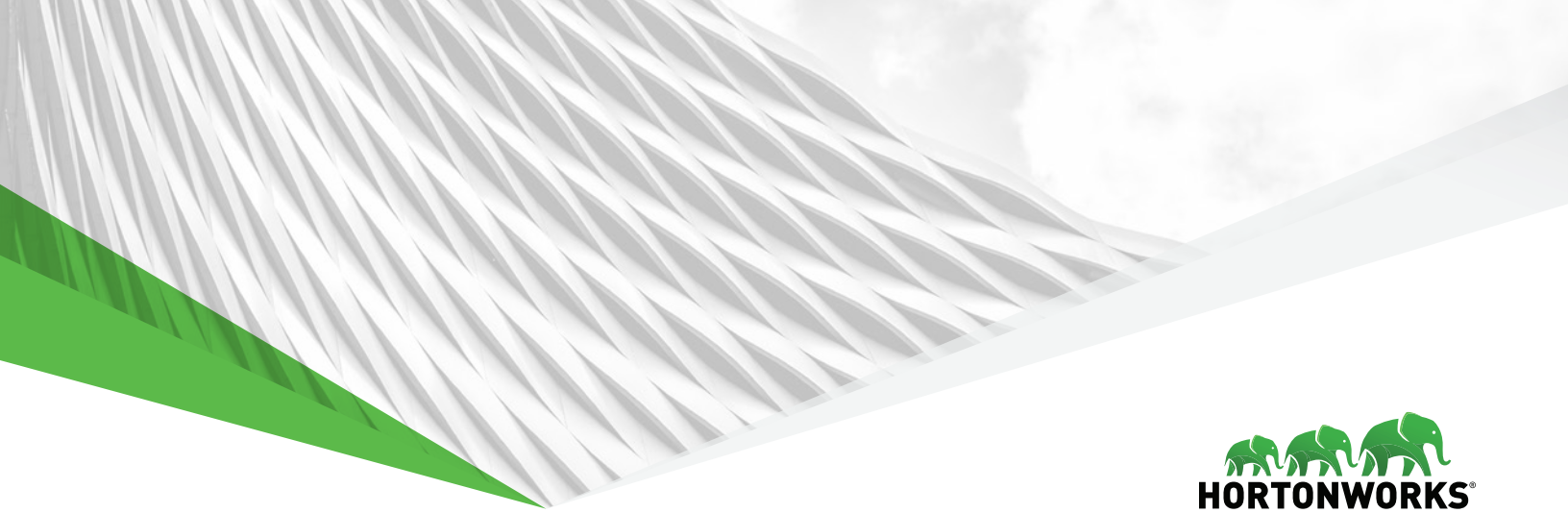

# QUATRE PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES MODERNE ET DE L'APPRENTISSAGE AUTOMATISÉ

LIVRE BLANC HORTONWORKS

MARS 2017

# Sommaire


| | |
|---|---|
| Sommaire | 2 |
| Introduction | 3 |
| Le Data Lake d'entreprise pour une plus grande réussite des initiatives relatives à la science des données | 4 |
| Fournir un environnement d'exploration de données sécurisé et riche pour les atouts de l'entreprise dans le Data Lake | 5 |
| Tirer parti d'une plate-forme flexible compatible avec les techniques émergentes et existantes | 6 |
| Accélérer les initiatives relatives à la science des données grâce à une plus grande autonomie et collaboration | 7 |
| Mettre en œuvre les initiatives relatives à la science des données avec une gestion du modèle tout au long de son cycle de vie | 8 |
| Conclusion | 9 |

# Introduction

Tandis que le nombre de données continue d'augmenter de façon exponentielle à l'échelle mondiale, les technologies du big data et les pratiques associées évoluent rapidement. À l'origine, les entreprises se concentraient sur la recherche d'une méthode rentable et évolutive pour stocker et gérer ces données. Grâce à des technologies telles qu'Apache Hadoop®, les entreprises pouvaient stocker des données structurées et non structurées dans un seul Data Lake tout en réduisant les coûts opérationnels des entrepôts et magasins de données. À l'heure actuelle, tandis qu'elles prennent conscience du potentiel stratégique de leurs ressources d'informations, les entreprises s'efforcent d'analyser toutes ces nouvelles données via de nouvelles méthodes ; c'est-à-dire qu'elles ne visent plus simplement à diminuer les coûts mais à mettre en place des initiatives relatives au big data qui leur permettent également d'accroître leur avantage compétitif, d'augmenter leur chiffre d'affaires et d'améliorer leur rentabilité.

Les entreprises cherchent maintenant à passer de l'analytique descriptive, qui permet d'expliquer ce qu'il s'est passé, à l'analytique prédictive, qui permet d'anticiper en prédisant ce qu'il va se passer. Les entreprises de média et de télécommunication veulent par exemple maintenant aller au-delà de l'analyse des tendances de perte de clients et être capable de prédire lorsqu'un client est sur le point de partir afin de pouvoir lui faire une offre en temps réel pour le retenir. Ce besoin entraîne l'essor de la science des données comme principal vecteur de la valeur prédictive basée sur des données. En combinant les mathématiques, les statistiques, la recherche opérationnelle et l'informatique, la science des données offre aux entreprises la possibilité de libérer le potentiel des données stockées dans leurs Data Lakes par le biais de méthodes telles que l'analyse prédictive, l'apprentissage automatisé et l'exploration de données.

# Le Data Lake d'entreprise pour une plus grande réussite des initiatives relatives à la science des données

Le Data Lake d'entreprise est une progression naturelle du passage au big data. En règle générale, une entreprise démarre cette transformation par le développement de nouvelles applications analytiques pour certains secteurs d'activité. Au fur et à mesure que de nouvelles applications sont créées et génèrent de la valeur à partir de nouveaux types de données tels que les capteurs, les machines, les journaux de serveurs, les parcours de navigation et autres sources de différents secteurs, le Data Lake prend forme. Il représente le référentiel central sécurisé des données d'entreprise avec des services partagés permettant d'obtenir des informations poussées grâce à plusieurs ensembles de données et applications de diverses natures. Si cette fonction est utilisée de façon rentable et à l'échelle, le Data Lake permet aux entreprises de générer davantage de valeur à partir de leurs données. En collaborant avec des clients et des utilisateurs à toutes les étapes du passage au big data, nous avons constaté que le concept du Data Lake d'entreprise avait une forte résonance en raison des avantages qu'il procure. L'écosystème d'Apache Hadoop de logiciels open source est en train de devenir la norme du secteur pour créer 'un Data Lake d'entreprise avec HDFS pour le stockage de données et YARN pour le traitement de données. En permettant un accès plus rapide à un plus grand nombre et une plus grande variété de données, le Data Lake d'entreprise peut également être à la base des initiatives relatives à la science de données au sein de l'entreprise.

Dans ce livre blanc, nous présentons l'approche d'Hortonworks des initiatives relatives à la science des données au sein d'une organisation. Selon nous, les entreprises peuvent grandement améliorer le taux de réussite de ces initiatives et en tirer le plus de valeur possible en suivant ces quatre recommandations :

- Fournir un environnement d'exploration de données sécurisé et riche pour les atouts de l'entreprise dans le Data Lake
- Tirer parti d'une plate-forme flexible compatible avec les techniques émergentes et existantes
- Accélérer les initiatives en matière de science des données grâce à une plus grande autonomie et collaboration
- Mettre en œuvre les initiatives relatives à la science des données avec une gestion du modèle tout au long de son cycle de vie.

# Fournir un environnement d'exploration de données sécurisé et riche pour les atouts de l'entreprise dans le Data Lake

Le Data Lake basé sur Hadoop représente un changement de paradigme pour les initiatives relatives à l'apprentissage automatisé et à la science des données dans l'entreprise. Les économies d'échelle permises par le Data Lake permettent de disposer d'historiques plus longs et que les données brutes stockées et traitées à des fins analytiques et d'apprentissage automatisé soient totalement fidèles. L'augmentation du volume des diverses données se traduit généralement par une amélioration des capacités prédictives des pipelines existantes relatives à la science des données qui, historiquement, utilisent des données sur des périodes plus courtes et seulement un échantillon des données disponibles pour créer des modèles prédictifs. Des données couvrant des historiques plus longs de comportements clients pourraient contribuer à atténuer les effets saisonniers dans le cadre de la prédiction des pertes de client en permettant aux scientifiques des données de former leurs modèles sur la période prenant en compte cette fluctuation.

Le Data Lake permet également aux organisations de créer leurs applications en utilisant la quasi-totalité des données disponibles en cassant les silos existants. En utilisant une plus grande variété de données pouvant être comparées et recoupées, les entreprises peuvent maintenant obtenir de nouvelles informations et trouver des réponses à des questions complexes que personne ne savait comment poser ou auxquelles personne ne savait comment répondre. Dans un contexte de perte de clients, les signaux sociaux des données marketing peuvent être combinés avec des journaux de données de l'assistance client afin d'améliorer l'efficacité de la prédiction de perte de clients.

Les processus relatifs à la science des données débutent souvent par une exploration des données afin de savoir quelles données sont disponibles pour créer le modèle et d'avoir une meilleure idée des métadonnées de ses contenus, par exemple les types de colonnes et les résumés statistiques. Le lignage des données joue également un rôle essentiel pour les scientifiques des données en leur fournissant des informations sur quand et comment un ensemble de données a été produit et de quelle manière il est utilisé. Apache Atlas, dans le cadre de la plate-forme de données Hortonworks Data Platform (HDP®), peut fournir ces informations aux scientifiques des données et les aider à identifier les ensembles de données adéquats pour leurs projets et à savoir les utiliser. Les informations relatives aux lignages stockées dans Atlas permettent non seulement de voir les modèles qui seraient affectés par une modification des données en amont mais également d'identifier les modèles existants basés sur des ensembles de données pouvant être utilisés comme point de départ de nouveaux projets relatifs à la science des données.

L'accès aux données contenues dans le Data Lake doit toujours être autorisé de façon sécurisée et audité par une couche de gouvernance des données partagées afin d'éviter les fuites de données sensibles. Dans HDP, Apache Ranger offre un contrôle d'accès centralisé et un audit pour tous les actifs de données ainsi que d'autres fonctionnalités telles que le masquage de données contextuelles afin de permettre l'accès personnalisé à des données sensibles telles que les numéros de sécurité sociale ou de téléphone.

HDP offre le Data Lake fondamental sur lequel peut reposer la prochaine génération d'applications avancées relatives à la science des données. Grâce à la sécurité centralisée et l'audit via Ranger et la gestion des métadonnées et des données partagées à l'aide d'Atlas, les actifs du Data Lake peuvent être utilisés de façon sécurisée et efficace par les scientifiques des données.

# Tirer parti d'une plate-forme flexible compatible avec les techniques émergentes et existantes

La science des données et l'apprentissage automatisé est une discipline très ancienne mais qui évolue rapidement et avec une riche histoire de kits d'outils existants tels que R, SciKit-Learn ainsi que de techniques émergentes telles que TensorFlow™ et MxNet pour un apprentissage approfondi. Afin de tirer parti de ce kit d'outils, les entreprises ont besoin d'une plate-forme informatique suffisamment flexible pour faire fonctionner les technologies existantes et émergentes dans le Data Lake. YARN, dans le cadre d'Apache Hadoop, et les projets open source tels qu'Apache Slider et Apache Spark™ offrent une telle plate-forme permettant d'intégrer les systèmes d'apprentissage automatisé au Data Lake. Dans le cadre de cette approche, TensorFlow peut être déployé de manière native via Apache Slider et R peut être intégré à l'aide de SparkR. La dynamique communauté contribuant à ces projets open source Apache et leur importance dans l'écosystème du big data garantissent que les projets importants tels que SciKit ainsi que les nouvelles technologies d'apprentissage automatisé telles que TensorFlow continueront d'être intégrés à YARN ou Spark.

Avant qu'un scientifique des données ne puisse entreprendre de mettre au point un algorithme d'apprentissage automatisé, il/elle doit préparer des données, ce qui passe notamment par le nettoyage, la standardisation et la conversion du format des données. Cette transformation de données est une étape essentielle du processus de création d'un modèle. Les données peuvent être à différents formats tels qu'au format ORC, Parquet ou texte et peuvent être traités de différentes façons telles que par lots, en temps réel, en streaming, en mémoire, etc. Il est donc essentiel pour les scientifiques des données d'avoir accès aux meilleurs moteurs de traitement des données tels qu'Apache Hive™ ou Spark afin de transformer ces données de façon efficace. Avec tous les principaux moteurs de traitement des big data intégrés à YARN dans le cadre de HDP, cela offre la plate-forme idéale pour nettoyer, standardiser et intégrer les big data avant de créer les algorithmes d'apprentissage automatisé. Avec le soutien récent des technologies logicielles telles que Docker et des matériels tels que les unités de traitement graphique (GPU), YARN constitue la plate-forme de gestion informatique offrant la flexibilité et le support des écosystèmes nécessaires pour une valeur temps efficace pour les initiatives relatives à la science des données.

Il va sans dire que la science des données, ce n'est pas seulement le big data mais également le big compute car les techniques d'apprentissage automatisé sont gourmandes en calculs. Les clusters des Data Lakes offrent non seulement d'importantes capacités de stockage mais également d'importantes capacités de calcul. YARN offre une plate-forme efficace permettant d'utiliser et de partager de façon optimale cette capacité de calcul dans le cadre de la science des données et dans les autres équipes. De plus, l'intégration des kits d'outils relatifs à la science des données à YARN pourrait offrir la possibilité de paralléliser des actions séquentielles telles que le balayage des paramètres ou la sélection des algorithmes. En tirant profit des calculs en parallèle via YARN, le processus d'apprentissage automatisé pourrait être révolutionné.

Nous pensons qu'une plate-forme de données moderne telle qu'HDP permet aux scientifiques des données de travailler sur les logiciels et les matériels de leur choix avec un minimum d'effort. La tâche de création de la plate-forme de big data prise en charge par l'écosystème d'Hadoop, les scientifiques de données peuvent se concentrer de façon plus productive sur la création d'applications prédictives. HDP avec YARN, Spark et Hive en plus d'autres moteurs de données tels qu'Apache Storm offrent une plate-forme complète qui permet aux scientifiques des données de tirer profit de la totalité des outils, bibliothèques, matériels et données à leur disposition.

# Accélérer les initiatives relatives à la science des données grâce à une plus grande autonomie et collaboration

La science des données est par nature un processus exploratoire et créatif car il n'existe en règle générale aucune réponse définitive à un problème ni une approche unique pour atteindre un objectif. Les scientifiques des données explorent les données, visualisent les tendances et utilisent leur expérience et jugement pour choisir des paramètres et processus pouvant être pertinents par rapport au problème devant être résolu. Cela rend le partage et la collaboration importants afin que les équipes de scientifiques des données puissent s'appuyer sur les connaissances de chacun pour atteindre les meilleurs résultats possible.

Tandis que la science des données évolue avec le big data, de nouvelles techniques et technologies émergent. Cette évolution se reflète dans la formation des scientifiques des données dans les organisations. Il existe une vaste gamme de langages et de kits d'outils (R et Python), de logiciels commerciaux (SAS et SPSS) et de technologies open source telles que Spark auxquelles les scientifiques des données peuvent avoir été formés. Pour que les initiatives relatives à la science des données réussissent, les entreprises doivent permettre aux scientifiques des données de travailler efficacement, sans les limiter en fonction leur formation, et d'utiliser la meilleure technique ou technologie possible pour traiter le problème en question. À cet égard, les blocs-notes relatifs à la science des données tels qu'Apache Zeppelin apportent une importante valeur ajoutée. En raison de sa conception extensible, Zeppelin peut facilement être connecté à différents types de plateformes. La visualisation universelle de Zeppelin lui permet également de s'intégrer à des bibliothèques de pointe dont matplotlib, ggplot et Bokeh, entre autres.

À l'heure actuelle, les scientifiques des données passent beaucoup de temps à créer leur environnement analytique. Ce processus consiste à identifier les données, déplacer les données de différentes sources vers leur environnement et ensuite y effectuer les expérimentations. Apache Zeppelin permet aux scientifiques des données de se connecter directement aux sources des données dans le Data Lake avec une préparation minimale. Une fois connectés aux sources des données, les scientifiques peuvent simplement utiliser Zeppelin pour exploiter la puissance de traitement du cluster en utilisant le meilleur support disponible pour Spark ou les technologies d'apprentissage automatisé de leur choix.

Les blocs-notes offrent également un environnement collaboratif agrémenté par des visualisations dans lequel les scientifiques des données experts dans différentes techniques et technologies peuvent partager leurs résultats. Les équipes peuvent non seulement partager leur code mais également regrouper la totalité du bloc-note Zeppelin dans un seul environnement reproductible afin que les autres puissent commencer rapidement sans avoir à effectuer de préparatifs supplémentaires. L'intégration de première qualité de Zeppelin à Apache Spark et Livy dans HDP permet aux scientifiques des données de non seulement partager le code mais également des ensembles de données en temps réel grâce aux fonctions de génération de rapports de Zeppelin. Ces paradigmes collaboratifs avancés supportés par Zeppelin favorisent non seulement l'apprentissage et la pollinisation croisée d'idées mais permettent également aux équipes avec différentes expertises de travailler conjointement sur un modèle prédictif et par conséquent d'améliorer sa solidité et son pouvoir prédictif.

HDP ainsi que les projets d'écosystèmes tels que Spark, Livy et Zeppelin peuvent rendre les données et les puissantes capacités de calcul facilement accessibles aux scientifiques des données. Ces capacités permettent le partage de code, d'ensembles de données et de résultats intermédiaires avec d'autres membres d'équipes à des fins de développement et d'amélioration collaboratifs. Tous ces facteurs combinés permettent de rendre les scientifiques autonomes, d'améliorer l'efficacité de leurs modèles et, plus important encore, d'obtenir des informations plus rapidement.

# Mettre en œuvre les initiatives relatives à la science des données avec une gestion du modèle tout au long de son cycle de vie

Les organisations récoltent véritablement les bénéfices de leur investissement dans un modèle lorsque ce dernier est mis en pratique dans le cadre d'un processus commercial. Les modèles sont intégrés aux systèmes de production en les incorporant à des applications en ligne ou en les utilisant hors ligne dans le cadre d'analyses par lots. Ces applications tirent parti de différents flux de données en tant qu'informations permettant de réaliser des prédictions ou de prendre des décisions telles qu'effectuer des recommandations de produits ou signaler des transactions de cartes de crédit frauduleuses, pour ne citer que quelques exemples.

Le déploiement d'un modèle en production inclut la surveillance de ses performances, ses défaillances et ses temps d'arrêt. Les modèles doivent également être étalonnés afin de faire face à un plus grand nombre de données ou de répondre aux besoins de cotation plus importants des applications commerciales. Il est possible que les modèles doivent être retirés ou mis à niveau après un certain temps. À l'heure actuelle, le déploiement d'un modèle en production est une tâche ennuyeuse souvent réalisée ponctuellement et reposant essentiellement sur les équipes informatiques, ce qui peut avoir pour conséquence des interruptions de projets critiques. Nous pensons qu'une plate-forme de gestion de modèles est nécessaire afin d'automatiser la plupart de ces tâches de façon à ce que les principaux aspects des déploiements en production soient solides, fiables et reproductibles tout en réduisant la charge des tâches ennuyeuses pesant sur les scientifiques des données.

Une fois qu'un modèle d'apprentissage automatisé est prêt, il peut être publié en tant que Model-as-a-Service (MaaS) pouvant automatiser ou semi-automatiser plusieurs tâches de gestion du cycle de vie du modèle. Les modèles pourraient être rendus opérationnels dans un API REST pour une intégration facilitée aux applications en ligne. À l'aide de l'intégration au YARN, le service API pourrait être étalonné en fonction de la demande. Le service pourrait également mesurer les performances d'un nouveau modèle et le mettre en place étape par étape ou revenir à un ancien modèle si les mesures définies ne sont pas atteintes. Les défaillances pourraient être automatiquement traitées en redémarrant les serveurs API. En les intégrant aux plateformes de lignage telles qu'Apache Atlas, les MaaS pourraient offrir la version, la reformation et mettre à jour l'historique des modèles à des fins de réutilisation et de suivi.

# Conclusion

Dernière évolution de la technologie et de la stratégie du big data, la science des données joue déjà un rôle essentiel afin de libérer le potentiel des données d'entreprise, et elle va continuer à prendre de l'importance. Il est essentiel de trouver la bonne approche en matière d'architecture et d'outils relatifs au big data afin de permettre aux scientifiques des données de tirer le plus de valeur possible des données de façon à accroître l'avantage compétitif, d'augmenter le chiffre d'affaires et d'améliorer la rentabilité de leur entreprise. En permettant aux initiatives relatives à la science des données de reposer sur un Data Lake d'entreprise, les organisations peuvent casser les silos et ainsi utiliser l'ensemble de leurs données tout en maintenant un contrôle et un suivi centralisés par le biais d'Apache Ranger et Atlas. En fournissant une plate-forme destinée à la science des données flexible basée sur YARN et Spark, les entreprises peuvent permettre aux scientifiques des données de se concentrer sur la création d'applications prédictives plutôt que de devoir recommencer chaque initiative à zéro, et tirer parti de la puissance collective de tous les outils et données à leur disposition. Apache Zeppelin peut aider les scientifiques des données de tous les horizons à collaborer, à partager des ressources et à devenir plus autonomes et ainsi obtenir plus rapidement des informations. Une approche sous forme de Model-as-a-service basée sur une plate-forme peut faciliter la mise en œuvre de modèles pour un véritable impact commercial. Avec cette solide base flexible et évolutive en place, l'organisation est prête à maximiser la valeur de ses initiatives relatives au big data aujourd'hui et à l'avenir.

# À propos d'Hortonworks

Hortonworks est une société innovante, chef de file de son secteur, spécialisée dans la création, la distribution et l'assistance dédiées aux plateformes de données connectées et ouvertes, et d'applications de données modernes destinées aux entreprises qui fournissent des informations exploitables issues de toutes les données possibles : données au repos et en mouvement. La société Hortonworks est résolument tournée vers l'innovation au sein de communautés open source telles qu'Apache Hadoop, Apache NiFi et Apache Spark. Avec plus de 1 800 partenaires, Hortonworks offre une expertise, une formation et des services permettant à nos clients d'exploiter la valeur transformationnelle de leurs organisations, quel que soit leur secteur d'activité.

**Contact**

Pour en savoir plus, rendez-vous sur www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
Numéro international: +44 (0) 20 3826 1405

HORTONWORKS