volumes massifs. Les données entrantes peuvent avoir une valeur faible ou nulle sous forme individuelle ou de petits groupes d'enregistrements. Mais des volumes élevés et des historiques plus longs peuvent être analysés pour établir des modèles et utilisés pour des applications analytiques avancées.

La croissance d'Apache Hadoop

Indépendamment des problèmes de capture et de stockage, l'amalgame des données d'entreprises existantes et la valeur ainsi apportée par ces nouvelles données ont été plébiscités par de nombreuses entreprises issues de différents secteurs, qu'il s'agisse de la vente au détail, des soins de santé, de la publicité ou de l'énergie.

La technologie émergente qui paraît capable de relever ce défi et de créer de la valeur avec le « big data » est Apache Hadoop, dont le développement a été jugé « irrésistible » par Forrester Research dans Forrester Wave™: Big Data Hadoop Solutions, Q1 2014.

La maturation d'Apache Hadoop au cours des dernières années a élargi ses capacités : l'application est passée d'un simple traitement de vastes ensembles de données à une plate-forme de données à part entière contenant tous les services nécessaires à l'entreprise – sécurité, gestion opérationnelle et autres.

Découvrez d'autres informations sur ces nouveaux types de données sur ***Hortonworks.com***

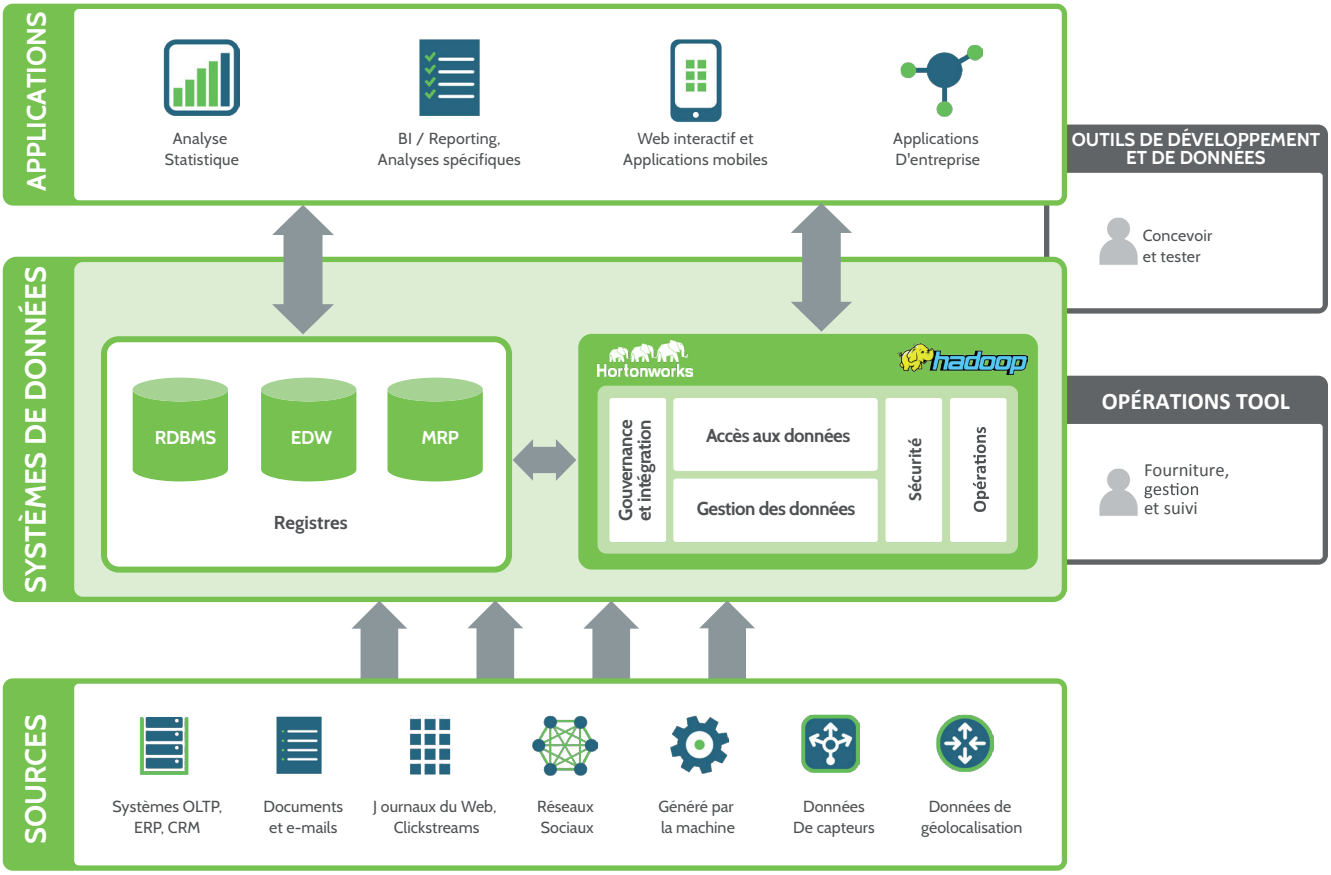
- *Clickstream*
- *Réseaux sociaux*
- *Journaux de serveur*
- *Géolocalisation*
- *Machine et capteur*

Qu'est-ce que Hadoop ?

Apache Hadoop est une technologie open source issue de l'expérience d'entreprises grand public présentes sur le Web – comme Yahoo, Facebook et d'autres – qui ont été les premières à être confrontées au besoin de stocker et traiter des quantités massives de données numériques.

Hadoop et vos systèmes de données existants : Une architecture de données moderne

Sur le plan de l'architecture, l'utilisation de Hadoop en tant que complément aux systèmes de données existants est particulièrement séduisante : il s'agit d'une technologie open source conçue pour fonctionner sur de multiples serveurs standard. Hadoop adopte une approche « scale-out » à faible coût pour le stockage et le traitement des données et peut s'adapter aux besoins des plus importantes applications Web dans le monde.



Ill. 1
Une architecture de données moderne avec l'intégration d'Apache Hadoop dans les

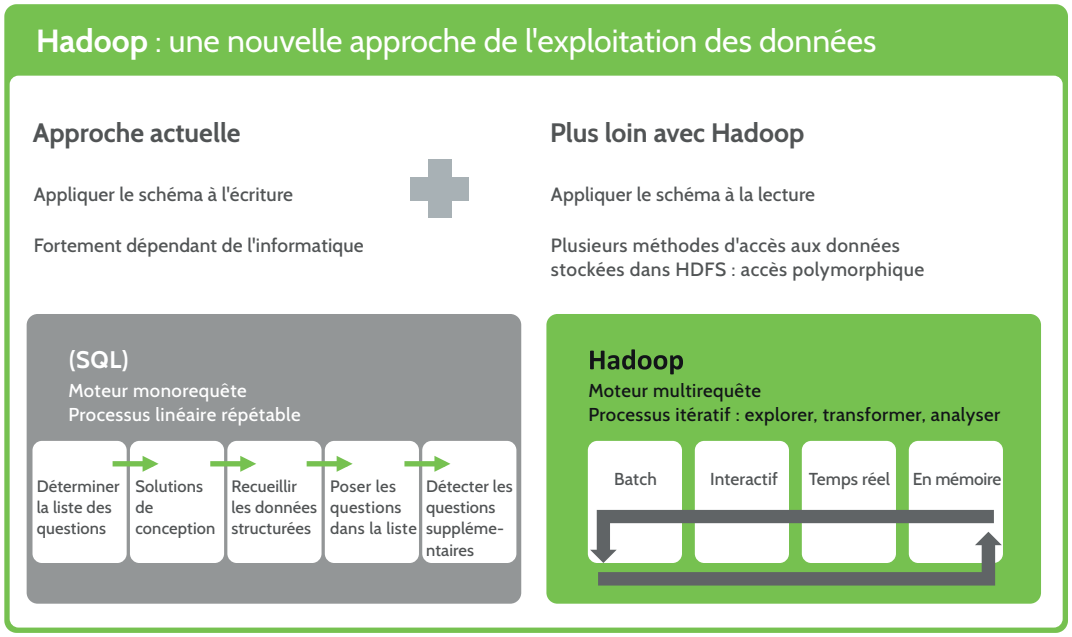
Hortonworks consacre tous ses efforts à faire de Hadoop un composant essentiel des centres de données. Après avoir intensément travaillé avec certains des plus importants fournisseurs d'entrepôts de données, nous avons pu identifier plusieurs opportunités et avantages essentiels que Hadoop apporte à l'entreprise.

Nouvelles opportunités pour l'analytique

L'architecture de Hadoop offre de nouvelles opportunités pour l'analytique de données :

Schéma à la lecture Contrairement à un EDW, dans lequel les données sont transformées selon un schéma spécifique au moment de leur chargement dans l'entrepôt – ce qui nécessite un « schéma à l'écriture » –, Hadoop permet aux utilisateurs de stocker des données sous forme brute ; les analystes peuvent ensuite créer le schéma qui convient aux besoins de leur application au moment qu'ils choisissent pour analyser les données, donnant ainsi naissance au « schéma à la lecture ». Cela permet de résoudre le problème du manque de structures et d'investissements dans le traitement des données lorsque la valeur initiale des données d'entrée peut être mise en doute.

Supposons par exemple qu'une application existe et qu'elle combine des données CRM et Clickstream pour obtenir une vision unique de l'interaction avec le client. Lorsque de nouveaux types de données pertinents deviennent disponibles (p. ex., journal de serveur ou données de sentiment), ils peuvent être ajoutés pour enrichir la vision du client. La différence essentielle est qu'au moment où les données ont été stockées, il n'a pas été nécessaire de déclarer leur structure et leur liaison avec une application particulière.



Ill. 2

Traitement des données multiutilisation, multicharge. En supportant de multiples méthodes d'accès (batch, temps réel, streaming, en mémoire, etc.) à un ensemble commun de données, Hadoop permet aux analystes de transformer et de visualiser les données de multiples manières (au moyen de différents schémas) pour obtenir une analytique en boucle fermée, ce qui rend le temps d'accès aux données plus proche que jamais du temps réel.

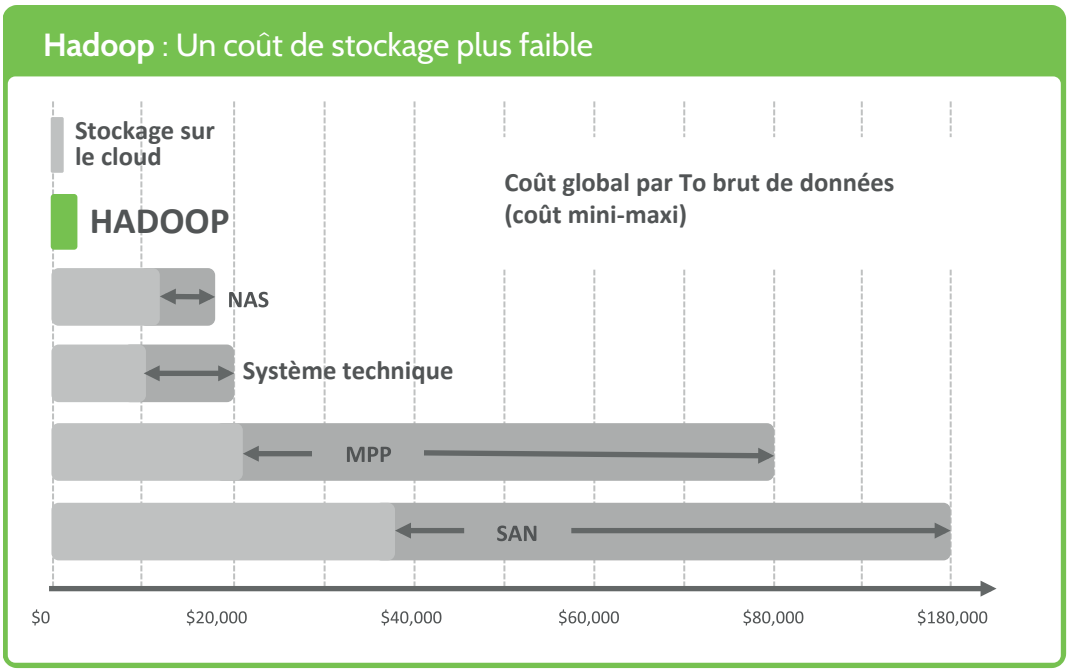
Par exemple, une usine pourrait choisir de réagir à des données de capteurs entrantes par un traitement en temps réel, en permettant aux analystes de vérifier les journaux de données pendant la journée à l'aide d'un traitement interactif et en lançant une série de processus batch pendant la nuit. Hadoop rend ce scénario possible avec un seul cluster de ressources partagées et des versions uniques de données.5

Des gains d'efficacité pour l'architecture des données

Par-delà les opportunités offertes pour l'analyse des données volumineuses, Hadoop accroît l'efficacité de l'architecture des données :

Un coût de stockage plus faible. Par conception, Hadoop fonctionne sur des serveurs standard à faible coût et des périphériques de stockages locaux, ce qui permet de baisser conséquemment le coût global du stockage. En particulier, quand on la compare aux réseaux de stockage haut de gamme (SAN) proposés par des fournisseurs comme EMC, l'option d'utiliser des moyens de calcul et de stockage standard extensibles avec Hadoop est une alternative très séduisante – elle permet à l'utilisateur de faire évoluer son matériel uniquement lorsque la croissance de ses données le justifie. Cette approche dynamique en matière de coût permet de stocker, traiter, analyser et retrouver plus de données que jamais auparavant

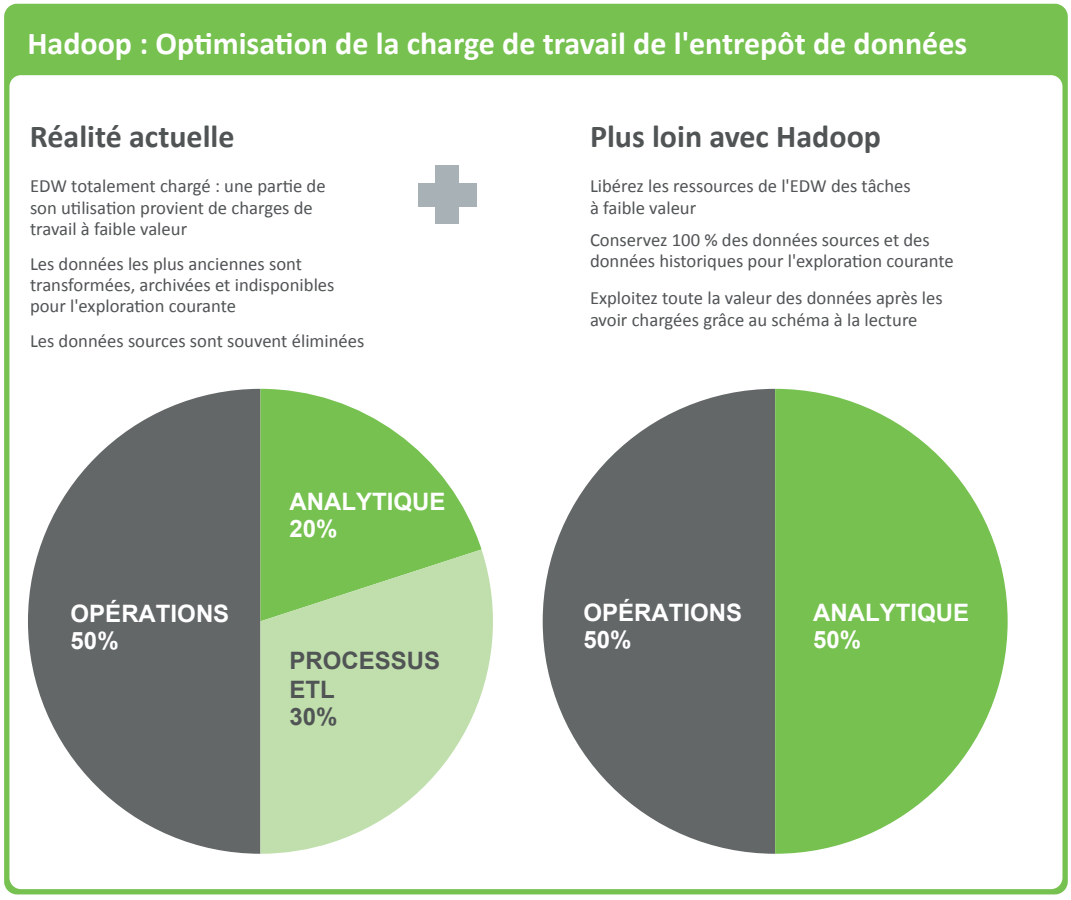
Par exemple : dans une application traditionnelle de business intelligence, vous pourriez être limité à une seule année de stockage de données après leur transformation à partir de leur format original, tandis qu'en ajoutant Hadoop, il devient possible de stocker dix années de plus dans l'entrepôt de données, y compris dans leur format original. Cela débouche sur des applications beaucoup plus riches et un contexte historique bien plus important.



Ill. 3
Source : Juergen Urbanski, membre du Conseil Big Data et analytique,

Optimisation de la charge de travail de l'entrepôt de données.. L'étendue des tâches réalisée par l'EDW s'est considérablement accrue avec la fonction ETL (extraction, transformation et chargement), l'analytique et les opérations. La fonction ETL est une charge de travail de calcul à valeur relativement faible qui peut être réalisée de manière beaucoup moins coûteuse. Beaucoup d'utilisateurs délèguent cette fonction à Hadoop, qui va extraire et transformer les données et charger les résultats dans l'entrepôt de données.

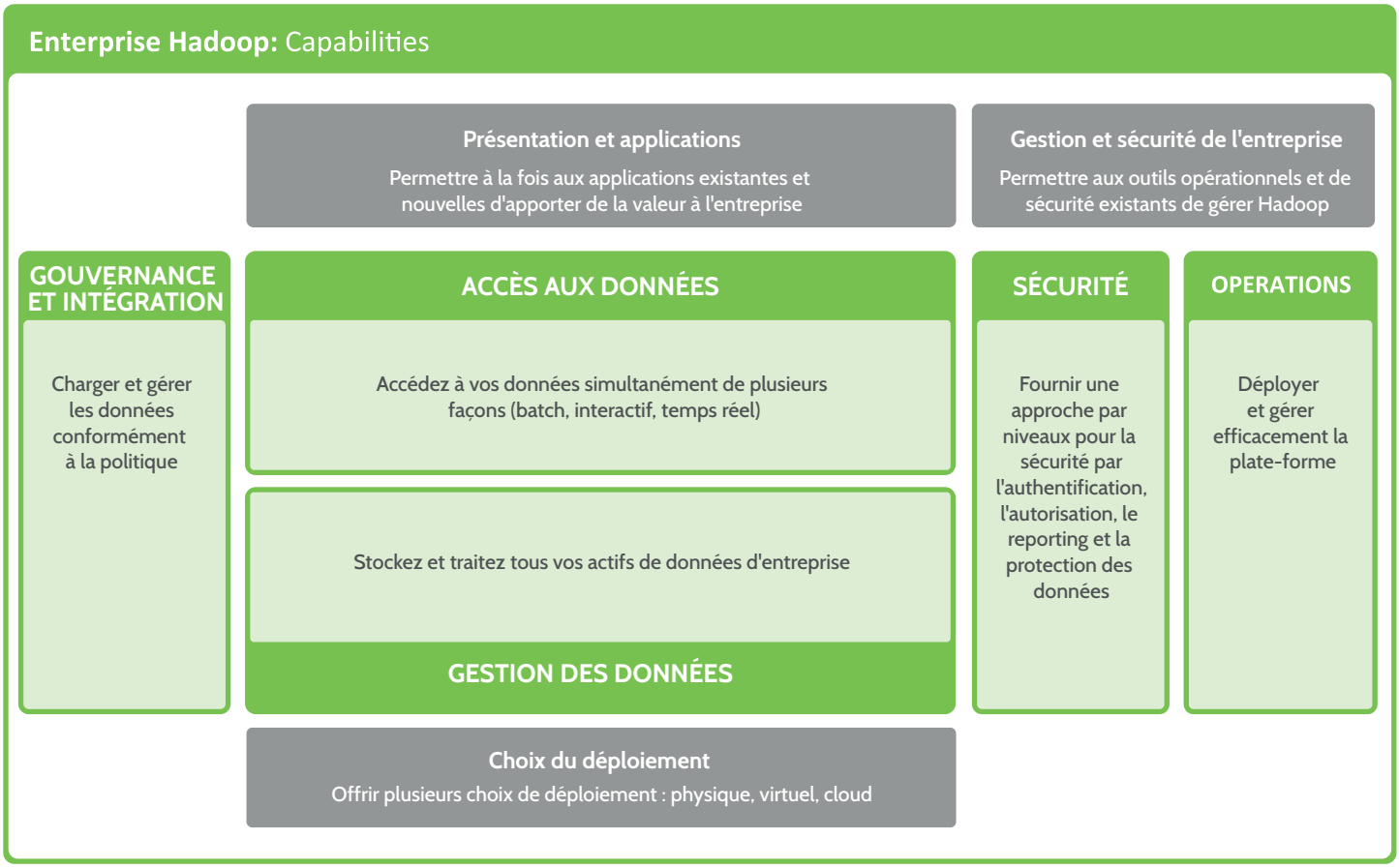
Ainsi, des cycles critiques de CPU et de l'espace de stockage peuvent être libérés de l'entrepôt de données, ce qui lui permet d'exécuter les fonctions à haute valeur ajoutée – l'analytique et les opérations – qui exploitent le mieux ses capacités avancées.



Ill. 4

Un plan d'action pour Enterprise Hadoop

Au fur et à mesure qu'Apache Hadoop obtenait des succès dans les architectures de données des entreprises, les capacités de la plate-forme se sont considérablement élargies en réponse aux exigences de celles-ci. Ainsi, au tout début, les composants de base pour le stockage (HDFS) et le calcul (MapReduce) représentaient les éléments clés d'une plate-forme Hadoop. Bien que ces composants restent essentiels aujourd'hui, un grand nombre de projets de soutien ont été ajoutés sur la Fondation Apache Software (ASF) par les fournisseurs et les utilisateurs, ce qui permet de transformer Hadoop en une plate-forme de données d'entreprise beaucoup plus large.



Ill. 5

Ces capacités d'Enterprise Hadoop satisfont aux domaines fonctionnels suivants qui constituent une exigence fondamentale pour toute technologie de plate-forme :

Gestion des données . Stocker et traiter une grande quantité de données dans une couche de stockage évolutive.

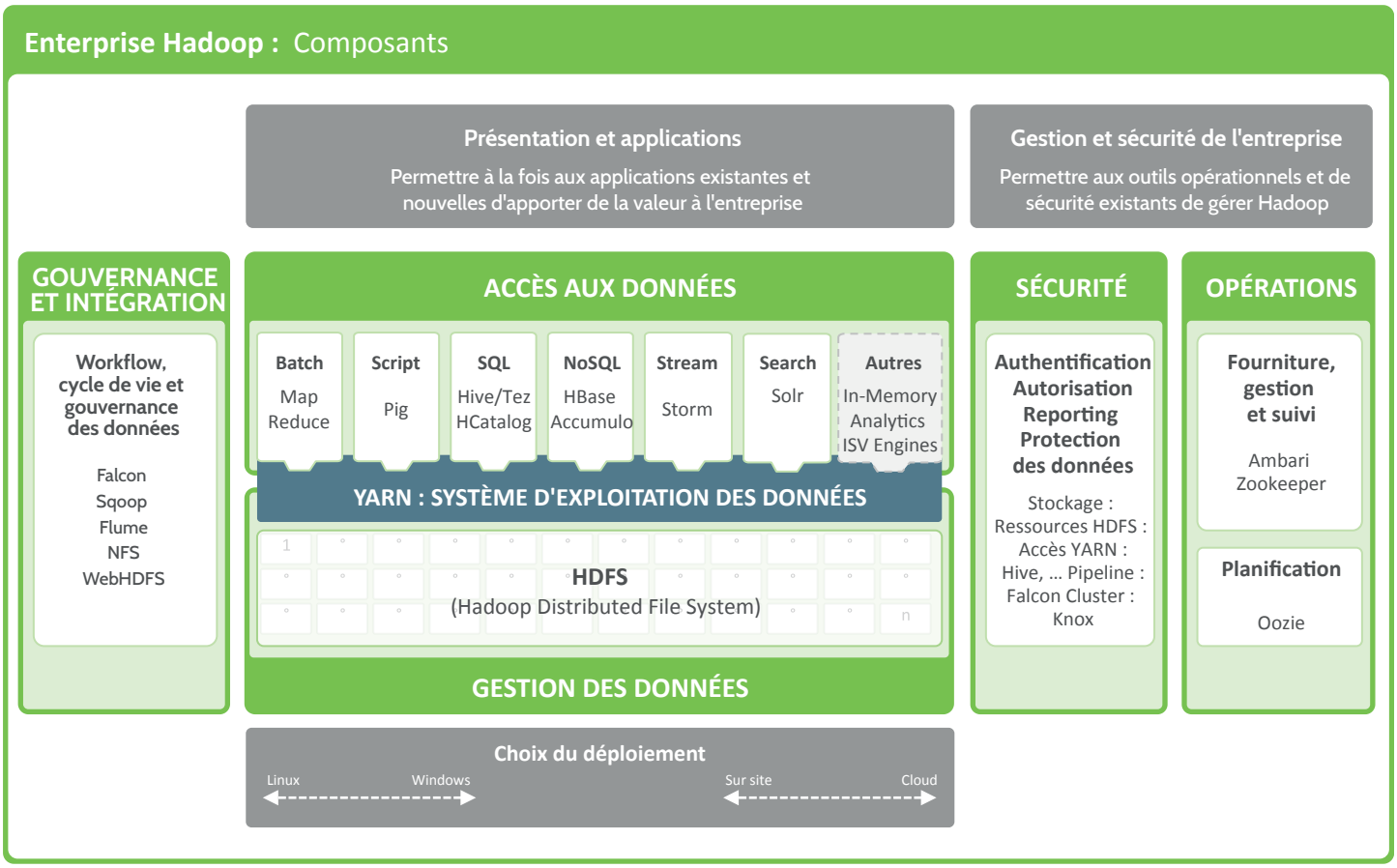
Accès aux données.. Accédez à et traitez vos données de différentes façons : batch, interactif, streaming et temps réel.

Gouvernance et intégration des données Charger rapidement et facilement les données, et les gérer conformément à la politique

Sécurité. Prendre en compte les exigences liées à l'authentification, l'autorisation, le reporting et la protection des données.

Opérations. Établissez, gérez, surveillez et exploitez pleinement les clusters Hadoop.

Les projets Apache qui assurent cet ensemble de fonctions sont détaillés dans le diagramme suivant. Cet ensemble de projets et de technologies représente le cœur d'Enterprise Hadoop. Les locomotives technologiques que sont Microsoft, SAP, Teradata, Yahoo!, Facebook, Twitter, LinkedIn – ainsi que beaucoup d'autres – contribuent en permanence à renforcer les capacités de la plate-forme open source, chacune d'entre elles apportant ses propres capacités¬ et cas d'utilisation. C'est pourquoi le rythme d'innovation d'Enterprise Hadoop surpasse encore notre propre rythme.



Ill. 6

Gestion des données : Hadoop Distributed File System (HDFS) est la technologie de base pour une couche de stockage évolutive et efficace, et elle est conçue pour fonctionner sur des matériels standard de faible coût. Apache Hadoop YARN est un prérequis pour Enterprise Hadoop, car cette application dispose de la gestion de ressources et de l'architecture insérée nécessaires à l'utilisation d'un grand nombre de méthodes d'accès aux données, et elle permet d'exploiter les données stockées dans Hadoop avec des niveaux de performance et de services prévisibles.

Accès aux données : Bien qu'il y ait de nombreux moteurs spécialisés, Apache Hive est la technologie d'accès aux données la plus largement adoptée. Par exemple, Apache Pig offre des capacités de scénarisation, Apache Storm un processus en temps réel, Apache HBase un stockage NoSQL sous forme de colonnes et Apache Accumulo un contrôle d'accès au niveau de la cellule. Tous ces moteurs peuvent fonctionner avec un ensemble de données et ressources grâce à des cadres de programmation comme le traitement en cascade.

Gouvernance et intégration des données : Apache Falcon dispose de workflows pour la gouvernance basés sur la politique choisie, tandis que Apache Flume et Sqoop facilitent l'acquisition de données, de même que les interfaces NFS et WebHDFS, vers HDFS.

Sécurité: La sécurité est assurée sur chaque couche de l'empilage Hadoop, depuis HDFS et YARN jusqu'à Hive et les autres composants d'accès de données, sur la totalité du périmètre du cluster via Apache Knox.

Opérations : Apache Ambari propose l'interface et les API (interfaces de programmation d'applications) nécessaires pour fournir, gérer et suivre les clusters Hadoop et pouvoir les intégrer avec les autres logiciels de la console de gestion.

Un écosystème riche

Au-delà de ces composants de base et dans le cadre de son processus d'innovation – dont YARN est un exemple –, Apache Hadoop dispose d'un riche écosystème de fournisseurs qui proposent des capacités et/ou des points d'intégration supplémentaires. Ces partenaires contribuent à Hadoop et le développent avec certaines fonctionnalités, et cette combinaison de fonctions de base et d'écosystème aboutit à des solutions séduisantes pour les entreprises, quels que soient leurs cas d'utilisation. Parmi les exemples d'intégrations en partenariat figurent :

La Business Intelligence et l'analytique : Tous les principaux fournisseurs de BI proposent une intégration avec Hadoop, et les fournisseurs d'analytique spécialisée proposent des solutions de niche pour des types de données et des cas d'utilisations spécifiques.

Gestion des données et outils : De nombreux partenaires proposent des solutions verticales et horizontales de gestion de données¬ en parallèle de Hadoop, et il existe une grande quantité d'outils – depuis les kits de développement logiciel jusqu'aux interfaces IDE complètes – pour développer les solutions Hadoop.

Infrastructure : Bien que Hadoop soit conçue pour des matériels standard, elle peut aussi fonctionner comme une appliance et être facilement intégrée dans d'autres solutions de stockage, de données et de gestion, à la fois sur site et dans le cloud.

Intégrateurs de systèmes : : Naturellement, en tant que composants de l'architecture des données d'une entreprise, les SI (intégrateurs de systèmes) de toute taille acquièrent des compétences pour favoriser le développement de l'intégration et de solutions.

Comme la plupart de ces fournisseurs sont déjà présents dans les entreprises et proposent des capacités comparables pour un EDW, la mise en œuvre présente peu de risques, car les équipes peuvent s'appuyer sur les outils et compétences existants issus des charges de travail de l'EDW.

Il existe aussi un riche écosystème de nouveaux fournisseurs en train d'émerger pour servir la plate-forme Enterprise Hadoop. Ces nouvelles sociétés tirent profit des API ouvertes et des nouvelles capacités des plates-formes pour créer une génération entièrement nouvelle d'applications. Les applications qu'ils conçoivent s'appuient à la fois sur les types de données existants et nouveaux, et utilisent de nouveaux types de traitement et d'analyse qui n'étaient, techniquement ou financièrement, pas envisageables avant l'émergence de Hadoop. Par le fait, ces nouvelles sociétés parviennent à maîtriser la croissance impressionnante des données et créent des opportunités pour améliorer, entre autres, la connaissance des clients, la recherche médicale et la fourniture de soins de santé, l'efficacité des activités d'exploration et de production d'énergie, les prévisions policières.

Hortonworks dispose d'un large écosystème de partenaires et a noué des relations stratégiques avec les principaux fournisseurs de centres de données :

- *HP*
- *Microsoft*
- *Rackspace*
- *Red Hat*
- *SAP*
- *Teradata*

Le chemin vers le data lake

La mise en place de Hadoop dans une architecture de données est une décision importante pour toute entreprise. Bien que le développement de Hadoop soit « irrésistible », son adoption constitue un cheminement à partir des applications à instance unique vers un data lake à part entière. Ce cheminement a pu être observé de nombreuses fois dans notre base de clients.

Nouvelles applications analytiques

L'utilisation de Hadoop naît généralement de la volonté de créer de nouvelles applications analytiques alimentées par des données qui n'étaient préalablement pas capturées. Bien que chaque application soit toujours spécifique à un secteur d'activité ou une entreprise, il existe de nombreuses similitudes entre les types de données.

Ci-dessous figurent quelques exemples d'applications analytiques pour différents secteurs :

INDUSTRIE	CAS D'UTILISATION	TYPE DE DONNÉES								
		Capteur	Journaux de serveur	Texte	Social	Géographique	Machine	Clickstream	Structuré	Non structuré
Services Financiers	Dépistage des risques liés aux nouveaux comptes		✓	✓						
	Risques de trading		✓							
	Souscription d'assurances	✓		✓		✓				
Télécommunications	Enregistrements détaillés des appels (CDR)					✓	✓			
	Investissements en infrastructures		✓				✓			
	Attribution de bande passante en temps réel		✓	✓	✓					
Vente au détail	Profilage complet des clients			✓				✓		
	Promotions localisées et personnalisées					✓				
	Optimisation de sites Web							✓		
Fabrication	Chaîne d'approvisionnement et logistique	✓								
	Assurance qualité sur ligne de montage	✓								
	Assurance qualité participative				✓					
Soins de santé	Utilisation des données	✓							✓	
	Génomiques dans les essais médicaux									
Secteur pharmaceutique	Pharmaceutique Recrutement et conservation Des patients pour les tests de médicaments				✓			✓		
	Amélioration du respect des ordonnances				✓	✓				✓
Pétrole et gaz	Standardisation des données d'exploration et de production	✓				✓				✓
	Suivi de la sécurité des puits en temps réel	✓								✓
Gouvernement	Réponse aux pressions budgétaires du Gouvernement par délestage des tâches ETL								✓	
	Analyse de sentiments pour le gouvernement Programmes				✓					

Ill. 7

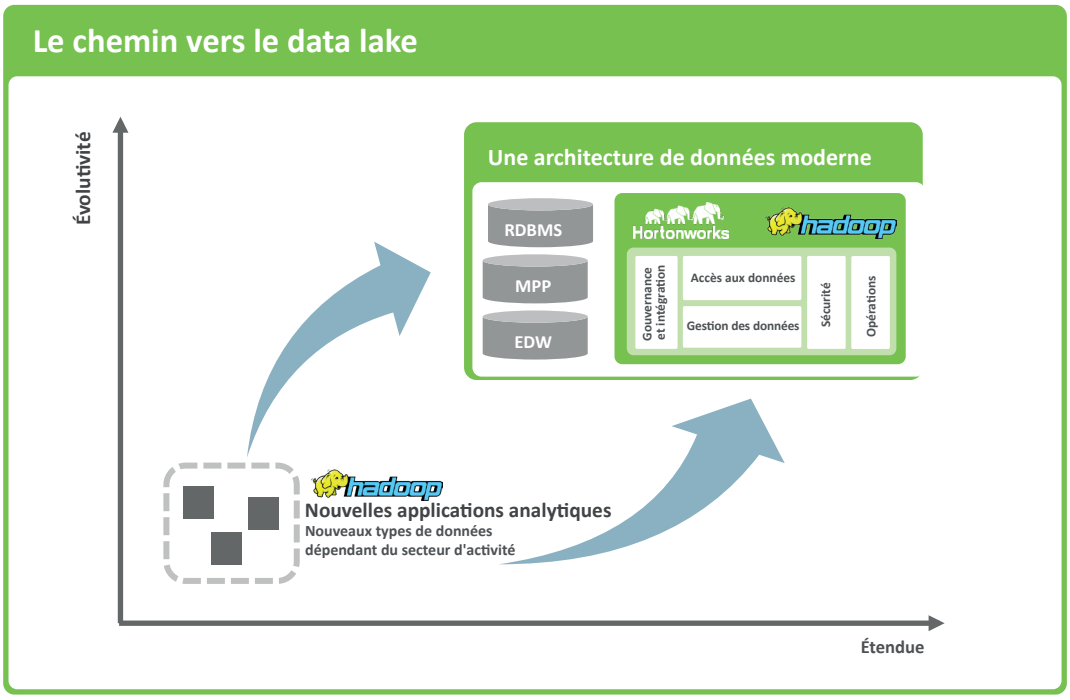
Enterprise Hadoop
Découvrez d'autres cas d'utilisation dans l'entreprise

- *Soins de santé*
- *Télécommunications*
- *Vente au détail*
- *Fabrication*
- *Services financiers*
- *Pétrole et gaz*
- *Publicité*
- *Gouvernement*

Développement de l'étendue et de l'évolutivité

Au fur et à mesure que Hadoop prouve sa valeur sur une ou plusieurs instances d'application, les données et les opérations sont enrichies en étendue et en évolutivité. L'architecture de données résultante permet d'assister progressivement une entreprise sur de nombreuses applications.

Les études de cas décrites plus loin dans le document précisent le chemin vers le data lake emprunté par des clients des secteurs de la vente au détail et des télécommunications.



Ill. 8

Vision d'un data lake

Avec la progression continue des applications analytiques en étendue et en évolutivité liée à l'utilisation de Hadoop et d'autres sources de données, la vision d'un data lake d'entreprise commence à devenir une réalité.

De manière pratique, un data lake se caractérise par trois attributs principaux :

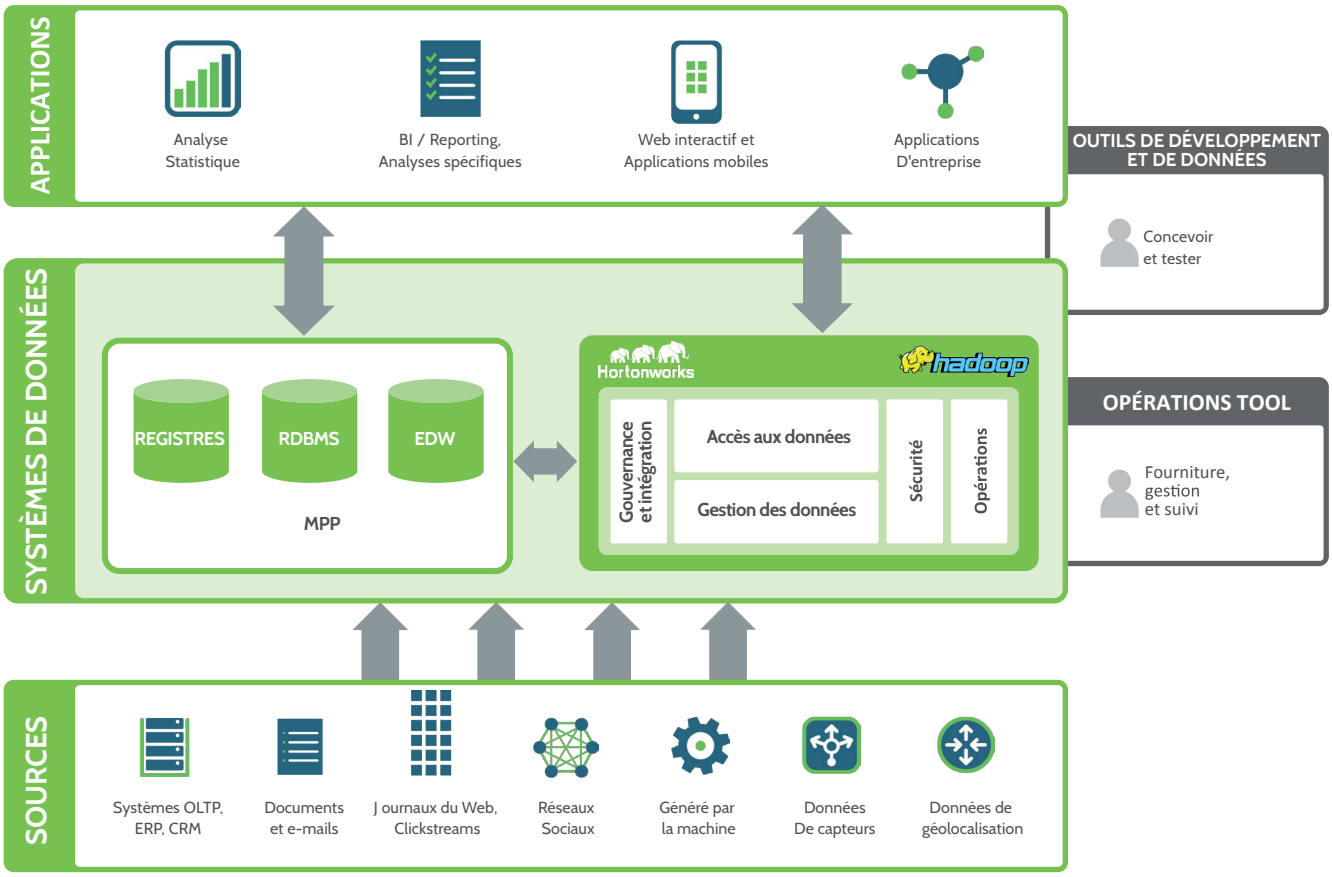
La collecte de toutes les données. Un data lake contient toutes les données provenant de sources brutes sur de longues périodes, mais aussi toutes les données traitées.

Une exploration exhaustive. Un data lake permet à des utilisateurs répartis sur plusieurs divisions d'affiner, d'explorer et d'enrichir les données selon leurs besoins.

Un accès flexible. Un data lake offre plusieurs solutions d'accès aux données à travers une infrastructure batch, interactive, online, search, in-memory and other processing engines.

Le résultat : Un data lake assure une évolutivité et une visibilité maximales avec le minimum d'interférences et au coût le plus faible possible.

Alors que les données continuent à croître exponentiellement, un investissement dans Enterprise Hadoop et dans un EDW peut alimenter une stratégie d'efficacité grâce à une architecture de données moderne et des opportunités issues du data lake d'entreprise.



Ill. 9
Une architecture de données moderne avec l'intégration d'Apache Hadoop dans les systèmes de données existants

Étude de cas 1 :

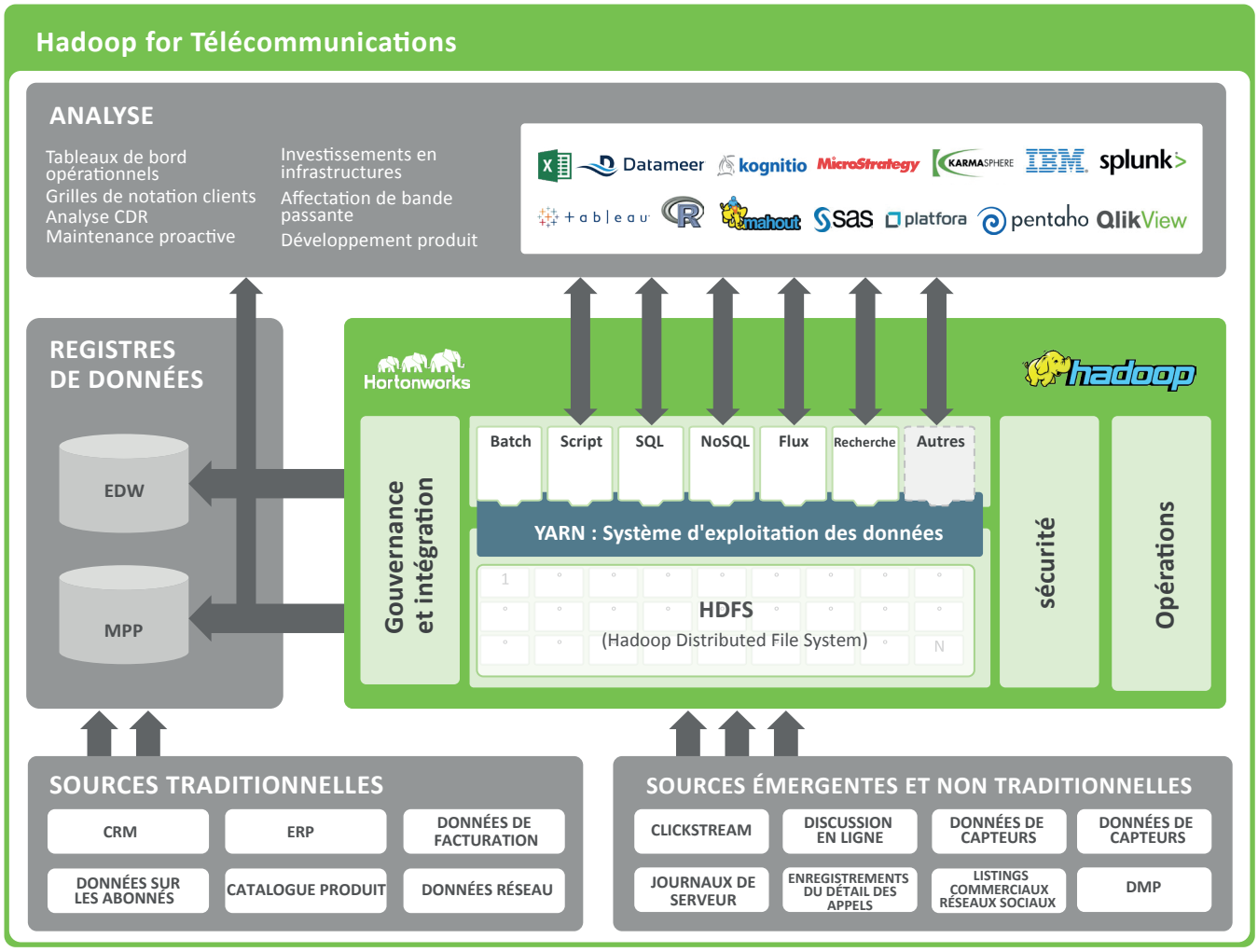
Une entreprise de télécommunications crée un profilage complet de ces clients

Pour l'industrie des télécommunications, un foyer comprend souvent plusieurs individus qui ont chacun signé un contrat pour différents types de produits auprès d'un fournisseur de services particulier et qui sont servis par différentes entités organisationnelles à travers le même prestataire.

Ces clients communiquent avec le prestataire par différents canaux en ligne et hors ligne pour des questions concernant les ventes ou le service et, ce faisant, s'attendent à ce que le fournisseur de services dispose de toutes les informations issues des différents points de contact.

Dans notre cas, pour une grande entreprise de télécommunications américaine, il devenait trop difficile d'absorber la croissance rapide du volume et du type de données clients qu'elle recevait et, en définitive, elle ne parvenait plus à avoir une vision cohérente des problèmes de ses clients. Les précieuses données clients étaient largement fragmentées, à la fois dans de multiples applications et dans différents centres de stockage de données, comme les EDW.

Apache Hadoop 2.0 a permis à ce fournisseur de services d'avoir une vision cohérente de tous les foyers qu'il servait en utilisant l'ensemble des différents canaux de données pour la transaction, l'interaction et l'observation, ce qui lui a fourni un profilage complet sans précédent de ses clients. De plus, Hadoop 2.0 a permis à ce fournisseur de créer, à un prix économique, un data lake de plusieurs pétaoctets pour l'ensemble de l'entreprise et de disposer de la vision nécessaire pour améliorer significativement le service aux clients.



Ill. 10

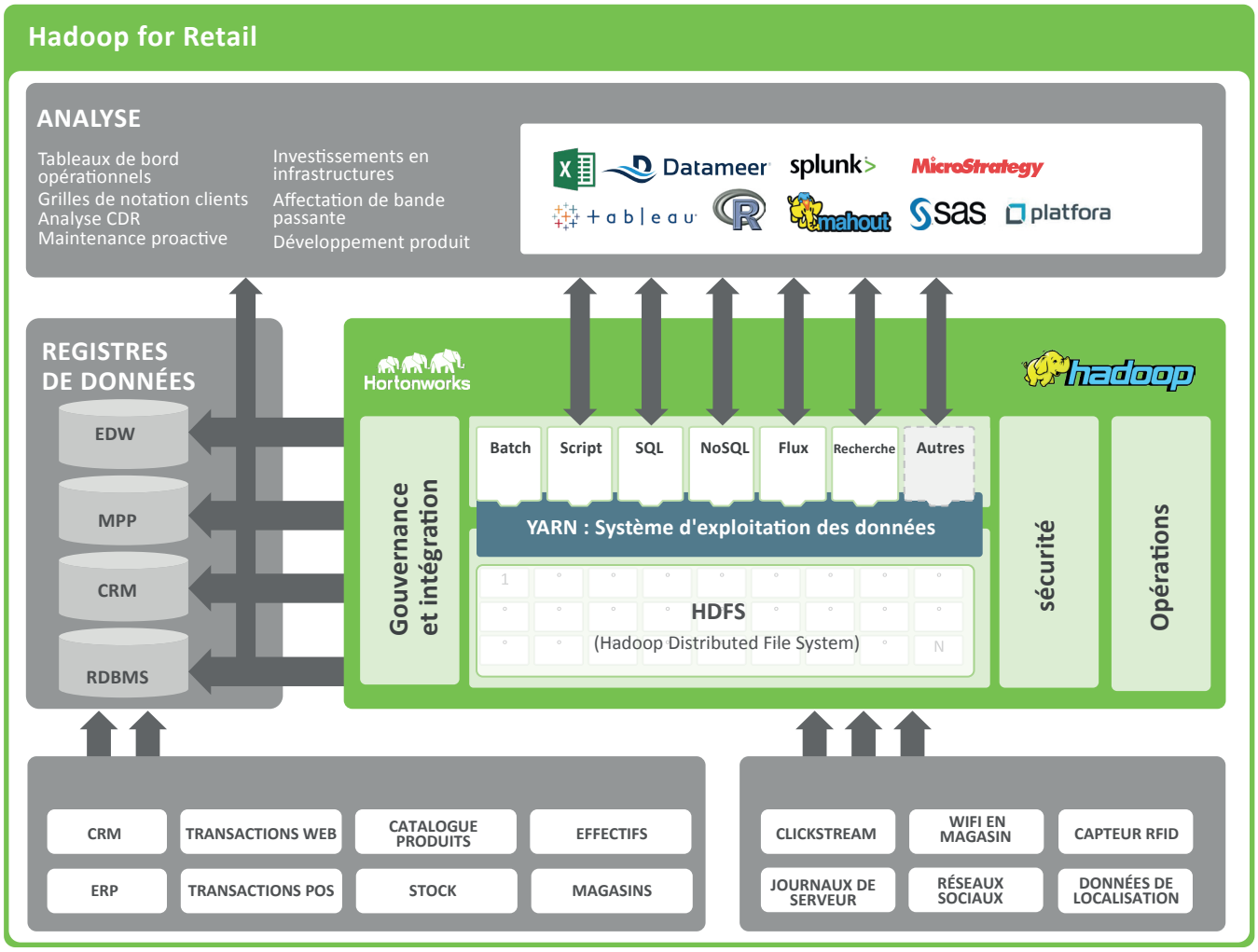
Étude de cas 2 :

Un détaillant de produits d'amélioration résidentielle renforce sa performance commerciale

Pour un détaillant important de produits d'amélioration résidentielle basé aux États-Unis qui dépensait annuellement plus d'un milliard de dollars sur son marché¬, il n'était pas facile de dépenser mieux et d'avoir des messages marketing plus pertinents vis-à-vis de ses clients, notamment parce que les solutions existantes répondaient très mal à ce besoin.

Bien que les 100 millions d'interactions annuelles entre ce détaillant et ses clients représentaient 74 milliards de dollars d'achats annuels, les données correspondantes restaient cloisonnées et empêchaient l'entreprise de faire un lien entre les données des transactions, les différentes campagnes de marketing et les comportements de navigation en ligne des clients. De plus, le rassemblement de ces données fragmentées et cloisonnées dans une structure de base de données relationnelle était considéré comme gourmand en temps, textrêmement onéreux et techniquement difficile. Ce dont ce détaillant important avait besoin était un « enregistrement en or » qui homogénéise les données

La plate-forme de données Hortonworks a permis de créer cet enregistrement, en fournissant des informations essentielles que l'équipe de marketing du détaillant a ensuite utilisées pour lancer des campagnes parfaitement ciblées vers les clients, celles-ci incluant des bons d'achat, des promotions et des e-mails personnalisés. Comme Hadoop 2.0 a été utilisé pour dimensionner correctement son entrepôt de données, l'entreprise a économisé des millions de dollars chaque année et, aujourd'hui encore, l'équipe de marketing découvre un usage inattendu et unique du profilage complet du comportement d'achat de ses clients.



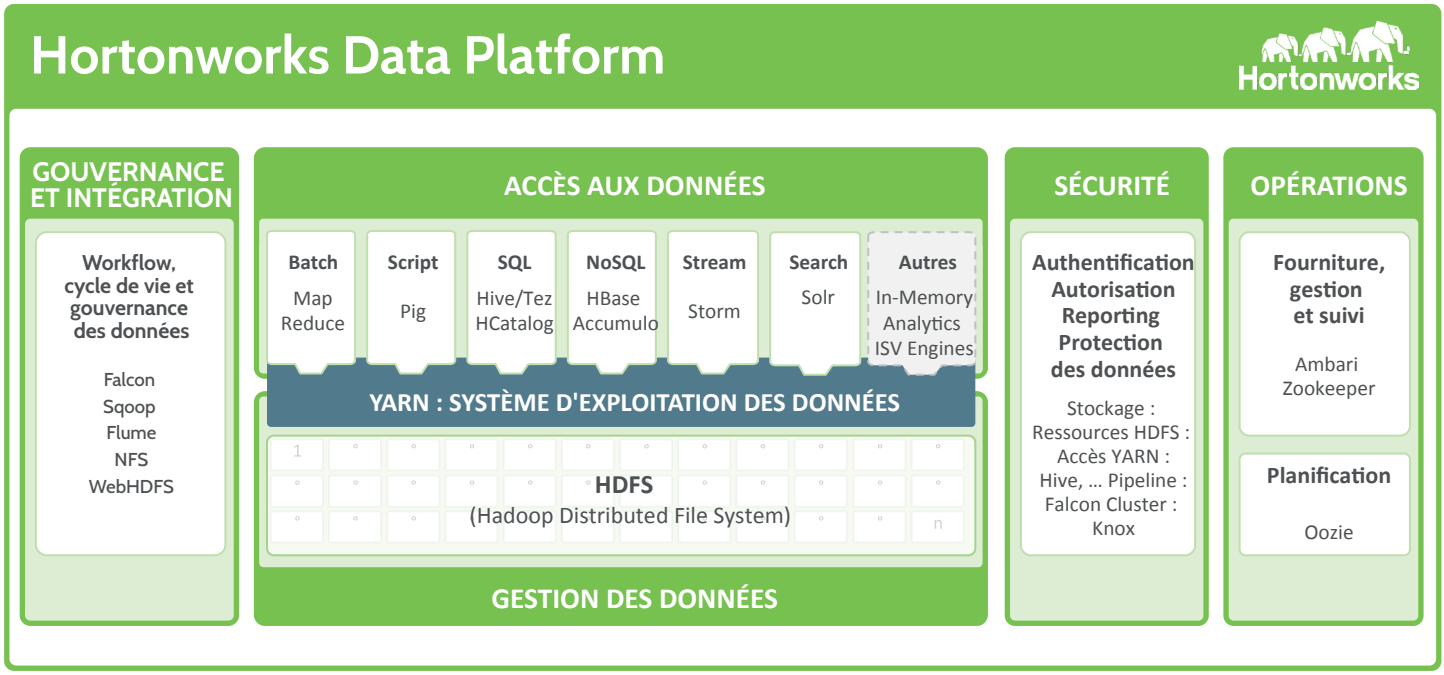
Ill. 11

Bâtir une architecture de données moderne avec Enterprise Hadoop

Pour obtenir un bon retour sur votre investissement dans le big data, planifiez l'intégration de Enterprise Hadoop avec votre EDW et les systèmes de données associés. En bâtissant une architecture de données moderne, votre entreprise pourra stocker et analyser de façon massive les données les plus importantes pour votre activité, extraire les informations critiques provenant de toutes les données et sources pour, en final, améliorer votre position concurrentielle sur le marché, fidéliser vos clients et maximiser votre chiffre d'affaires. Apprenez-en plus sur http://hortonworks.com/hdp

La plate-forme de données Hortonworks supporte Enterprise Hadoop

La plate-forme de données Hortonworks (HDP) est administrée à 100 % par la solution open source Apache Hadoop. HDP fournit tous les projets liés à Apache Hadoop qui sont nécessaires à l'intégration de Hadoop en parallèle avec un EDW dans le cadre d'une architecture de données moderne.



Ill. 12

HDP assure trois valeurs essentielles à toute entreprise :

Développement Totalement libre

HDP supporte Apache Hadoop pour les entreprises, avec un développement totalement libre et le recours à la meilleure expertise technologique

HDP incorpore les innovations communautaires les plus répandues ; elle est testée sur les suites Hadoop les plus avancées et sur des milliers de nœuds.

HDP est développée et supportée par des ingénieurs ayant la connaissance la plus approfondie d'Apache Hadoop.

Fondamentalement Polyvalente

HDP est conçue comme une plate-forme unique qui répond aux besoins nouveaux du traitement des données volumineuses, tout en assurant une plate-forme complète pour la gouvernance, la sécurité et les opérations.

HDP supporte tous les scénarios de big data : batch, interactif, temps réel et streaming.

HDP offre une couche polyvalente d'accès aux données grâce à YARN qui est au cœur d'Enterprise Hadoop, et elle permet ainsi aux nouveaux moteurs de traitement d'être intégrés lorsqu'ils sont prêts à être utilisés par l'entreprise.

HDP assure toutes les capacités de sécurité, de gouvernance et opérationnelles pour la mise en œuvre de Hadoop dans l'entreprise.

Totalement Intégrée

HDP est conçue pour fonctionner avec n'importe quel centre de données et s'intègre dans tout système existant.

HDP peut être déployée avec tout type de scénario : de Linux à Windows, sur site ou dans le cloud.

HDP s'intègre parfaitement avec les principales plates-formes des fournisseurs de technologie : Red Hat, Microsoft, SAP, Teradata et autres.

Options de déploiement pour Hadoop

HDP offre de multiples options de déploiement :

Sur site. HDP est la seule plate-forme Hadoop qui fonctionne avec Linux et Windows.

Dans le cloud. HDP peut fonctionner dans le cadre d'une IaaS (infrastructure en tant que service) et supporte aussi Big Data Cloud de Rackspace, HDInsight Service de Microsoft, CSC et bien d'autres.

Appliance.HDP fonctionne par défaut sur du matériel standard et peut également être achetée en en temps qu'appliance auprès de Teradata.

Components of Enterprise Hadoop

Découvrez-en davantage sur les composants individuels de Enterprise Hadoop.

Gestion des données

hdfs
yarn

Accès aux données

mapreduce
pig
hive
tez
hbase
accumulo
storm
hcatalog

Gouvernance et intégration des données

falcon
flume

Sécurité

knox
security

Opérations

ambari

Pourquoi Hortonworks pour Hadoop ?

Fondée en 2011 par 24 ingénieurs venant de la société Yahoo! d'origine constituée d'équipes de développement et d'exploitation Hadoop, Hortonworks est l'entreprise qui regroupe le maximum d'expérience sur Hadoop. Nos collaborateurs sont des participants et des leaders actifs dans le développement de Hadoop : ils conçoivent, bâtissent et testent le cœur de la plate-forme Hadoop. Nous avons cumulé des années d'expérience opérationnelle sur Hadoop et nous sommes les mieux armés pour supporter votre projet critique Hadoop. Apprenez-en plus sur http://hortonworks.com/why

Leadership ouvert

Hortonworks a un projet et un engagement particuliers pour conduire l'innovation exclusivement dans l'environnement libre via le processus de la Fondation Apache Software.

Hortonworks est le concepteur des principales avancées de base sur les codes sources qui permettent à Apache Hadoop de constituer une véritable plate-forme de données d'entreprise.

Approbation de l'écosystème

Hortonworks se consacre à l'intégration totale de Hadoop avec les technologies existantes de centres de données et l'expertise des équipes.

Hortonworks a établi des relations stratégiques durables avec des partenaires renommés en centres de données, notamment Microsoft, SAP, Teradata, Rackspace, et bien d'autres. Hortonworks offre, en complément de sa vaste

Rigueur d'entreprise

Expérience sur les déploiements importants de Hadoop, un support aux entreprises et une organisation de services de classe mondiale.

Hortonworks développe et certifie Apache Hadoop dans l'objectif de servir les entreprises, et elle réalise ses tests avec rigueur sur des applications réelles dans les clusters Hadoop les plus importants du monde.

Pour obtenir une analyse indépendante de la plate-forme de données Hortonworks, vous pouvez télécharger le rapport *Forrester Wave™: Big Data Hadoop Solutions, Q1 2014* de Forrester Research.

À propos de Hortonworks

Hortonworks développe, distribue et supporte la seule plate-forme de données Apache Hadoop 100 % open source. Notre équipe comprend les plus importants effectifs de concepteurs et d'architectes de l'écosystème Hadoop : au sein de cette communauté, ils représentent et pilotent les exigences de l'entreprise au sens large. La plate-forme de données Hortonworks supporte une plate-forme libre qui s'intègre parfaitement avec les investissements informatiques existants et sur laquelle les entreprises peuvent bâtir et déployer des applications basées sur Hadoop. Hortonworks a des relations approfondies avec les principaux fournisseurs de centres de données stratégiques, ce qui permet à nos clients de tirer le meilleur profit de Hadoop.
Pour plus d'informations, consultez www.hortonworks.com.