Informatica

Livre blanc

Intelligence artificielle pour l'entreprise intelligente Data-Driven

Découvrez comment les innovations basées sur le Machine Learning de CLAIRE permettent de réaliser de nouvelles avancées en matière de gestion des données

À propos d'Informatica

La transformation digitale fait évoluer les attentes : meilleurs services, livraisons plus rapides, à moindre coût. Les données sont la clé de la réussite des entreprises, et ces dernières doivent évoluer pour rester compétitives.

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. Informatica vous donne la possibilité de devenir plus agile, de saisir de nouvelles opportunités de croissance ou de créer de nouvelles inventions. Nous nous concentrons sur les données afin de vous offrir la polyvalence nécessaire pour réussir.

Découvrez nos solutions et libérez tout le potentiel de vos données en vue de la prochaine révolution intelligente.

Table des matières

« Les responsables de l'analyse et des données sont confrontés à la complexité du paysage des données. Nos prévisions en matière de solutions de gestion des données mettent en avant les développements clés et la demande croissante en termes de capacités Cloud, d'architectures de données connectées, de métadonnées et d'automatisation des tâches routinières et non routinières grâce à l'application de l'IA. »[1]

— Gartner

L'importance de l'intelligence artificielle (IA)

L'intelligence artificielle (IA) et le Machine Learning (ML) sont aujourd'hui au cœur de la transformation digitale que connaissent les secteurs du monde entier. En tant que stratégie de transformation des entreprises, l'IA est l'une des priorités des membres des conseils d'administration. Et elle est devenue omniprésente dans l'amélioration de notre vie quotidienne, des films que nous regardons aux voitures que nous conduisons. L'IA et le ML sont indispensables à la découverte de nouveaux traitements en sciences de la vie, à la réduction des fraudes et risques dans les services financiers, et à l'offre d'expériences clients véritablement personnalisées.

Pour les dirigeants d'entreprise, l'IA et le ML peuvent sembler un peu magiques — bien que leur impact potentiel soit clair, les dirigeants peuvent ne pas tout à fait le comprendre ou ne pas savoir comment exploiter au mieux ces puissantes innovations. L'IA et le ML constituent la technologie sous-jacente à de nombreuses nouvelles solutions métiers — qu'il s'agisse des mesures les plus efficaces, du suivi de la satisfaction client, de l'efficacité des opérations ou de produits innovants. En général, le Machine Learning, et en particulier le Deep Learning, est un processus consommant de grandes quantités de données. Pour obtenir la précision requise, le ML nécessite d'importants volumes de données. Ces données doivent refléter précisément l'état actuel de l'activité. Une IA formée à partir de données limitées ou de mauvaise qualité aura un impact désastreux sur les initiatives métiers, au point de produire un effet inverse au résultat souhaité.

Pour obtenir une IA efficace, dans laquelle on utilise et on forme les fonctionnalités appropriées, nous devons exploiter une grande variété de données, externes et internes à l'entreprise. Ces données doivent être regroupées de manière à pouvoir créer et former un modèle de ML. Cela nécessite une certaine gestion des données. Il s'agit non seulement de gérer l'ampleur et la complexité, mais également de faire confiance. Les données utilisées pour former le modèle proviennent-elles des bons systèmes ? Avons-nous supprimé les informations d'identification personnelle et respecté toutes les réglementations ? Sommes-nous transparents et pouvons-nous prouver la traçabilité des données utilisées par le modèle ? Pouvons-nous documenter et être prêts à montrer aux organismes de réglementation ou aux enquêteurs qu'il n'y a pas de biais dans les données ? Tout cela nécessite un bon contrôle et une base de gestion des données. Sans une solide base de gestion des données, l'IA est incompréhensible et peu fiable — en d'autres termes, sans gestion des données, l'IA peut constituer une boîte noire aux conséquences imprévues.

L'IA requiert des données

La réussite de l'IA dépend de l'efficacité des modèles conçus par les data scientists pour la former et l'adapter. Et la réussite de ces modèles repose sur la disponibilité de données fiables et opportunes.

Pourquoi les data scientists chargés de créer des modèles d'IA et de ML ont-ils besoin de données de haute qualité ? Prenons, par exemple, un modèle prédictif chargé d'anticiper le comportement d'un consommateur. L'emplacement du client, indiqué par le code postal, peut constituer une fonctionnalité utile pour un tel modèle. Mais que se passe-t-il si les données du code postal sont manquantes, incomplètes ou inexactes ? Le comportement du modèle de formation est affecté négativement pendant la formation et le déploiement, ce qui peut entraîner des prévisions incorrectes et réduire la valeur de l'ensemble de l'effort. En outre, un code postal précis, complet et vérifié peut également aider à prévoir la segmentation du marché, la catégorie de revenus, l'âge, l'espérance de vie, etc. d'un individu — et bien plus encore. Nous devons nous attendre à ce que l'IA « transparente » devienne un prérequis réglementé, et non plus une simple option. Sans une traçabilité basée sur les métadonnées, les applications et les informations basées sur l'IA ne peuvent pas être déployées en production.

[1] Gartner, Predicts 2020: Data Management Solutions, Rick Greenwald, Donald Feinberg, Mark Beyer, Adam Ronthal, Melody Chien, 5 décembre 2019.

L'IA nécessite une gestion intelligente des données pour trouver rapidement toutes les fonctionnalités du modèle ; pour transformer automatiquement les données afin de répondre aux besoins du modèle d'IA (mise à l'échelle des fonctionnalités, normalisation, etc.) ; pour dédupliquer les données et fournir des données de référence fiables sur les clients, les patients, les partenaires et les produits ; et pour fournir une traçabilité de bout en bout des données, y compris au sein du modèle et de ses opérations. La réussite de l'IA dépend de l'efficacité des modèles conçus par les data scientists pour la former et l'adapter. Et la réussite de ces modèles repose sur la disponibilité de données fiables et opportunes.

Les données requièrent l'IA

L'IA et le ML jouent également un rôle essentiel dans l'évolution des pratiques de gestion des données. En raison des volumes massifs de données nécessaires à la transformation digitale, les entreprises doivent découvrir et cataloguer leurs données et métadonnées les plus pertinentes pour certifier la pertinence, la valeur et la sécurité — et pour garantir la transparence. Elles doivent nettoyer et maîtriser ces données. Et elles doivent également les gouverner et les protéger efficacement. Si les données ne sont pas gérées efficacement— et à l'échelle — les modèles d'IA et de ML connaîtront le même sort que toutes les initiatives de data warehousing traditionnelles au cours des 30 dernières années : ils utiliseront des données de mauvaise qualité et fourniront des informations non fiables.

Selon une étude récente, le volume global du trafic des datacenters devrait atteindre 20,6 zettaoctets en 2021, tandis que le nombre d'appareils et de connexions connectés devrait atteindre plus de 25 milliards d'ici 2022.[2] Toutes ces données doivent être traitées et rendues utilisables et fiables tout en respectant les politiques de gouvernance. En outre, il est nécessaire d'agir rapidement et de réagir aux changements de stratégie et de processus métiers. Les efforts impliqués dans la préparation des données pour les initiatives de transformation digitale ont gagné en complexité, en même temps que la croissance des données. Selon LinkedIn, le poste de data scientist est l'un des emplois les plus prometteurs aux États-Unis.[3] Et le nombre de data engineers recherchés par les sociétés a récemment connu une augmentation de 96 % d'une année sur l'autre.[4] Toutefois, le recrutement seul ne suffit pas à gérer l'augmentation du volume de données.

Une approche linéaire ne peut pas répondre à un défi exponentiel

Nous ne pouvons pas résoudre ces problèmes en y consacrant simplement plus d'ingénieurs et de développeurs — ils ne peuvent pas être résolus à une échelle linéaire et humaine. Les approches traditionnelles sont truffées d'inefficacités. Les projets sont implémentés en silos, avec une visibilité réduite des métadonnées de bout en bout et une automatisation limitée. L'apprentissage est absent, le traitement coûteux, et les étapes de gouvernance et de confidentialité sont répétées à maintes reprises. Comment les entreprises peuvent-elles évoluer au rythme des activités, activer le libre-service, mieux servir leurs clients, augmenter leur efficacité opérationnelle et innover rapidement ?

[2] Cisco, Global Cloud Index Forecast and Complete Visual Networking Index Forecast
[3] LinkedIn, « LinkedIn's Most Promising Jobs of 2019. »
[4] Datanmi, « Data Engineering Continues to Move the Employment Needle. »

C'est là que l'IA entre en scène. L'IA peut automatiser et simplifier les tâches liées à la gestion des données — à travers la découverte, l'intégration, le nettoyage, la gouvernance et la maîtrise des données. Les méthodes de Machine Learning peuvent apprendre et prendre en charge les tâches routinières et répétitives, ce qui permet aux développeurs et aux utilisateurs de travailler sur des projets innovants à forte valeur ajoutée. L'IA améliore la compréhension des données et identifie les anomalies de confidentialité et de qualité des données. L'IA est un partenaire idéal pour les développeurs, les analystes, les gestionnaires et les utilisateurs métiers. Elle accélère les tâches grâce à l'automatisation et à l'augmentation, via des recommandations et des mesures efficaces.

L'IA est plus efficace lorsque vous réfléchissez à la manière dont elle peut vous aider à accélérer les processus de bout en bout dans l'ensemble de votre environnement de données. C'est pourquoi nous considérons l'IA comme essentielle à la gestion des données, et la raison pour laquelle Informatica® a concentré ses investissements en matière d'innovation sur le moteur CLAIRE®, notre capacité d'IA basée sur les métadonnées. CLAIRE exploite toutes les métadonnées unifiées de l'entreprise pour automatiser et faire évoluer les tâches routinières de gestion et de prise en charge des données.

Les quatre principaux avantages de l'IA pour la gestion des données
En général, l'IA présente quatre principaux avantages pour les équipes de gestion des données : amélioration de la productivité des professionnels des données, augmentation de l'efficacité des opérations, expérience de données guidée plus intelligemment et compréhension approfondie, et accélération des processus de gouvernance des données. Voici quelques exemples de ce qui est possible aujourd'hui.

Productivité : Un système recommandé pour l'intégration de données aide les data engineers à créer rapidement des mappings afin d'extraire, de transformer et de livrer des données. Le système de recommandation apprend à partir des mappings existants, comprend le contenu métier des bases de données et des systèmes de fichiers, et suggère des transformations appropriées pour la normalisation et le nettoyage des données, avant de les livrer aux systèmes cibles et aux consommateurs de données.

Efficacité : Dans une entreprise type, des milliers de processus d'intégration de données sont exécutés chaque jour. La surveillance de ces processus est en grande partie passive, les outils d'administration enregistrant simplement le temps passé et la consommation de processeur et de mémoire. L'IA peut tirer les leçons des valeurs historiques des données de séries chronologiques dans les fichiers journaux et de surveillance et signaler de manière proactive les valeurs aberrantes, et prévoir les difficultés pouvant survenir si elles ne sont pas traitées à l'avance.

Expérience de données : Lorsqu'une entité réelle (par exemple, un dossier patient ou une commande) est stockée dans une base de données ou un ensemble de fichiers, ses données sont déchiquetées et distribuées dans plusieurs tables ou fichiers — ce qui optimise leur stockage et leurs performances. L'IA peut détecter les relations entre les données et reconstituer rapidement l'entité d'origine. Les utilisateurs n'ont pas besoin de se souvenir ou de rechercher des documents obsolètes sur les relations clés primaires/étrangères, ni de joindre manuellement les différents ensembles de données. En outre, l'IA peut identifier des ensembles de données similaires et formuler des recommandations en fonction des modèles d'utilisation, de la qualité des données et de la collaboration commune.

Gouvernance des données : Une étape courante mais fastidieuse de la gouvernance de données consiste à associer les termes métiers aux éléments de données physiques afin d'établir le contexte métier et la pertinence des éléments de données, et de rendre les données compréhensibles aux utilisateurs. Dans de nombreux cas, l'IA peut lier automatiquement les termes métiers aux données physiques à l'aide d'une combinaison de techniques de traitement du langage naturel (NLP) et d'identification du type d'entreprise. Cela peut réduire considérablement la corvée que représente cette tâche sujette aux erreurs. À l'ère du Cloud, il est important de savoir que cette approche fonctionne également pour les applications SaaS. Les métadonnées peuvent être intégrées à partir des applications SaaS telles que Salesforce et Workday, et ajoutées au catalogue de l'entreprise.

Gestion des données basée sur l'IA : exemple dans le secteur bancaire
Pour illustrer pourquoi l'IA requiert une gestion des données et pourquoi les données requièrent une IA, penchons-nous sur cet exemple du secteur bancaire.

En appliquant l'IA à de plus en plus de données afin d'obtenir des analyses avancées, prédictives et en temps réel, les banques peuvent :
- proposer des services plus personnalisés qui augmentent la fidélisation des clients ;
- réduire les transactions frauduleuses sur les points de vente ;
- augmenter les résultats des investisseurs particuliers tout en réduisant le coût des conseillers en gestion de patrimoine ;
- réduire le coût de la conformité réglementaire liée aux projets.

Du point de vue de la gestion des données, l'IA peut détecter et cataloguer automatiquement tous les types de données pertinentes, comme les ERP, les CRM, les applications Cloud et Web, les fichiers machine, les journaux, les données tierces, etc. Les data scientists ont ainsi une longueur d'avance pour accéder à toutes les données dont ils ont besoin pour mener des centaines d'expériences à la recherche de modèles qui révèlent des informations sur le comportement des consommateurs, les activités frauduleuses, les opportunités d'investissement correspondant à la propension au risque des consommateurs, et bien plus encore.

Concernant la gestion des données, l'IA peut automatiquement enrichir une vue à 360° des clients et des personnes concernées en détectant les relations existant entre les données des clients et en faisant correspondre les informations avec des personnes spécifiques. Cela permet aux entreprises de mieux interagir avec leurs clients grâce à des offres plus pertinentes, et de fournir une expérience fluide sur différents canaux, que ce soit en ligne, sur mobile ou téléphone. Une vue à 360° des personnes concernées aide les banques à détecter les schémas et les réseaux d'activités frauduleuses beaucoup plus rapidement, ce qui peut potentiellement permettre d'économiser des millions.

De plus, l'IA peut automatiser et guider les tâches d'intégration des données et de qualité des données pour combiner et nettoyer les données provenant de centaines de sources de données, augmentant ainsi la puissance prédictive des modèles et algorithmes analytiques. Il a été prouvé que des données plus nombreuses et de meilleure qualité, combinées à l'IA et au ML et à l'analyse avancée, produisent des résultats significatifs, tels que l'amélioration des meilleures offres et l'identification des fraudes.

L'IA favorise également la gouvernance des données, qui garantit que les stratégies ne sont pas seulement documentées, mais réellement appliquées. Cela permet aux professionnels de la sécurité des informations de se conformer aux réglementations en matière de confidentialité des données, telles que le Règlement général sur la protection des données (RGPD), la loi Sarbanes-Oxley (SOX), Bâle II et Bâle III, etc.

Informatica CLAIRE : « L'intelligence » dans l'Intelligent Data Management Cloud

L'approche d'Informatica pour améliorer la productivité de la gestion de données avec le Machine Learning est la suivante :

1. L'Intelligent Data Management Cloud™ : Nous proposons une plate-forme intégrée et Cloud native de gestion de données de bout en bout, pour une productivité maximale. Grâce à ses fonctionnalités de connectivité unifiée et de gestion des métadonnées et des opérations, la plate-forme accélère le développement et le déploiement des nouveaux projets de gestion des données. Elle offre un ensemble puissant et cohérent de capacités permettant de gérer les données provenant de sources on-premise, Cloud, multi-Cloud et multi-hybrides. Nous avons nommé cette plate-forme de gestion de données unifiée Intelligent Data Management Cloud.

C'est une plate-forme modulaire : Vous pouvez démarrer avec n'importe quel outil et progresser à votre propre rythme :

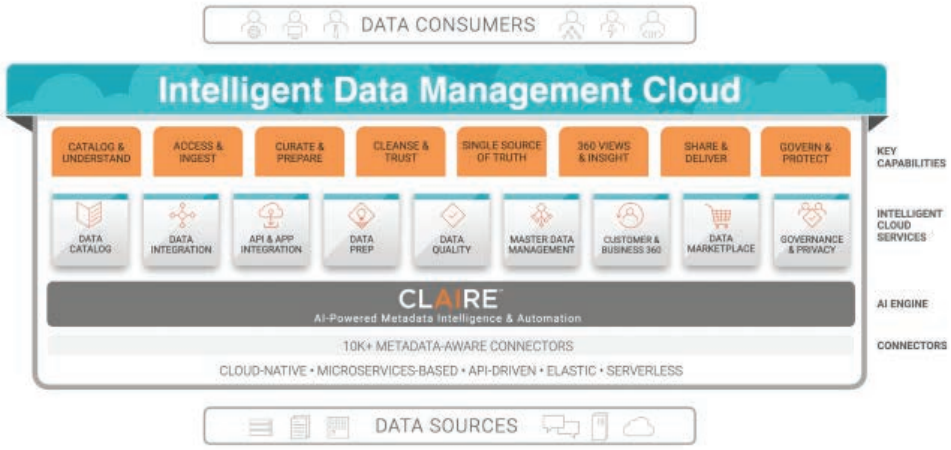

Figure 1 : L'Intelligent Data Management Cloud intègre des capacités de gestion des données avec la connectivité partagée, les analyses opérationnelles et l'intelligence basée sur les données et les métadonnées.

2. Métadonnées : Informatica est depuis longtemps reconnue pour être un leader dans la gestion des métadonnées techniques et métiers. L'entreprise a depuis accru ses fonctionnalités dans ce domaine, avec la collecte d'un plus large ensemble de métadonnées dans toute l'entreprise, incluant :
   - les métadonnées techniques, telles que les tables de base de données, les informations de colonne, les statistiques de profil de données, les scripts et la traçabilité des données ;
   - les métadonnées métiers, qui capturent le contexte des données, sa signification, sa pertinence et son importance par rapport aux divers processus et fonctions métiers ;
   - les métadonnées opérationnelles sur les systèmes et l'exécution des processus pour répondre à des questions telles que : la date de la dernière mise à jour des données la date de la dernière exécution de processus de chargement ou les données les plus consultées ;
   - les métadonnées d'utilisation relatives aux activités de l'utilisateur, y compris les ensembles de données et les résultats de recherche consultés, les classements et les commentaires.

Cette collecte plus large de métadonnées est essentielle au Machine Learning. Elle fournit les ensembles de données utilisés pour former les algorithmes du Machine Learning et permet à ces derniers de s'ajuster et de produire de meilleurs résultats.

3. Intelligence : Avec CLAIRE, Informatica propose une combinaison intégrée de métadonnées et de Machine Learning/IA.

   Les métadonnées collectées par l'Intelligent Data Management Cloud offrent un vaste ensemble d'informations que les algorithmes de CLAIRE peuvent utiliser pour assimiler l'environnement de données de l'entreprise. Ces connaissances permettent à CLAIRE de faire des recommandations intelligentes, d'automatiser le développement et la surveillance des projets de gestion des données, et de s'adapter aux changements dans et hors de l'entreprise. CLAIRE permet d'alimenter l'intelligence de toutes les capacités de gestion des données de l'Intelligent Data Management Cloud.

CLAIRE aide de nombreux utilisateurs :

- Les data engineers bénéficient de l'automatisation partielle ou totale de nombreuses tâches d'implémentation
- Les analystes de données peuvent localiser et préparer les données dont ils ont besoin plus facilement
- Les utilisateurs métiers peuvent identifier rapidement les données qui doivent être soumises à la gouvernance des données et aux contrôles de conformité recommandés
- Les experts en données comprennent les données plus rapidement
- Les gestionnaires de données visualisent plus facilement la qualité des données
- Les professionnels de la sécurité et de la confidentialité des données peuvent détecter plus facilement les utilisations abusives de données, protéger les données sensibles et prouver que les contrôles appropriés sont en place
- Les administrateurs et les opérateurs bénéficient de toute la puissance de la maintenance prédictive et de l'optimisation des performances des processus de gestion de données

Voici quelques exemples d'utilisation de l'intelligence de CLAIRE.

CLAIRE pour le catalogage des données

La découverte et la compréhension des données dont vous disposez constituent la première étape de toute initiative axée sur les données. CLAIRE fournit un moteur de découverte basé sur le Machine Learning pour analyser et cataloguer les ressources de données au sein de l'entreprise. Un catalogue de données intelligent optimisé par CLAIRE peut aider les data scientists, les analystes et les data engineers à trouver et à recommander les données dont ils ont besoin, ce qui réduit considérablement le temps consacré à la découverte et à la préparation des données.

Découverte avancée des relations

L'une des principales tâches de catalogage et de modélisation des données consiste à documenter les relations entre les ensembles de données. CLAIRE utilise des techniques de Machine Learning pour identifier automatiquement les clés primaires et uniques, ainsi que les jonctions entre les ensembles de données structurés. Des mois d'efforts de documentation sont ainsi réduits à quelques minutes. CLAIRE améliore en permanence sa capacité à identifier les relations en incluant des êtres humains dans le processus de conservation des données — par exemple, les utilisateurs peuvent accepter ou rejeter les relations déduites et CLAIRE apprend de ces actions.

Par exemple, l'analyste de données d'une banque qui crée un rapport sur les clients les plus susceptibles de répondre à une campagne marketing doit pouvoir trouver des informations sur les produits existants et les prêts pour tous les clients. Cependant, étant donné la nature cloisonnée des données au sein de l'entreprise, il est difficile de trouver de tels ensembles de données au sein des services et des data stores. CLAIRE utilise des jonctions documentées dans les bases de données, des jonctions effectuées dans d'autres outils tels que BI et ETL, ainsi que des statistiques dérivées des valeurs des données pour déduire et recommander des jonctions à l'analyste de données. Cela permet d'élargir l'analyse de l'utilisateur et d'utiliser toutes les informations disponibles pour trouver le public cible approprié pour la campagne.

CLAIRE combine plusieurs techniques pour la découverte des clés et des jonctions. Concernant les clés, les statistiques de profilage telles que l'unicité, les nombres nuls, les métadonnées de colonne (par exemple, les noms de colonne contenant « ID »), entre autres, sont utilisées pour découvrir les clés primaires et uniques. Les jonctions et l'inférence des clés de jonction utilisent ensuite une combinaison de techniques de Machine Learning, telles que l'analyse des signatures de colonne, pour découvrir les jonctions à grande échelle au sein de nombreux ensembles de données potentiels.

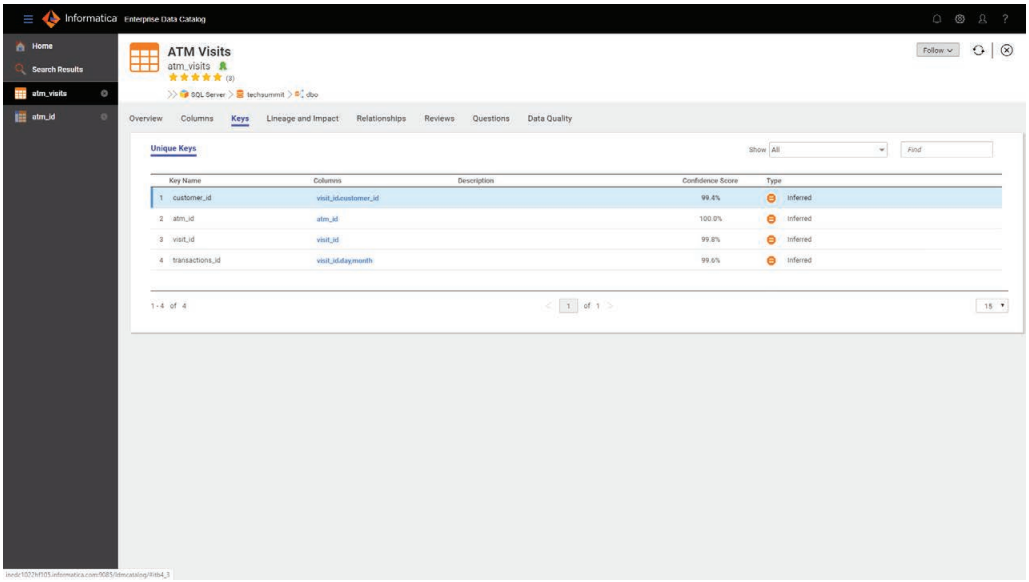

Figure 2 : Découverte de clés uniques grâce à l'inférence à l'aide de techniques de Machine Learning.

Similarité intelligente des données

CLAIRE utilise les techniques de Machine Learning telles que le clustering pour détecter les similarités entre les données réparties dans des milliers de bases de données et d'ensembles de fichiers. La similarité intelligente des données est une des capacités clés utilisées pour de multiples objectifs comme l'identification des données, la détection des doublons, l'association des champs de données individuels dans les entités métiers, la propagation des balises dans les ensembles de données et la recommandation d'ensembles de données aux utilisateurs.

La similarité des données calcule dans quelle mesure les données de deux colonnes sont identiques. L'utilisation d'une approche manuelle pour tenter de comparer les colonnes par paires dans un ensemble de données d'entreprise (par exemple, 100 millions de colonnes) serait trop coûteuse en ressources de calcul. Pour sa part, la similarité des données utilise des techniques de Machine Learning pour regrouper les colonnes similaires et identifier les correspondances potentielles.

Le processus fonctionne en plusieurs étapes. Tout d'abord, les colonnes sont regroupées en fonction de leurs caractéristiques. Puis les chevauchements de données sont traités pour déterminer les valeurs uniques de chaque cluster. Enfin, les paires les plus prometteuses sont sélectionnées pour rechercher les similarités de données à l'aide des coefficients Bray-Curtis et Jaccard.

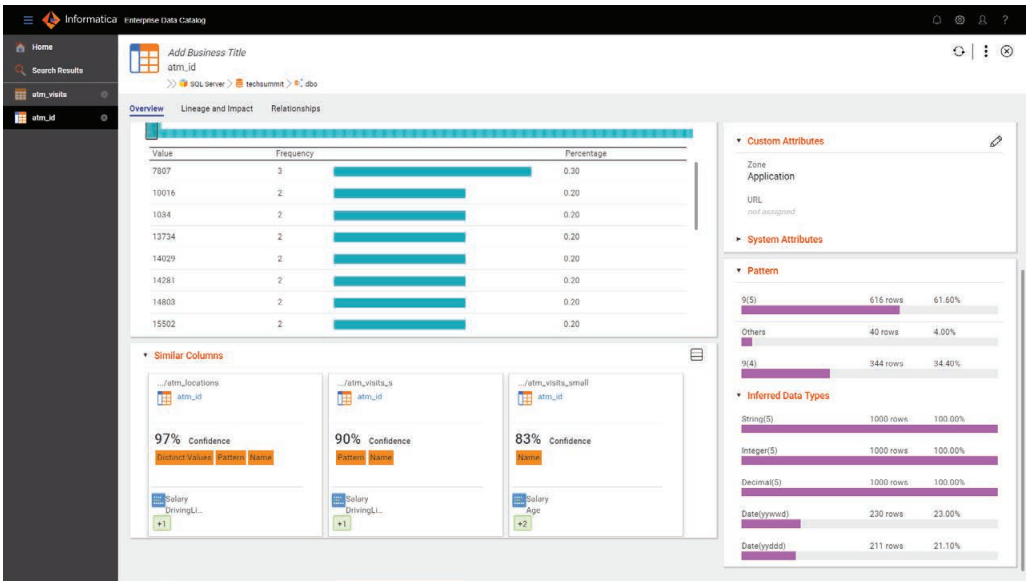

Figure 3 : Identification de colonnes similaires à l'aide du regroupement et des coefficients Bray-Curtis et Jaccard.

Découverte intelligente des domaines à l'aide des balises

CLAIRE peut classer les champs de données en appliquant des étiquettes sémantiques à chaque colonne. Ces étiquettes sémantiques sont appelées des domaines de données.

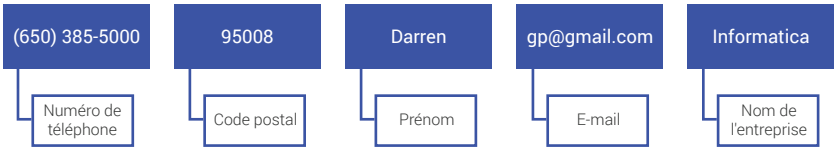

Figure 4 : CLAIRE classe automatiquement les champs de données et applique des étiquettes sémantiques appelées balises.

Généralement, les étiquettes sémantiques sont appliquées en évaluant des règles basées sur des expressions régulières, des tables de références ou autres logiques complexes codées manuellement. La définition et la maintenance de milliers de règles telles que celles-ci peuvent s'avérer laborieuses.

CLAIRE utilise le concept des balises pour simplifier au maximum le processus de découverte et d'étiquetage des champs de données. Pour les colonnes qui n'ont pas été classées, l'utilisateur doit simplement attribuer une balise (par exemple, « Date de paiement demandée ») indiquant le contenu de la colonne. Le système apprend par association, puis propage automatiquement ces balises aux colonnes identiques. La « reconnaissance faciale » pour les technologies de données est similaire à celle utilisée pour identifier les gens sur une photo Facebook, avec un avantage : les mêmes personnes sont identifiées simultanément sur des millions d'autres photos.

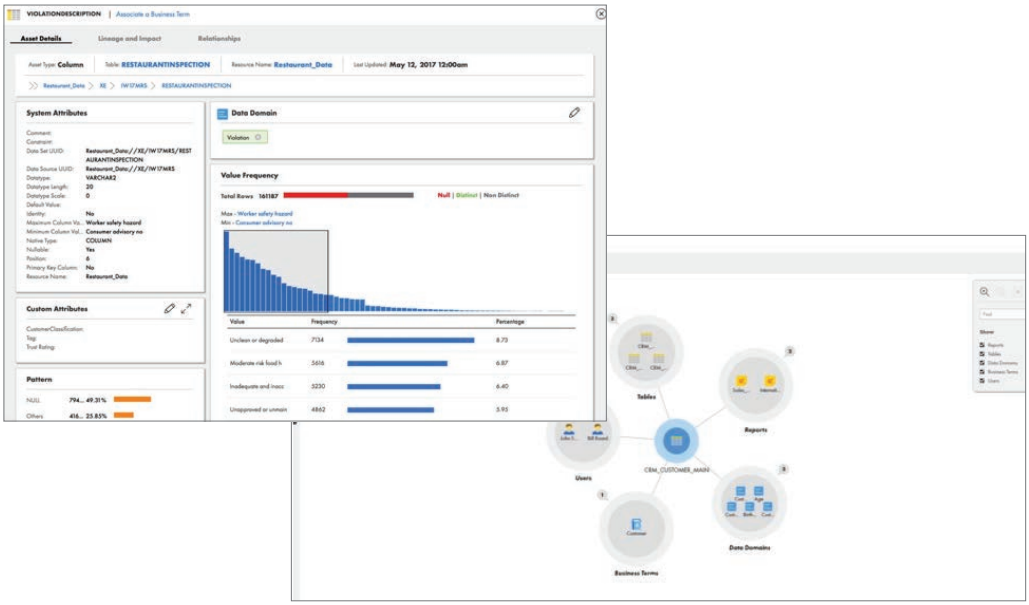

Figure 5 : Classification automatique des données.

Découverte intelligente des entités

Une fois les domaines des colonnes identifiés, CLAIRE peut assembler ces champs individuels dans des entités métiers de plus haut niveau. L'exemple ci-dessous montre comment créer une entité appelée Bon de commande en combinant les champs identifiés comme Client et Produit. La découverte d'entités apprend en se fondant sur la façon dont les utilisateurs ont assemblé des champs de données disparates dans leurs processus d'analyse ou d'intégration de données, et applique cet apprentissage pour créer des entités dans tout l'environnement de données de l'entreprise.

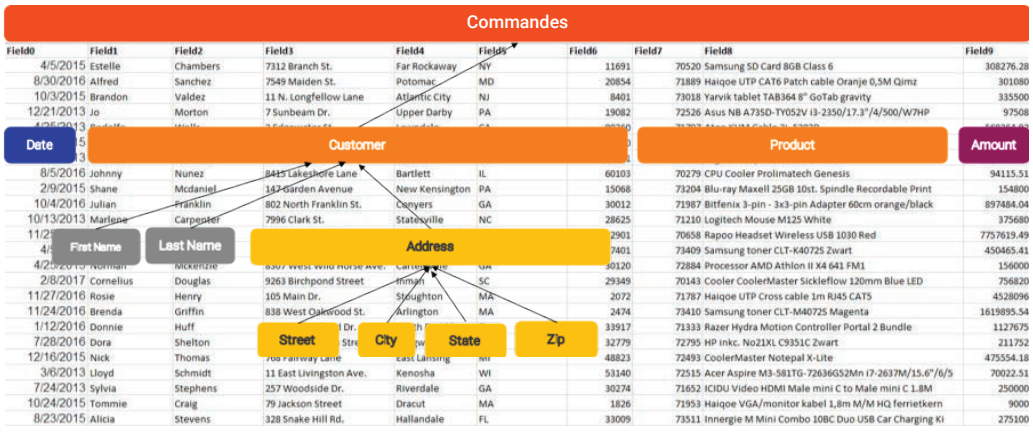

Figure 6 : Combinaison des domaines de données pour détecter des entités à partir des tables et des fichiers.

CLAIRE pour l'analyse

L'automatisation et l'intelligence basées sur CLAIRE accélèrent considérablement les informations et les processus analytiques, augmentent la disponibilité des données et rationalisent la préparation des données pour l'analyse. CLAIRE améliore la productivité des data engineers grâce à des recommandations de pipeline de données et à la possibilité d'analyser automatiquement des données complexes et multi-structurées.

Recommandations de transformation

Bouclez la boucle de conception et améliorez la productivité des data engineers grâce à la création automatisée de mappings d'intégration des données avec des prévisions pour les prochaines transformations et expressions. Lorsqu'une entreprise choisit de recevoir des recommandations basées sur CLAIRE, les métadonnées anonymes provenant des pipelines de données de l'entreprise sont analysées et l'IA et le ML sont appliqués pour proposer des recommandations de conception. Ces métadonnées sont utilisées pour générer des recommandations de transformation et d'expression. CLAIRE s'améliore à chaque utilisation — acceptation ou rejet de la recommandation. Cela accélère le développement, automatise les tâches répétitives et permet à un plus grand nombre d'utilisateurs de se connecter et d'intégrer rapidement les données.

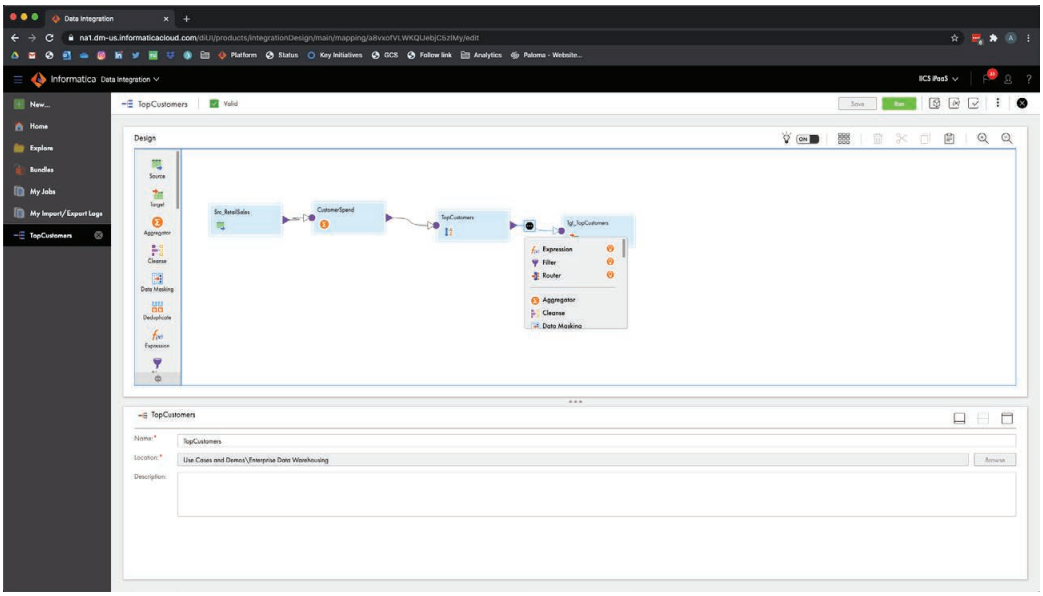

Figure 7 : CLAIRE recommande les transformations les plus adaptées lors de la création de pipelines de données.

Exécution optimisée des processus à grande échelle

CLAIRE utilise diverses méthodes d'optimisation pour améliorer les performances d'intégration dans l'ensemble du pipeline de données. Un optimiseur intelligent choisit le moteur de traitement le plus approprié à l'exécution d'une charge de travail Big Data en fonction des caractéristiques de performances ; les recommandations de mapping sont présentées aux data engineers en fonction des activités passées des utilisateurs, et un optimiseur basé sur les coûts et l'heuristique modifie intelligemment l'ordre de jonction dans un pipeline de données pour des performances optimales. Voici quelques exemples de la manière dont CLAIRE optimise les pipelines de données.

Recommandations relatives aux colonnes de jonction

CLAIRE suggère automatiquement des colonnes de jonction (c'est-à-dire des clés de jonction) lorsqu'un utilisateur choisit l'action de combiner deux ensembles de données. Cela permet aux analystes de données d'économiser des centaines d'heures de travail manuel visant à identifier la meilleure façon de fusionner des ensembles de données en un ensemble composite à des fins d'analyse. CLAIRE commence par les relations de clés primaires et étrangères (Pk-FK) définies dans les systèmes sources d'origine (par exemple, les bases de données relationnelles telles qu'Oracle) des ensembles de données importés dans le data lake. Si les mêmes ensembles de données sont joints dans d'autres projets, ces informations de colonne de jonction seront également utilisées pour les recommandations. Toutes ces informations sont traitées et classées par CLAIRE pour suggérer les meilleures colonnes de jonction entre deux ensembles de données. De plus, en fonction de l'échantillonnage des ensembles de données, le pourcentage de chevauchement des données entre les colonnes suggérées est également affiché.

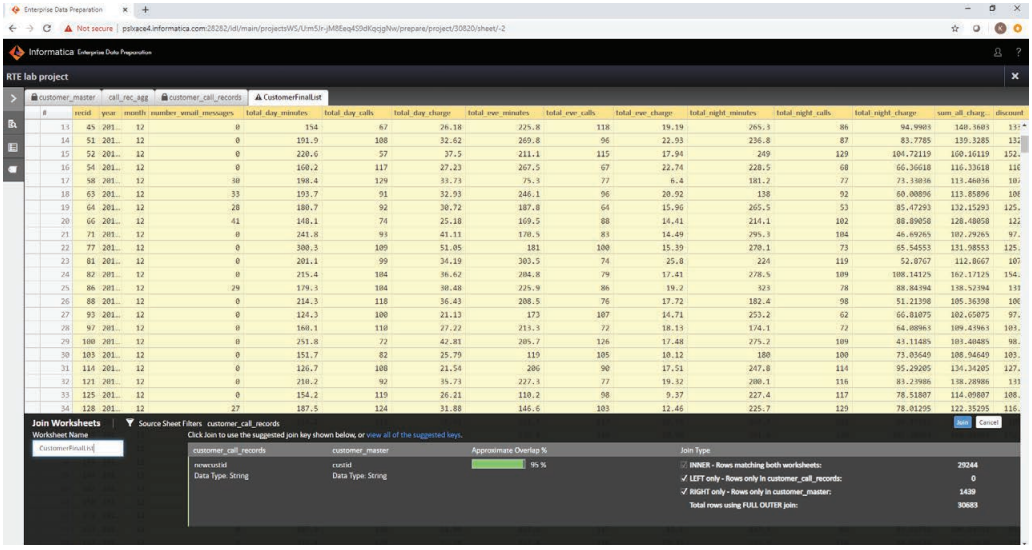

Figure 8 : Suggestions automatiques de colonnes de jonction lors de la combinaison de deux ensembles de données.

Recommandations de visualisation Apache Zeppelin

Informatica Enterprise Data Preparation utilise Apache Zeppelin pour afficher les feuilles de route sous la forme d'un bloc-notes contenant des graphiques et des tableaux. Lorsque l'utilisateur ouvre le bloc-notes d'une publication, il peut voir les recommandations de visualisation de CLAIRE. Lorsque l'utilisateur ouvre le bloc-notes pour la première fois après sa publication, il voit des histogrammes de colonnes numériques dérivées. Si la publication ne contient pas de colonnes numériques dérivées, l'utilisateur voit une requête de table « SELECT * FROM » dans le premier paragraphe du bloc-notes.

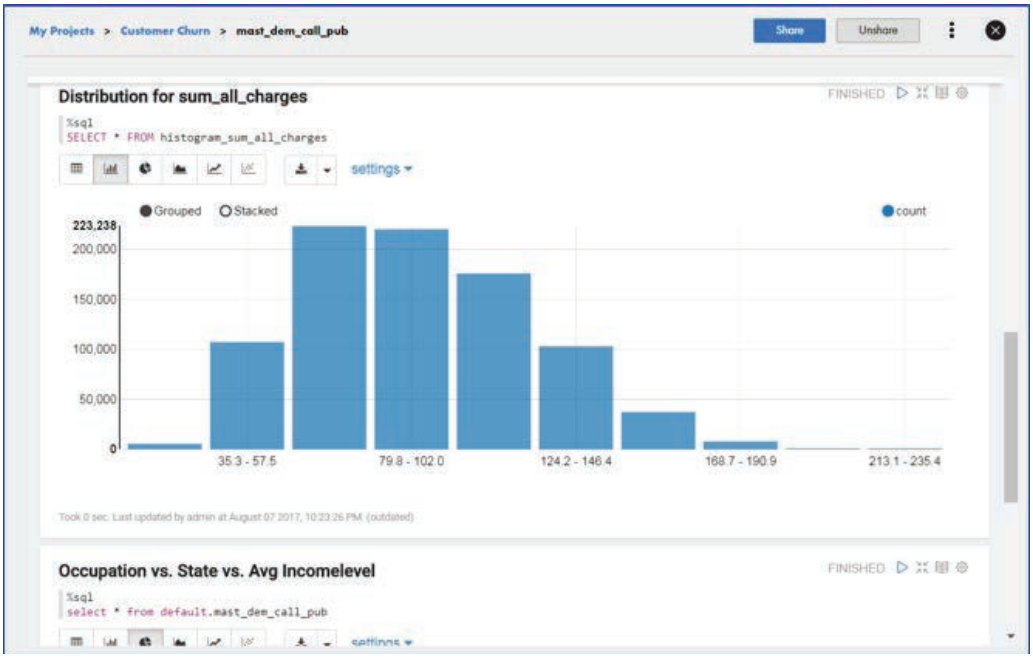

Figure 9 : Visualisation recommandée dans le bloc-notes Apache Zeppelin.

Recommandations intelligentes de données

CLAIRE fournit aux analystes de données et aux data scientists des suggestions sur les ensembles de données à utiliser pour leurs projets. CLAIRE étudie les ensembles de données que les utilisateurs ont sélectionnés et suggère des ensembles similaires ou mieux classés, ou des ensembles supplémentaires pour compléter ceux dont les utilisateurs se servent déjà. Les recommandations intelligentes de données permettent d'éviter aux utilisateurs de répéter des tâches déjà effectuées par leurs collègues. Les recommandations comprennent :

- une version préparée des données identiques (données substituables) ;
- une autre table contenant le même type d'enregistrements (données unifiables) ;
- une table susceptible d'être adjointe pour enrichir les données avec des attributs supplémentaires (données adjoignables).

Les recommandations de données utilisent des techniques de filtrage basées sur le contenu pour fournir des suggestions sur les ensembles de données supplémentaires. Les caractéristiques (termes) utilisées pour les ensembles de données comprennent les informations de traçabilité, le classement des utilisateurs et la similarité des données. Plusieurs mesures de similarité sont utilisées pour noter les équivalences entre les différents ensembles de données. Ces notes sont ensuite utilisées pour recommander des ensembles de données possédant des propriétés similaires. Des recommandations d'éléments complémentaires sont effectuées via la recherche dans le graphique des métadonnées afin de trouver des ensembles de données couramment utilisés ensemble par différents utilisateurs.

Découverte intelligente des structures

De plus en plus de données sont générées et collectées sur des machines, des entreprises et des applications au format non structuré ou non relationnel. Ces types de données se caractérisent non seulement par les grands volumes, mais également par leur vitesse, leur variété et leur variabilité. Le terme « data drifting » est aujourd'hui couramment utilisé pour décrire la fluctuation du format, du rythme et du contenu des données dans ces nouveaux types de données.

La solution Informatica Intelligent Structure Discovery (ISD), dotée de CLAIRE, est conçue pour automatiser le processus d'ingestion et d'intégration de fichiers afin que les entreprises puissent découvrir et analyser des fichiers complexes. ISD fournit une prise en charge prête à l'emploi de divers formats de fichiers de données, notamment les flux de clics, les journaux d'Internet des objets, les fichiers CSV, les fichiers texte délimités, les fichiers XML, JSON, Excel, ORC, Parquet, Avro, les formulaires PDF et les fichiers de table Word. CLAIRE peut automatiquement extraire la structure de ces fichiers, ce qui les rend plus faciles à comprendre et à utiliser. En utilisant une approche basée sur le contenu pour analyser les fichiers, CLAIRE peut s'adapter aux fréquentes modifications de fichiers sans affecter leur traitement.

ISD utilise un algorithme génétique pour automatiser la reconnaissance des schémas dans les fichiers. Cette approche utilise le concept d'« évolution » pour améliorer les résultats. Chaque solution candidate comprend un ensemble de propriétés qui peuvent être automatiquement modifiées puis testées pour déterminer si elles fournissent une meilleure solution. Ces structures sont ensuite notées en fonction de plusieurs facteurs, tels que la couverture de saisie et les domaines dérivés. Les structures les mieux notées entrent dans une phase de « mutation » pendant laquelle plusieurs changements sont apportés aux structures, par exemple, en combinant des sous-structures pour voir si la note augmente. Le processus se termine lorsque la structure est jugée adaptée aux données.

ISD utilise également des mécanismes personnalisés de NER (reconnaissance d'entités nommées) et de NLU (compréhension du langage naturel) basés sur le ML pour identifier les champs et les types de champs, ce qui simplifie les intégrations suivantes et permet aux applications externes d'utiliser ISD comme moteur NER/NLU sous-jacent. Par exemple, ISD est utilisé pour détecter les informations d'identification personnelle dans la charge utile API entrante et sortante et facilite la conformité réglementaire et la sécurité de l'entreprise. ISD est également utilisé dans les cas d'usage de découverte, d'ingestion et de streaming de données.

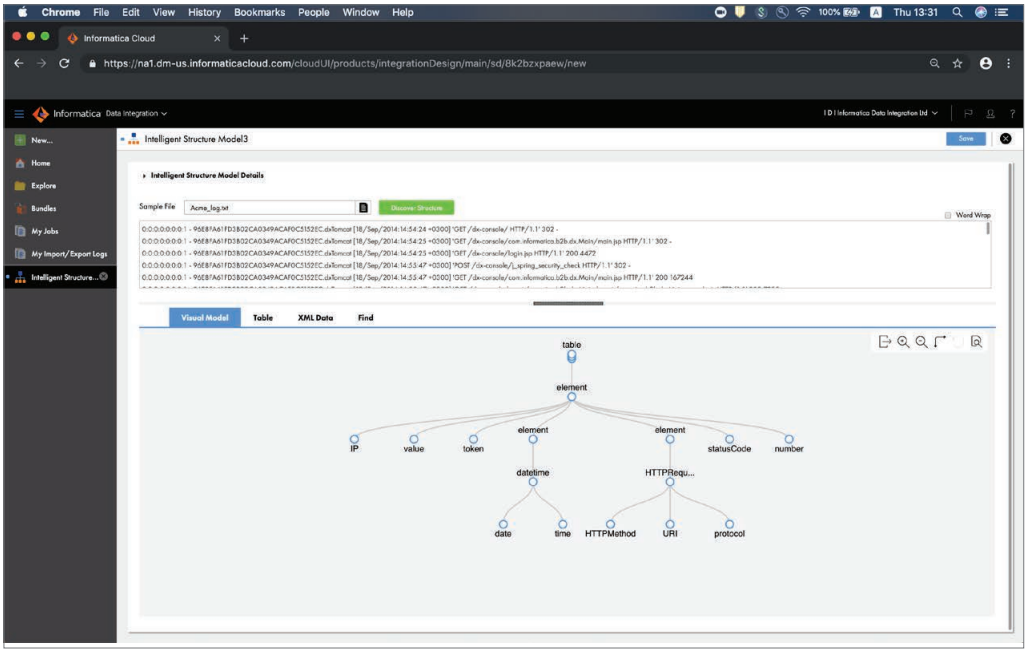

Figure 10 : Recherche intelligente des structures dans des fichiers de données non structurées

CLAIRE pour la gestion des données de référence

L'automatisation et l'intelligence basées sur CLAIRE à l'aide de l'IA et du Machine Learning avancés enrichissent et améliorent la précision des vues à 360° de l'entreprise pour les clients, les produits, les fournisseurs et d'autres domaines. Une variété de techniques d'IA et de ML mixtes allant des algorithmes déterministes, heuristiques et probabilistes à la mise en correspondance de synthèse contextuelle et à la mise en correspondance d'entités d'apprentissage actif sont utilisées pour fournir une mise en correspondance et un enrichissement rapides, évolutifs et extrêmement précis des enregistrements des données de référence.

Mise en correspondance de synthèse

La synthèse est une technique de mise en correspondance de nouvelle génération qui permet, par exemple, de faire correspondre les prospects aux clients, d'associer les interactions et les données non structurées aux clients, et de découvrir des relations suspectes. Pour ce faire, elle utilise des « attributs contextuels », le Machine Learning, le NLP et une combinaison de mise correspondance probabiliste et de règles déclaratives.

Les attributs démographiques (par exemple, nom, adresse et numéro de téléphone), les attributs d'interaction (par exemple, e-mail, webchats) et les attributs contextuels (par exemple, quand, quoi, où, qui) permettent de mettre en correspondance les données relatives aux clients selon un niveau de confiance donné. Le NLP peut extraire les « attributs contextuels » provenant de textes non structurés afin de fournir de nombreux points de données supplémentaires à utiliser dans le processus de mise en correspondance. Un algorithme de ML peut être très efficace dans la mise en correspondance lors de l'utilisation d'une approche de formation supervisée dans laquelle les gestionnaires de données et les experts du domaine déterminent si les paires de correspondances d'un ensemble dûment sélectionnées sont, ou non, des correspondances. Ces paires de correspondances étiquetées forment un ensemble de formation utilisé pour produire un algorithme de correspondance.

La synthèse permettra d'obtenir une vue complète à 360° des clients, composée de données démographiques, de comptes, de transactions, d'interactions et de données non structurées. Les algorithmes de mise en correspondance traditionnels fusionnent les enregistrements pour former une vue client unique, tandis que la mise en correspondance de synthèse gère toutes les données clients dans un graphique. Les données sont associées à des niveaux de confiance, où il est alors possible de fournir plusieurs vues, ou « perspectives », relatives à un client.

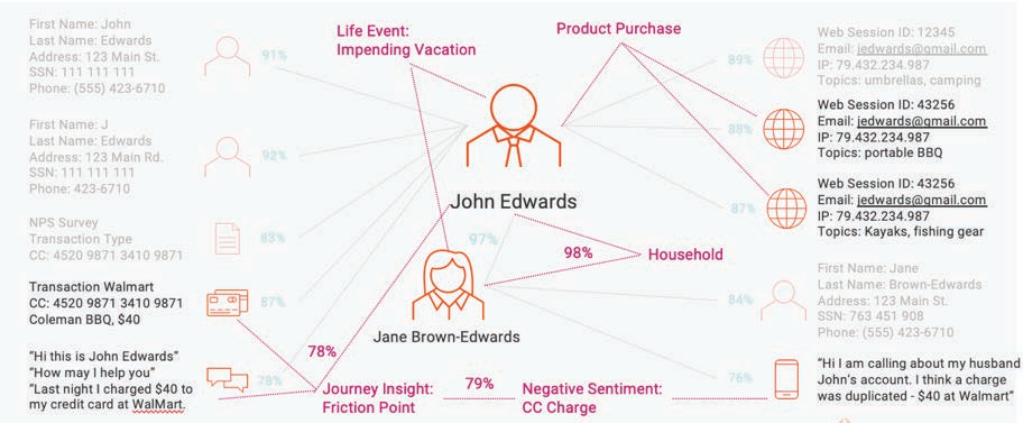

Figure 11 : La mise en correspondance de synthèse et le raisonnement permettent de déduire des renseignements qui sont ensuite stockés dans Customer 360.

Rapprochement d'identités

Le rapprochement d'identité NAME3 de CLAIRE offre plus de 30 ans de formations et de réglages à l'aide de diverses techniques, telles que la génération de clés intelligentes pour l'indexation et le blocage, la stabilisation sémantique du texte et la comparaison des données relatives aux parties et à la localisation, les listes d'édition et les règles de stabilisation du texte pour 80 populations, ainsi qu'une pondération intelligente de l'importance des fonctionnalités pour différents objectifs. Ces techniques puissantes permettent l'indexation et le blocage de plusieurs champs, des règles de correspondance et d'anti-correspondance définies par le client selon les exigences, ainsi que des règles de correspondance et d'anti-correspondance définies par l'implémentation pour compléter d'autres règles d'IA.

Mise en correspondance des entités

La mise en correspondance d'entités recherche les enregistrements de données qui font référence à la même entité réelle (par exemple, clients, produits, etc.). Les enregistrements de données peuvent être non structurés (par exemple, des informations clients masquées dans un chat en ligne) et structurés. La classification des correspondances compare une paire de correspondances et détermine s'il existe une correspondance, une correspondance éventuelle ou une non-correspondance, ainsi qu'un niveau de confiance. Certaines techniques utilisent des règles configurées par l'utilisateur (règles déclaratives) ou des règles configurées par l'IA (configuration provenant du Machine Learning). La combinaison de ces deux techniques permet d'obtenir de meilleurs résultats.

Les règles déclaratives, créées par des experts du domaine, complètent les puissantes règles d'IA utilisées par CLAIRE sous la forme d'un classificateur de forêt aléatoire appris. CLAIRE utilise l'apprentissage actif supervisé (par opposition à l'apprentissage participatif ou multi-utilisateur) pour accélérer le processus de formation de l'IA, où des micro-lots de 10 ou 20 paires de correspondances sont présentés à un utilisateur afin d'être étiquetés (en tant que correspondance, correspondance éventuelle, non-correspondance). Une fois que ces paires de correspondances sont étiquetées, CLAIRE réentraîne le classificateur de forêt aléatoire et identifie les meilleures paires de correspondances à présenter à un utilisateur dans le cadre de ce processus d'étiquetage itératif. CLAIRE utilise les paires étiquetées pour inférer les règles de blocage (c'est-à-dire supprimer les non-correspondances évidentes), effectuer le blocage, former un modèle et effectuer la mise en correspondance des entités.

CLAIRE utilise une combinaison de comparaisons/similarités de chaînes telles que Jaccard, des règles déclaratives dérivées du profilage des données, des ensembles de données stabilisés (fichiers de population, surnoms, comparaisons sémantiques, etc.) et des règles définies par l'utilisateur pour la gestion des exceptions. Ces règles déclaratives traitent les lacunes et les exceptions et aident à accélérer le processus de formation actif (c'est-à-dire à réduire le nombre de paires de correspondances requises pour l'apprentissage), à accélérer la création de fonctionnalités de règles d'IA et à améliorer la précision des correspondances. Par exemple, chaque fois que le nom, la date de naissance et le numéro de sécurité sociale sont fortement comparables, la règle considère qu'il s'agit d'une correspondance. Ce mélange de règles déclaratives et de règles d'IA accélère la formation et améliore la précision de la mise en correspondance.

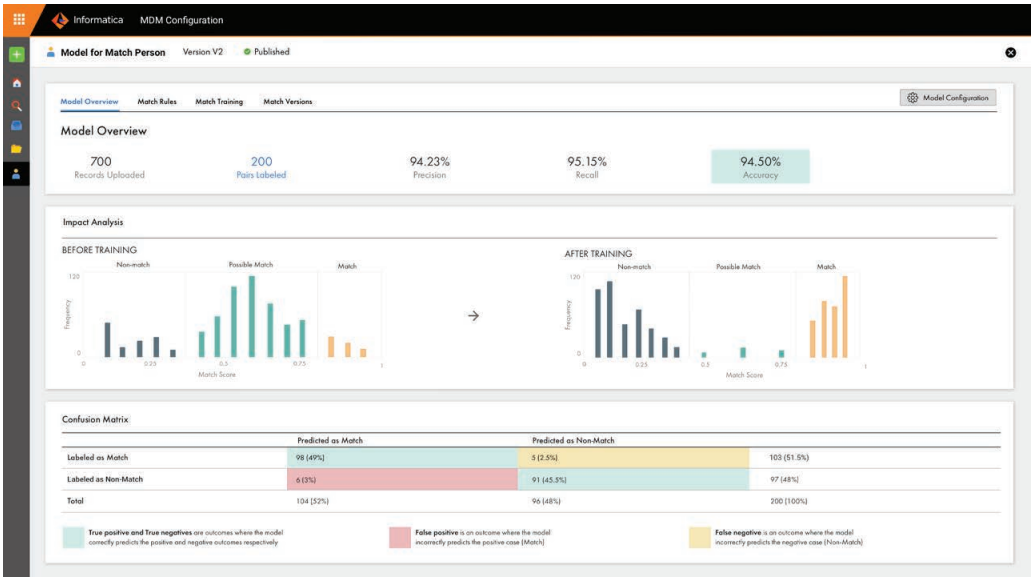

Figure 12 : Mise en correspondance des entités

CLAIRE pour la gouvernance et la conformité des données

L'IA et le Machine Learning sont essentiels pour automatiser les tâches de gouvernance des données les plus difficiles aujourd'hui : trouver des données, mesurer leur qualité et permettre la collaboration afin de les gérer. CLAIRE génère automatiquement des règles de stratégie (par exemple, la qualité des données) et relie la sémantique de l'entreprise aux métadonnées techniques. Elle aide également les utilisateurs à trouver les données les plus pertinentes et les plus fiables pour leurs besoins métiers.

Enrichissement automatique de la qualité des données

CLAIRE utilise une approche NLP basée sur Stanford NER pour analyser et extraire des entités de texte non structuré. Généralement, pour extraire des entités à partir de chaînes (par exemple, du code produit), les utilisateurs finissent par écrire des règles d'analyse à l'aide de tables de référence et d'expressions régulières. La quantité de données, la complexité et les modèles ne cessent d'augmenter ; écrire toutes les règles possibles afin de correspondre à chaque entrée n'est ni pratique, ni évolutif. Au lieu de cela, CLAIRE utilise des modèles pré-formés pour identifier et extraire les entités basées sur Stanford NER.

CLAIRE utilise le Machine Learning pour classer le texte entrant, par exemple : Langue, Type de produit et Problème de support technique. La méthodologie de Machine Learning utilisée est appelée apprentissage supervisé avec Naïve-Bayes et Max Entropy (régression logistique multinomiale). L'apprentissage supervisé est utilisé pour former les modèles et attribuer des étiquettes. Par la suite, le modèle formé peut être déployé pendant le traitement des données pour étiqueter, acheminer et traiter différentes classes d'entrée — par exemple, traiter les « problèmes de moteur » séparément des « problèmes de configuration » ayant des significations similaires et distinguer les utilisations des mots possédant plusieurs significations. CLAIRE automatise le balisage et la classification des images en exploitant les modèles NLP et ML pour la classification des produits et l'extraction des balises d'image.

Une grande entreprise internationale du secteur de la santé disposait d'un employé à temps plein pour le mapping de 21 000 ressources techniques avec 6 000 termes métiers, un processus qui a pris deux mois. Avec Axon Data Governance et Enterprise Data Catalog, CLAIRE a automatisé le mapping de 18 000 ressources techniques avec une précision de 99 % en 8 minutes.

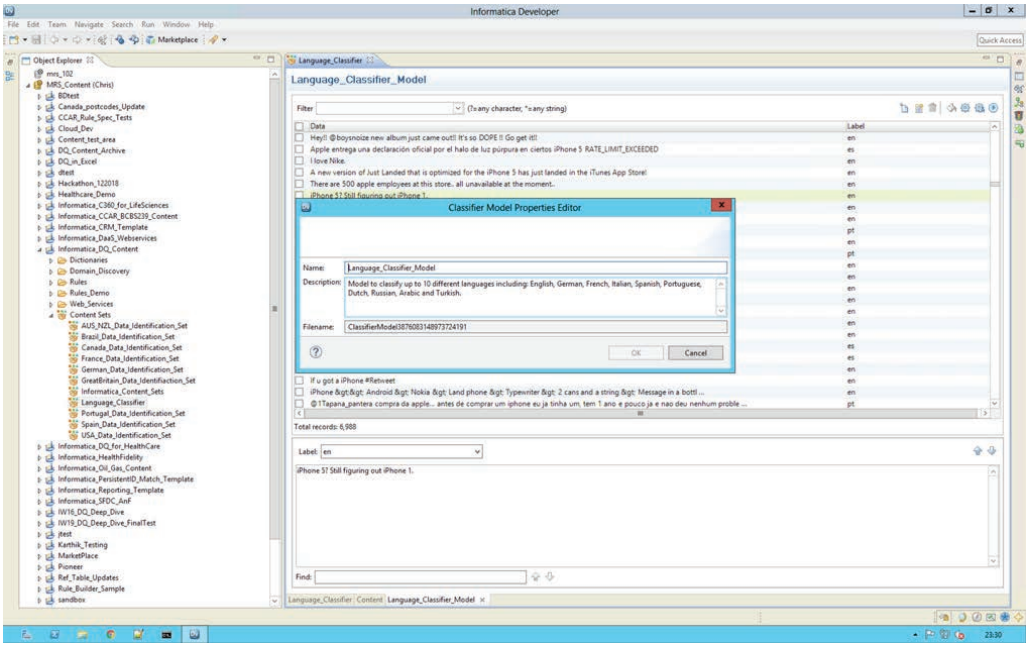

Figure 13 : Le NLP du Machine Learning classe le texte et extrait les entités.

Associer automatiquement les termes métiers aux ensembles de données physiques

La gouvernance de données nécessite la documentation des artefacts métiers, des définitions, des parties prenantes, des processus, des politiques, etc. Pour obtenir une vue véritablement alignée, les utilisateurs doivent pouvoir associer les définitions et les vues métiers aux implémentations techniques sous-jacentes au sein de leur système de données. En général, cette tâche est lente, laborieuse et sujette aux erreurs — elle repose en effet sur des personnes clés pour communiquer et aligner manuellement les manifestations techniques une par une — et peut prendre plusieurs jours, plusieurs semaines, voire plusieurs mois.

Informatica Axon Data Governance, grâce à une intégration étroite avec Informatica Enterprise Data Catalog, peut raccourcir ce processus. CLAIRE fournit aux utilisateurs des recommandations sur les éléments de données pertinents et appropriés à lier au fur et à mesure que les analyses de métadonnées sont terminées. Cela réduit la tâche de recherche, de validation et de liaison d'éléments de données, ce qui permet aux gestionnaires de données et au bureau de gouvernance des données de se concentrer sur leurs tâches critiques. Au fur et à mesure de l'avancement des implémentations, le processus peut être entièrement automatisé.

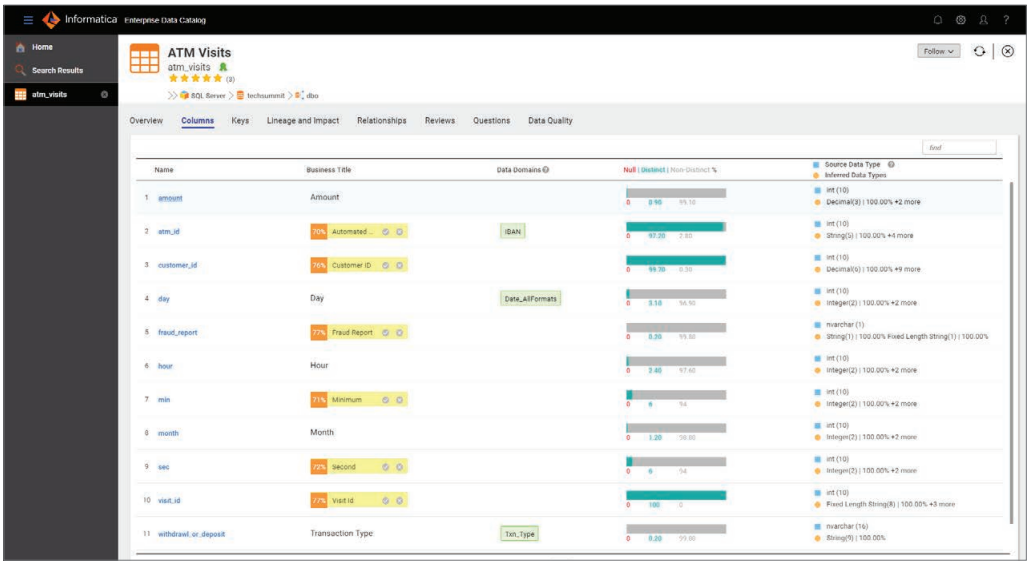

Figure 14 : Association automatique de termes métiers avec des ensembles de données physiques.

Évaluer automatiquement la qualité des données

La qualité des données au sein d'un système qui prend en charge un processus, soutient des politiques, etc. est un indicateur de performance clé (KPI) pour la gouvernance de données. Le bureau de gouvernance des données doit s'assurer que les données sont complètes, exactes, cohérentes, valides, etc. En bref, elles doivent être fiables et suffisamment bonnes pour soutenir les opérations métiers. À mesure que les implémentations de gouvernance des données se développent, l'évaluation de la qualité pour un nombre croissant de systèmes et de champs dans l'ensemble du paysage des données, des bases de données aux data lakes, prend de plus en plus de temps.

Grâce à CLAIRE, Axon Data Governance — en coordination avec Informatica Data Quality et Informatica Enterprise Data Catalog — peut automatiser l'application des mesures de qualité des données dans toute l'entreprise, ce qui permet d'économiser des milliers d'heures de travail. L'équipe de gouvernance des données associe les règles de qualité des données pour différentes dimensions de qualité des données aux termes métiers et aux éléments de données critiques. Le système sous-jacent génère ensuite les contrôles de qualité des données requis sur les différents systèmes et rapporte les mesures au bureau de gouvernance.

Cette automatisation est rendue possible par la combinaison de trois éléments d'information clés :

1. Connaissance des éléments métiers critiques et des règles de qualité des données requises par Axon
2. Règles de qualité des données portables et exécutables, et moteur d'exécution flexible d'Informatica Data Quality
3. Détails des métadonnées provenant des ressources de données physiques d'Enterprise Data Catalog

CLAIRE combine ces informations pour générer des tâches d'exécution des règles de qualité des données dans Informatica Data Quality par rapport aux ressources de données physiques d'Enterprise Data Catalog. CLAIRE conserve également le contexte de l'utilisateur métier d'Axon pour s'assurer que les résultats sont affichés dans les tableaux de bord appropriés et dans les vues agrégées pour être utilisés par le bureau de gouvernance.

L'automatisation permet aux programmes de gouvernance de s'adapter plus rapidement que jamais, éliminant des milliers d'heures de travail manuel associées à la création d'évaluations de la qualité des données et à leur mise en relation avec le contexte de gouvernance, une par une. CLAIRE s'assure également que la qualité des nouvelles ressources actives identifiées est automatiquement évaluée. En outre, de nouveaux domaines sont découverts à l'aide de l'extraction d'entités nommées ou du classificateur dans les règles de qualité des données.

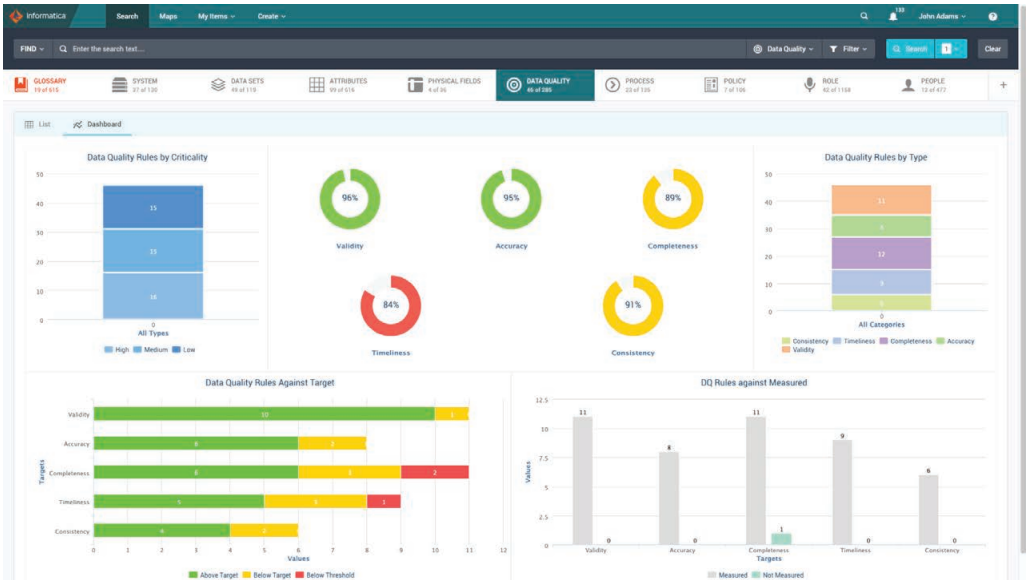

Figure 15 : Les évaluations automatiques de la qualité des données sur l'ensemble du système de données permettent d'économiser des milliers d'heures de travail manuel.

Identification et règle de qualité des données assistées par ML/NLP

La qualité des données est un impératif essentiel pour tout programme de gouvernance des données, et les grandes implémentations peuvent comporter de nombreuses règles de qualité des données. Pour aider les gestionnaires de données à identifier les règles à utiliser, CLAIRE peut non seulement aider à identifier les règles, mais également générer les règles manquantes. Un utilisateur Axon Data Governance peut spécifier ses exigences de règle en texte brut (par exemple : « Les identifiants clients doivent comporter huit caractères et commencer par un C ») et faire appel à CLAIRE pour l'aider. Grâce aux techniques ML et NLP, CLAIRE analyse les besoins de l'utilisateur et les traduit en une représentation technique. En fonction de cette représentation, ainsi que des métadonnées associées (par exemple : nom du terme dans le glossaire), CLAIRE effectuera une recherche dans les règles d'Informatica Data Quality et identifiera les candidats potentiels. L'utilisateur peut alors choisir une règle existante correspondante ou (si aucune règle n'est applicable) demander à CLAIRE de générer une nouvelle règle de qualité des données.

Si aucune règle applicable n'a été trouvée, CLAIRE génère automatiquement une nouvelle règle de qualité des données pour satisfaire aux exigences du référentiel Informatica Data Quality et la relie au contexte Axon Data Governance. En outre, CLAIRE associe automatiquement les règles de qualité des données aux profils Cloud basés sur Microsoft Common Data Model (CDM) et sur les sources Salesforce. À mesure que les utilisateurs créent de nouveaux profils pour les objets de base à partir de l'une de ces sources, CLAIRE suggère automatiquement des règles de qualité des données conformes aux meilleures pratiques à appliquer à la mesure.

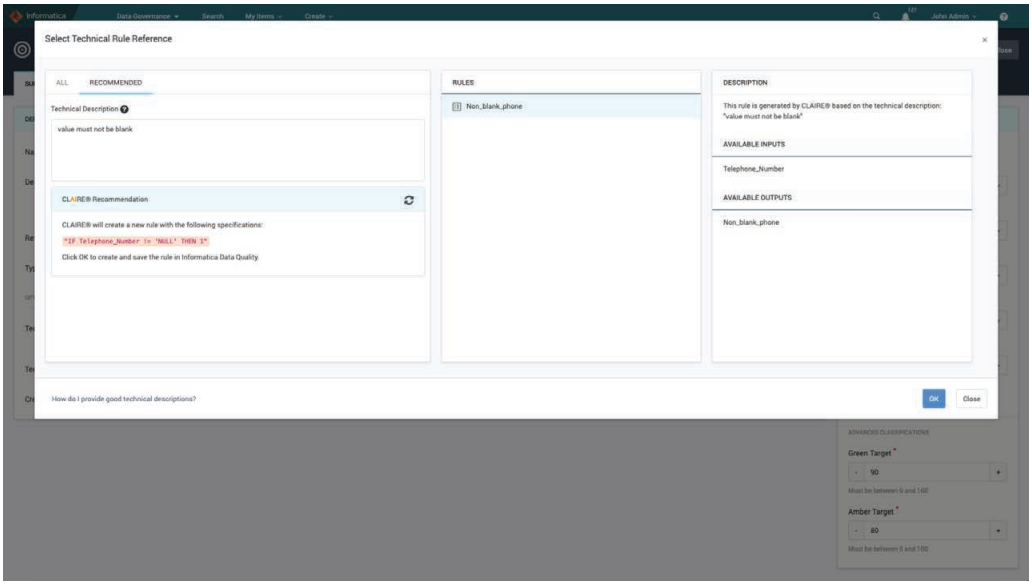

Figure 16 : Identification automatique des règles de qualité des données à l'aide du NLP.

CLAIRE pour la confidentialité et la protection des données

Grâce aux solutions intelligentes de confidentialité des données dotées de CLAIRE, les entreprises peuvent bénéficier d'une vue et d'une analyse à l'échelle de l'entreprise des informations d'identification personnelle au sein des ressources de données. L'automatisation basée sur l'IA vous permet de découvrir les données personnelles et sensibles, de comprendre le mouvement des données, de lier les identités, d'analyser les risques et de résoudre les problèmes.

Mapping de l'identité du registre des sujets

CLAIRE détermine la corrélation d'identité avec les données sensibles qui fournit un mapping des données pour la conformité en matière de confidentialité et la création de rapports sur l'accès aux données des personnes concernées. CLAIRE évalue et note les données qui, une fois combinées, permettent d'identifier les personnes concernées. En plus de la correspondance exacte, diverses techniques avancées, notamment la reconnaissance d'entités nommées (NER), sont utilisées pour améliorer les résultats généralement obtenus lorsque des données sont combinées à partir de différentes sources.

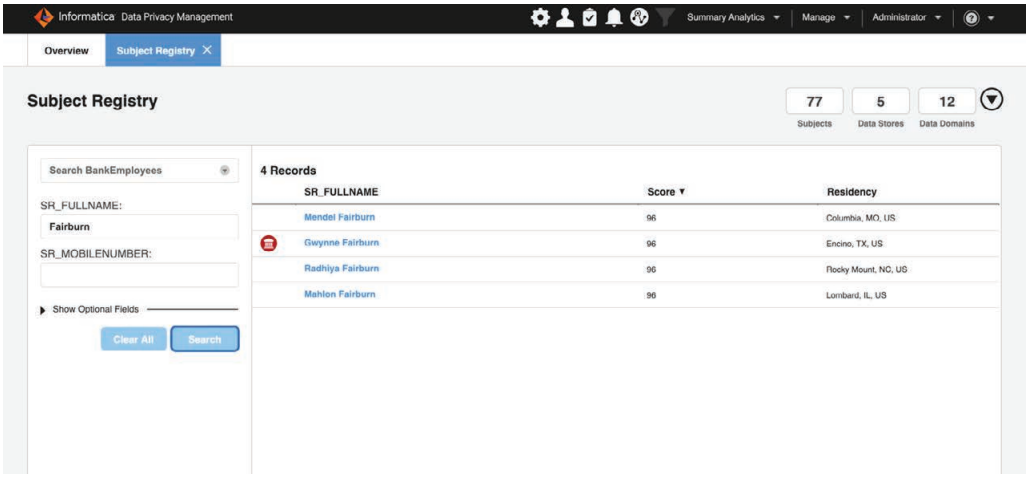

Figure 17 : Mapping de l'identité du registre des sujets pour la conformité en matière de confidentialité et la création de rapports sur l'accès aux données des personnes concernées.

Mapping et mouvement des données sensibles

CLAIRE exploite et étend les capacités de traçabilité mentionnées ci-dessus pour identifier également la manière dont les données sensibles prolifèrent dans les repositories afin de prendre en charge les exigences de conformité en matière de sécurité et de confidentialité. CLAIRE détermine à la fois les mouvements en amont et en aval ainsi que les métadonnées associées, telles que le type de données, le processus, l'état de protection et l'emplacement des données, afin d'évaluer si des violations se sont produites. Par exemple, une violation peut se produire si les données personnelles sont transmises d'une source à une cible au-delà des frontières géographiques, ou si les données intégrées pour les processus de facturation sont désormais propagées à d'autres services ou sites pour des processus marketing qui peuvent enfreindre les réglementations en matière de confidentialité. Les parties prenantes de la politique ou du processus peuvent alors être informées pour résolution.

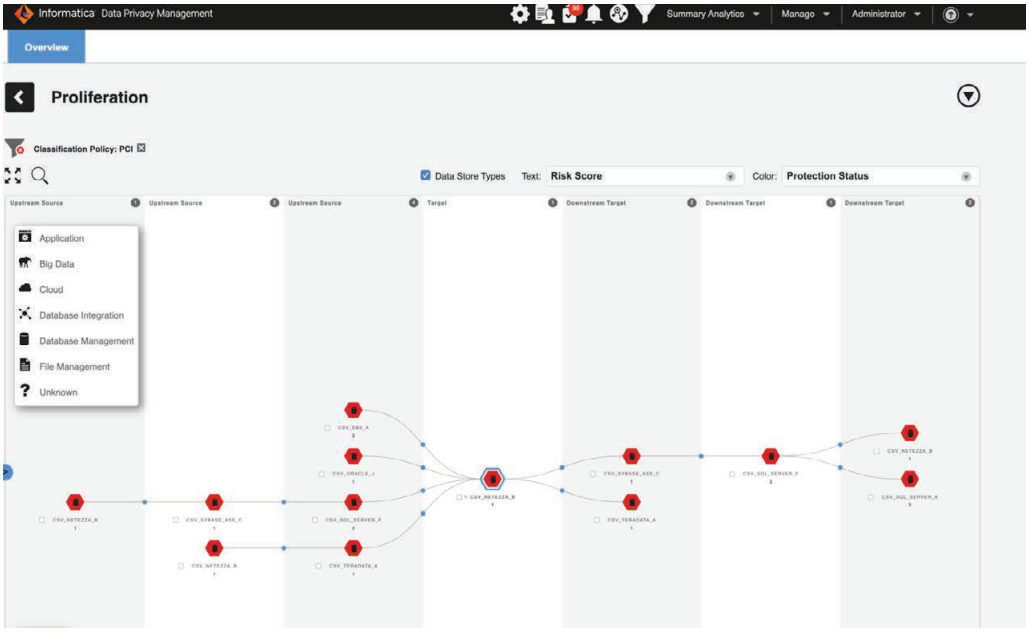

Figure 18 : Identifiez et suivez le mouvement des données sensibles entre les repositories.

Plans de simulation des risques

Les réglementations en matière de confidentialité exigent de plus en plus que les entreprises disposent de plans de protection des données. CLAIRE peut aider les entreprises à simuler les impacts de ces plans de protection afin d'assurer un meilleur retour sur investissement et de faciliter les processus budgétaires. CLAIRE évalue les techniques de protection appliquées à un ou plusieurs domaines de données, puis calcule la modification de la valeur de risque, l'exposition des données sensibles et le coût de risque résiduel pour chacun des data stores sélectionnés, ainsi que l'impact global pour l'entreprise à l'aide d'un modèle d'utilité attendue.

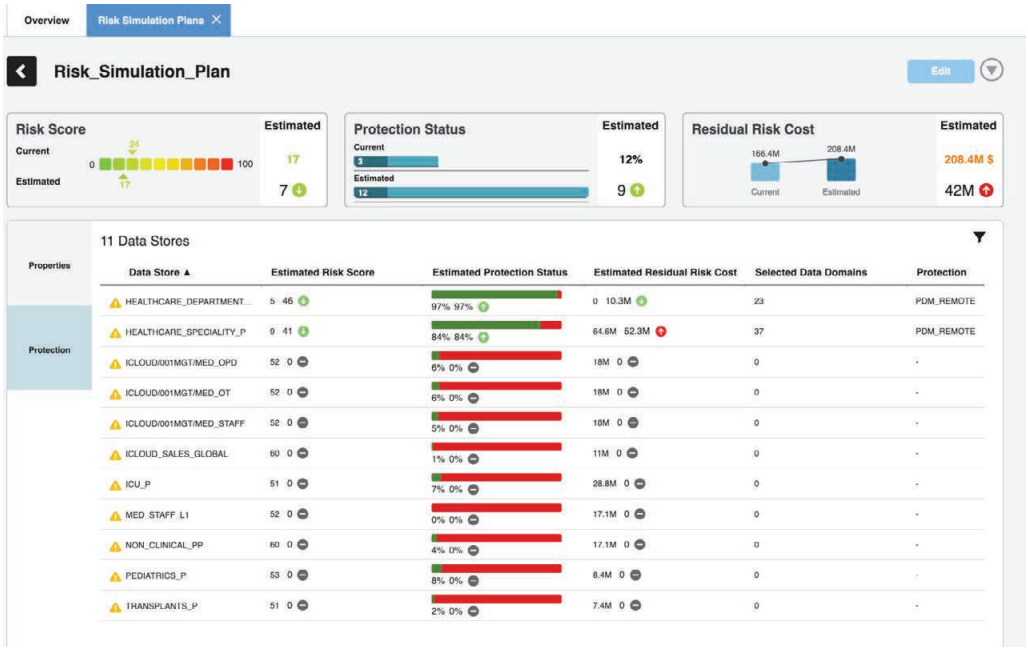

Figure 19 : CLAIRE évalue les techniques de protection appliquées aux domaines de données afin de déterminer les risques.

Détection intelligente des anomalies

CLAIRE utilise des approches de Machine Learning et de statistique, pour détecter les valeurs hors normes et les anomalies de données. La fonction d'analyse du comportement des utilisateurs (UBA) détecte les schémas de comportement utilisateur susceptibles de représenter un risque et d'exposer l'entreprise à une utilisation abusive des données. L'UBA est capable de détecter les emprunts d'identité, les piratages d'informations d'identification et les attaques d'escalade des privilèges.

L'UBA applique le Machine Learning non supervisé à un modèle multidimensionnel d'activités utilisateurs, qui inclut le nombre de data stores consultés par l'utilisateur, le nombre de requêtes effectuées et le nombre d'enregistrements concernés au sein des différents systèmes. L'analyse du composant principal est appliquée à ce modèle pour la réduction de la dimensionnalité. La technique BIRCH est appliquée pour le clustering hiérarchique non supervisé, afin de rechercher des utilisateurs dont le comportement était différent sur une période donnée. Pour valider le comportement anormal, des méthodes de détection des valeurs hors normes basées sur la densité et la distance sont utilisées, et le test statistique de Grubbs pour les valeurs hors normes est effectué, afin de confirmer que les objets indiqués par les deux premières méthodes sont effectivement des valeurs hors normes dans le système du cluster.

Figure 20 : Analyse du comportement des utilisateurs pour détecter automatiquement les anomalies utilisateurs pouvant indiquer un usage abusif des données.

Protection des données API en temps réel

Protégez les données sensibles (par exemple, les informations d'identification personnelle) en temps réel en identifiant les fuites de données personnelles dans les API, en bloquant et en masquant les données. Informatica API Management intègre des bibliothèques de protection des données pour bloquer les données sensibles lors des appels d'API entrants et sortants, réduisant ainsi le risque d'exposition des données sensibles.

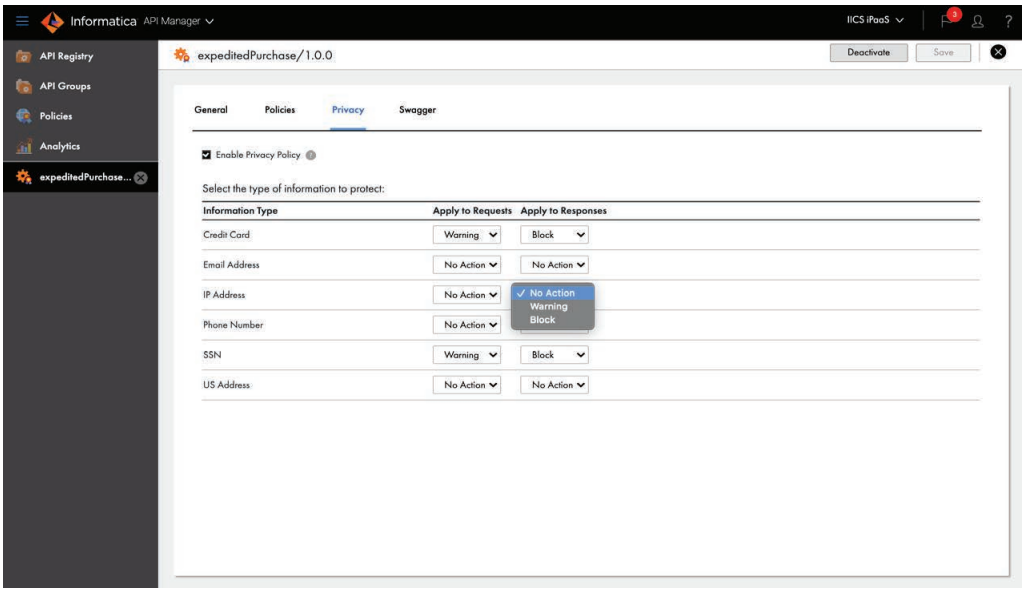

Figure 21 : Bloquez l'accès aux données sensibles lors des appels API entrants et sortants.

CLAIRE pour DataOps

Grâce à CLAIRE, les entreprises peuvent accélérer les pipelines de traitement des données, en automatisant de nombreux aspects de la gestion des données pour l'intégration continue (CI) et la distribution continue (CD) liée aux DataOps.

Analyses prédictives et pertinentes pour les environnements de gestion des données

L'analyse opérationnelle permet de comprendre l'utilisation actuelle des projets et des ressources existants et de planifier les capacités futures. Elle offre des paramètres pour la création de modèles de refacturation tout en prenant en charge plusieurs départements d'entreprise sur une seule plate-forme de gestion des données. Sur la base d'une observation continue des tendances d'utilisation des ressources, des projections de traitement des volumes de données sont proposées pour faciliter la planification de la capacité. CLAIRE passe à l'étape suivante en proposant une mise à l'échelle automatique des ressources d'exécution de la gestion des données.

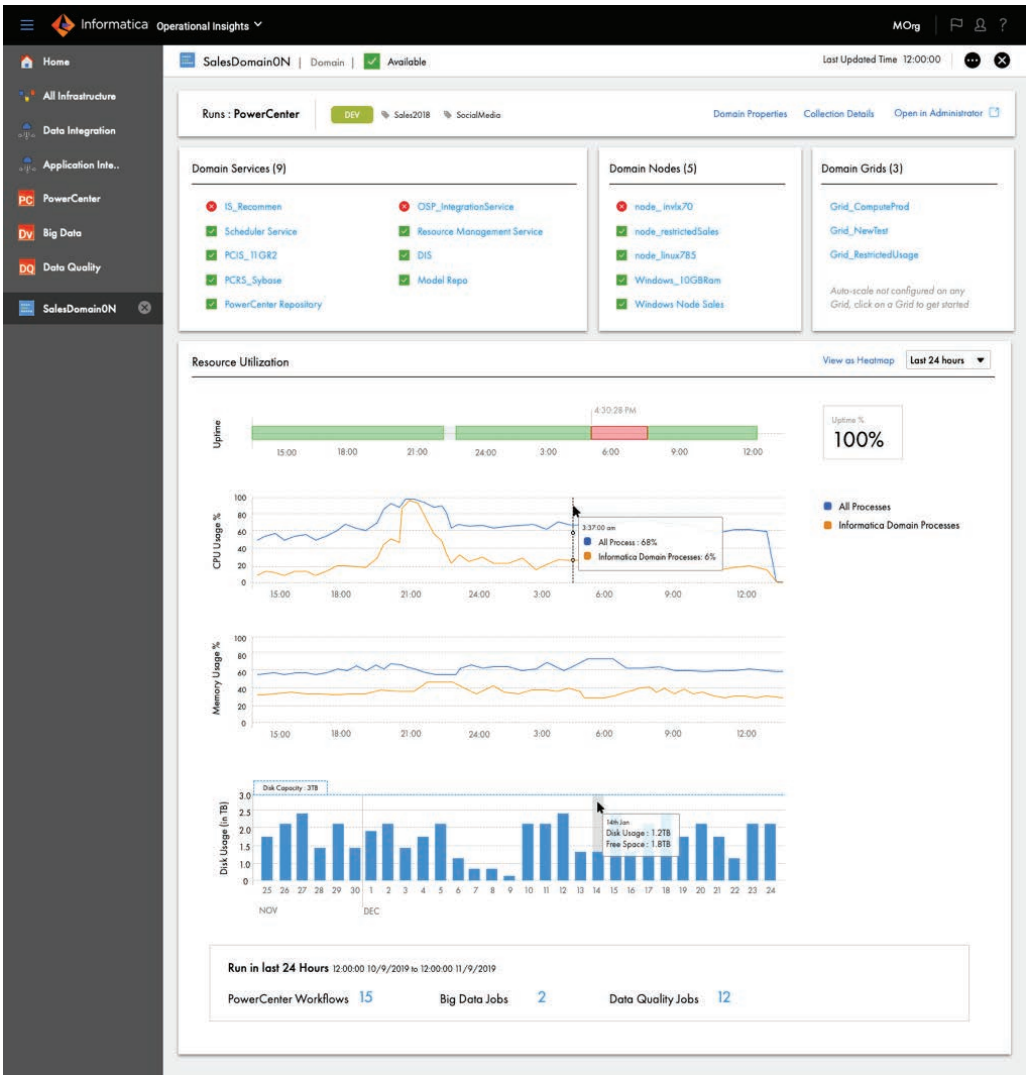

Figure 22 : Utilisation des ressources de vision opérationnelle pour les processus de domaine Informatica.

Détection des anomalies dans les exécutions des tâches

CLAIRE détecte automatiquement les anomalies liées aux temps d'exécution des tâches, aux données traitées, aux données chargées, aux ressources consommées, au débit, etc. La détection automatique de ces anomalies aide le service informatique à résoudre de manière proactive les problèmes liés aux tâches d'intégration des données avant qu'ils n'affectent les processus métiers en aval. L'algorithme Seasonal Hybrid ESD est utilisé pour détecter les anomalies dans le comportement d'exécution des tâches. Cet algorithme prend en compte la saisonnalité (pic de charge en fin de mois, période des fêtes, etc.) et élimine les tâches présentant des aberrations attendues induites par les cycles d'activité.

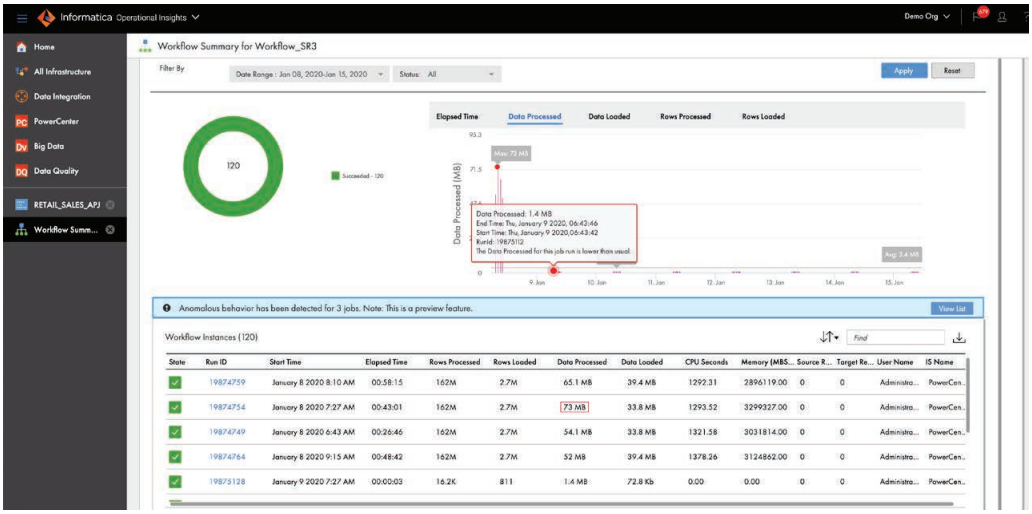

Figure 23 : CLAIRE détecte automatiquement les anomalies liées au traitement des tâches et des données Informatica.

CLAIRE dans le futur

Au fur et à mesure que CLAIRE se développe, elle va continuer à augmenter la productivité et l'efficacité, permettant aux responsables des données d'exploiter l'automatisation intelligente pour obtenir des informations plus rapides et plus pertinentes et une gestion plus efficace des données. Les futures capacités incluent :

1. **Intégration automatique :** Intégration automatique des nouvelles données entrantes dans les processus d'intégration de données. Identification des données, localisation des schémas d'intégration qui traitent des données similaires, transformation et déplacement automatiques des données grâce à l'apprentissage à partir de millions de mappings existants et d'actions d'utilisateurs.
2. **Assistance au développement :** Recommandations aux utilisateurs et suggestion des prochaines actions à privilégier durant le processus de développement, notamment :
   - Transformation automatique
   - Recommandations de modèles
   - Suggestions de type masquage pour les données sensibles
   - Suggestions sur la qualité des données pour le nettoyage et la standardisation
   - Optimisation automatique des performances
3. **Mapping automatique :** Détection des entités de données de référence dans toute l'entreprise et mapping automatique de ces données au modèle de données de référence en appliquant les transformations requises et les règles de qualité.

4. **Réparation automatique :** Gestion fluide des problèmes externes au système, tels qu'une mémoire faible ou un problème de puissance de calcul. Par exemple, ajout de ressources de calcul supplémentaires (« débordement vers le Cloud ») pour gérer les pics de données.
5. **Réglage automatique :** Prévision et ajustement des calendriers et des ressources informatiques en fonction des informations de l'historique, des volumes actuels de données et des ressources système disponibles, afin de respecter les critères de performance.
6. **Sécurisation automatique :** Détection automatique des données sensibles et masquage de ces données avant qu'elles ne quittent la zone sécurisée.

Conclusion

Les stratégies métiers actuelles orientées sur les données sont conçues en se fondant sur les données. Pour les exploiter au mieux, vous devez disposer des compétences de gestion de données qui vous permettent de libérer le potentiel de ces données. Avec tous les défis que représente la gestion de données dans des circonstances ordinaires, les approches traditionnelles ne peuvent répondre aux exigences actuelles — ou futures. Une des méthodes employées pour exploiter les données dans des projets d'innovation consiste à les standardiser au sein d'une plate-forme de gestion de données de bout en bout utilisant la puissance des données, des métadonnées, le Machine Learning et l'IA pour améliorer la productivité de tous les utilisateurs de la plate-forme : services techniques, opérationnels, métiers et, en particulier, libre-service professionnel.

Pour en savoir plus sur l'utilisation de CLAIRE et de l'Intelligent Data Management Cloud, et maîtriser la puissance de vos données, contactez-nous.

Informatica

Informatica France Tour CB 21, 16 Place de l'Iris, 92040 Paris La Défense Cedex Tél : +33 01 42 04 89 00
Numéro gratuit aux États-Unis : 1.800.653.3871

IN09_0521_03328