POURQUOI UNE STRATÉGIE CONNECTÉE EST INDISPENSABLE POUR ASSURER LA PÉRENNITÉ DE VOS DONNÉES ?

Un livre blanc sur l'avenir de la Data pour les DSI, Directeurs Technique, experts IT et Data

UN LIVRE BLANC HORTONWORKS
MARS 2016

Sommaire

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Abstract

L'avènement du big data a révolutionné le secteur de l'analytique et de la data science et a donné le jour au concept de nouvelles plateformes de données, permettant aux entreprises de stocker, d'analyser et d'accéder à de gros volumes de données historiques. Le monde du big data était né. Mais les plateformes de données doivent encore évoluer pour faire face au tsunami de données produites en temps réel par l'internet of Anything (IoAT).

Pour suivre l'évolution des consommateurs et mettre en place des actions toujours plus personnalisées et efficaces en termes de marketing, de vente ou de support client, les entreprises doivent aller au-delà de la compréhension du passé et se focaliser sur les événements à venir. Nous avons besoin d'une nouvelle approche pour tirer le meilleur parti de nos données. Nous devons améliorer notre capacité à intégrer le tsunami de données dynamiques dans une plateforme de données tout en analysant les tendances dégagées en temps réel. Nous avons besoin d'une stratégie de données au service de l'entreprise et prête à relever les défis futurs. Ce livre blanc présente les pré-requis et les éléments clés pour déployer une plateforme de données connectées capable de gérer à la fois les données statiques et les données dynamiques grâce à une stratégie flexible et interactive.

L'évolution des plateformes de données

Depuis seulement quelques années, les entreprises ont commencé à mettre en place des entrepôts de données pour mieux exploiter et comprendre les données historiques issues des logiciels ERP, CRM et des systèmes de stockage de données. Cette démarche s'est rapidement avérée trop coûteuse, laborieuse à mettre en place et pas suffisamment agile pour gérer les nouvelles données structurées ou non structurées telles que les fichiers log, les données relatives aux volumes de clics ou issues des réseaux sociaux. Le Big Data et Apache Hadoop sont nés pour apporter une réponse à ces problématiques et les premières plateformes de données ont vu le jour dans la foulée.

La capacité de la plateforme de données à stocker et analyser la data à un faible coût, ainsi que la possibilité de traiter tous les types de données a laissé entrevoir aux organisations les nouvelles chaînes de valeur potentiellement exploitables — par exemple l'amélioration de la qualité et de la rapidité de la recherche web, l'exploitation plus rationnelle de la publicité online ou l'analyse des interactions clients et des comportements d'achat. Ces nouvelles opportunités ont suscité des réflexions sur les possibilités offertes par les données.

Cette chaîne de valeur est bel est bien réelle et continue à se développer. Selon le cabinet McKinsey[1], les entreprises qui font partie du classement « Fortune 1000 » (liste des 1000 plus grandes entreprises américaines classées par chiffre d'affaires) et qui se sont engagées dans une démarche de digitalisation peuvent multiplier leur activité par cinq. Si vous faites de la vente en ligne, vous pouvez augmenter votre productivité de 15%. La digitalisation de votre stratégie d'entreprise peut vous permettre de multiplier vos marges par trois. Ces performances ne seraient pas envisageables sans Apache Hadoop et la mise en place d'un data lake grâce à une plateforme capable de gérer dans un espace de stockage unique de gros volumes de données, en y associant des services essentiels pour l'entreprise tels que la sécurité ou le management opérationnel.

1. http://www.mckinsey.com/business-functions/business-technology/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

Le tsunami de l'Internet of Anything reste encore à venir

Dans le passé, le volume de données à l'échelle mondiale doublait chaque siècle. Désormais, il double tous les deux ans. Cette croissance est accélérée par l'internet des objets qui génère toujours plus de données via Internet, les appareils mobiles, les logs de connexion aux serveurs, la géolocalisation, les capteurs intelligents et autres systèmes embarqués. L'influence des objets connectés, des capteurs sensoriels ou de l'intelligence artificielle ne cesse de s'amplifier d'année en année. Ce phénomène a poussé les plateformes de données à évoluer afin de pouvoir gérer ces données dynamiques.

Aujourd'hui, nous dépassons toutes les prédictions émises il y a seulement quelques années sur la croissance du volume de données créées, stockées, traitées et analysées. Avec la 5G, un smartphone peut potentiellement délivrer un volume impressionnant de 1Tbps[2] de données ! Dans deux ans, il y aura plus d'appareils mobiles que d'habitants sur la planète. A l'heure actuelle, 35% des américains possèdent au moins un appareil connecté autre qu'un téléphone, comme par exemple un thermostat, un réfrigérateur ou une montre — ces appareils générant également des données.[3] On prévoit que le nombre d'objets connectés atteindra 6,4 milliards d'ici 2020 et 21 milliards pour les appareils mobiles.

La déferlante des données et leur capacité à se multiplier de façon exponentielle nous conduit en envisager un univers digital qui passera de 4 zettabytes de données à 44 zettabytes au cours de ce siècle. 1,7 megabytes d'informations nouvelles vont voir le jour chaque seconde pour chaque humain présent sur cette planète, un tiers de ces informations passant par le cloud.[4]

Les types de données sont également en constante évolution. Evidemment, il ne s'agit plus de données simplement classées dans des lignes et des colonnes, mais des images, toutes sortes de données en streaming, des coordonnées géospatiales et des séries chronologiques. Le million et demi de membres actifs sur Facebook représente plus de 140 milliards de connexions potentielles, 265 milliards de photos téléchargées, 62 millions de titres audio écoutés 22 milliards de fois.[5]

Le cabinet Gartner mentionne que 32% des entreprises qui ont réalisé leur transformation digitale se définissent désormais comme des « entreprises digitales ».[6] Nous avons toutes les raisons de penser que ces entreprises digitales feront de la data leur atout le plus précieux.

2. http://www.trustedreviews.com/news/5g-researchers-crack-1tbps-data-transfer-at-uk-university
3. https://www.truste.com/about-truste/press-room/35-of-americans-now-own-at-least-one-smart-device-other-than-a-phone/
4. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
5. http://www.ge.com/docs/chapters/Industrial_Internet.pdf
6. http://www.gartner.com/technology/research/digital-business/

Selon Forbes :

- 59% des entreprises considèrent la data et l'analytique comme des ressources « vitales » pour leur bon fonctionnement, et 29% les qualifient de « très importantes »
- 69% évoquent de nouvelles pistes pour exploiter les données et utiliser leur valeur dans le cadre de projet d'entreprise
- 83% déclarent que les données permettent de rendre les produits et services proposés plus rentables.
- 60% déclarent que leurs données génèrent des revenus au sein de l'organisation

Mais 48% ont le sentiment que leur organisation a, dans la passé, échoué à tirer parti des opportunités d'exploiter leurs données. La révolution Internet ouvre de nouvelles perspectives pour optimiser la productivité, réduire les tâches sans valeur ajoutée, et améliorer le travail et l'expérience de vie humaine. Quel que soit le nom qu'on lui donne, la phase que nous vivons actuellement accélère ce phénomène. S'appuyer sur des technologies disparates pour apporter une solution à chaque problème, individuellement, n'est pas la solution ; la clé se trouve dans des technologies capables de faire interagir les données statiques et les données dynamiques afin de partager les mêmes opérations ainsi que les mêmes règles de gouvernance et de sécurité. C'est précisément cette démarche qui constitue le cœur des plateformes de données connectées.

Les plateformes de données doivent évoluer

Cette déferlante des données ne fait qu'augmenter et les plateformes de données doivent évoluer pour répondre aux nouveaux besoins. Une plateforme de données moderne ne peut plus se contenter de gérer uniquement les données statiques et les traitements par lots (batch processing). Elle doit être connectée à l'internet des objets.

Les plateformes de données connectées doivent également être capables de gérer des données statiques et dynamiques réparties entre différents départements, serveurs et emplacements physiques, le tout dans un environnement sécurisé, en maîtrisant les coûts et en prenant en compte les aspects de bande passante et de connectivité.

Les plateformes de données connectées ne doivent pas simplement se reposer sur Apache Hadoop. Elles ont besoin de renforcer leurs capacités en termes de routage des données, de transformation, et dans la mise en place de systèmes logiques de médiation tels que nous les voyons émerger de projets comme Apache NiFi[7] — qui prend en charge le traitement temps réel, la détection de pattern, le routage et l'analyse.

Par ailleurs, un grand nombre d'entreprises ont du mal à concevoir les richesses qui se trouvent dans leur plateforme de données car ils n'ont pas en interne les compétences techniques nécessaires pour réaliser les opérations d'analyse. Les moteurs de données distribuées tels que MapReduce ou Apache Spark ou d'autres outils de traitement de données permettant l'écriture de scripts et de requêtes comme Apache Pig et Apache Hive sont trop complexes pour être utilisés par le plus grand nombre. Ainsi, de nouvelles fonctionnalités et de nouveaux outils comme Apache Zeppelin sont indispensables pour rendre les plateformes de données plus accessibles au commun des mortels. Globalement, on peut constater que le degré d'abstraction prend de l'importance, de manière à simplifier et démocratiser l'analyse.

Enfin, les plateformes de données connectées ont besoin d'évoluer vers une meilleure prise en compte de leur utilisation au sein des entreprises. Elles doivent être prêtes pour un déploiement dans un environnement business, ce qui implique des performances prévisibles, des capacités de cryptage des données, un haut niveau de sécurité, une stratégie de gouvernance des données, une haute disponibilité, un plan de reprise d'activité et des fonctionnalités de débogage. Apache Hadoop est conçu pour être utilisé par la plupart des entreprises, et ces dernières recherchent de la qualité, de la disponibilité et une certaine tranquillité d'esprit au travers d'un support joignable 24h/24, 7j/7.

7. https://nifi.apache.org

Les applications modernes nécessitent des plateformes de données connectées

Les plateformes de données sont nécessaires pour capturer les informations éphémères issues des données temps réel tout en assurant la compréhension des informations historiques issues des données statiques. Les règles du jeu ont changé avec les fameux 3V du Big Data (Variété, Volume, Vélocité des données). Le niveau d'exigence étant de plus en plus élevé, on doit être capable de réaliser des analyses simultanées des données. Comprendre le passé n'est plus suffisant, nous devons pouvoir anticiper les événements futurs. Pour y parvenir, il est indispensable de s'appuyer sur des plateformes capables de connecter les données dynamiques et les données statiques.

Les voitures sans conducteur sont un excellent exemple : sur la route, elles doivent impérativement être capables d'éviter les autres véhicules et cela nécessite des calculs prédictifs en temps réel, à la vitesse de 55mph, sur la base d'événements historiques archivé combinés à l'analyse en temps réel de données en provenance de dizaines voire de centaines de capteurs. Le pari est que d'ici 2030 les voitures sans conducteur dominent les routes, et cela ne sera pas envisageable sans une plateforme de données connectées.

Sur un plan plus opérationnel, l'avenir du service client, du marketing et de la vente se trouve du côté des applications modernes d'exploitation des données. De plus en plus, les consommateurs préfèrent accéder à votre marque ou à vos services via des applications plutôt que de s'adresser à une personne réelle. Vous avez peut-être récemment séjourné dans un hôtel haut de gamme qui a fait le choix de remplacer le téléphone et la conciergerie par une tablette équipée d'une application… Le room service est désormais accessible grâce à une application dédiée ! En tant que consommateur, nous exigeons une expérience contextuelle sécurisée et personnalisée depuis nos applications mobiles. Pour délivrer cette expérience, les entreprises doivent pouvoir analyser de multiples données statiques et dynamiques simultanément, quasiment en temps réel, afin de comprendre le contexte, personnaliser leur offre et commencer à faire de l'étude prédictive.

De nos jours, les cas d'utilisation cités comme exemple reposent toujours sur des applications modernes de données capables de transformer les défis perçus hier comme impossibles en nouveaux produits, solutions innovantes ou simples commodités, et chacune de ces applications nécessite une stratégie connectée.

On peut également prendre l'exemple de l'automatisation des campagnes marketing, auparavant basées sur l'analyse du tracking par cookies ou sur les habitudes de navigation confrontées à des personas définis en amont sur la base des émotions et des comportements attendus des internautes.

Aujourd'hui, il est possible de s'appuyer sur des moteurs de recommandation automatisés qui associent les produits aux préférences des utilisateurs en quelques millisecondes grâce à l'étude de leur comportement, de leur localisation ou d'autres éléments contextuels : Où avez-vous cliqué il y a quelques secondes ? En quoi cela vient-il confirmer ou infirmer une tendance plus générale ? Quand êtes-vous entré et quand avez-vous quitté la boutique ? Sur quel réseau social étiez-vous connecté à ce moment-là ? Etes-vous passé à proximité d'un produit ou d'une balise bluetooth ? A quelle offre avez-vous répondu favorablement ? Que pouvons-nous vous offrir sur la base de ces différents éléments ?

Une mise en pratique avec Progressive Insurance

Progressive Insurance s'appuie sur une stratégie de données connectées visant à récompenser les conducteurs qui ont les comportements les plus sûrs et à favoriser la sécurité routière. Une de leur publicité met en avant un outil de traçage « snapshot » permettant d'analyser la conduite des assurés dans ses moindres détails. Via une application web, les clients peuvent observer leur conduite afin de réduire les comportements à risque en suivant les indications de la compagnie d'assurances.

Grâce aux données enregistrées sur plus de 10 milliards de kilomètres et stockées dans Hortonworks, Progressive Insurance est en mesure de prédire les risques inhérents à la conduite de chaque assuré et ainsi d'ajuster sa politique de pricing au cas par cas, permettant aux conducteurs les plus prudents d'obtenir les meilleurs tarifs.

De plus, l'accumulation des informations récoltées au fil des kilomètres dans la plateforme de données connectées affine peu à peu le modèle prédictif de la compagnie d'assurance et apporte une contribution sur d'autres aspects de l'activité, en analysant par exemple les déclarations d'assurance aux fins de contrôler leur exactitude.

Grâce à cette technologie, Progressive Insurance a pu accomplir la transformation de son modèle économique et réduire ses coûts dans des proportions très importantes. L'outil Snapshot couplé au modèle d'assurance basé sur les comportements de conduite ont permis à Progressive Insurance d'économiser plus 2,6 milliards en 2014.

Les six caractéristiques indispensables d'une plateforme de données connectées

Nous vous proposons de passer en revue six points essentiels pour bâtir votre stratégie de données connectées. Commencez par vous demander si votre plateforme de données répond à ces différentes exigences :

1 **Intégration des données sécurisée :** La plateforme permet-elle de récupérer facilement et de manière sécurisée des données disparates issues de l'Internet of Anything et de les intégrer rapidement, tout en détectant des patterns de données pertinents ?

2 **Intelligence opérationnelle :** votre plateforme peut-elle fournir en temps réel des informations exploitables issues des données statiques et des données temps réel ?

3 **Connectivité distribuée :** Etes-vous capable de connecter, relier ou mettre en relation tous les types de données afin de fournir une vision du contexte à 360° et ainsi réaliser des prédictions et des actions de personnalisation de vos applications sur l'ensemble de vos canaux ?

4 **Accessibilité à l'ensemble des utilisateurs :** Votre plateforme fournit-elle des outils permettant aux analystes et autres utilisateurs métier d'identifier des patterns et d'acquérir des informations exploitables ?

5 **Pérennité de la plateforme :** Votre plateforme est-elle bâtie sur une technologie 100% open source vous permettant de bénéficier d'innovations permanentes, quel que soit votre contexte économique ?

6 **Adaptation au contexte business :** Votre plateforme fournit-elle des solutions adaptées à l'entreprise en termes de sécurité, de haute disponibilité, de reprise d'activité après sinistre, etc… Vos données sont-elles sécurisées ? Existe-t-il un contrôle des accès ? Les règles de conformité sont-elles bien prises en compte ? Avez-vous mis en place une piste d'audit fiable ? Le cycle de vie des données est-il correctement contrôlé et managé ? Existe-t-il un support technique assuré par des experts ?

Ce que propose Hortonworks

Notre technologie a pour but de répondre à ces différentes exigences. Nous sommes convaincus que les plateformes de données connectées constituent la meilleure approche pour gérer et exploiter la valeur des données à la fois statiques et dynamiques.

La plateforme de données Hortonworks couvre tous les besoins en matière de données statiques. HDP™ est structuré, développé et conçu dans un environnement entièrement open source, mettant à disposition une plateforme de données prête à l'emploi et permettant aux entreprises de construire des applications modernes de données. Avec YARN pour cœur d'architecture, HDP fournit une plateforme capable de supporter de multiples workloads traités selon des méthodes variées — du traitement batch jusqu'au temps réel, incluant les fonctionnalités essentielles que l'on peut attendre d'une plateforme conçue pour l'entreprise — comprenant la gouvernance, la sécurité et le contrôle des opérations.

Hortonworks DataFlow (HDF™) fait partie de notre stratégie de données connectées. HDF fournit les capacités de streaming temps réel indispensables au traitement des données dynamiques. Cette plateforme représente une technologie centrale de l'Internet-of-AnyThing (IoAT) et de tout scénario d'utilisation de données disparates.

En combinant HDP™ et HDF™, on obtient la plateforme de données connectées la plus ouverte, innovante et opérationnelle du marché.

A propos d'Hortonworks

Hortonworks est un précurseur dans la création, la distribution et le développement de plateformes de données d'entreprise ouvertes et prêtes à l'emploi. Notre mission est d'optimiser la gestion et le management des données à l'échelle mondiale.

Nous focalisons notre expertise sur l'innovation dans les communautés open source telles qu'Apache Hadoop, NiFi et Spark. Notre plateforme de données connectées soutient des applications modernes capables d'exploiter à la fois les données statiques et dynamiques.

Avec plus de 1600 partenaires, nous fournissons l'expertise, la formation et les services qui permettent à nos clients de libérer le potentiel de leurs données, quel que soit le secteur d'activité. Nous construisons l'avenir des données™.

Contact

Pour plus d'information rendez-vous à l'adresse
www.hortonworks.com

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

HORTONWORKS