alteryx

Analytique en libre-service sur Google Cloud Platform :

5 enseignements tirés de la préparation des données pour garantir la réussite

Vous aspirez à relever vous-même les défis de l'analytique métier et cherchez à devenir plus productif et innovant. Pour y parvenir, vous avez besoin d'exploiter vos données.

Dans le paysage de données dynamique et complexe d'aujourd'hui, les solutions traditionnelles d'entrepôt de données et d'intelligence métier atteignent leurs limites. Même en utilisant des solutions comme Tableau ou Qlik pour mieux tirer parti du potentiel des entrepôts de données traditionnels, vous finissez par retomber dans les feuilles de calcul et par solliciter l'aide de l'équipe IT.

La solution : miser sur l'analytique en libre-service.

Ce livre blanc présente cinq enseignements qu'Alteryx a tirés en aidant des organisations comme la vôtre à adopter l'analytique en libre-service. Il se concentre sur Google Cloud Platform, une suite analytique complète et intelligente qui donne aux analystes la liberté d'exploiter toute forme de données, à n'importe quelle échelle et n'importe quelle vitesse.

1 Oubliez l'approche « lift-and-shift » lors de votre migration vers le cloud

Lorsqu'elles migrent leur environnement analytique vers le cloud, de nombreuses entreprises adoptent une approche « lift-and-shift », qui consiste à installer les applications analytiques sur une machine virtuelle hébergée dans le cloud. Bien que cette approche facilite la maintenance de l'infrastructure, elle n'améliore en rien les résultats de l'analytique, n'améliore pas les bénéfices et ne réduit pas les coûts d'exploitation. Le processus et les outils restent les mêmes, et les utilisateurs ne constatent aucune réelle valeur ajoutée sur le plan du libre-service.

Pour être véritablement en libre-service, les solutions analytiques doivent être en natif dans le cloud. Les processus d'ingestion, de stockage, de manipulation et de reporting des données doivent s'intégrer nativement aux systèmes conçus exclusivement pour le cloud.

Ils doivent prendre en charge des environnements cloud dynamiques, flexibles, évolutifs, et de plus en plus conteneurisés et orientés vers la fourniture de microservices. Plutôt que de recréer les plateformes de données rigides dont elles disposaient sur site, les entreprises les plus performantes préfèrent adopter l'agilité du cloud et utiliser des systèmes open source, tels que Kubernetes ou Docker, pour orchestrer les conteneurs et leurs flux de données.

Google Cloud Platform propose un environnement intéressant pour l'analytique en libre-service, car il fournit une suite analytique intelligente, en mode natif et serverless. Chaque composant est activé facilement, se met à l'échelle de manière dynamique ou réduit l'allocation des ressources en fonction de l'utilisation. Il n'est pas nécessaire de planifier les opérations de maintenance. Avec la capacité d'exploiter librement n'importe quel composant analytique, d'utiliser les ressources de manière flexible et de maîtriser leurs coûts, les organisations sont en mesure de se concentrer sur les données et sur la valeur qu'elles apportent à leur activité.

2
ETL et libre-service ne font pas bon ménage

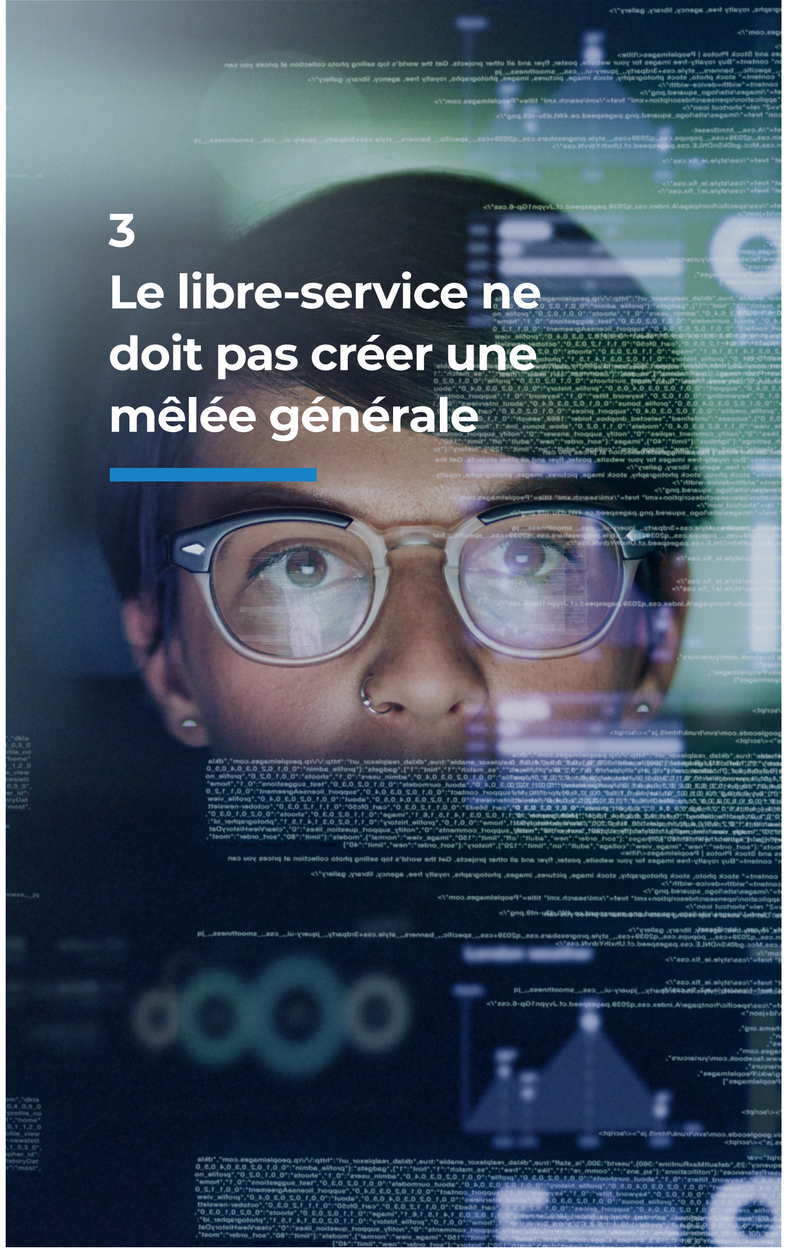
Les technologies d'intégration de données traditionnelles comme ETL (Extract-Transform-Load) peuvent être utilisées pour créer des pipelines afin de déplacer vers Google Cloud Platform les données stockées sur site et dans d'autres clouds. Les nouvelles technologies d'intégration et de streaming de données dédiées au cloud, telles que StreamSets, Fivetran ou Google Cloud Data Fusion, permettent également de faire circuler efficacement les données entre les systèmes et de les rendre disponibles à des fins d'analytique dans Google Cloud Platform.

Il peut être tentant de continuer à utiliser ETL après avoir alimenté un lac de données ou un entrepôt de données, mais ETL n'est tout simplement pas adapté à l'analytique en libre-service.

Il s'agit d'une technologie assez technique et complexe qui s'adresse à des ingénieurs de données habitués à intervenir sur des transformations en temps réel et par lots, bien définies et récurrentes, pour déplacer des données à grande échelle. Pour permettre aux professionnels métier experts des données d'accéder aux données et de les utiliser dans leurs projets analytiques, les données doivent faire l'objet d'une préparation supplémentaire.

Au sein de sa suite analytique intelligente, Google Cloud Platform offre une solution de préparation des données appelée Cloud Dataprep by Trifacta. Elle évalue la qualité des données, affine, standardise et nettoie les données, les combine et gère divers calculs de données.

La solution Cloud Dataprep by Trifacta est conçue pour aider un grand nombre de professionnels des données (ingénieurs de données, data analysts, analystes métier et autres professionnels data-driven) à exploiter un lac ou un entrepôt de données en interagissant avec le contenu des données pour les affiner de manière itérative et les rassembler afin d'alimenter l'analytique métier en aval.



3
Le libre-service ne doit pas créer une mêlée générale

Lorsque l'analytique en libre-service s'ouvre à un plus grand nombre de participants, elle est naturellement exploitée par davantage d'utilisateurs. Mais cela ne doit pas donner lieu à une mêlée générale. Le fait de permettre à davantage d'utilisateurs de profiter du libre-service impose une gouvernance rigoureuse de manière à :

- **Éviter une prolifération des données qui échappe à tout contrôle**
- **Se conformer aux exigences réglementaires**
- **S'assurer de la fiabilité des données utilisées pour prendre des décisions métier**

Il est important de trouver le juste équilibre entre la protection des actifs de données et la possibilité pour les utilisateurs de collaborer et d'exploiter les données à leur guise.

Les organisations doivent trouver le juste équilibre entre la protection des actifs de données (sur le plan de la gouvernance et de la sécurité) et la possibilité pour les utilisateurs de collaborer et d'exploiter les données à leur guise. Il existe trois façons d'atteindre cet équilibre.

Réduire les silos

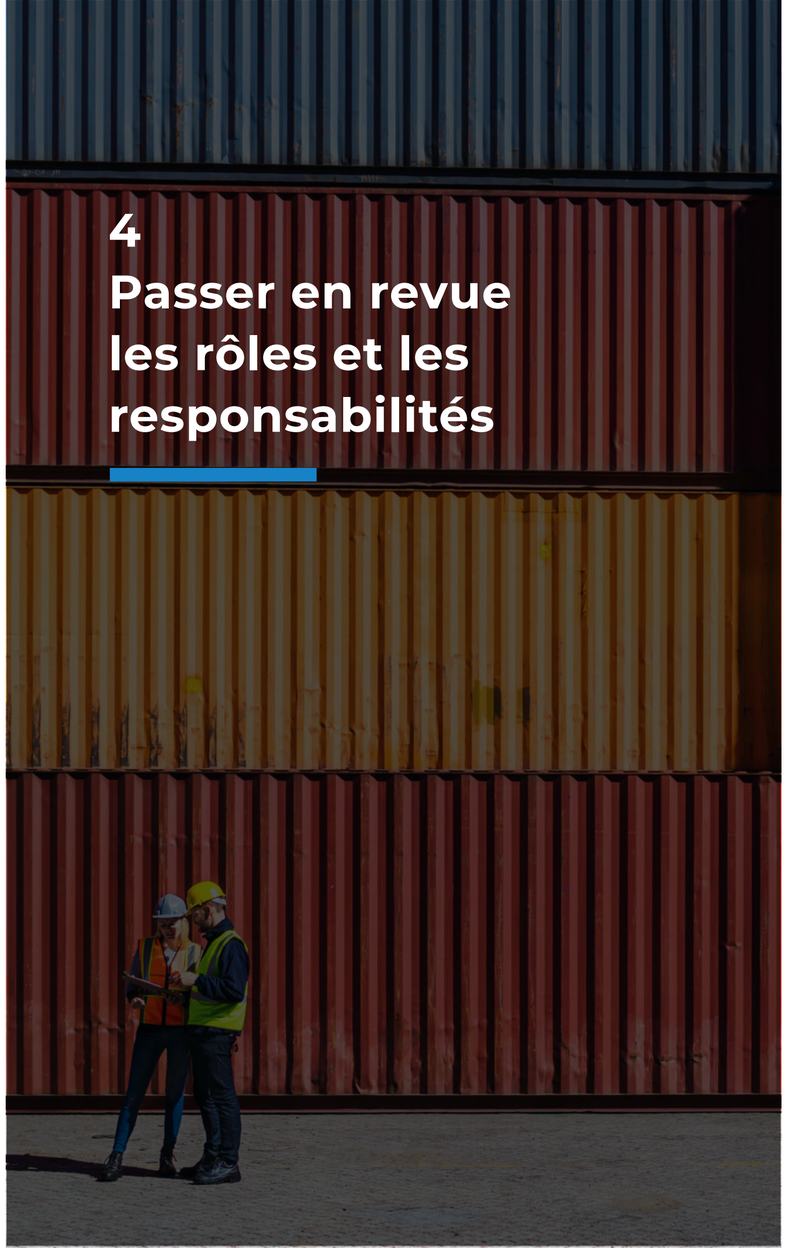
Lorsque le stockage et le traitement des données sont centralisés dans le cloud avec une évolutivité quasi illimitée, et lorsque les utilisateurs sont autorisés à introduire leurs propres données dans le cloud, les silos de données cessent de proliférer. Les utilisateurs collectent des extraits de données, exécutent leurs propres routines de préparation et créent leurs rapports dans et depuis le cloud, au lieu d'extraire des données et de les dupliquer dans des feuilles de calcul.

Utiliser un catalogue de données

Les ressources partagées (par exemple, un catalogue ou un glossaire central) qui gèrent les définitions de données, les métadonnées et les informations sur la traçabilité des données, permettent aux utilisateurs de trouver les données plus rapidement, tout en aidant les entreprises à contrôler les sources de données et à surveiller leur cycle de vie. Les solutions de machine learning (ML) et d'intelligence artificielle (IA) automatisent la collecte et la gestion des métadonnées et des connaissances sur les données associées.

Suivre et documenter la traçabilité des données

Pour le reporting et les audits réglementaires, il est important d'assurer la traçabilité des données, c'est-à-dire de comprendre la manière dont elles ont été utilisées et transformées, et par qui. La traçabilité est également importante pour les décideurs, qui doivent comprendre l'historique des données qui alimentent l'analytique, les visualisations et les recommandations prescriptives, ainsi que l'impact des nouvelles exigences sur le pipeline utilisé pour produire l'analyse.



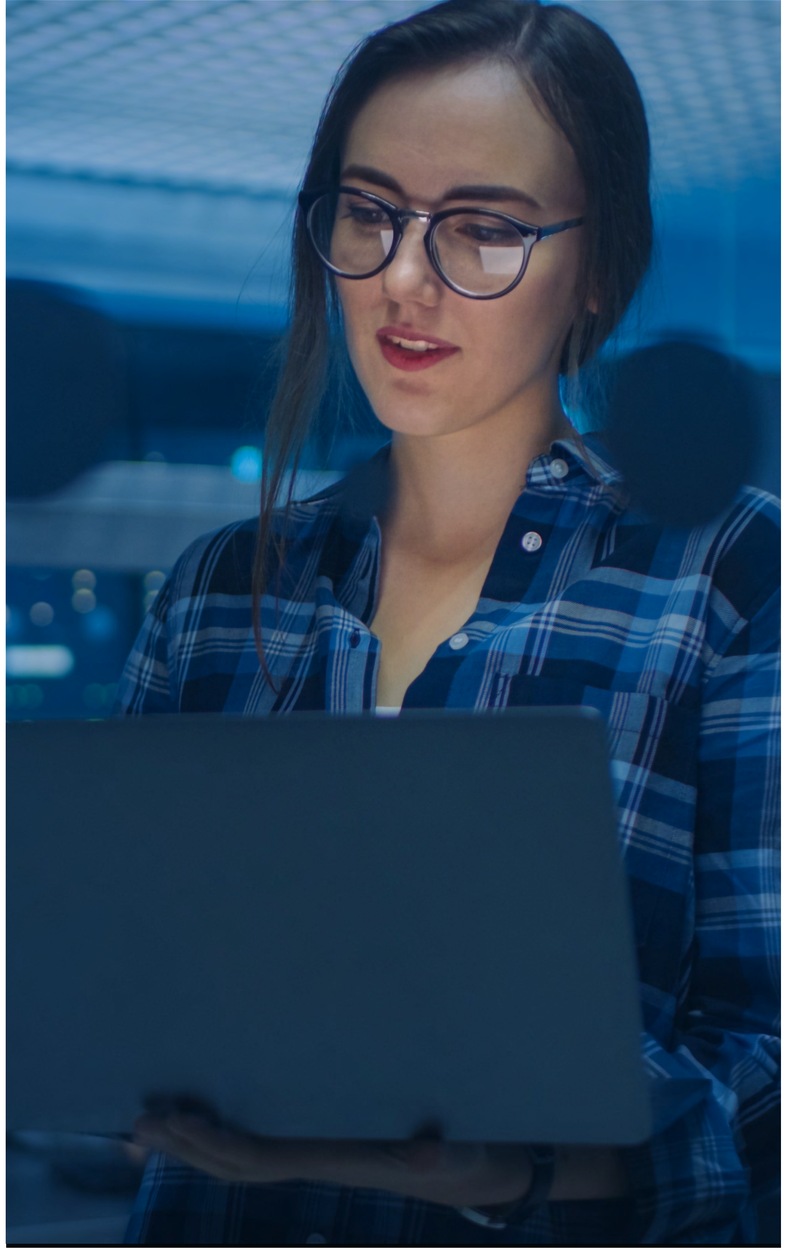
4
Passer en revue les rôles et les responsabilités

La préparation des données en libre-service est pour les organisations l'occasion d'examiner les rôles et les responsabilités des principales parties prenantes dans la chaîne de valeur de l'analytique des données. Elle les encourage également à repenser le processus de pipeline de données, jusque-là rigide, pour lui apporter davantage de flexibilité et d'agilité. Chaque partie prenante utilise la préparation des données en libre-service différemment, en fonction de ses besoins, tout en adoptant un cadre commun de conception, de communication et de collaboration pour produire des données analytiques.

La préparation des données en libre-service offre de nouvelles manières de collaborer aux ingénieurs de données, aux data analysts, aux data scientists, aux architectes de données et aux dirigeants et responsables métier qui privilégient l'analytique.

LES DATA ANALYSTS, en plus des analystes métier, des chefs de projet et des autres fonctions associées, agissent généralement en étroite collaboration avec les utilisateurs métier. Ils sont souvent amenés à explorer les données brutes pour identifier les informations potentiellement utiles afin de répondre aux questions de l'entreprise. Leur objectif est d'obtenir des réponses aussi rapidement et facilement que possible.

Cloud Dataprep by Trifacta est une solution idéale pour les data analysts, car elle est facile à mettre en œuvre et ne nécessite aucune expertise technique poussée en matière de codage. Les data analysts peuvent être impliqués dès le début du processus de préparation des données, en explorant et en réalisant des prototypes de données pour répondre à leurs besoins métier. Lorsqu'une automatisation est nécessaire pour des questions de durabilité et de répétabilité, ils peuvent collaborer avec les ingénieurs de données pour orchestrer le pipeline de données de bout en bout.



LES INGÉNIEURS DE DONNÉES conçoivent, développent et gèrent le traitement des données et l'architecture nécessaires pour soutenir l'analytique et la data science. Ils sont étroitement impliqués dans la transformation des données, et notamment dans l'exploration et le profilage des données brutes. Les ingénieurs de données ont pour principal objectif de rationaliser et d'automatiser les processus liés aux données afin de pouvoir en gérer davantage.

Cloud Dataprep by Trifacta est idéal pour les ingénieurs de données, car la solution leur permet d'opérationnaliser et de surveiller les différents flux de données qu'eux-mêmes ou les data analysts conçoivent. Elle permet aux ingénieurs de données de collaborer facilement avec toutes les parties prenantes afin de comprendre les exigences de l'infrastructure de données et de fournir des conseils aux utilisateurs pour améliorer leur façon d'explorer, d'analyser, de modéliser et de consommer les données.

LES DATA SCIENTISTS utilisent des connaissances et des compétences spécialisées pour concevoir et modéliser des algorithmes, en tirant parti du machine learning et de l'intelligence artificielle. Pour autant, ils consacrent jusqu'à 80 % de leur temps à des tâches courantes de préparation des données, ce qui laisse peu de temps à l'innovation.

Cloud Dataprep by Trifacta convient idéalement aux data scientists, car il s'agit d'une solution de préparation des données cloud native qui simplifie les tâches de préparation de routine, ce qui leur permet de les déléguer à des ressources plus facilement disponibles et moins coûteuses.

5
La qualité des données est la responsabilité de tous

Toutes les parties prenantes doivent avoir la garantie que les données qui alimentent leur analytique, leurs visualisations et leurs algorithmes sont parfaitement fiables. La qualité des insights tirés d'un modèle prédictif ou d'un projet analytique est intimement liée à celle des données qui les alimentent.

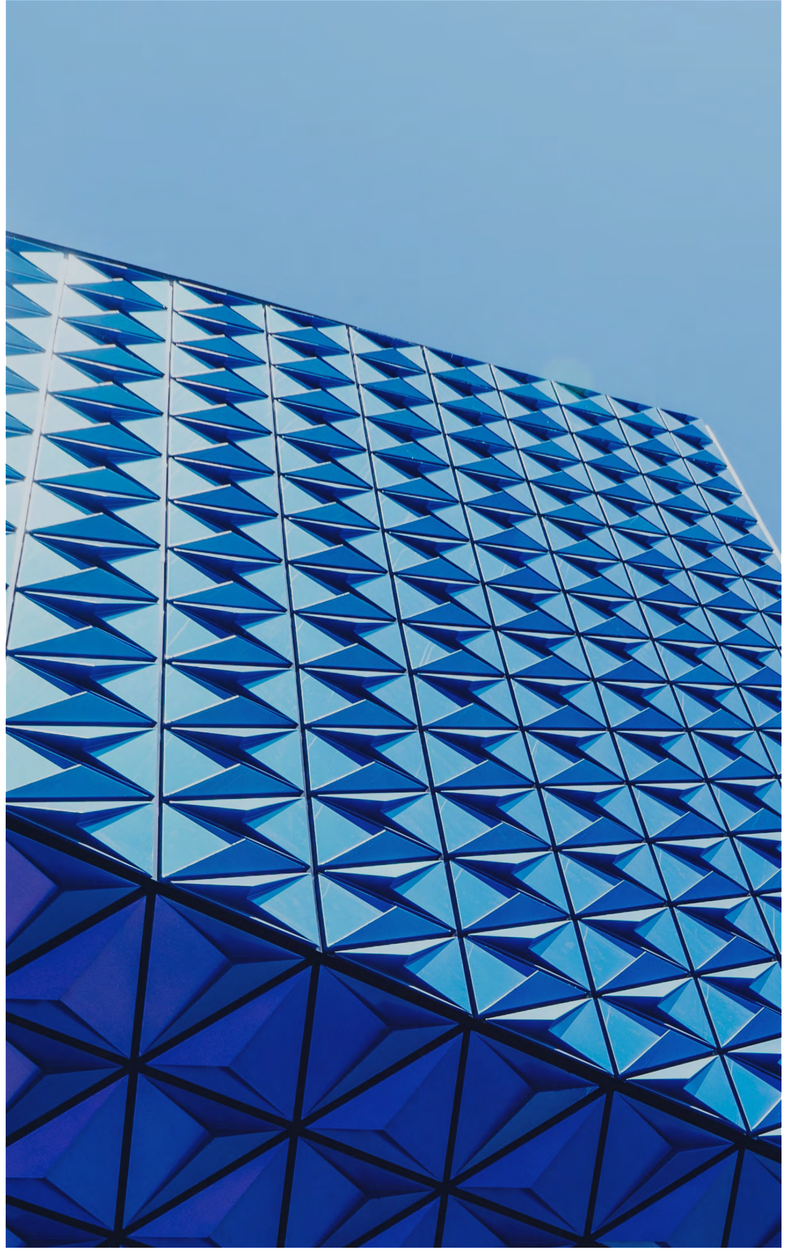
Chaque organisation doit s'engager en priorité à fournir des données « propres ».

Dans un monde idéal, la question de la qualité des données devrait être abordée au niveau de l'entreprise, mais cela n'est tout simplement pas réaliste. Au vu de l'immense volume de données et de la grande diversité de sources de données (internes et externes), de types et de contextes, il est difficile d'évaluer la qualité des données, de corriger les défauts et de surveiller les nouvelles données pour identifier d'éventuels problèmes de qualité. Les nouveaux référentiels de données, tels que les lacs de données dans le cloud et les entrepôts de données modernes, présentent des défis plus complexes en matière de qualité des données.

Il est donc préférable de gérer la qualité des données pendant l'étape de préparation des données. La solution Cloud Dataprep by Trifacta améliore la précision, la cohérence et l'exhaustivité des données en appliquant des techniques d'IA et de ML pour automatiser les procédures de nettoyage des données. L'automatisation gère l'étendue des très vastes référentiels de données et identifie rapidement les valeurs de données qui semblent incorrectes, non valides, manquantes ou incohérentes. Les valeurs aberrantes qui méritent une inspection plus rigoureuse sont automatiquement signalées.

Les routines automatisées de profilage et de nettoyage des données permettent de repérer les incohérences entre les sources intégrées dans un stockage de données cloud, de mettre en évidence les doublons de données probables et de produire des recommandations pour corriger visuellement les problèmes de qualité des données via des transformations automatisées et sans code.

La qualité des données est validée en continu à mesure que de nouvelles données structurées, non structurées ou semi-structurées sont ingérées et intégrées dans le cloud. Avec une « validation continue », les utilisateurs n'ont pas à attendre la fin du processus de validation pour visualiser et tester les résultats, un délai incompatible avec les méthodologies de développement Agile actuelles.



Conclusion

Alteryx accompagne de nombreuses entreprises dans le déploiement efficace de solutions d'analytique en libre-service dans le cloud. Nous avons tiré de cette expérience cinq enseignements majeurs :

1. L'utilisation d'une simple approche « lift-and-shift » pour migrer l'environnement analytique sur site vers le cloud constitue une stratégie inefficace.

Choisissez une plateforme cloud qui propose des composants analytiques spécialement conçus pour le cloud, comme Google Cloud Platform avec sa suite analytique intelligente et complète, en mode natif et serverless.

2. Les technologies ETL peuvent alléger les frictions liées au déplacement des données sur site vers des référentiels cloud, mais elles sont trop techniques et manquent de flexibilité pour une analytique en libre-service. Une meilleure approche consiste à adopter des solutions de préparation des données en libre-service, telles que Cloud Dataprep by Trifacta, pour permettre à tous les utilisateurs métier et techniques d'exploiter les données dont ils ont besoin pour leurs projets d'analytique.

3. Le libre-service n'est pas une zone de non-droit. Pour exploiter efficacement l'analytique en libre-service, les entreprises doivent trouver le juste équilibre entre gouvernance prudente et ouverture.

Pour réussir, il est important de comprendre la traçabilité des données et d'utiliser un glossaire de données métier.

4. Adoptez de nouveaux processus agiles qui privilégient la collaboration et l'adoption d'un langage commun entre toutes les parties prenantes afin de transformer les données en une ressource affinée pour leurs projets analytiques.

5. La qualité des données est la responsabilité de tous. La solution Cloud Dataprep by Trifacta offre une évaluation automatique de la qualité des données, ainsi qu'une validation et une résolution continues qui garantissent la fiabilité des données utilisées pour l'analytique en libre-service.

alteryx

À PROPOS D'ALTERYX

Alteryx, la société spécialisée dans l'automatisation analytique, cherche avant tout à aider tous les utilisateurs à transformer les données en avantage décisif. Alteryx unifie l'automatisation de l'analytique, de la data science et des processus métier dans une plateforme complète tout-en-un pour accélérer la transformation digitale et façonner l'avenir de l'automatisation analytique. Partout dans le monde, des entreprises de toutes tailles s'appuient sur Alteryx pour produire des résultats qui changent la donne et développer rapidement les compétences de leur personnel. Pour plus d'informations, rendez-vous sur www.alteryx.com/fr.

Alteryx Analytics Cloud

Déployez Alteryx dans le cadre de l'environnement Alteryx Analytics Cloud et profitez de pipelines de données automatisés pour accélérer et améliorer la consultation, le profilage, la transformation et la distribution de vos données. No-code ou low-code, nous répondons à tous vos besoins. Explorez visuellement la qualité des données, accélérez la préparation des données pour l'analytique, et créez et déployez des pipelines de données en quelques secondes.

EN SAVOIR PLUS >

Plateforme Alteryx Machine Learning

Déployez la Data Science dans toute l'entreprise grâce au machine learning automatisé (AutoML) et à l'ingénierie des caractéristiques, afin de permettre aux spécialistes et aux data scientists d'accélérer la production d'informations exploitables.

EN SAVOIR PLUS >

Alteryx Designer

Automatisez chaque étape de l'analytique, dont la préparation et la fusion des données, le reporting, l'analyse prédictive et la Data Science. Accédez aux sources de données, fichiers, applications ou types de données de votre choix, et expérimentez la simplicité et la puissance d'une plateforme en libre-service proposant plus de 300 blocs de construction pour l'automatisation. Commencez dès aujourd'hui à produire des résultats interactifs.

EN SAVOIR PLUS >

Alteryx Intelligence Suite

Que vous soyez novice ou expert, vous pouvez créer des modèles de machine learning innovants en quelques minutes. Alteryx Intelligence Suite est conçu pour vous aider à déceler les insights qui se cachent dans vos données. De l'analyse de données structurées à la découverte d'informations exploitables enfouies dans le texte et les documents, Intelligence Suite vous permet de résoudre les problèmes les plus épineux de votre entreprise.

EN SAVOIR PLUS >

Alteryx Server

Déployez à grande échelle, partagez et contrôlez vos processus, modèles et données de workflows analytiques avec Alteryx Server. Créez des workflows analytiques avec Alteryx Designer et publiez-les sur Alteryx Server pour planifier des rapports et des résultats automatisés. La gouvernance des données, la gestion centralisée de la sécurité et la haute disponibilité intégrées permettent aux entreprises d'étendre l'analytique dans les différents services et à grande échelle.

EN SAVOIR PLUS >