StreamSets

PAPIER BLANC

Données modernes

Intégration pour DataOps

Apporter vitesse, flexibilité, résilience et fiabilité à l'analyse

Aperçu

Les entreprises modernes doivent pouvoir pivoter sur un sou dans le monde d'aujourd'hui. Les données sont la seule chose qui peut nous dire ce qui est arrivé à nos canaux de vente et de marketing, à nos opérations et à la sécurité en ligne. Il nous permet de prendre des décisions éclairées et en temps réel dans un monde en constante évolution. Les données alimentent la science des données et l'apprentissage automatique qui nous permettent de mieux prédire les résultats futurs.

Afin de profiter des avantages de ces informations analytiques, vous devez d'abord concevoir la livraison des données à un nombre croissant d'équipes d'analyse, toutes avec des exigences différentes en matière de données . Ajoutant à la complexité, les écosystèmes de données d'entreprise et de cloud changent tout le temps, brisant les pipelines de données et orientant mal les informations.

Opérations de données est un ensemble de pratiques et de technologies qui opérationnalisent la gestion et l'intégration des données pour assurer la résilience et l'agilité malgré les changements constants. Il combine les principes DevOps de livraison continue avec la capacité d'exploiter **dérive des données** (modifications inattendues et non documentées des données) .

DataOps n'est pas seulement une plate-forme technologique, cela nécessite un changement d'état d'esprit et souvent un changement dans la façon dont vous mettez en place des équipes et des processus. Vous devez penser différemment la chaîne de valeur de la gestion et de l'intégration des données. Si vous pouvez faire évoluer les mentalités, DataOps fournit les données continues nécessaires pour conduire des analyses modernes et une transformation numérique .

Le cœur de la mise en œuvre de la pratique DataOps est un **plate-forme d'intégration de données moderne** qui fournit aux utilisateurs **vitesse, flexibilité, résilience, et fiabilité** . Les systèmes DataOps s'adaptent au changement en permettant à leurs utilisateurs d'adopter et de comprendre facilement de nouvelles plates- formes complexes afin de fournir les fonctionnalités commerciales dont ils ont besoin pour rester compétitifs.

Qu'est-ce que DataOps ?

Dans l'environnement difficile d'aujourd'hui, fournir des données pour prendre des décisions aussi rapidement que possible est plus critique que jamais. La seule façon de suivre le rythme de tous les changements qui se produisent dans le monde au sens large, dans votre organisation et à travers toutes les données dont vous avez besoin pour prendre des décisions estOpérations de données. L'analyse moderne nécessite la libre circulation des données à tous les niveaux de l'entreprise pour vous permettre de prendre les meilleures décisions possibles à tout moment, en utilisant les informations les plus récentes. DataOps peut aider à réduire le temps de cycle de l'analyse des données en alignement étroit avec les objectifs de l'entreprise.

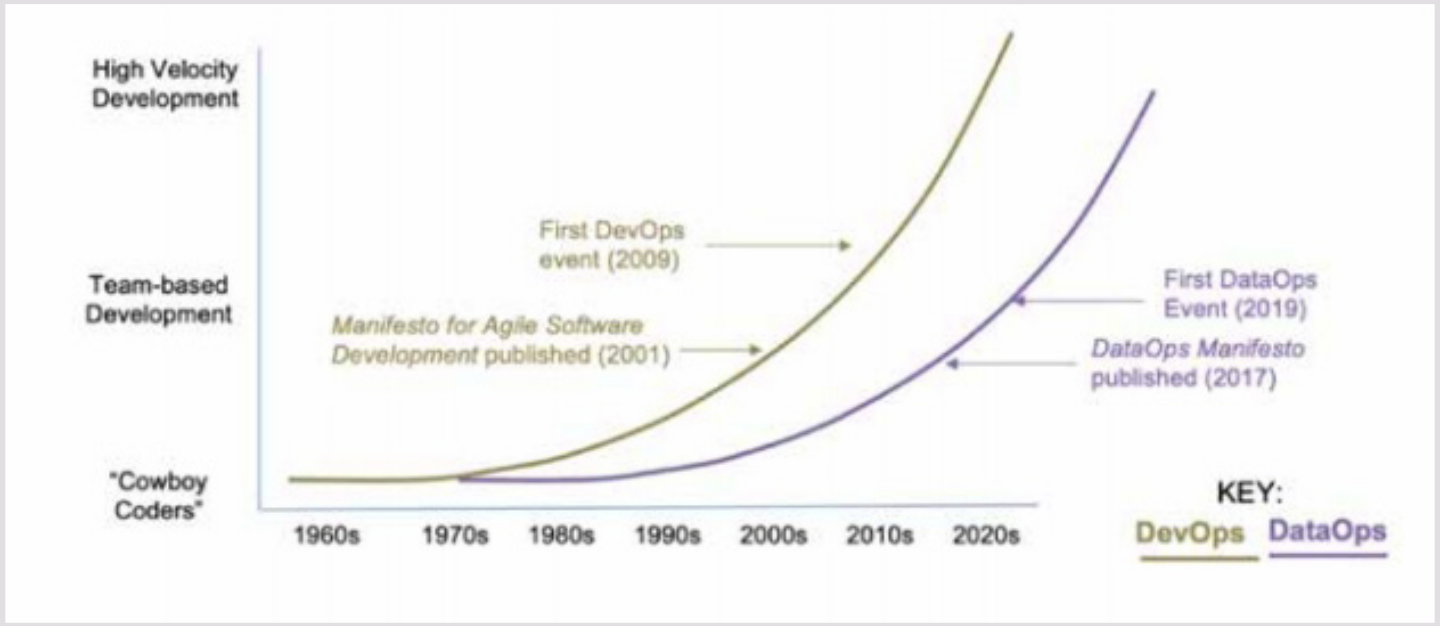
> *DataOps est un ensemble de pratiques et de technologies qui opérationnalisent la gestion et l'intégration des données pour assurer la résilience et l'agilité malgré les changements constants. Il combine les principes DevOps de livraison continue avec la capacité d'exploiter la dérive des données (modifications inattendues et non documentées des données).*

DataOps est la nouvelle façon de penser le travail avec les données, il offre aux praticiens (architectes, développeurs) la possibilité d'intégrer et de faire évoluer rapidement des projets de données tout en donnant aux opérateurs et aux dirigeants la visibilité et l'assurance que les moteurs sous-jacents fonctionnent bien. Il s'agit d'un changement d'état d'esprit fondamental qui nécessite des changements au niveau des personnes, des processus et des technologies de support. DataOps n'est pas seulement la façon dont vous modifiez votre stratégie de données actuelle, mais également la façon dont vous prévoyez des changements imprévus dans la stratégie. Il s'inspire de la philosophie d'agilité que DevOps a apportée au cycle de vie du développement logiciel et l'adapte aux défis uniques de travailler avec des données modernes.

Cependant, DataOps n'est pas seulement DevOps pour les données. Là où le plan de travail d'application d'entreprise est assez statique tout au long du cycle de développement et de déploiement, DataOps est de plus en plus complexe en raison du taux de changement constant appelé **Dérive des données**. DataOps est complémentaire et cohabite avec des philosophies telles que MLOps (gestion du cycle de vie complet du machine learning). Il permet aux praticiens d'alimenter en continu et de manière fiable les données dans les systèmes d'analyse et d'apprentissage automatique.

DataOps gagne du terrain depuis son entrée dans le Cycle de battage médiatique de Gartner en 2018. Eckerson Group a suivi l'émergence des DataOps dans **leschéma ci-dessous**. Le premier de l'industrie Guide des opérations de données, astuces et tendances en DataOps, et le tout premier Sommet DataOps ont tous été dévoilés en 2019. Également à la fin de 2019, John Schmidt et Kirit Basu ont publié DataOps : la première édition faisant autorité. Bien qu'il s'agisse toujours d'une pratique en croissance, l'accent et le développement des meilleures pratiques au cours de la dernière année ont montré des progrès notables. Alors que de nombreuses entreprises commencent à formaliser leur équipe et leurs processus DataOps, de nouveaux rôles et des architectures standardisées renforceront davantage l'adoption de DataOps par l'entreprise.

Figure 1. La trajectoire de DevOps et DataOps



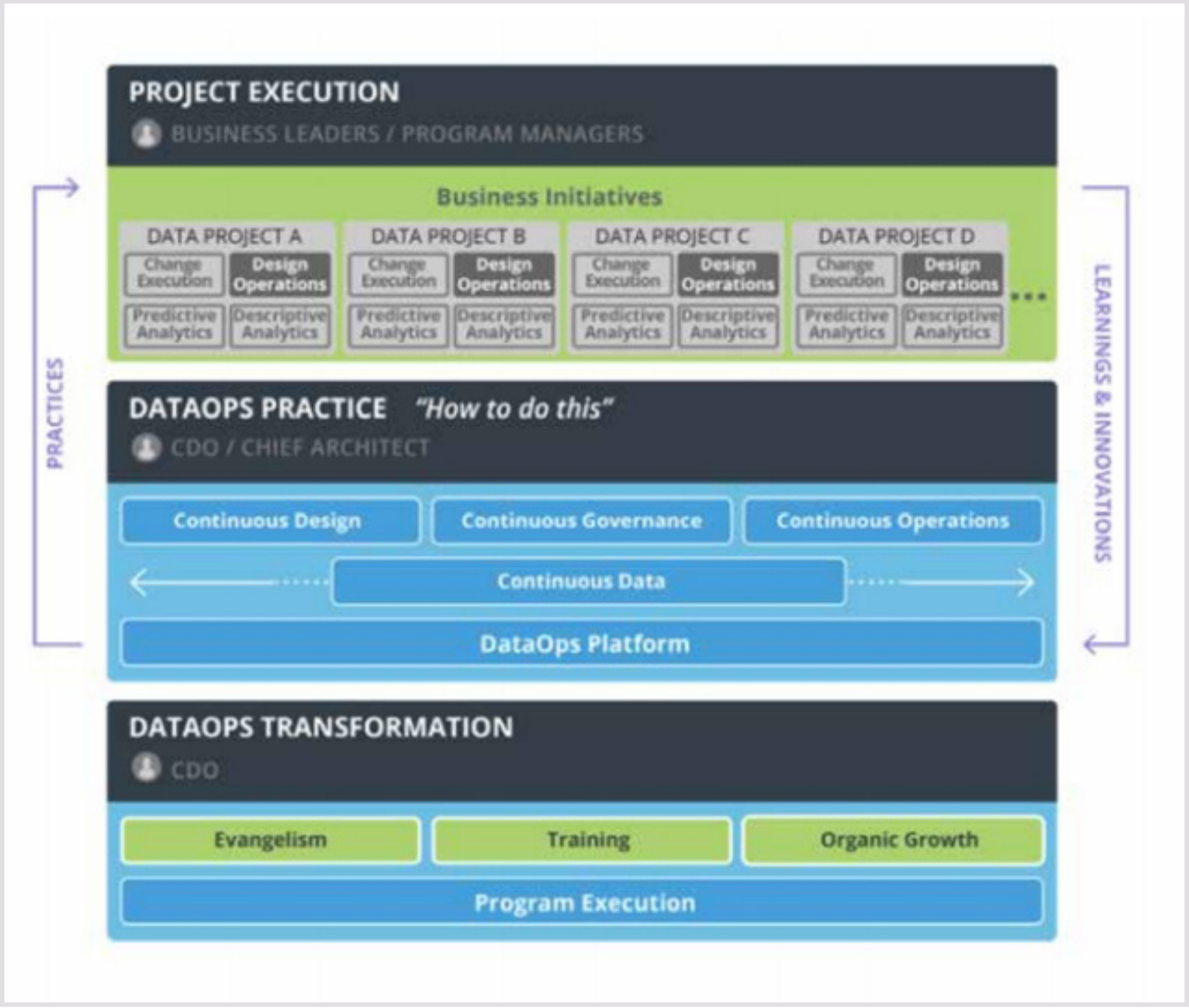
4 façons dont vous devez penser différemment : l'état d'esprit DataOps

DataOps peut fournir de nombreuses optimisations et accélérer la valeur à une variété d'initiatives d'analyse. Cependant, les processus et plates-formes hérités n'offrent pas la vitesse, la résilience et la flexibilité nécessaires pour effectuer une intégration de données moderne, ce qui peut vous aider à mieux tirer parti des sources de données en continu et non structurées. Lorsque vous déterminerez comment mettre en œuvre et mesurer les DataOps, vous devrez changer fondamentalement votre façon de penser à ces quatre choses.

- **En libre service** : Quels aspects du cycle de vie des données peuvent être automatisés ? Pouvons-nous donner aux utilisateurs finaux plus de pouvoir pour trouver, comprendre, ingérer, manipuler et transformer les données sans sacrifier la sécurité et la gouvernance ? L'ancien monde où les utilisateurs professionnels déposaient des tickets de demande et attendaient des semaines que le service informatique mette en œuvre leurs objectifs n'est plus acceptable. L'accès aux données en libre-service doit être mis en œuvre chaque fois que cela est sûr et possible. Cela donne à votre entreprise la vitesse et la flexibilité nécessaires pour fournir des activités d'analyse parallèles.
- **Visibilité et échelle** : Cela signifie une visibilité qui couvre les systèmes et les ensembles de données. Visibilité sur vos pipelines, vos transformations et ce que les autres utilisateurs font avec les données . DataOps vous aide à découvrir des modèles de conception et à mettre en œuvre des optimisations basées sur des modèles d'utilisation. Cette visibilité globale garantit que les systèmes de données évoluent pour répondre aux besoins actuels et futurs.
- **Gouvernance en temps réel** : Fini le temps de la mise en œuvre de la gouvernance avec de bonnes intentions. Des réglementations coûteuses, telles que le RGPD et le CCPA, et le fait que les données doivent passer par plusieurs systèmes avec des degrés variables de visibilité et de gouvernance. Ces lacunes dans la gouvernance introduisent de nouveaux domaines de risque. DataOps nécessite une approche de gouvernance des données continue où les politiques sont appliquées au moment de l'exécution ainsi que dans des zones d'analyse de confiance, y compris lorsque les données sont en mouvement, plutôt que par le biais de processus de gouvernance longs et onéreux qui entravent l'agilité de l'entreprise.
- **Concevoir pour le changement** : Les choses changent tout le temps : de nouveaux ensembles de données et schémas sont introduits, ceux qui existent sont modifiés ou utilisés de manières nouvelles et inattendues, les modèles de ML dérivent, les niveaux de qualité des données

fluctuent, les responsabilités de propriété et de gestion changent. L'objectif de DataOps est de détecter et de gérer les changements afin que les choses ne se cassent pas. Cela nécessite un degré élevé de surveillance, une détection sophistiquée des modifications et une gestion automatisée des modifications. Les architectures et les processus DataOps sont résilients aux changements inévitables et offrent une fiabilité à mesure que vous évoluez.

L'exécution de projets de données n'est pas seulement une question de garantir la réussite de la livraison des données, DataOps est une philosophie qui met le praticien au défi d'essayer constamment d'atteindre un état de tout en continu avec tous les aspects des données.



Continu Tout

Le cœur du changement de mentalité est un pivot pour penser les données non pas comme une ressource statique mais comme un processus continu. DataOps comprend quatre principes continus qui sont pris en compte dans la conception, le déploiement et l'exploitation des systèmes de données modernes.

- **Conception continue :** La fonction de conception continue permet de fournir des solutions de données sur une base continue plutôt que sous forme d'événements de projet discrets.
- **Opérations continues :** DataOps encourage une vue holistique où les opérateurs sont en mesure de voir une carte vivante de toutes les données fonctionnant ensemble pour répondre aux besoins en données des fonctions commerciales d'ordre supérieur.
- **Gouvernance continue :** La gouvernance continue est chargée d'établir un cadre de gouvernance de l'information, une méthodologie et des normes pour la gestion de l'information d'entreprise .
- **Données continues :** Les données continues sont responsables de la publication des données dans un hub unifié, maintenant les services de données, les niveaux de service et les performances pour les informations de source externe et générées en interne et leur utilisation par les utilisateurs ou les systèmes d'application.

En intégrant ces principes, DataOps permet de fournir les données continues nécessaires pour conduire des analyses modernes et une transformation numérique.

Données continues

Les systèmes d'entreprise modernes doivent fonctionner avec une très grande variété de données : les capteurs IoT émettent des lectures, les applications Web produisent des événements et des messages, les clients ou partenaires envoient des fichiers binaires et toutes les saveurs des systèmes de base de données (relationnelle, nosql, graphique, séries temporelles, etc.) contiennent différents types de données. Les informations ou les décisions en temps réel générées par les systèmes opérationnels et analytiques sont alimentées par des événements et des données provenant de ces systèmes variés ainsi que par des données historiques et contextuelles .

Les données continues convergent vers les paradigmes du traitement par flux et par lots. Le streaming de données est impératif pour les analyses prédictives et préventives en temps réel. Les processus par lots sont nécessaires pour enrichir ou fournir des données de formation pour ces systèmes analytiques prospectifs, ainsi que pour conduire des analyses descriptives critiques pour l'entreprise.

Conception continue

Les architectures de données modernes sont souvent très distribuées et complexes. Les applications et les cas d'utilisation construits avec un certain nombre de ces systèmes complexes présentent un degré élevé d'interconnexion.

La conception continue est un paradigme où l'intégration des données est toujours effectuée en contexte. La « vue d'ensemble » est gardée en vue et est souvent le point de départ du processus de conception . Les diagrammes d'architecture de données ne sont pas seulement des images sur un jeu de diapositives, mais la carte vivante de ce qui est réel dans le système. Lorsqu'un changement se produit dans les systèmes sous-jacents, par ex . les sources de données ou les destinations sont ajoutées ou supprimées, cela est immédiatement reflété dans cette carte vivante. Un pipeline de flux de données conçu pour déplacer des données d'une source à une destination apparaît dans cette carte et est globalement visible pour toute personne interagissant avec l'environnement. Lorsque le prochain projet souhaite accéder aux mêmes données, les développeurs se tournent vers la carte dynamique ou la topologie pour réutiliser les pipelines existants .

Opérations continues

Les grands flux de données interconnectés sont souvent surveillés de manière isolée ; un opérateur surveillant un pipeline dans une zone de l'application ne saura pas comment il se rapporte à une autre zone de l'application. Les effets en cascade d'une défaillance sont impossibles à comprendre lorsque la seule vue des opérateurs est une liste tabulée de pipelines individuels .

DataOps encourage une vue holistique de l'ensemble de l'architecture. Les opérateurs qui sont en mesure de voir une carte vivante de tous les pipelines travaillant ensemble pour répondre aux besoins de données de la fonction commerciale d'ordre supérieur sont en mesure de mieux gérer le système dans son ensemble. Pourtant, ils doivent toujours être en mesure d'explorer les points problématiques selon les besoins.

Gouvernance continue

Les architectures modernes sont souvent hybrides. Les données proviennent, sont traitées ou stockées dans un nombre illimité de systèmes - à la périphérie, au sein du centre de données ou dans le cloud. Les données effectuent souvent plusieurs sauts, étant traitées à la volée ou dans des systèmes de calcul éphémères au fur et à mesure de leur déplacement. Les solutions de lignage de données traditionnelles ont été conçues pour les flux de données point à point et ne sont pas en mesure de capturer ou d'interpréter le lignage de bout en bout . Une pratique DataOps utiliserait des systèmes conçus pour ces architectures et s'attendrait à ce que les données elles- mêmes fournissent des métadonnées fines et un lignage de bout en bout sur elles-mêmes.

DataOps appelle un cadre de vigilance et de protection constante. La gouvernance continue consiste à définir une politique de sécurité et à l'appliquer automatiquement chaque fois que des données circulent dans l'entreprise.

La promesse de tout en continu en tant que principe de conception et de mise en œuvre de base nous donne la latitude d'utiliser les données de la manière dont les différentes équipes ont besoin pour véritablement activer une culture de libre-service et de DataOps. Une fois cette solution continue créée, l'automatisation peut être utilisée pour garantir que les données continues sont quelque chose que l'entreprise peut se sentir en confiance pour garantir des accords de niveau de service à ses parties prenantes .

Alors si tout en continu est si idéal, pourquoi de nombreuses entreprises sont-elles encore si loin de réaliser cet état ? La vérité est qu'il existe des complexités cachées à l'exécution d'une conception et d'un développement continus qui sont liées aux limitations des approches traditionnelles de la gestion et de l'intégration des données.

La complexité cachée
des opérations de données

Alors que les entreprises ciblent et acquièrent de nouvelles sources de données et visent à fournir de nouveaux facteurs de forme d'analyse de données, une bonne quantité de capital créatif est dépensée pour concevoir ces nouveaux modèles et solutions . Sans surprise, de riches outils ont vu le jour qui donnent un contrôle visuel sur les analyses et, plus récemment, sur le mouvement des données . Les ingénieurs de données disposent désormais d'outils très intuitifs pour contrôler l'intégration des données dans leur entreprise. Un ingénieur avisé peut énormément renforcer sa marque interne en réalisant une nouvelle capacité avant-gardiste. Cependant, comment peuvent-ils trouver le temps alors qu'ils sont si souvent embourbés dans la tâche de maintenir les idées existantes saines, modernes et en production ?

La vérité est qu'ils passent souvent beaucoup de temps à gérer les opérations qui assurent le fonctionnement et le fonctionnement des pipelines de données pour répondre aux exigences des projets en aval. Il est largement compris parmi les ingénieurs de données, mais pas largement reconnu, comme une activité qui sera récompensée par des éloges et des distinctions . Mais lorsque les choses tournent mal, la douleur ressentie par les ingénieurs de données est réelle. Qu'il s'agisse d'un modèle de science des données qu'un analyste a convaincu un ingénieur de prendre en charge ou du tableau de bord quotidien soutenant l'équipe de vente, lorsque ces capacités se brisent, l'amitié et un dossier impeccable de nouvelles idées ne vous gagneront que tant de grâces.

Cependant, les opérations sur les données ne doivent pas nécessairement être un jeu à somme nulle. Vous pouvez créer un système qui offre résilience et flexibilité, même à grande échelle. Réfléchir intelligemment à la façon dont vous évoluez et réagissez aux besoins opérationnels de vos pipelines de données et de votre traitement de données peut être fastidieux au début, mais cela portera ses fruits à mesure que les charges de travail et les projets de données s'accumuleront.

> *« Dans un monde de données statiques, la productivité initiale des développeurs compte plus que les opérations. Dans un monde de données en continu, les opérations sont primordiales.*
>
> *- Kirit Basu, auteur de DataOps : l'édition faisant autorité*

Alors, comment créons-nous des opérations pour un monde de données en continu ? Les composants suivants sont essentiels pour fournir la fonctionnalité DataOps .

Automatisation

À mesure que les entreprises évoluent et accélèrent leur pratique des données, l'automatisation et l'intégration avec les outils d'automatisation deviennent primordiales pour évoluer rapidement. Aucun aspect du libre-service ne peut être fourni sans un certain niveau d'automatisation. Lorsque les ingénieurs sont capables de travailler sur des tâches d'automatisation, l'impact peut être amplifié sur plusieurs charges de travail. Dans l'écosystème DataOps, les utilisateurs voudront automatiser tout ce qu'ils peuvent, tout en restant fiables. Les systèmes propriétaires offrent souvent une faible extensibilité, ce qui les rend compliqués à automatiser et à intégrer aux outils d'automatisation. C'est pourquoi DataOps se concentre souvent sur des solutions ou plates- formes ouvertes qui offrent une extensibilité des API et une intégration programmable avec l'infrastructure et les plates-formes informatiques.

Visibilité

Considérez un pipeline de données unique comme un onglet sur votre navigateur Internet. À petite échelle, le basculement entre les onglets de votre navigateur est relativement gérable (mais pas idéal) . Selon la taille de votre organisation, vous disposez probablement de plusieurs outils (anciens et modernes) pour créer des pipelines de données. Ces outils auront tous un degré de contrôle et de granularité différent pour comprendre à la fois la progression et la santé de vos pipelines de données. Les équipes de données passent souvent une bonne partie de leur temps à gérer les limites des outils.

Mais qu'en est-il lorsque vous avez 30, 100, 1K pipelines ? Plonger dans 100 onglets de navigateur Internet avec différents degrés d'utilité ne produira probablement que des retards et causera une quantité croissante de douleur à mesure que les connexions se développent. Avec DataOps, l'objectif est d'avoir une carte complète et vivante de tous vos travaux de mouvement et de traitement de données. Lorsque des pipelines ou des étapes d'un pipeline tombent en panne, les erreurs se regroupent en un seul point de correction visuelle . De cette façon, les ingénieurs de données comprennent le problème et réagissent de manière à ne pas nuire aux projets en aval et à porter un coup à la marque de l'ingénieur. Cette carte doit être aussi utile et réactive pour les charges de travail de données d'aujourd'hui que pour gérer les charges de travail de demain.

Surveillance

La surveillance active est un élément important de la résilience et de la fiabilité. De nombreux outils visent à surveiller les opérations au sein de leur portefeuille de produits, mais DataOps exige un niveau de surveillance qui persiste en dehors d'un système ou d'une charge de travail unique . La surveillance continue nécessite que les systèmes partagent des métadonnées et des informations opérationnelles afin que les utilisateurs puissent voir un plus large éventail de défis. Ces capacités de surveillance aident également les entreprises à définir, affiner et fournir des SLA de données en aval. Cela garantit que les analyses peuvent être fournies en toute confiance et que l'entreprise peut faire évoluer l'art du libre-service, ce qui éloigne davantage l'ingénieur de données du risque. La surveillance ne doit pas seulement s'attaquer à alerter une équipe en cas de panne, mais, dans un scénario DataOps, elle doit activement surveiller les précurseurs de problèmes potentiels.

Gérer un paysage en évolution

Quel est le coût d'un système de données rigide et fragile ? Cela vous coûtera-t-il votre position concurrentielle? Le paysage des données d'aujourd'hui évolue à un rythme effréné et le tribut que cela représente pour les professionnels des données est impitoyable . Votre stratégie de données doit non seulement être compétitive par rapport aux exigences d'aujourd'hui, mais également être tournée vers l'avenir pour envisager la transformation de votre entreprise au cours des cinq à dix prochaines années.

La gestion et l'évaluation de ce paysage en évolution rapide nécessitent une flexibilité, architecturée sur des outils et des plates-formes qui peuvent s'affranchir de la dépendance à une plate-forme de données ou à une solution d'analyse unique. La logique de conception des pipelines doit être transférable, quelles que soient la source et la destination, permettant un changement en fonction des besoins de l'entreprise plutôt que de gérer les limites du système. Dans DataOps, le changement est une évidence. Les systèmes DataOps s'adaptent au changement en permettant à leurs utilisateurs d'adopter et de comprendre facilement de nouvelles plateformes complexes afin de fournir les fonctionnalités commerciales dont ils ont besoin pour rester compétitifs. DataOps adopte le cloud et aide les entreprises à créer des solutions cloud hybrides qui pourraient un jour vivre nativement dans le cloud.

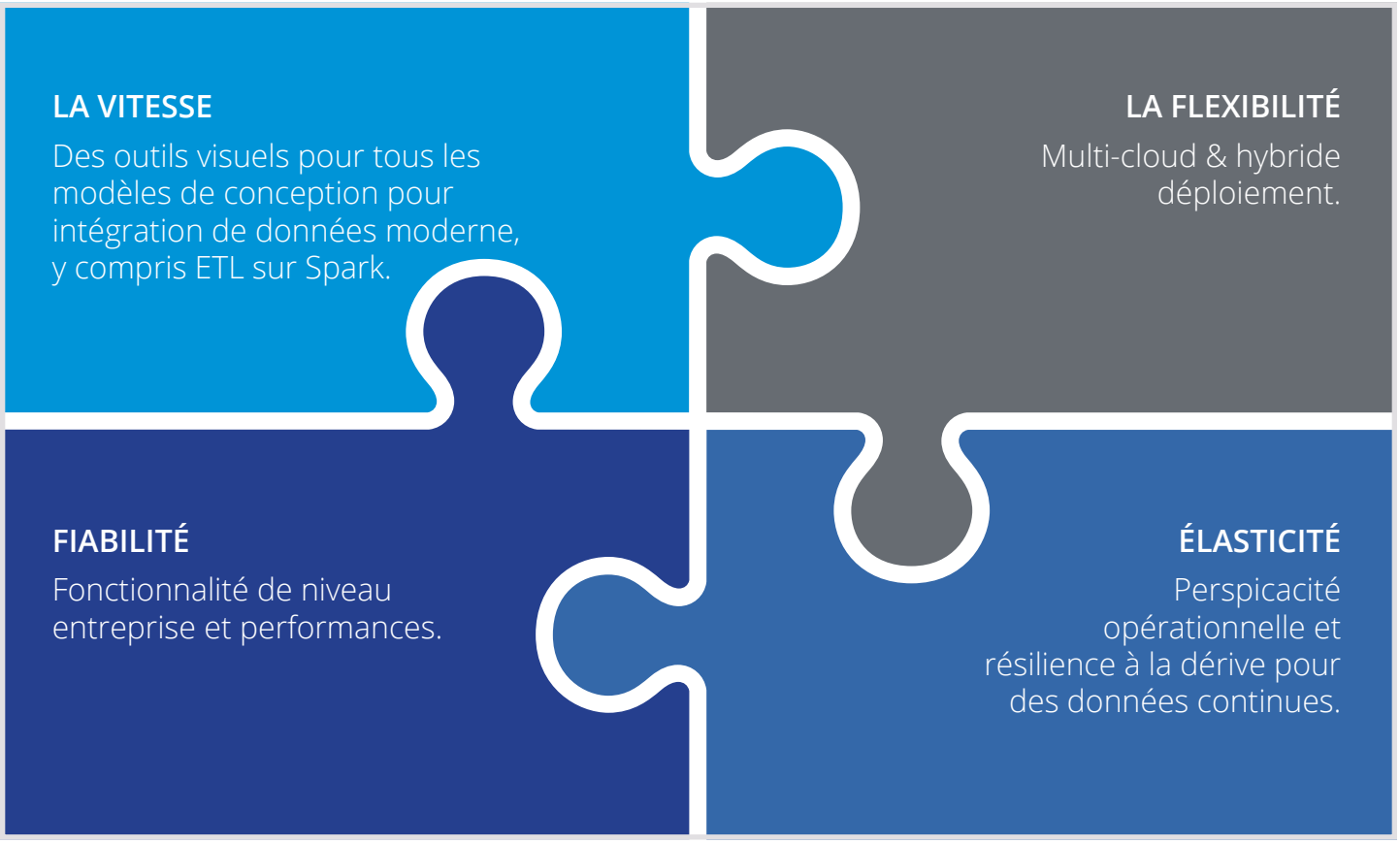
L'architecture du changement

Les données ne sont plus une ressource statique. Les valeurs des données mutent au fil du temps, la sémantique dérive et les architectures évoluent . Un processus ou une plate-forme qui suppose que les données agiront de la même manière chaque fois qu'elles sont déplacées ou traitées est vouée à l'échec. Étonnamment, le meilleur scénario possible est qu'un pipeline ou un processus de données se brise, l'effet de ce petit changement peut faire des ravages sur les systèmes en aval. Cependant, parfois, le pipeline ne se brise pas et ne fait finalement que corrompre les données en aval et tous ceux qui travaillent sur ces données sont parfaitement inconscients. Pendant des années, les professionnels des données ont géré la dérive des données dans le but d'atténuer l'impact sur les équipes en aval. DataOps embrasse la dérive des données et suppose que les données, la sémantique et l'infrastructure vont changer . Les pipelines et les processus DataOps sont faiblement couplés et flexibles aux variations de schéma.

Surmonter la complexité croissante des opérations de données dépend fortement du choix d'outils d'intégration et d'ingestion qui offrent le même degré de vitesse, de flexibilité, de fiabilité et de résilience qui est essentiel pour faire évoluer cet état d'esprit clé.

Une intégration de données moderne

Plateforme pour les opérations de données



DataOps est une pratique qui implique des personnes, des processus et des technologies, et au cœur de la mise en œuvre des notions de tout en continu et de la construction d'une base pour votre pratique DataOps est une plate-forme d'intégration de données moderne. L'accent mis sur la modernité signifie un large éventail de formats de données et l'interopérabilité entre les systèmes d'entreprise est requise. Le système doit gérer à la fois la sémantique des flux et des lots et s'exécuter de manière optimale là où se trouvent vos données.

4 Considérations d'évaluation

Lors de l'évaluation d'une plate-forme de données moderne, vous devez prendre en compte les forces et les faiblesses d'une solution en fonction de sa capacité à offrir vitesse, flexibilité, fiabilité et résilience. Cette section identifie les fonctionnalités recommandées pour répondre à ces exigences.

La vitesse

La vitesse peut être mesurée à la fois en termes finis et utilisée comme terme pour décrire une accélération de haut niveau. Par exemple, pour un développeur, la vitesse peut signifier des outils faciles à utiliser pour concevoir et exploiter des pipelines de tous les modèles, y compris ETL sur Spark, l'ingestion de streaming et CDC. Ces capacités peuvent aider les développeurs à concevoir des pipelines en quelques minutes et à fonctionner en continu sans trop de difficultés.

Pour les responsables d'équipe de données, la vitesse peut être la possibilité d'utiliser un seul outil visuel pour tous les modèles de conception (streaming, CDC, batch) sans aucun codage requis, même sur Spark . Cela aiderait les équipes à réutiliser les artefacts et à appliquer leurs compétences à une grande variété de cas d'utilisation.

Pour l'informatique, la vitesse peut être un terme utilisé pour décrire la capacité à résoudre les problèmes. Vous apprécierez peut-être un outil unique pour l'exploitation de pipelines de données de tous types, sur toutes les plates-formes, qui fournirait une gestion simplifiée des opérations et réduirait les frais généraux opérationnels et le temps de cycle pour les nouveaux projets. Et pour un CDO ou un responsable exécutif des données, la vitesse peut signifier la possibilité de déployer le développement Apache Spark pour le plus grand nombre, en accélérant les projets d'IA et d'apprentissage automatique pour l'entreprise.

La flexibilité

La vitesse et la flexibilité multiplient souvent les forces qui peuvent avoir un impact important sur l'entreprise lorsqu'elles sont appliquées à travers les équipes et les rôles. Lors de la création de pipelines de données et de la planification de processus par lots et ETL volumineux, le traitement Apache Spark disponible n'importe où (sur site, cloud public ou cloud privé virtuel) donne aux concepteurs la possibilité de sélectionner la plate-forme offrant les meilleures performances.

Lorsque les outils ont une prise en charge prédéfinie des destinations cloud et s'exécutent de manière native sur les services managés Apache Spark, cela signifie que les équipes obtiennent une livraison plus rapide de leurs principaux cas d'utilisation. Un avantage supplémentaire est la portabilité et la réutilisation sur ces plates-formes, à la fois sur site et dans le cloud. Cela peut donner à l'équipe d'exploitation la possibilité de changer rapidement de plate-forme et d'éviter le

verrouillage du fournisseur. Pour l'entreprise, vous pouvez être sûr d'obtenir un support hybride et multi-cloud et garder des options ouvertes pour évoluer avec l'élan de votre secteur et de vos exigences. La flexibilité ultime est atteinte lorsque des événements imprévus ne sont pas en mesure d'entraver la directive principale de l'entreprise.

Élasticité

La résilience face au changement est l'objectif ultime des DataOps. Cependant, la résilience a beaucoup de complexité cachée. Les détails impliqués dans la construction de grands systèmes d'entreprise interconnectés qui sont vraiment résilients nécessitent beaucoup de réflexion et d'efforts coordonnés.

Pour que les pipelines de données soient résilients, ils doivent être entièrement instrumentés afin que les utilisateurs puissent comprendre ce qui se passe à chaque étape du pipeline . Ce niveau de visibilité et d'instrumentation permet un débogage plus rapide. Si la dérive des données est toujours présente, des pipelines résilients à la dérive et entièrement instrumentés devraient réduire les risques et faciliter le dépannage. Cela peut éviter les pannes des systèmes d'analyse.

Pour gérer les pipelines en continu, une vue opérationnelle des flux de données en temps réel donne à l'équipe informatique la possibilité d'avoir une vue d'ensemble et d'explorer en détail le cas échéant. Si, au niveau de l'entreprise, vous pouvez avoir une vue unique des flux de données, vous pouvez assurer la visibilité et le contrôle, réduisant encore les silos de données et les opérations redondantes. La véritable résilience nécessite une conception axée sur l'intention. Cela signifie que même si les sources et les destinations changent, l'intention du pipeline est conservée .

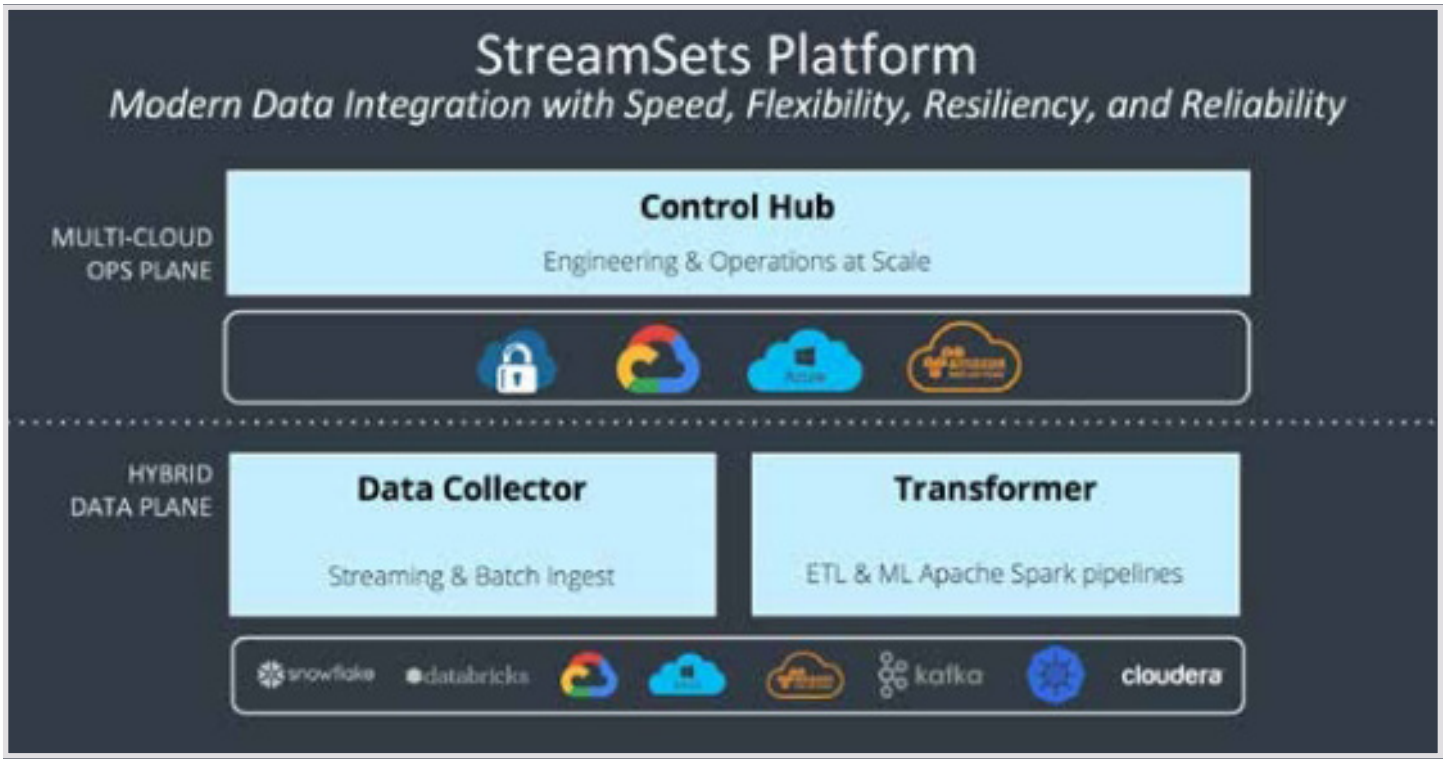
Fiabilité

La fourniture de fonctionnalités et de performances de niveau entreprise face à des changements constants est presque insurmontable. Trop souvent, les entreprises sont obligées de faire des compromis entre tirer parti de la solution la plus rapide à un problème (c. Afin d'atténuer le risque des nouvelles plates-formes, les ingénieurs de données doivent disposer de fonctionnalités avancées de niveau entreprise. Cela permet de réduire le temps de configuration, de réglage et de maintenance des pipelines de données. Les entreprises doivent rechercher des fournisseurs de

DataOps qui ont fait leurs preuves. Cela en fait une valeur sûre sur laquelle standardiser lorsque vous construisez et faites évoluer vos systèmes. La fiabilité signifie que vous pouvez dire « oui » au niveau exécutif aux nouvelles demandes de données, en étant confiant que les développeurs, les équipes de données et les équipes informatiques peuvent exécuter la vision.

La plateforme StreamSets

StreamSets a construit le premier Plateforme d'intégration de données moderne pour DataOps. La plate- forme est conçue pour répondre aux capacités detout en continu et permet aux entreprises de va vite et sois confiant. La plateforme offre aux professionnels des données vitesse, flexibilité, résilience et fiabilité au niveau du pipeline unique et sur l'ensemble de la fabrique de données d'entreprise. À mesure que les pipelines évoluent, ils sont entièrement instrumentés, offrant aux utilisateurs un degré élevé de visibilité et permettant aux utilisateurs de partager des composants de pipeline . Ils détectent activement lesDérive des données et leur conception axée sur l'intention signifie que lorsque les choses changent, les pipelines ne se cassent pas.



Au plus haut niveau, la plate-forme StreamSets DataOps est une plate-forme d'intégration de données moderne, combinant des moteurs d'exécution hautes performances et des outils visuels à cycle de vie complet pour concevoir, exploiter, gérer et optimiser les pipelines de données dans votre entreprise. Avec une instrumentation et une visibilité de bout en bout, la plate-forme fournit des informations opérationnelles en temps réel sur tous vos pipelines, où qu'ils se trouvent, sur site ou dans le cloud .

Les pipelines résilients à la dérive de la plate-forme réduisent le risque de pannes ou de données perte due à la dérive des données. Et la solution est conçue pour être indépendante de la plate-forme, permettant une portabilité facile des pipelines entre les plates-formes, que ce soit sur site ou dans le cloud, garantissant une flexibilité pour répondre aux besoins changeants. La plate-forme DataOps est une base technologique clé pour votre pratique DataOps, vous permettant de fournir rapidement des données de manière continue à l'entreprise, dans un monde en constante évolution.

La plateforme se compose de deux puissants moteurs d'exécution et d'un hub de gestion.

Collecteur de données StreamSets

StreamSets Data Collector est un moteur d'exécution moderne et facile à utiliser pour ingestion rapide des données et transformations légères. Les outils visuels faciles à utiliser de Data Collector vous permettent de concevoir, déployer et exploiter des pipelines de streaming, de capture de données modifiées (CDC) et de données par lots sans codage manuel. La gamme complète de sources de données telles que Kafka, S3, Snowflake, Databricks, JDBC, Hive, Salesforce, Oracle et bien d'autres sont disponibles prêtes à l'emploi. Des pipelines de données « intelligents » entièrement instrumentés vous permettent de surveiller les données en vol et sont conçus pour gérer la dérive des données avec une détection et une gestion intégrées. Les pipelines Data Collector sont conçus pour être indépendants de la plate-forme, vous pouvez donc vous adapter au besoin et éviter le verrouillage du fournisseur.

Data Collector donne aux utilisateurs la possibilité de charger rapidement des données dans les systèmes d'entreprise avec des sources et des destinations prédéfinies et une interface graphique intuitive qui leur permet de concevoir des pipelines de données en quelques minutes. Les pipelines

peuvent être portés sur plusieurs environnements sur site ou dans le cloud sans refonte lourde. Les pipelines sont par défaut des pipelines « intelligents » résilients à la dérive, entièrement instrumentés, aidant les utilisateurs à réduire les risques avec un dépannage plus facile et moins de pannes. Data Collector équilibre les fonctionnalités et les performances de l'entreprise dans un seul outil pour tous les types de modèles d'ingestion.

Transformateur StreamSets

StreamSets Transformer est un moteur de transformation moderne conçu pour tout développeur ou ingénieur de données pour créer des transformations de données qui s'exécutent sur Apache Spark. À l'aide d'une interface utilisateur simple par glisser-déposer, les utilisateurs peuvent créer des pipelines pour effectuer des opérations ETL, de traitement de flux et d'apprentissage automatique. Il permet à tout le monde, pas seulement au développeur Spark averti, d'utiliser pleinement la puissance d'Apache Spark sans avoir à coder dans Scala ou PySpark. Les pipelines Transformer sont instrumentés pour offrir une visibilité inégalée sur l'exécution des applications Spark avec des aperçus intégrés et un dépannage facile. Et Transformer est conçu pour fonctionner sur toutes les principales distributions Spark afin de vous garantir la flexibilité de fonctionner sur la plate-forme de votre choix ou de changer de plate-forme lorsque vos besoins changent.

Transformer permet de mettre Apache Spark entre les mains de tout type d'ingénieur de données. Grâce à Transformer, les entreprises peuvent accélérer leur adoption d'Apache Spark sans investissements lourds en compétences qui ne nécessitent aucune montée en puissance pour les projets ETL et d'apprentissage automatique.

Transformer s'exécute là où se trouvent vos données et peut s'exécuter de manière native sur des plates-formes telles que Hadoop YARN, EMR, HDInsight, Databricks et dans des environnements Spark conteneurisés tels que Microsoft SQL Server 2019 Big Data Cluster et Kubernetes Cluster. Le moteur offre une visibilité approfondie sur l'exécution de Spark, permettant aux utilisateurs de dépanner au niveau du pipeline et à chaque étape de la progression du pipeline. Transformer offre aux utilisateurs les fonctionnalités d'entreprise et l'agilité qu'ils obtiennent des outils ETL hérités, tout en révélant toute la puissance et les opportunités d'Apache Spark .

Les pipelines StreamSets Data Collector et Transformer peuvent être déployés sur site, sur des clouds publics et sur des services managés dans le cloud. Toutes les instances de moteur peuvent être déployées, gérées et affichées directement dans le hub central pour un contrôle à l'échelle du système. Ce hub organise les pipelines en cartes graphiques appelées topologies .

Centre de contrôle

StreamSets Control Hub est un hub unique pour la conception, le déploiement, la surveillance, la gestion et l'optimisation de tous vos pipelines de données et tâches de traitement de données. Le système nerveux central de la plate-forme DataOps, Control Hub permet à votre équipe de collaborer pour gérer les pipelines de données et les travaux exécutés sur Data Collector et Transformer, permet la réutilisation des pipelines et vous offre une vue de bout en bout en temps réel de tous flux de données dans votre entreprise.

Control Hub simplifie et centralise également la gestion des moteurs StreamSets Data Collector et Transformer eux-mêmes pour optimiser votre environnement StreamSets global. Et enfin,

l'architecture hybride/multi-cloud de Control Hub fournit une surveillance et une gestion centralisées des sources de données et des plates-formes sur site et dans le cloud, afin que vous puissiez ajouter ou modifier des sources de données ou des plates-formes de données sans perdre en visibilité ou en contrôle.

Control Hub fournit une console unifiée pour la collaboration et la visibilité à travers toutes les étapes du cycle de vie, tous les modèles de conception, tous les moteurs. Cela permet d'accélérer le développement grâce à une utilisation agile et à la réutilisation des compétences et des actifs. Le hub offre une surveillance et une gestion centralisées des sources de données et des plates-formes sur site et dans le cloud, et vous offre la possibilité d'ajouter ou de modifier des sources ou des plates-formes sans perdre en visibilité ou en contrôle. Les cartes de données en direct appelées topologies donnent des informations opérationnelles en temps réel pour réduire les risques avec une large visibilité pour détecter et prévenir les problèmes.

StreamSets Data Collector, StreamSets Transformer et Control Hub travaillent de concert pour répondre aux exigences de DataOps à une grande variété de cas d'utilisation d'entreprise courants, y compris le traitement de flux, l'ingestion, l'ETL, l'apprentissage automatique et l'alimentation d'applications en temps réel.

Conclusion

DataOps est une pratique bien préparée pour devenir le modèle de la façon dont les entreprises avant-gardistes conçoivent des systèmes de données qui gèrent le changement et la complexité des écosystèmes en évolution. En alignement étroit avec les objectifs commerciaux, ces pratiques évoluent pour répondre aux besoins d'aujourd'hui et évoluent pour répondre aux besoins de la prochaine plate-forme, de l'environnement d'analyse et des tendances du secteur. Lors de la création d'un système DataOps, vous devez permettre une conception continue, des données continues, des opérations continues et une gouvernance continue . Une plate-forme d'intégration de données moderne conçue pour DataOps doit fournir ces capacités et donner à votre entreprise vitesse, flexibilité, résilience et fiabilité.

À Propos Des Streamets

StreamSets a construit le premier multi-cloud de l'industrie Plate-forme DataOps pour une intégration de données moderne, aidant les entreprises à transférer en continu des données volumineuses, en continu et traditionnelles vers leurs applications de science des données et d'analyse de données. La plate-forme gère de manière unique la dérive des données, ces modifications fréquentes et inat-tendues des données en amont qui cassent les pipelines et endom-magent l'intégrité des données. LePlateforme DataOps StreamSets permet l'exécution de pipelines n'importe où, le traitement ETL et l'apprentissage automatique avec un portail d'opérations natif du cloud pour l'automatisation et la surveillance continues de topolo-gies multi-pipelines complexes.

Essayez Maintenant

Soyez opérationnel avec StreamSets en quelques minutes . Rendez-nous visite à :
www.streamsets.com