Hortonworks DataFlow

Accélération de la collecte de big data et de la gestion des flux de données

Livre blanc Hortonworks
DÉCEMBRE 2015

Sommaire

Qu'est-ce que Hortonworks DataFlow ?

Hortonworks DataFlow (HDF), administré par Apache™ NiFi, est la première plateforme intégrée qui simplifie et solutionne les problèmes liés à la collecte et au transport de données en provenance de sources multiples, qu'elles soient volumineuses ou restreintes, rapides ou lentes, connectées en permanence ou disponibles uniquement par intermittence.

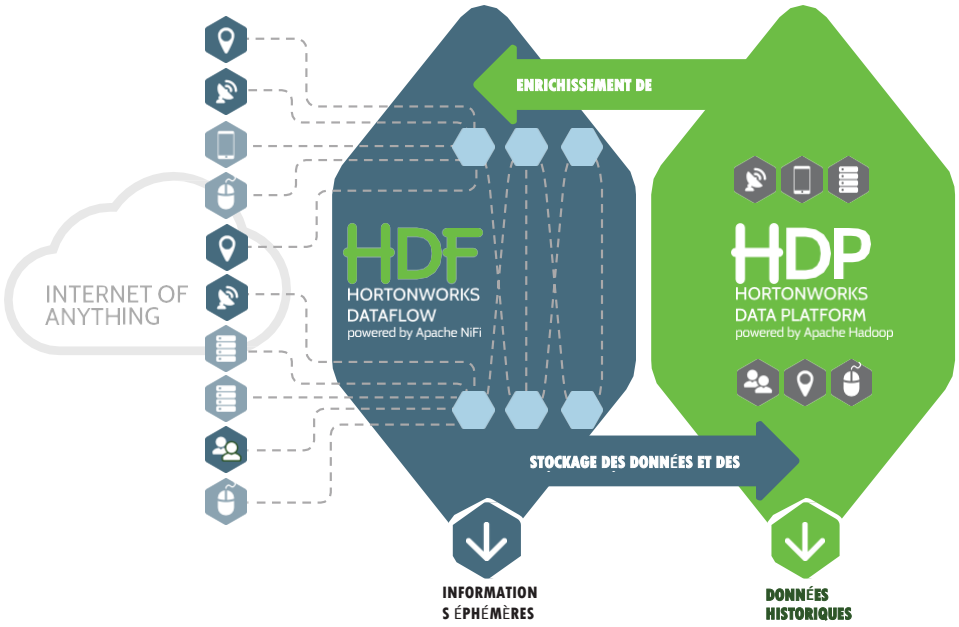
Hortonworks DataFlow est une plateforme combinée unique dédiée à l'acquisition de données, au traitement d'événements simples, au transport et à la mise à disposition d'informations, conçue pour gérer les flux de données complexes et très variés qui sont générés par l'ensemble des personnes, systèmes et objets connectés.

Solution idéale pour l'IoAT, l'Internet of Any Thing (Internet des objets, des personnes et des systèmes connectés), HDF permet l'acquisition simple et rapide de données, le transport sécurisé de données, la priorisation des flux de données et une traçabilité précise des données depuis la périphérie de votre réseau jusqu'au cœur de votre centre de données. Grâce à une interface visuelle intuitive, à un mécanisme d'accès et d'autorisation extrêmement fiable et à un framework de traçabilité (provenance des données) toujours « actif », HDF est le complément parfait de HDP pour rassembler les données historiques et les informations éphémères dans l'intérêt de votre entreprise.

Une plateforme intégrée unique pour l'acquisition de données, le traitement d'événements simples, le transport et la mise à disposition d'informations depuis la source jusqu'au stockage.

Hortonworks DataFlow permet la collecte et le traitement en temps réel des informations éphémères.

Hortonworks DataPlatform peut être utilisée pour enrichir le contenu et gérer les modifications des flux de données en temps réel.



Hortonworks DataFlow est conçue pour collecter et transporter, en toute sécurité, des données provenant de sources de données très hétéroclites, qu'elles soient volumineuses ou restreintes, rapides ou lentes, connectées en permanence ou disponibles uniquement par intermittence.

Figure 1 : Hortonworks DataFlow

Hortonworks DataFlow repose sur Apache NiFi, la technologie créée en 2006 par la NSA, l'agence de sécurité nationale des États-Unis, pour parvenir à automatiser les flux de données entre les systèmes, quel que soit leur type. C'est également le problème auquel sont confrontées les entreprises aujourd'hui.
À l'automne 2014, après huit ans de développement et d'utilisation à grande échelle, le programme Technology Transfer de la NSA a confié le NiFi à la fondation Apache.

Avantages de Hortonworks DataFlow

Par essence, DataFlow a été conçue pour relever un défi concret, à savoir réussir à collecter des données provenant de sources de données multiples et disparates, et cela avec efficacité et en toute sécurité au sein d'un réseau dispersé géographiquement et potentiellement fragmenté. Cette technologie a été éprouvée sur le terrain car la NSA a rencontré une grande partie des difficultés auxquelles sont confrontées les entreprises aujourd'hui. Par conséquent, elle intègre des capacités en termes de sécurité, d'évolutivité, d'intégration, de fiabilité et d'extensibilité, et a fait ses preuves en termes de facilité d'utilisation et de déploiement.

HORTONWORKS DATAFLOW PERMET AUX ENTREPRISES DE :

Optimiser l'efficacité opérationnelle

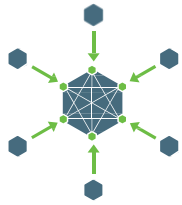
- Accélérer le ROI du big data grâce à une collecte de données simplifiée et une interface de gestion des flux de données visuelle et intuitive
- Réduire considérablement le coût et la complexité de la gestion, du maintien et du développement des flux de données
- Assurer la traçabilité et vérifier la valeur des sources de données pour de futurs investissements
- S'adapter rapidement aux nouvelles sources de données grâce à une plateforme extrêmement extensible et évolutive

Prendre de meilleures décisions pour l'entreprise

- Prendre de meilleures décisions pour l'entreprise grâce à des politiques de partage de données extrêmement détaillées
- Se focaliser sur l'innovation en automatisant le routage et la gestion des flux de données ainsi que la résolution des problèmes, sans avoir recours au codage
- Permettre une prise de décision immédiate et opportune en tirant parti des flux de données bidirectionnels en temps réel
- Renforcer la flexibilité de l'entreprise grâce à la définition de niveaux de priorités pour la collecte de données

Renforcer la sécurité des données

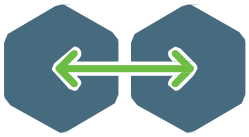
- Assurer une sécurité des données sans précédent et pourtant très simple à mettre en œuvre, de la source au stockage
- Améliorer la conformité et réduire le risque grâce à des politiques d'accès, de partage et d'utilisation des données extrêmement granulaires
- Créer un écosystème de flux de données sécurisé capable d'assurer le même niveau de sécurité et de chiffrement pour les sources de données Java de petite échelle que pour les centres de données des grandes entreprises

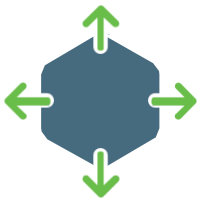


Accélération du ROI du big data grâce à une plateforme de collecte unique et agnostique en ce qui concerne les sources de données

Réduction du coût et de la complexité grâce à une interface utilisateur visuelle en temps réel

Sécurité des données sans précédent et pourtant très simple à mettre en œuvre, de la source au stockage





Meilleures décisions pour l'entreprise grâce à des politiques de partage de données extrêmement granulaires

Réaction en temps réel en tirant parti des flux de données bidirectionnels et des sources de données associées à un niveau de priorité

Adaptation aux nouvelles sources de données grâce à une plateforme extrêmement extensible et évolutive

Fonctionnalités de Hortonworks DataFlow

COLLECTE DE DONNÉES

Collecte de données intégrée à partir de sources dynamiques, disparates et distribuées de différents formats, schémas, protocoles, vélocités et volumes, telles que les machines, les dispositifs de géolocalisation, les clickstreams, les fichiers, les flux sociaux, les fichiers journaux et les

DÉCISIONS EN TEMPS RÉEL

Évaluation en temps réel de toutes les informations éphémères pour déterminer si elles sont pertinentes ou non et prendre, en conséquence, la décision d'envoyer, d'abandonner ou de stocker localement ces données en fonction des besoins

EFFICACITÉ OPÉRATIONNELLE

Interface efficace et rapide utilisant l'outil glisser-déposer pour la création, la gestion et l'ajustement des flux de données, ainsi que la résolution des problèmes, permettant la création et l'ajustement des flux de données en cinq minutes maximum, sans avoir recours au codage

SÉCURITÉ ET PROVENANCE

Routage de bout-en-bout sécurisé, de la source à la destination, grâce à un système d'autorisation utilisateur distinct et à une chaîne de traçabilité et des métadonnées visuelles (provenance des données) en temps réel

FLUX DE DONNÉES BIDIRECTIONNELS

Définition de niveaux de priorité et transport des données en temps réel en toute fiabilité en tirant parti des flux de données bidirectionnels pour une adaptation dynamique aux fluctuations du volume de

COMMANDE ET CONTRÔLE

Possibilité immédiate de créer, modifier, ajuster, visualiser, démarrer, arrêter, retracer, analyser, filtrer, réunir, combiner, transformer, dévier, cloner ou reproduire des flux de données grâce à une interface

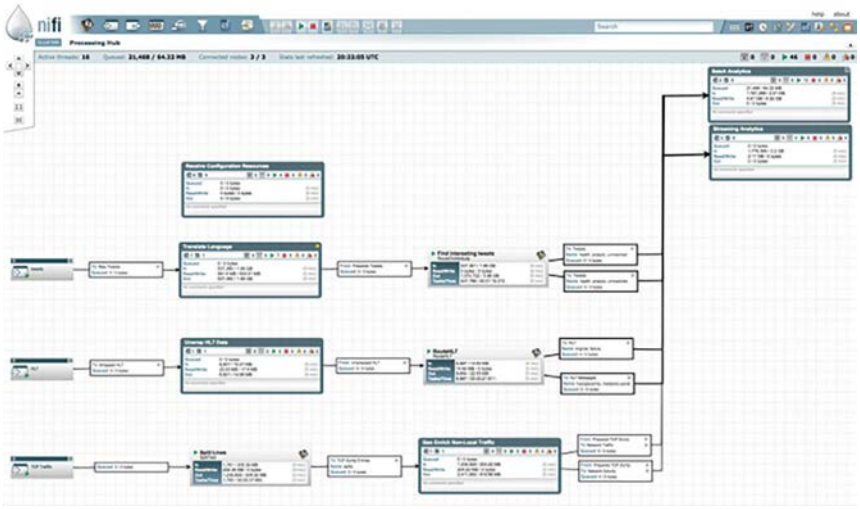


Figure 2 : interface utilisateur visuelle en temps réel Apache NiFi

Applications courantes de Hortonworks DataFlow

Hortonworks DataFlow accélère l'accès aux informations en permettant la mise en œuvre d'une programmation standard basée sur le flux pour l'infrastructure du big data et en réduisant la complexité actuelle liée à l'acquisition sécurisée, l'ingestion et l'analyse en temps réel des données provenant de sources de données diverses et distribuées.

Framework idéal pour la collecte de données et la gestion des flux de données, Hortonworks DataFlow est surtout utilisé pour simplifier et rationaliser l'ingestion des big data, renforcer la sécurité de la collecte et du partage de données grâce à des métadonnées et une chaîne de traçabilité extrêmement fiables, et en tant qu'infrastructure sous-jacente pour l'Internet des objets.

Cas 1 : collecte de données accélérée et efficacité opérationnelle

Ingestion des big data rationalisée

Hortonworks DataFlow accélère le bus d'ingestion des big data grâce à une interface visuelle unique, intégrée et facilement extensible, pour l'acquisition et l'ingestion, en temps réel, de données provenant de sources de données différentes, disparates et distribuées. La simplification et l'intégration de la création, du contrôle et de l'analyse des flux de données permettent d'accélérer le retour sur investissement des projets big data et d'augmenter l'efficacité opérationnelle.

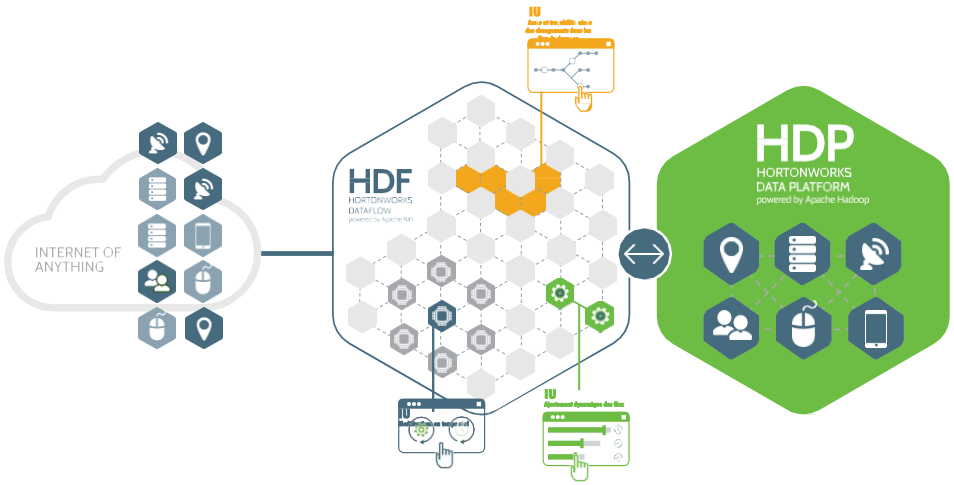


Figure 2 : une plateforme de collecte intégrée et agnostique en ce qui concerne les sources de données

QU'EST-CE QU'UNE INTERFACE DE COMMANDE ET DE CONTRÔLE ?

Une interface de commande et de contrôle permet de manipuler le flux de données en temps réel de sorte que les données contextuelles en cours puissent être renvoyées au système pour modifier immédiatement les résultats. Cela contraste avec une approche axée sur la conception et le déploiement, qui implique la programmation statistique d'un système de flux de données avant que les données n'y soient enregistrées puis le retour à une phase de programmation statique pour procéder à des ajustements avant de relancer une nouvelle fois le flux de données. Une analogie pourrait être faite avec la différence entre l'impression 3D, qui nécessite une planification préalable avant son exécution, et le modelage d'une sculpture en argile, qui réagit de façon immédiate et permet l'ajustement de la production finale en temps réel.

Pour en savoir plus sur Hortonworks DataFlow, rendez-vous sur http://hortonworks.com/hdf

Pourquoi les systèmes actuels de collecte de données ne sont pas idéaux

Les outils actuels de collecte et d'ingestion des big data sont conçus pour des usages spécifiques et sont trop sophistiqués car ils n'ont tout simplement pas été créés dans un souci d'utilisation universelle et d'efficacité opérationnelle. Cela a généré une architecture complexe composée d'outils d'acquisition, de messagerie et de transformation disparates et souvent personnalisés qui rendent l'ingestion des big data complexe, chronophage et coûteuse tant au niveau du déploiement que de la maintenance.

En outre, le décalage lié à la ligne de commande, et les outils dépendants du codage entravent l'accès aux données et ne permettent pas la prise de décisions opérationnelles opportune qu'exige aujourd'hui le monde des entreprises.

Figure 3 : les solutions actuelles d'ingestion des big data sont complexes et inefficaces sur le plan opérationnel

Cas 2 : sécurité accrue et chaîne de traçabilité sans précédent

Renforcement de la sécurité et des informations sur la provenance avec Hortonworks DataFlow
La sécurité des données devient de plus en plus importante dans le monde des appareils connectés en permanence et la nécessité de respecter des règles de conformité et de sécurité des données est actuellement difficile, complexe et coûteuse. La vérification de l'accès aux données et de leur utilisation est difficile, chronophage et implique souvent un processus manuel de réunion des différents systèmes et rapports pour vérifier la source des données, comment elles sont utilisées, qui les a utilisées et à quelle fréquence.

Les outils actuellement utilisés pour le transport des données électroniques ne sont pas conçus pour les futures exigences en termes de sécurité. Il est difficile, voire pratiquement impossible, que les outils actuels partagent des bits de données distincts et encore moins de façon dynamique et c'est un problème qui devait être trouvé sa solution dans l'environnement Apache NiFi sous la forme d'une plateforme de flux de données utilisée par les agences gouvernementales.

Hortonworks DataFlow satisfait les besoins en termes de sécurité et d'informations sur la provenance des données dans un monde électronique de gestion en temps réel des flux de big data distribués. Hortonworks DataFlow améliore les systèmes existants grâce à une plateforme d'ingestion de big data sécurisée, fiable, simplifiée et intégrée, qui garantit la sécurité des données en provenance de toutes les sources, qu'elles soient centralisées, qu'il s'agisse de centres de données volumineux ou que l'on gère des sources de données distribuées et distantes sur des liens de communication géographiquement dispersés. HDF offre différentes fonctionnalités liées à la sécurité, dont la provenance des données de bout-en-bout : une chaîne de traçabilité pour les données. Au-delà de la capacité à satisfaire les règles de conformité, la provenance constitue une méthode pour retracer le parcours des données depuis leur point d'origine jusqu'à n'importe quel point dans le flux de données afin de déterminer quelles sont les sources de données les plus utilisées et les plus précieuses.

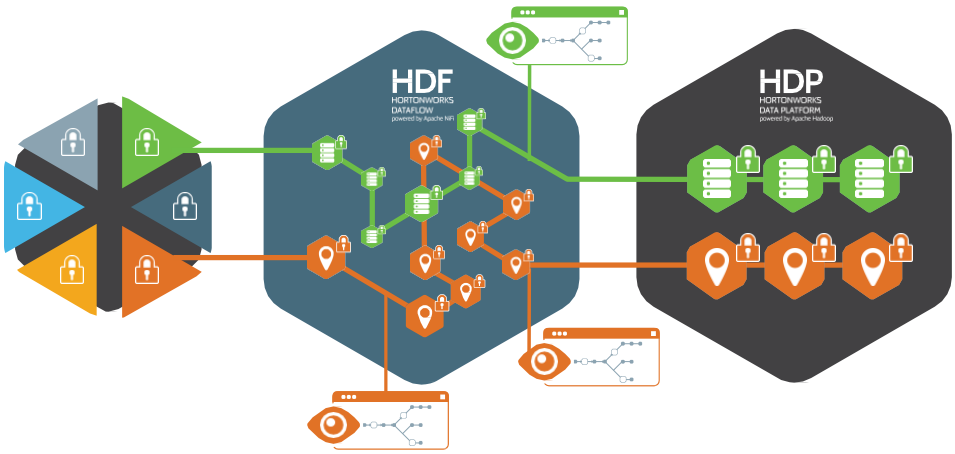


Figure 4 : sécurisation de la source au stockage grâce à des informations sur la provenance d'une grande fiabilité

QU'EST-CE QUE LA PROVENANCE DES DONNÉES ?

La provenance se définit par le lieu d'origine de quelque chose ou sa première information historique connue.
Dans le contexte d'un flux de données, la provenance des données est la capacité à retracer le parcours d'une donnée au sein d'un flux de données, depuis le lieu de sa création jusqu'à sa destination finale.
Au sein de Hortonworks DataFlow, la provenance des données permet d'avoir la possibilité de vérifier visuellement d'où proviennent les données, comment elles ont été utilisées, qui les a visualisées, et si elles ont été envoyées, copiées, transformées ou reçues. Tout système ou personne ayant consulté une donnée spécifique génère l'enregistrement de l'ensemble des informations le concernant en termes de temps, de date, d'action, d'historique et de variables dépendantes dans le but d'obtenir une vue complète de la chaîne de traçabilité pour cette donnée au sein d'un flux de données.
Ces métadonnées indiquant la provenance sont utilisées pour contribuer au respect des exigences de conformité pour le partage des données, et à des fins de résolution de problèmes et d'optimisation.

Démocratisation des données avec un niveau de sécurité sans précédent

Hortonworks DataFlow ouvre de nouvelles perspectives en termes d'informations d'entreprise en permettant d'accorder des accès sécurisés sans entraver l'analyse car les données très spécifiques peuvent être partagées ou non. Marie pourrait, par exemple, être autorisée à accéder aux données étiquetées « finance » au sein d'un flux de données alors que Jacques pourrait être autorisé à accéder au même flux de données mais uniquement aux données étiquetées « 2015 » et « chiffre d'affaires ». Cela supprime les inconvénients liés à l'accès aux données lié à la fonction, qui peut incidemment créer des risques en termes de sécurité, tout en permettant la démocratisation des données pour une analyse complète et une prise de décisions éclairée.

La capacité inhérente de Hortonworks DataFlow, à gérer des données et métadonnées granulaires indiquant la provenance grâce à son processus de collecte, de transport et d'ingestion, permet de fournir les informations complètes et détaillées qui sont nécessaires à des fins d'audit et de conversion, une compétence qu'aucun autre système d'ingestion de données existant ne peut fournir aujourd'hui.

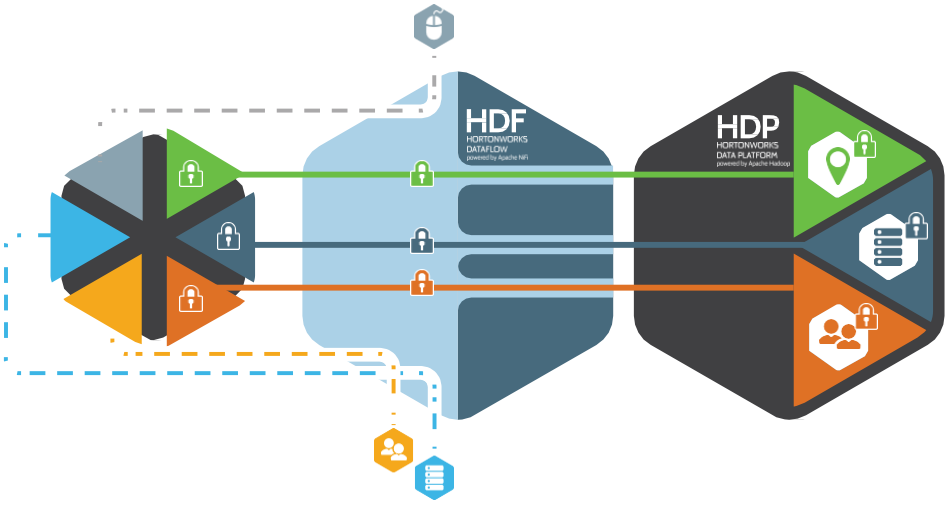


Figure 5 : accès aux données et contrôle granulaires

Cas 3 : L'Internet of Any Thing (IoAT)

L'Internet of Any Thing (IoAT) avec Hortonworks DataFlow

Conçue sur le terrain, où les ressources sont limitées (puissance électrique, connectivité, bande passante), Hortonworks DataFlow est une plateforme évolutive éprouvée pour l'acquisition et l'ingestion de l'Internet des objets (IoT, Internet of Things), et même plus largement, l'Internet des objets, des personnes et des systèmes connectés (IoAT, Internet of Any Thing).

Adaptable aux contraintes de ressources

La création d'un Internet des objets toujours connecté et pourtant physiquement dispersé entraîne de nombreuses difficultés. Les sources de données sont souvent distantes, le matériel peut être limité, et il est fréquent que la puissance électrique et la bande passante soient à la fois variables et restreintes. Le manque de fiabilité de la connexion perturbe la communication et entraîne des pertes de données, et les insuffisances, en termes de sécurité, de la plupart des capteurs du monde créent des risques pour les entreprises.

Parallèlement, les appareils produisent plus de données que jamais. Une grande partie des données générées sont des données en mouvement et il est essentiel de débloquer la valeur commerciale de ces données pour transformer l'économie moderne.

Pourtant, la transformation des entreprises repose sur un accès aux données sécurisé et précis depuis la source jusqu'au stockage. Toutes ces contraintes existant dans le monde réel, à savoir la limitation de la puissance électrique, la fluctuation de la connectivité, la sécurité des données, la traçabilité des données, la diversité des sources de données et la distribution géographique, ont été prises en compte lors de la conception de Hortonworks DataFlow afin de permettre une prise de décision objective et opportune.

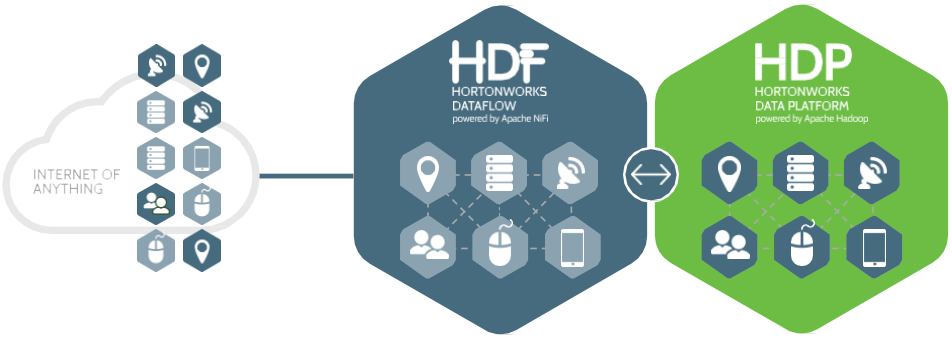


Figure 6 : une plateforme éprouvée pour l'Internet des objets

Sécurisation de la collecte de données

Hortonworks DataFlow répond aux besoins de l'Internet des objets en termes de sécurité grâce à une plateforme de collecte des big data, qui est sécurisée, fiable, simplifiée et intégrée, et qui garantit la sécurité des données provenant de sources de données distribuées sur des liens de communication géographiquement dispersés. Les fonctionnalités liées à la sécurité de HDF incluent la provenance des données de bout-en-bout : une chaîne de traçabilité pour les données. Cela permet aux systèmes de l'IoT de vérifier l'origine des flux de données et de résoudre les problèmes entre le point d'origine et le point de destination, et leur donne la capacité de déterminer quelles sont les sources de données les plus fréquemment utilisées et les plus précieuses.

Hortonworks DataFlow est capable d'assurer la sécurité et le chiffrement à la fois pour les sources de données Java de petite échelle et pour les centres de données des grandes entreprises. Cela dote l'Internet des objets d'une plateforme de collecte et de transport des données qui est sécurisée, fiable et simplifiée, et qui assure un retour d'informations en temps réel en vue d'améliorer immédiatement et constamment les algorithmes et les analyses pour une prise de décisions objective, éclairée et opportune.

Priorisation du transfert de données et boucles de rétroaction bidirectionnelles

Hortonworks DataFlow permet la définition de niveaux de priorité au sein d'un flux de données car la connectivité et la bande passante disponible peuvent fluctuer et que le volume de données générées par une source peut dépasser celui que peut recevoir la destination. Cela signifie qu'en cas de contraintes de ressources, les sources de données peuvent recevoir l'ordre de favoriser automatiquement les informations les plus importantes qui seront envoyées en premier, et de conserver les données moins importantes qui passeront par de futures fenêtres de transmission ou ne seront pas envoyées du tout.

Si un appareil distant devait subir une coupure de courant, par exemple, il est essentiel d'envoyer les données « les plus importantes » de cet appareil en premier dès la résolution de la panne et le rétablissement de la communication. Une fois que les informations « les plus importantes » ont été envoyées, elles peuvent être suivies des données en attente, qui sont moins prioritaires et moins essentielles pour la prise de décisions immédiates mais néanmoins vitales pour l'analyse historique.

Hortonworks DataFlow permet de prendre des décisions ultimes concernant la nécessité d'envoyer, d'abandonner ou de stocker localement les données, en fonction des besoins et de l'évolution des conditions. Par ailleurs, grâce à une interface de commande et de contrôle granulaire, les files d'attente de données peuvent être ralenties ou accélérées afin de trouver un équilibre entre les exigences de la situation en cours et la disponibilité et le coût actuels des ressources.

Avec la capacité de s'adapter facilement aux contraintes de ressources en temps réel, de garantir une collecte de données sécurisée et d'appliquer des niveaux de priorité au transfert de données, Hortonworks DataFlow a prouvé qu'elle était une plateforme idéale pour l'Internet des objets.

Pourquoi Hortonworks pour Apache™ Hadoop®?

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

Pour obtenir une analyse indépendante de la plateforme de données Hortonworks et vous faire confirmer sa prééminence auprès des fournisseurs d'Apache Hadoop, vous pouvez télécharger le rapport Forrester Wave™: Big Data Apache Hadoop Solutions, Q1 2014 de Forrester Research.

À propos de Hortonworks

Hortonworks développe, distribue et supporte 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 Apache 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 Apache Hadoop. Hortonworks entretient des relations étroites avec des partenaires stratégiques de centres de données, qui permettent à nos clients de profiter des meilleures opportunités de Apache Hadoop.

Pour plus d'informations, rendez-vous à l'adresse : www.hortonworks.com.