Informatica

Livre blanc

Le guide du directeur des données pour une gestion intelligente des Data Lakes

Neuf principes pour fournir des informations exactes et cohérentes

À PROPOS D'INFORMATICA

La transformation digitale fait évoluer nos attentes : un service amélioré, des livraisons plus rapides, davantage de convivialité, le tout à moindre coût. Les entreprises doivent évoluer pour rester compétitives. Bonne nouvelle : les données sont la clé de la réussite.

En tant que leader mondial dans la gestion des données Cloud d'entreprise, nous sommes prêts à vous guider de manière intelligente, quel que soit le secteur, la catégorie ou la niche. Notre but est de vous permettre de gagner en flexibilité, de concrétiser des opportunités de croissance ou d'innover. Nous nous concentrons sur les données afin de vous offrir la polyvalence nécessaire pour réussir.

Plus de 7 000 entreprises du monde entier font appel aux solutions de données d'Informatica.

Table des matières

Rapport de synthèse 4

La promesse d'une gestion intelligente des Data Lakes 5

Pourquoi la gestion de données est importante pour votre Data Lake 5

Comment les Data Lakes apportent une valeur métier ajoutée en pratique 6

Les neuf principes de conception fondamentaux pour la gestion des Data Lakes 8

Conclusion 12

Rapport de synthèse

Les entreprises qui révolutionnent leur secteur et les leaders progressistes en matière de données s'appuient sur les Data Lakes pour obtenir de puissantes informations métiers et garantir des résultats en matière de transformation digitale axée sur les données. Grâce à l'efficacité et à l'évolutivité des environnements de Data Lake, votre entreprise peut rapidement obtenir de nouvelles informations et rendre possible ce qui n'avait jamais existé auparavant.

De nouvelles capacités technologiques et pratiques organisationnelles, également connues sous le nom de gestion intelligente des Data Lakes, constituent la base permettant de maximiser la valeur des Data Lakes. En abordant la gestion de données des Big Data de manière systématique, vous gérez un projet de Data Lake entièrement automatisé avec beaucoup moins de ressources manuelles. Une approche systématique de la gestion de données vous garantit non seulement des informations cohérentes et exactes, mais permet également aux entreprises de gagner le capital politique nécessaire pour se voir allouer des budgets et étendre la portée de vos efforts.

Il est temps de se remettre en question et de repenser votre point de vue concernant les Big Data. Ce livre blanc fournit aux directeurs des données et autres leaders de la gestion de données, comme vous, les conseils utiles pour tirer le meilleur parti des Data Lakes. Nous vous présentons les principes fondamentaux de conception pour gérer les Data Lakes et obtenir des résultats – pas seulement une fois, mais de façon durable.

Qu'est-ce qu'un Data Lake ?

Un Data Lake vous permet de stocker et de traiter l'ensemble de vos données (y compris les Big Data) provenant de plusieurs sources de données (Cloud, sur site et hybrides) sans avoir à les préstructurer. Vous pouvez remplir votre Data Lake avec tous les types de données — structurées, non structurées ou multistructurées — ce qui signifie que vos dirigeants et analystes métiers peuvent réaliser des analyses plus novatrices à partir d'un plus grand nombre de données.

La promesse d'une gestion intelligente des Data Lakes

Les principes de gestion de données sont la base pour fournir des données fiables aux personnes appropriées, au moment opportun. Ils garantissent que tous les processus essentiels de gestion de données, de la collecte des données à leur préparation et à leur administration, sont entièrement automatisés pour le Data Lake. En outre, une approche de la gestion de données basée sur l'apprentissage automatique intelligent permet aux entreprises de traiter un plus grand volume de données variées avec nettement moins de ressources humaines que les approches traditionnelles. Votre entreprise peut désormais transformer plus de sources de données brutes en sources fiables et opportunes d'informations précieuses qui alimenteront des analyses plus pertinentes qu'auparavant.

Pourquoi la gestion de données est importante pour votre Data Lake

La gestion de données est une source de préoccupation encore plus importante dans l'environnement de Data Lake que dans les environnements de données plus traditionnels, et ce pour deux raisons.

Le besoin d'immédiateté

- À mesure que les pressions liées à la concurrence augmentent et que le rythme de l'activité s'accélère, il est nécessaire de pouvoir accéder aux données plus rapidement qu'auparavant.
- De nouveaux types de données sont intrinsèquement limités dans le temps, par exemple les données relatives à la détérioration du service client, et leurs fenêtres temporelles pour une correction efficace sont très courtes.
- Les processus en cascade utilisés autrefois sont trop lents pour répondre aux besoins modernes des entreprises.
- Les processus complexes de collecte des exigences et les longs cycles de développement ne font que retarder l'obtention des informations utiles aux départements d'entreprise.
- Les méthodes obsolètes de codage manuel qui nécessitent le recrutement, la formation continue et la rétention de viviers importants et croissants d'effectifs nuisent à la maintenabilité à long terme et accablent les équipes de tâches de remaniement et de renouvellement.

Le besoin de libre-service

- Le contexte métier des ensembles de données est intrinsèquement plus distribué qu'auparavant.
- Les consommateurs de données au sein des différents départements d'entreprise sont également producteurs de données et fournissent leurs propres ensembles de données pour les projets d'entreprise.
- Le contrôle superflu des services informatiques retarde l'accès des départements d'entreprise aux données dont ils ont besoin.
- Avec la transformation rapide des conditions métiers, le contexte métier est de plus en plus nécessaire pour les projets de données.
- Les équipes qui gèrent et utilisent les données sont souvent distribuées.

Du fait des paradigmes de gestion de données traditionnels et de l'extrême efficacité des Data Lakes, les entreprises peuvent être confrontées à une prolifération de « bourbiers » qui ne conduisent pas à une valeur durable pour l'entreprise en général. Les directeurs de données doivent adopter une plate-forme systématique et reproductible et une série de processus pour faciliter aussi bien la flexibilité que la collaboration. L'approche systématique est le seul moyen qui permet de transformer de manière cohérente les sources de Big Data en informations utiles.

Comment les Data Lakes apportent une valeur métier ajoutée en pratique

Les Data Lakes sont actuellement utilisés par les secteurs d'activité et les fonctions métiers pour transformer des données brutes en nouvelles sources de rentabilité. Si elles connaissent les possibilités offertes par les Data Lakes dans certaines situations concrètes, vos équipes comprendront mieux le contexte métier du projet et créeront un Data Lake capable d'apporter une valeur réelle aux décideurs métiers.

Data Lakes et fraude

Les banques qui s'intéressent aux données se servent de Data Lakes pour détecter les fraudes ou les risques de blanchiment d'argent en analysant différentes formes de menaces. Elles recueillent les données de grand livre et d'autres données financières depuis les systèmes mainframe pour un accès plus aisé. Elles les combinent ensuite avec des ensembles de données externes, comme des listes de surveillance. Les établissements financiers identifient les anomalies et les tendances à l'aide de méthodes d'analyse avancées de détection des fraudes. Lorsque des capacités de sécurité et de gouvernance sont intégrées à l'architecture du Data Lake anti-fraude, les analystes de données obtiennent, sans encourir de risques supplémentaires, des informations fiables d'un plus grand nombre de données. Ils obtiennent ainsi un panorama complet des risques à l'échelle de l'entreprise, ce qui contribue à rendre leurs services plus sûrs et plus rentables.

Data Lakes et marketing

Les données marketing sont disséminées depuis longtemps dans les divers départements de l'entreprise, ce qui empêche les équipes marketing de bien cerner leurs prospects et leurs clients. Avec un Data Lake marketing, celles-ci comblent enfin ces lacunes en accédant en libre-service à un panorama des données marketing intégré de bout en bout, quels que soient la plate-forme, le silo ou le canal. Tout ce que les équipes marketing veulent savoir sur les relations entre l'entreprise et ses clients et prospects est analysé afin d'en dégager des tendances et des constantes — pour faciliter la vente croisée et la vente incitative de nouveaux produits aux clients. Ces Data Lakes sont de plus en plus utilisés pour prédire « l'étape suivante la mieux adaptée » pour fidéliser les clients.

Data Lakes et santé

Les organismes de santé possèdent plus de données et de types de données que la plupart des autres secteurs d'activité ; ils doivent respecter des réglementations strictes, notamment en matière de confidentialité des données (p. ex. la loi HIPAA). La gouvernance et la sécurité des données sont essentielles à tout organisme de santé ayant à gérer de gros volumes de données sensibles au sein de ses propres départements et avec ses partenaires. Un Data Lake du secteur de la santé géré par la solution de gestion intelligente des Data Lakes favorise la collaboration et, par conséquent, la gestion, des innombrables données utilisées dans les analyses de Big Data pour prédire le résultat des traitements, réduire les dépenses de santé et optimiser les effectifs et les chaînes logistiques. En collectant les données sur les patients provenant des hôpitaux et en les associant à des données d'assurance, les organismes de santé composent pour chaque patient un panorama complet qui facilite la prise de décisions et contribue à réduire les dépenses de santé tout en améliorant le résultat des traitements.

Data Lakes et industrie

Pour devenir plus rentables et atténuer leurs risques, les entreprises industrielles doivent rendre leurs chaînes logistiques plus efficaces. Les données émises par des capteurs fixés sur les machines contribuent à cerner de façon plus unifiée l'état de fonctionnement du matériel pour prédire les pannes éventuelles et intervenir à temps pour les réparer, d'où une réduction considérable des coûts de maintenance.

Data Lakes et conformité

Les établissements de services financiers sont tenus de démontrer leur conformité à un grand nombre de dispositions réglementaires, par exemple le RGPD. L'ingestion dans un Data Lake de toutes les données liées à la conformité des données permet aux experts du domaine de garantir et de prouver la conformité aux exigences réglementaires de façon plus globale.

Les neuf principes de conception fondamentaux pour la gestion des Data Lakes

En tant que directeur des données chargé de planifier un projet de Data Lake, vous devez prendre en considération neuf principes de conception pour maximiser la valeur de vos environnements de Data Lake. Voyons plus en détail comment procéder.

1. **Exploiter les capacités d'une équipe centrale réduite, agile et interfonctionnelle, dont les membres proviennent des équipes chargées du développement, des opérations et des aspects métiers**

   On parle beaucoup de développement agile, mais la composante de taille manquant dans ce débat est l'importance des équipes interfonctionnelles. Le recours à des équipes interfonctionnelles dans les projets de Data Lake présente plusieurs avantages. Premièrement, la possibilité d'intégrer des connaissances de domaines fonctionnels provenant de plusieurs sources. Les projets de Data Lake demandent des connaissances en implémentation (fournies par les ingénieurs de données), un contexte métier (maîtrisé par les gestionnaires de données), ainsi que l'expertise analytique des spécialistes en science et analyse des données. Ces perspectives multiples favorisent la création d'informations métiers précises et cohérentes qui répondent aux besoins métiers. Elles font aussi en sorte que tout le monde ait la même compréhension des données disponibles. En règle générale, les projets de Data Lake dotés d'une équipe interfonctionnelle réalisent leurs objectifs plus rapidement que les projets sans équipe interfonctionnelle, et rencontrent moins de problèmes de qualité des données.

2. **Habiliter les spécialistes en science des données à obtenir rapidement les données dont ils ont besoin pour aider à la préparation des données**

   Les outils de visualisation en libre-service (Tableau, Qlik, Kibana, Zoomdata) ont gagné en popularité au fil des années, car de plus en plus d'analystes métiers cherchent à accéder directement aux données. Cependant, avec la visualisation seule, les utilisateurs métiers se retrouvent à attendre que le service informatique leur transmette les données dont ils ont besoin. C'est là que la préparation des données en libre-service entre en jeu. Elle permet en effet d'autoriser les analystes métiers expérimentés à fusionner, transformer et nettoyer les données pertinentes afin de disposer d'éléments plus fiables et certifiés pour l'analyse. Des outils sophistiqués aident les utilisateurs à publier leurs ensembles de données préparés sur des espaces de travail collaboratifs, de sorte que plusieurs participants des services métiers puissent accéder aux données et les préparer ensemble. De plus, les techniques d'apprentissage automatique intégrées à ces outils guident les analystes métiers dans l'exploration et la découverte des données du Data Lake.

3. **Utiliser le bon sens collectif grâce au crowdsourcing et le balisage pour gouverner les ressources de données**

   Alors que les entreprises adoptent les Data Lakes pour traiter des données sensibles, comme les données de patients ou de consommateurs, des méthodes efficaces de gouvernance de données sont nécessaires. Les données étant de plus en plus distribuées dans toute l'entreprise, les méthodes en libre-service appliquées par les utilisateurs de données eux-mêmes doivent optimiser les approches traditionnelles axées sur l'informatique pour la gouvernance de données. Le crowdsourcing de gouvernance de données permet à l'entreprise de puiser dans la sagesse des utilisateurs métiers pour tirer parti des connaissances, du contexte et de l'expertise qui collectivement améliorent la qualité des données. Dans un environnement en libre-service, chaque utilisateur a la capacité d'appliquer son expertise en la matière pour améliorer la qualité et la structure des données. À titre d'exemple, les analystes métiers devraient être en mesure

d'apporter leurs connaissances, via des étiquettes et d'autres classifications, afin que la qualité des ressources de données et des éléments de données clés augmente continuellement. La collaboration devient alors un mécanisme destiné à permettre aux analystes métiers de s'aider mutuellement pour atteindre l'objectif commun dans toute l'entreprise de fournir des ressources de données fiables.

L'apprentissage automatique est aussi une approche destinée à automatiser la découverte de domaines de données à l'aide d'algorithmes de classification. Il peut également utiliser une analyse de regroupement pour découvrir de manière proactive des similitudes entre les ensembles de données. Cela permet au système de traiter automatiquement les nouvelles données de la même manière que les données antérieures, dispensant ainsi les professionnels de la gestion de données de travaux répétitifs. L'analyse des connaissances et des comportements les plus répandus accroît aussi considérablement l'efficacité de la gouvernance de données.

4. **Automatiser la collecte et la transformation des données**

   L'ingestion et la transformation manuelles des données sont des processus complexes aux nombreuses étapes, qui donnent des résultats incohérents et impossibles à reproduire. Il n'y a aucun avantage métier à ralentir la collecte et la transformation par des processus manuels ou complexes. Les entreprises performantes tirent parti de connecteurs prédéfinis et de plates-formes d'ingestion des données ultra-rapides pour charger et transformer des ensembles de données dans le Data Lake. Cela permet aux Data Lakes de prendre en charge rapidement de nouveaux types de données et de s'adapter aux volumes croissants de données entrantes. Qui plus est, l'automatisation accélère l'itération et la flexibilité nécessaires à l'agilité, car les modifications peuvent être apportées très rapidement aux processus automatisés, sans aucun risque de bogues.

5. **Tirer parti de la validation et de la notation des données en fonction de règles pour identifier rapidement les problèmes de qualité des données**

   Comme les dirigeants le savent, les problèmes qui ne sont pas détectés assez tôt peuvent faire boule de neige en aval. Il en va de même avec les Data Lakes, où les erreurs de qualité des données qui ne sont pas identifiées suffisamment tôt ont de graves répercussions sur les informations métiers en raison d'inexactitudes ou d'incohérences entre les différents ensembles de données. L'intelligence artificielle (IA) appliquée aux profils de métadonnées et de données automatise les règles métiers et les processus de qualité des données. Les Data Lakes avec validation de données basée sur des règles, dotés de l'intelligence artificielle, détectent et corrigent automatiquement les données incomplètes, inexactes ou incohérentes. La détection et la correction précoces de ces anomalies ont une incidence considérable sur l'exactitude et la cohérence des informations métiers.

   Un système de règles permet de définir le profil des données et de filtrer celles-ci lorsqu'elles sont collectées et transformées dans le Data Lake. Lorsque les règles identifient des données qui dépassent les seuils prédéfinis et ne peuvent pas être corrigées automatiquement, ces problèmes peuvent être triés et transmis pour suivi par les ingénieurs et analystes de données. En mettant en lumière les domaines prioritaires dans lesquels les données peuvent avoir un impact métier important, ce type de validation et de notation des données en fonction de règles permet aux membres de l'équipe de faire le meilleur usage possible du temps limité qui leur est imparti. Ainsi, les tableaux de bord et les fiches d'évaluation de la qualité des données donnent de la visibilité à l'équipe et l'aident à mieux comprendre sur quels aspects les tâches manuelles doivent se concentrer.

6. **Laisser à l'intelligence artificielle et à l'apprentissage automatique le soin de découvrir les données, d'en assurer la sécurité des données et de prendre en charge la qualité de données**

   Avec l'explosion des volumes de données, l'une des difficultés les plus importantes pour un directeur des données est simplement d'obtenir une réelle visibilité sur les ressources de données disponibles. Si la création d'un Data Lake permet de centraliser les principales ressources de données dans un environnement unique, il se pose toujours la question de la découverte des ressources à collecter en premier lieu dans le Data Lake.

   L'intelligence artificielle permet de découvrir automatiquement la structure des données non structurées. Cette compréhension peut ensuite être utilisée pour intégrer automatiquement d'autres données non structurées similaires. En conséquence, la productivité d'une tâche très chronophage est dopée.

   Semblables aux moteurs de recherche Web qui analysent et indexent le contenu Web, les analyseurs de données automatisés sont utilisés pour rechercher et indexer de façon proactive de nouvelles ressources de données dans toute l'entreprise. Les techniques d'apprentissage automatique déterminent ensuite les corrélations et les similitudes entre les différentes ressources de données. Cela contribue à établir une vision complète des ressources de données pour la sécurité des données et la prise en charge de la qualité de données. Au lieu de s'appuyer sur des approches manuelles pour détecter toute prolifération indésirable ou tout échec du respect des réglementations en matière de données, une approche basée sur l'apprentissage automatique surveille et détecte de façon proactive toutes les données au sein de l'entreprise pour garantir le niveau maximal de protection et de conformité.

   En outre, une vision globale des ressources de données permet de constituer un catalogue intelligent de toutes les ressources de données et de déduire les relations qui existent entre elles. Les consommateurs de données, tels que les analystes métiers, utilisent alors ce catalogue pour identifier de nouvelles ressources susceptibles de les intéresser — en fait, certains catalogues vont jusqu'à recommander des ressources de données en fonction des techniques d'apprentissage automatique.

7. **Une conception au service d'une source unique de référence à la disposition d'une entreprise fédérée**

   Les Data marts obsolètes des départements hantent toujours de nombreuses entreprises. Ils ont sans doute conduit à la création de bourbiers dans lesquels les équipes des départements créent des Data Lakes cloisonnés qui sont incohérents et redondants avec d'autres environnements de l'entreprise.

   Le principe de colocalisation est essentiel pour maximiser les bénéfices d'un Data Lake. Il est conseillé de chercher à avoir un nombre limité d'environnements de Data Lake importants organisés entièrement autour de domaines métiers essentiels. Ce principe de colocalisation garantit que les Data Lakes reflètent véritablement des références uniques, améliorent la capacité d'analyse prédictive et réduisent la duplication inutile dans toute l'entreprise.

« Au moment où nous mettons en service une nouvelle technologie, elle est déjà dépassée. La plate-forme Informatica nous isole des nombreuses modifications sous-jacentes qui ont lieu. »
— Chief Data & Analytics Officer, Ford

En outre, les approches de gestion des Data Lakes qui exploitent le partage des données, le marquage des données et les espaces de travail des projets facilitent la collaboration qui est essentielle entre analystes de données et spécialistes en science des données. Il est conseillé aux consommateurs de données de prendre en compte les travaux de leurs collègues lors de transitions analytiques, dans le cadre desquelles le travail d'un analyste du Data Lake est publié et partagé avec d'autres analystes pour une réutilisation ultérieure.

8. Normaliser le processus et améliorer la cohérence de l'architecture

Les équipes font souvent face au dilemme que pose la répétition des mêmes problèmes de gestion de données maintes et maintes fois. L'absence de normalisation nuit de façon permanente aux efforts des Data Lakes, tandis que les exigences s'intensifient, car les environnements ne sont tout simplement pas construits pour évoluer. La normalisation et la cohérence sont essentielles à l'évolutivité à long terme.

Se pose également un problème de réutilisation. Vous souhaitez que vos professionnels de la gestion de données et vos analystes métiers réutilisent les approches existantes de gestion de données au lieu d'en créer d'autres. Le problème qui se pose est toujours le même : il est plus facile de créer un nouveau « code » que de trouver un « code » existant. Grâce à l'intelligence artificielle intégrée, la plate-forme d'intégration de données peut recommander de façon proactive un « code » existant (c'est-à-dire des règles, une logique, des politiques) à réutiliser.

Un processus normalisé et une architecture homogène garantissent que vos analystes métiers et spécialistes en science des données se consacrent à l'innovation et à l'analyse, et non à la gestion de données. Plus les intervenants des services informatiques et des départements d'entreprise se concentrent sur la gestion de données, moins ils se consacrent aux innovations axées sur les données qui sont si précieuses pour votre entreprise.

9. Établir des politiques, taxonomies et classifications afin que toutes les équipes soient en harmonie

L'un des plus grands goulets d'étranglement qui nuisent à la rapidité, à la flexibilité et à la collaboration est l'absence de politiques et d'un langage ou d'un glossaire de termes communs pour fournir un contexte et une signification métiers. Si tout le monde dans l'entreprise ne respecte pas des politiques standards ni ne reconnaît les ressources de données de façon homogène, des malentendus cloisonnés de données apparaissent. En retour, des problèmes d'utilisation dans toute l'entreprise se produisent. De plus, les consommateurs de données comme les spécialistes en science des données signalent souvent qu'ils passent trop de temps à nettoyer les incohérences des données — au lieu de se consacrer à l'analyse source de valeur ajoutée.

Les processus, taxonomies et glossaires normalisés permettent de garantir que tous les membres de l'équipe chargée du projet parlent le même langage. La création de procédures simples au début du processus pour établir quelles sont les principales ressources de données — et comment elles seront gérées et référencées — élimine tâches de renouvellement et frustration. Les taxonomies et politiques normalisées simplifient radicalement les audits et la traçabilité pour des raisons de conformité, de sorte que vous pouvez toujours connaître la provenance des données et protéger de façon proactive les données sensibles.

Conclusion

Les Data Lakes offrent une occasion unique de fournir des informations métiers radicalement nouvelles très rapidement et efficacement. Les meilleures pratiques décrites dans le présent livre blanc vous aideront à éviter nombre des pièges courants qui entravent le succès et à mettre en route votre environnement de Data Lake de façon appropriée.

La gestion intelligente des Data Lakes d'Informatica est la solution de bout en bout intégrée la plus complète du secteur d'activité destinée à la transformation digitale axée sur les données. Informatica propose des solutions qui permettent aux sociétés d'exploiter tout le potentiel des Big Data afin de gagner en flexibilité et de concrétiser de nouvelles opportunités de croissance via l'innovation, avec à la clé des évolutions positives du marché.

Grâce à la gestion intelligente des Data Lakes d'Informatica, vous pouvez trouver, enrichir, préparer, cataloguer, maîtriser, gérer et protéger les Big Data dont vous avez besoin pour fournir des informations exactes et cohérentes — ce qui permet de prendre des décisions métiers plus rapides. S'appuyant sur la technologie d'intelligence artificielle unique d'Informatica fondée sur les métadonnées, connue sous le nom de moteur CLAIRE™, les entreprises trouvent systématiquement des données, découvrent les relations entre les données qui comptent, et préparent et partagent rapidement les données appropriées avec les personnes qui conviennent, au bon moment. Et, en fin de compte, elles fournissent des données innovantes, plus pertinentes, en temps opportun et personnalisées.

Les meilleures pratiques communiquées dans ce livre blanc doivent vous aider à libérer la valeur de votre prochain projet de Data Lake.

Étapes suivantes

Accédez à www.informatica.com/bigdataready pour y consulter les ressources, notamment les eBooks, rapports des analystes et webinaires, qui décrivent les meilleures pratiques pour gérer votre Data Lake.

Informatica

Siège mondial, 2100 Seaport Blvd., Redwood City, CA 94063, États-Unis Téléphone : +33 1 42 04 89 00 (France)
Numéro gratuit aux États-Unis : 1.800.653.3871
www.informatica.com/fr www.informatica.com linkedin.com/company/informatica twitter.com/InformaticaFr

IN09_0218_3401