DÉVELOPPEMENT D'APPLICATIONS D'ANALYSE EN TEMPS RÉEL AVEC HORTONWORKS DATAFLOW

Donnez aux utilisateurs la possibilité d'extraire des informations à partir des flux de données sans avoir à écrire une seule ligne de code

UN LIVRE BLANC DE HORTONWORKS
JUIN 2017

 ©2017 Hortonworks

Contenu

Rapport de synthèse

Pour réaliser tout le potentiel des applications de données modernes, les entreprises doivent être en mesure de capturer des informations historiques et enrichies à partir des données au repos, et des informations périssables à partir des données en mouvement. Bien que les outils de gestion des flux soient disponibles pour faciliter la collecte, l'acheminement, le filtrage et la transformation des données provenant de n'importe quelle source, les entreprises n'ont pas eu à leur disposition des outils équivalents pour créer les applications d'analyse nécessaires pour extraire des informations utiles de leurs flux de données. Hortonworks répond désormais à ce besoin en mettant à leur disposition Streaming Analytics Manager et Schema Registry, deux composants de Hortonworks DataFlow 3.0.

Streaming Analytics Manager permet aux développeurs de créer facilement des applications d'analyse en temps réel sans avoir à écrire une seule ligne de code. Les analystes d'affaires peuvent créer des tableaux de bord et des visualisations de données efficaces pour l'analyse descriptive des flux de données, tandis que les équipes de l'exploitation informatique peuvent gérer l'intégralité du cycle de vie des applications de lecture des données en temps réel depuis leur création jusqu'à leur suppression. Schema Registry améliore encore plus le développement d'applications de lecture des données en temps réel en fournissant un référentiel partagé des schémas qui peuvent désormais être partagés et réutilisés avec plus d'efficacité. Des services professionnels complémentaires prennent en charge le provisionnement, la gestion, la surveillance, la sécurité, l'audit, la conformité et la gouvernance sur l'ensemble de la plateforme.

Ce livre blanc présente les avantages des applications d'analyse en temps réel, les problèmes que leur développement peut poser, et les méthodes que les entreprises peuvent utiliser pour faciliter la conception, le développement, le test, le déploiement et la maintenance par les parties prenantes de ces applications sans avoir recours à des compétences ou des formations particulières.

Applications d'analyse en temps réel : pourquoi sont-elles utiles et pourquoi sont-elles si difficiles à créer ?

Les données en mouvement peuvent fournir des informations utiles aux entreprises de toutes sortes. En effectuant l'analyse des Big Data en temps réel sur des flux de données issues d'applications destinées à la clientèle, de systèmes propriétaires, d'outils de surveillance de l'infrastructure, de capteurs et d'autres sources, les entreprises peuvent extraire de précieux renseignements pour en savoir plus sur le comportement de leurs clients, orienter la prise de décisions, faciliter l'automatisation et offrir un meilleur service. La valeur des applications d'analyse en temps réel peut être constatée dans les nombreux exemples d'utilisation dans lesquels elles peuvent être appliquées, par exemple :

- La détection proactive des défaillances ou des défauts potentiels dans des machines industrielles afin que ceux-ci soient gérés avant qu'ils ne perturbent la production
- La surveillance d'un parc de véhicules pour les violations des politiques et des réglementations sur la vitesse, les périodes de repos et l'itinéraire pour assurer la conformité et réduire les risques
- L'analyse des comportements de navigation sur un site Web pour optimiser le placement du contenu en fonction des intérêts et préférences présumées de l'utilisateur
- L'évaluation et le rééquilibrage d'un portefeuille d'actions en fonction des fluctuations des prix pour respecter les objectifs de placement et la tolérance au risque d'un client

Toutefois, la création d'applications prenant en charge ces exemples d'utilisation n'est pas chose simple, loin de là. Tout d'abord, les entreprises doivent se décider entre une multitude d'options technologiques pour le traitement et l'analyse des données, y compris entre des services sur site et/ou cloud, et entre des plateformes Open Source et propriétaires. Une fois leur décision prise, ces technologies doivent être intégrées et déployées de façon transparente dans une pile complète afin de transmettre, stocker et analyser les données à grande échelle, un défi technique en soi potentiellement important.

Le déploiement d'une pile d'analyse en temps réel n'est que le début. Le développement d'applications d'analyse en temps réel nécessite à la fois une grande compréhension d'une multitude de systèmes, des connaissances approfondies du code et des compétences spécialisées rares dans l'industrie, et totalement absentes dans la plupart des entreprises. Même lorsque les compétences requises sont disponibles — la plupart du temps à un coût élevé, car ces talents sont très demandés — la mise en production de ces applications implique des processus fastidieux pour leur conception, programmation, test, optimisation et déploiement. Même les processus de lecture des données en temps réel qui consistent à associer et à séparer des flux, à agréger des données sur des périodes différentes, et à déterminer des tendances, sont très difficiles à mettre en œuvre. La réalisation d'une capacité d'analyse en temps réel d'entrée de gamme demeure un processus requérant de nombreux talents sur le long terme, sans parler du développement d'applications à haute valeur ajoutée plus sophistiquées et plus puissantes.

Bien qu'une poignée d'anciens fournisseurs apportent des solutions à certains de ces problèmes, leurs produits ont des coûts prohibitifs, restent exclusifs et ne sont pas évolutifs, ces caractéristiques étant aux antipodes des principes de l'Open Source qui est au cœur de la révolution du Big Data. Par conséquent, de nombreuses entreprises se retrouvent avec des compétences Big Data incomplètes, qui leur permettent d'extraire des informations à partir de données au repos, mais non à partir de données en mouvement.

Création d'applications d'analyse en temps réel avec Hortonworks DataFlow

Des applications de bout en bout pour les données en mouvement doivent être conçues pour recueillir et acheminer les données nécessaires, quel que soit leur emplacement, puis des informations exploitables doivent être extraites de celles-ci aussi rapidement et facilement que possible. À cette fin, deux types d'outils sont nécessaires :

- **Gestion des flux** : a pour objectif de fournir un moyen simple, sûr et fiable de transférer les données depuis n'importe quelle source (interne, cloud ou datacentre) vers un système en aval, capable d'effectuer intelligemment toutes les tâches relatives au routage, à la transformation, à la provenance et à la communication bidirectionnelle.
- **Analyse en temps réel** : a pour objectif de fournir des informations immédiates et continues sur les données en temps réel ou en quasi temps réel à l'aide d'agrégations réalisées sur plusieurs périodes, de l'association des modèles, de l'analyse prédictive et normative, et ainsi de suite.

Pour répondre aux besoins métiers en matière de Big Data, les organisations dépendent aussi de **services d'entreprise** dédiés au provisionnement, à la gestion, à la surveillance, à la sécurité, à l'audit, à la conformité et à la gouvernance.

De nombreux fournisseurs de Big Data n'utilisant aucune solution Hortonworks ont désormais pris l'initiative de répondre à ces besoins d'analyse en temps réel. Hortonworks DataFlow (HDF™) est une plateforme de données en mouvement permettant de créer des applications de gestion des flux de données et d'analyse en temps réel. Elle permet aux clients de collecter, d'organiser ou d'analyser les données en mouvement et d'agir sur celles-ci dans les périphériques et capteurs connectés au datacentre, au cloud et au réseau interne.

La solution Hortonworks comprend trois composants principaux : Flow Management, Stream Processing et Enterprise Services. **Flow Management** intègre Apache NiFi et Apache MiniFi et assure l'automatisation des flux de données entre les systèmes. Les données peuvent être obtenues auprès de diverses sources, y compris les systèmes transactionnels et opérationnels, les données du Web et des logs, les capteurs et périphériques propriétaires, et les plates-formes de médias sociaux tels que Twitter. Le composant Flow Management fournit également des fonctionnalités simples de transformation, d'acheminement et de livraison des données, la garantie de provenance des données étant intégrée de bout en bout.

Stream Processing intègre Apache Kafka, Apache Storm et le nouvel outil Hortonworks Stream Analytics Manager. Ce composant agit en tant que courtier de données évolutives pour les applications de lecture de données en temps réel et fournit un outil interactif pour concevoir, développer, déployer et gérer en quelques minutes des applications d'analyse en temps réel complexes, sans avoir à écrire une seule ligne de code

Enterprise Services intègre Apache Ambari, Apache Ranger et le nouvel outil Hortonworks Schema Registry. Ce composant offre la possibilité de surveiller, de gérer et de provisionner l'ensemble du cluster HDF, le tout grâce à une sécurité, une gouvernance et une prise en charge intégrée du multitenancy.

Les composants Streaming Analytics Manager et Schema Registry, qui ont été intégrés à HDF 3.0 pour compléter la solution Hortonworks dédiée aux applications d'analyse en temps réel, sont décrits plus en détail ci-dessous.

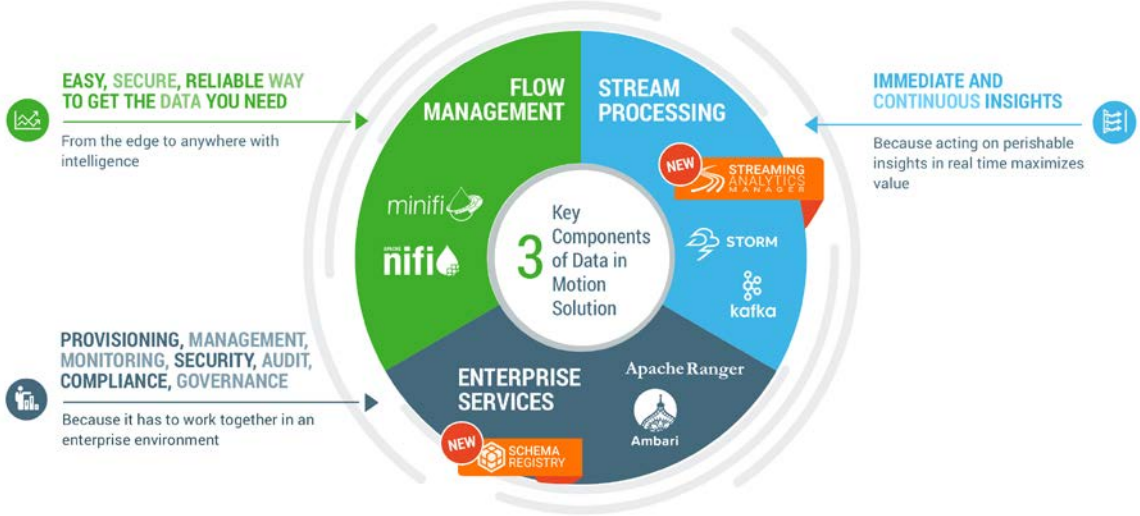

Figure 1: Hortonworks DataFlow

Hortonworks Streaming Analytics Manager

Nouvel élément de la pile HDF, Streaming Analytics Manager permet de créer des applications complexes d'analyse en temps réel sans écrire une seule ligne de code. Éliminant le besoin de compétences spécialisées, Streaming Analytics Manager fournit un modèle de programmation graphique avec une interface de type glisser-déposer pour créer des applications de lecture de données en temps réel pour la corrélation des événements, l'enrichissement contextuel, la correspondance des modèles et des agrégations d'analyse complexes, le tout accompagné par des alertes automatisées pour vous avertir de la découverte d'informations précieuses. Offrant le même type d'expérience enrichie pour la création d'applications d'analyse en temps réel dont les clients Hortonworks profitent déjà pour créer des applications de gestion des flux, Streaming Analytics Manager permet de mettre ses applications sur le marché beaucoup plus rapidement, et ce à moindre coût, pour accélérer leur délai de rentabilisation et leur impact stratégique.

Streaming Analytics Manager fournit des outils puissants pour répondre aux besoins des trois principaux intervenants du secteur des Big Data : les développeurs, les analystes d'affaires et les équipes chargées de l'exploitation informatique.

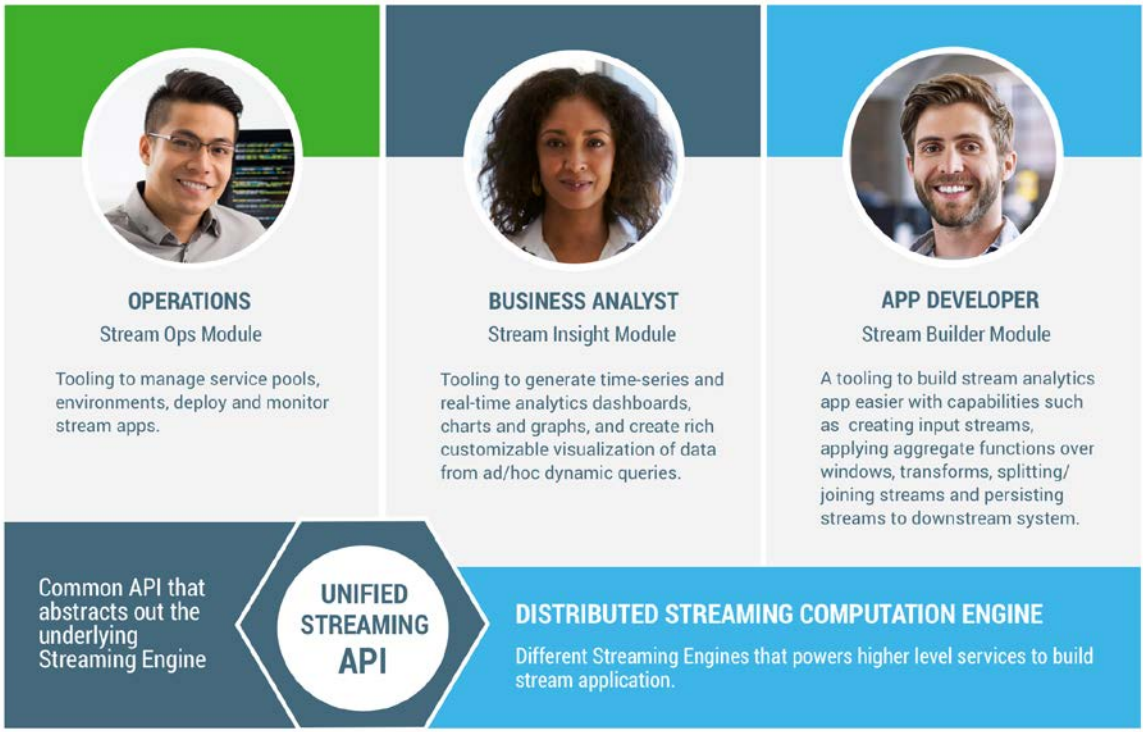

Figure 2: Hortonworks Streaming Analytics Manager

DÉVELOPPEURS — MODULE STREAM BUILDER

Les développeurs d'application ont besoin d'un moyen simple pour concevoir, développer, déboguer et déployer des applications d'analyse en temps réel sans avoir besoin de compétences spécialisées. Le module Stream Builder simplifie le développement de ces applications d'analyse en temps réel, tout en éliminant le besoin de maîtriser les subtilités des multiples systèmes internes ou même d'écrire une seule ligne de code. Son interface de type glisser-déposer fournit une page vierge où les développeurs peuvent facilement appliquer des fonctions telles que la création de flux d'entrée, l'agrégation des fonctions via des fenêtres, l'exécution des transformations, la séparation et l'association de flux, et la persistance des flux dans les systèmes en aval.

Stream Builder offre plusieurs blocs de construction pour créer des applications d'analyse en temps réel : sources, processeurs, récepteurs et composants personnalisés.

Des sources sont utilisées pour créer des flux de données, notamment Kafka, Azure Event Hub et HDFS.

Des processeurs sont utilisés pour manipuler des événements dans le flux. Voici les processeurs disponibles avec Streaming Analytics Manager :

- **Join** — Associe deux flux de données en fonction d'un champ provenant de chaque flux, conformément à une fenêtre configurée par l'utilisateur sur une base temporelle ou numérique.
- **Rule** — Fournit aux développeurs un menu déroulant permettant de configurer les conditions de règles qui acheminent les événements vers des flux différents. Les règles sont traduites automatiquement au format SQL pour pouvoir être appliquées au flux, et les développeurs peuvent aussi écrire directement des règles dans un format SQL complexe.
- **Aggregate** — Exécute des fonctions sur plusieurs fenêtres d'événements. Ces fenêtres peuvent être à bascule ou glissantes, et peuvent être basées sur un intervalle de temps ou un critère numérique. Les fonctions de fenêtre standard prêtes à l'emploi, notamment stddev, stddevp, variance, variancecep, avg, min, max, sum et count, peuvent être complétées avec des fonctions personnalisées.
- **Projection** — Applique des transformations aux événements contenus dans les flux. Un ensemble complet de fonctions incluses dans OpenOffice.org peut être complété avec des fonctions personnalisées.
- **Branch** — Réalise une construction if-else standard pour l'acheminement des données.

Des récepteurs sont utilisés pour envoyer des événements à d'autres systèmes. Streaming Analytics Manager prend en charge les récepteurs suivants :

- Kafka
- Hive
- Cassandra
- Druid
- JDBC
- Solr
- HDFS
- OpenTSDB
- HBase
- Notification

Des composants personnalisés peuvent être créés à l'aide du kit de développement (SDK) inclus pour répondre aux exigences spécifiques et créer des fonctions personnalisées pour les agrégations fenêtrées.

Indépendamment du moteur sous-jacent, Streaming Analytics Manager prend en charge plusieurs substrats de lecture de données en temps réel, tels que Storm, Spark Streaming et Flink. Les entreprises peuvent déployer des applications d'analyse en temps réel sur le moteur de lecture de données de leur choix, et les gérer à l'aide de métriques spécifiques aux applications.

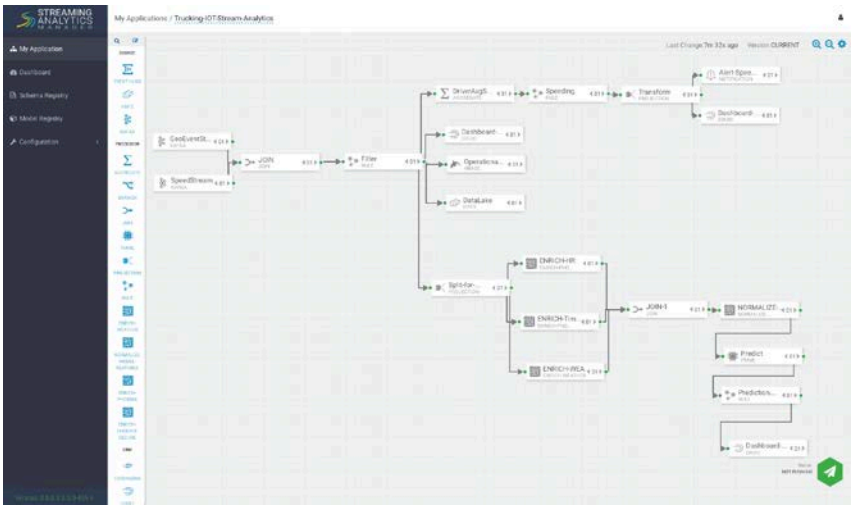

Figure 3 : Stream Builder : création par glisser-déposer d'applications de lecture de données en temps réel

ANALYSTES D'AFFAIRES — MODULE INSIGHT STREAM

Pour réaliser la pleine valeur des applications d'analyse en temps réel, les analystes d'affaires doivent être en mesure d'accéder aux flux de données et d'effectuer des analyses descriptives aussi facilement et intuitivement qu'avec d'autres outils décisionnels. Le module Stream Insight permet aux analystes d'affaires d'accéder eux-mêmes aux flux de données, en quasi temps réel — sans avoir à attendre pendant des heures que quelqu'un les leur fournisse — et d'effectuer des analyses descriptive à l'aide d'une puissante plateforme d'exploration. Son interface utilisateur interactive vous permet de créer des analyses de séries, ainsi que des tableaux de bord et des graphiques d'analyse en temps réel. Vous pourrez également générer de riches visualisations personnalisables de données à partir de requêtes dynamiques ad hoc. Le moteur d'analyse Stream Insight est optimisée par Druid, un magasin de données Open Source conçu pour les requêtes OLAP sur des données d'événements. Plus de 30 graphiques de visualisation prêts à l'emploi, combinés avec une fonctionnalité de création de bibliothèques personnalisées pour l'analyse ad hoc, permettent aux analystes d'affaires d'atteindre des niveaux supérieurs d'efficacité et de productivité.

Stream Insight a été conçu autour de quatre concepts : le moteur d'analyse, la source de données, la tranche et le tableau de bord.

- **Moteur d'analyse** — Les données peuvent être lues en temps réel dans le moteur d'analyse via le récepteur Druid ou celui du moteur d'analyse utilisé par les développeurs pour créer l'application de lecture de données en temps réel. Le moteur d'analyse peut ensuite diffuser les données dans de nouveaux cubes informatifs ou des cubes existants.
- **Source des données informatives** — Une source de données informatives, optimisée par Druid, représente la zone de stockage des flux de données. Le cube peut être interrogé pour effectuer des intégrations, des agrégations et d'autres analyses puissantes.
- **Tranche** — Une tranche est une visualisation qui peut être créée en posant des questions à la source de données. Ensuite, les informations extraites peuvent être ajoutées au tableau de bord.
- **Tableau de bord** — Créé par un analyste d'affaires pour réaliser une analyse descriptive, un tableau de bord se compose d'une série de tranches.

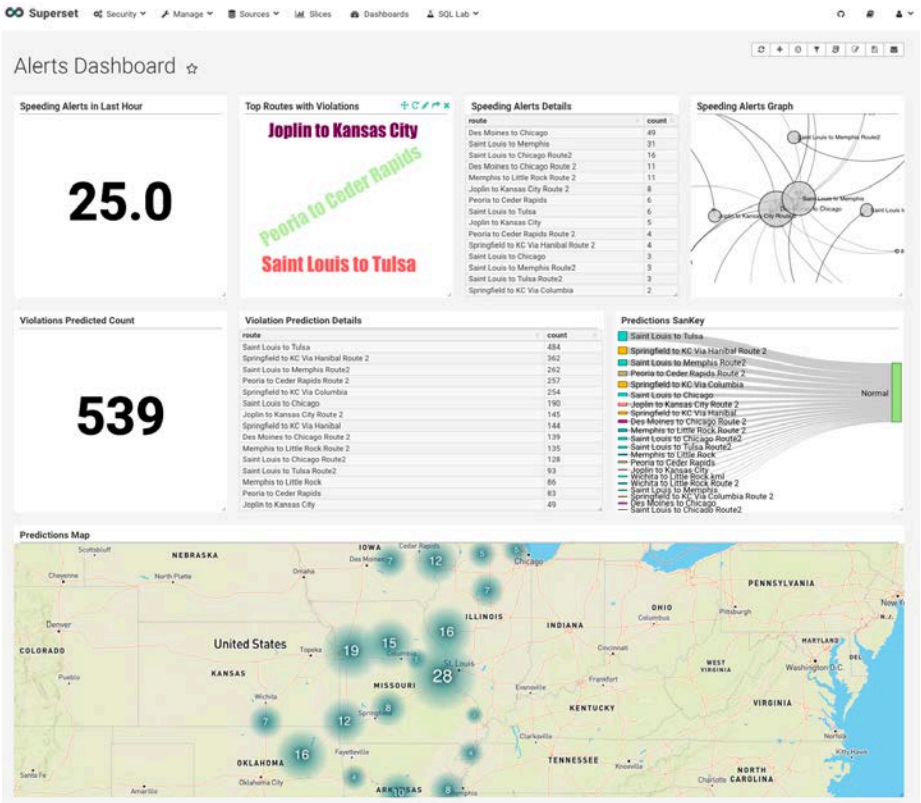

Figure 4 : Stream Builder : création par glisser-déposer d'applications de lecture de données en temps réel

EXPLOITATION INFORMATIQUE — MODULE STREAM OPS

Avant de créer une application d'analyse en temps réel, les développeurs s'appuient sur l'équipe de l'exploitation informatique pour configurer un environnement qui se soustrait aux services nécessaires pour la gestion des Big Data. Ces environnements peuvent prendre différentes formes selon les besoins de l'application devant être créée. Le module Stream Ops permet aux développeurs et aux équipes DevOps de créer une variété de pools de services intégrant à la fois Kafka, Storm, Apache HBase, Apache Hive, Apache Spark, Apache Zookeeper et Apache Hadoop HDFS. Ensuite, l'équipe de l'exploitation informatique peut facilement créer des environnements intégrant des des services provenant d'un ou plusieurs pools afin de les mettre à disposition des développeurs souhaitant créer des applications d'analyse en temps réel.

Une fois l'application créée, les utilisateurs dépendent des outils fournis par les équipes de l'exploitation informatique pour déployer, suivre et gérer les applications de lecture de données en temps réel. Streaming Analytics Manager simplifie le déploiement et la surveillance grâce à des outils entièrement intégrés ayant pour objectif de suivre et de gérer la totalité du cycle de vie de ces applications, de leur création à leur suppression en passant par leur modification et leur surveillance.

Chacun des trois principaux modules de Streaming Analytics Manager repose sur une API commune et unifiée de lecture de données en temps réel qui se soustrait au moteur de flux sous-jacent, qui alimente des services de niveau supérieur pour l'application d'analyse en temps réel.

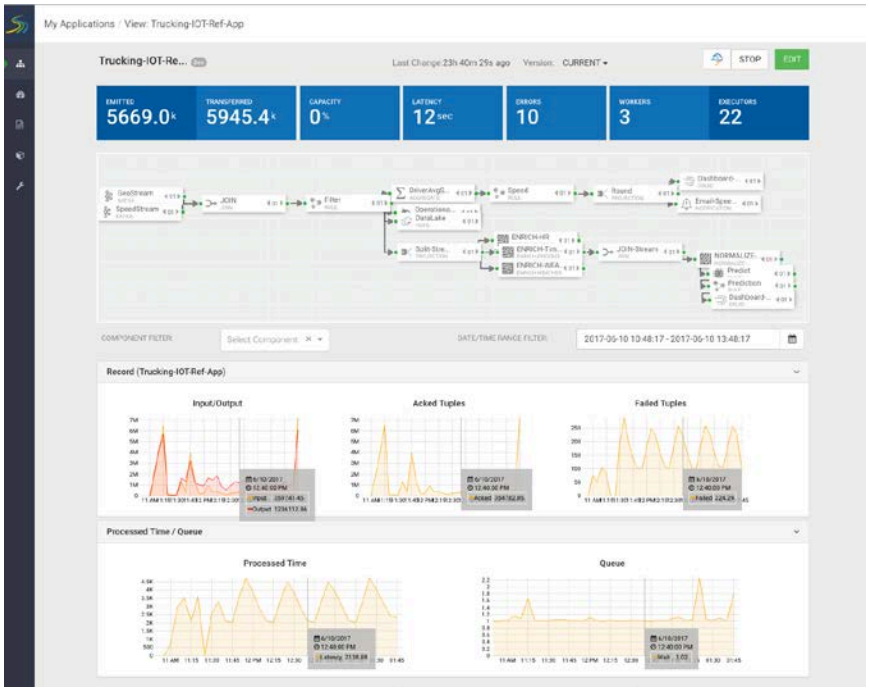

Figure 5 : Stream Ops : créez des pools de services et des environnements d'application partagés

Hortonworks Schema Registry

Les services de gestion des flux et de traitement de trains de données qui alimentent la plateforme de données en mouvements HDF sont complétés par des services d'entreprise dédiés au provisionnement, à la gestion, à la surveillance, à la sécurité, à l'audit, à la conformité et à la gouvernance. Les composants, tels qu'Apache Ambari et Apache Ranger, permettent à l'équipe de l'exploitation informatique de gérer l'ensemble du cluster HDF efficacement et de façon exhaustive. Désormais, ces services d'entreprise intègrent aussi Schema Registry, qui offre un moyen simple pour valider le schéma, assurer la conversion des formats et permettre au producteur et un consommateur de données d'évoluer à des rythmes différents.

Contrairement aux applications de gestion des flux créées avec NiFi, les applications d'analyse en temps réel exigent un schéma pour pouvoir fonctionner. Les développeurs dépendent de ce schéma pour appliquer des fonctions, telles que la connexion à une source de flux (par exemple, Kafka), appliquer des règles de filtrage, effectuer des agrégations sur une période de temps ou appliquer des transformations. Toutefois, Kafka, le principal mécanisme pour la connexion d'une application d'analyse en temps réel à un flux de données, ne possède aucun schéma. Par conséquent, les développeurs codent généralement en dur le schéma et sa sérialisation/dématérialisation dans l'application de lecture de données en temps réel elle-même, une approche inefficace qui empêche les schémas d'être réutilisés et qui ne parvient pas à répondre aux besoins de gouvernance et opérationnels de la plupart des entreprises.

Schema Registry améliore la gouvernance des données de bout en bout et l'efficacité opérationnelle en fournissant un registre centralisé, prenant en charge la gestion des versions et permettant la validation du schéma.

- **Registre centralisé** — Un référentiel partagé des schémas élimine la nécessité de joindre le schéma pour chaque élément de données. Les applications peuvent interagir simplement entre elles afin de sauvegarder ou de récupérer des schémas pour les données auxquelles elles doivent accéder. Entièrement intégré avec le composant de gestion des flux de HDF, y compris NiFi, Schema Registry permet aux schémas créés à l'aide de NiFi d'être facilement gérés et réutilisés par l'ensemble de la plateforme.
- **Gestion des versions** — En définissant les relations entre les schémas et en autorisant le partage des schémas entre les applications et composants HDF, Schema Registry prend en charge l'évolution du schéma afin qu'un consommateur et un producteur puissent comprendre différentes versions du même schéma tout en étant incapable de lire toutes les informations qu'ils partagent entre eux.
- **Validation du schéma** — Schema Registry prend en charge la validation du schéma en autorisant la conversion des formats génériques et le routage générique afin d'assurer la qualité des données.

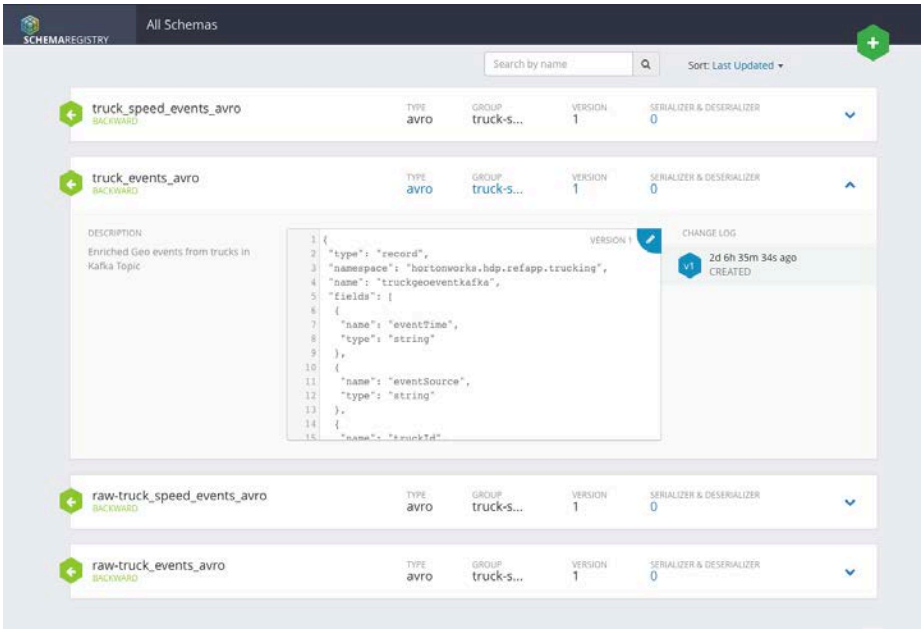

Figure 6: Schema Registry: Establish end-to-end data governance and improve operational efficiency

Conclusion

À mesure que les entreprises cherchent à exploiter toute la valeur stratégique de leurs données, celles-ci ont besoin d'une plateforme simple, efficace et évolutive pour créer des applications d'analyse en temps réel et extraire des informations utiles de leurs données en mouvement.

Hortonworks DataFlow permet aux développeurs de créer des applications d'analyse en temps réel sans avoir à écrire une seule ligne de code, éliminant ainsi les processus fastidieux qui dépendent de compétences spécialisées.

Les analyses d'affaires peuvent, au quotidien, explorer facilement des données, effectuer des analyses descriptives, créer des tableaux de bord et générer de riches visualisations pour obtenir des informations utiles. Les équipes de l'exploitation informatique peuvent contribuer plus facilement au développement des applications afin de fournir et de gérer les services sur lesquels les applications d'analyse en temps réel seront exécutées. Avec les composants Streaming Analytics Manager et Schema Registry HDF 3.0, les entreprises peuvent désormais parachever leur stratégie Big Data en créant la prochaine génération d'applications de lecture de données en temps réel.

Veuillez vous rendre sur la page www.hortonworks.com/hdf pour obtenir de plus amples informations.

À propos d'Hortonworks

Hortonworks est l'entreprise la plus innovante du secteur. Elle crée, distribue et soutient des plateformes de données connectées et des applications de données modernes ouvertes et prêtes à l'emploi dans l'entreprise. Ces outils exploitent tous types de données (données au repos et en mouvement) pour vous fournir des informations utiles. Hortonworks se concentre sur la promotion de l'innovation au sein de communautés open source telles que Apache Hadoop, Apache NiFi et Apache Spark. Avec plus de 1 800 partenaires, Hortonworks fournit le savoir-faire, la formation et les services permettant aux clients de débloquer de la valeur transformationelle pour tous les secteurs d'activité de leur entreprise.

Contact

Pour en savoir plus, consultez
www.hortonworks.com

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

HORTONWORKS

© 2011-2017 Hortonworks Inc. Tous droits réservés.
Politique de confidentialité | Conditions générales