LE LANGAGE DE LA DATA QUALITY

Maîtrisez les fondamentaux pour une gestion sans failles de vos données

Tale of Data

taleofdata.com

Data Quality : contexte, défis et solutions

Les entreprises s'appuient sur des données fiables pour prendre des décisions éclairées, optimiser leurs opérations et anticiper les tendances du marché. Le marché de la donnée en constante évolution, transforme les informations en un véritable levier stratégique. Pour tirer pleinement parti de ce potentiel, il est crucial que tous les acteurs de l'entreprise, des dirigeants aux équipes techniques, puissent communiquer efficacement sur les enjeux de la qualité des données. C'est dans cette optique que nous avons créé **"Le Langage de la Data Quality**". Ce glossaire a été conçu pour harmoniser les pratiques, fournir un vocabulaire commun et faciliter la collaboration au sein des équipes, permettant ainsi de maximiser l'exploitation des données.

"Le Langage de la Data Quality" n'est pas simplement un glossaire, mais un véritable guide pratique, conçu pour aider les entreprises à naviguer dans les complexités de la qualité des données. En offrant un langage commun et des méthodes éprouvées, il vise à renforcer la performance des organisations dans un environnement de plus en plus axé sur la qualité des données.

La gestion de la qualité des données représente un ensemble de défis importants que les entreprises doivent relever pour maximiser la valeur de leurs informations.

Défi technique

Les données sont souvent éparpillées dans différents systèmes, sans qu'une personne ou une équipe spécifique en soit réellement responsable. Ce manque de centralisation crée des incohérences et des doublons, rendant leur gestion et leur partage complexes et souvent inefficaces.

Défi culturel

Dans de nombreuses entreprises, il n'existe pas de processus clair pour gérer les données de manière cohérente. Les équipes peuvent être hésitantes à adopter de nouvelles méthodes, préférant s'en tenir à des habitudes bien ancrées. Cette résistance freine l'amélioration de la qualité des données et empêche l'adoption de pratiques plus efficaces.

Défi réglementaire

Se conformer aux différentes réglementations et normes en vigueur exige des entreprises de trouver un équilibre délicat entre la qualité des données et le respect des exigences de confidentialité et de sécurité, ce qui nécessite une gouvernance solide et constante.

Défi organisationnel

La qualité des données est une préoccupation partagée par tous, mais les initiatives en la matière sont souvent freinées par un manque de ressources et de coordination entre les équipes. Cela rend difficile la démonstration de l'impact positif des efforts en matière de Data Quality, ce qui peut limiter l'engagement et les investissements nécessaires pour avancer.

Tale of Data est une plateforme de Data Quality alimentée par l'IA qui révolutionne la manière dont les entreprises organisent, fiabilisent et industrialisent leurs données.
Notre plateforme est adoptée par des entreprises leaders dans leurs industries pour assurer la qualité de leurs données et les préparer pour retraitement dans leurs logiques métiers. Tale of Data permet la gouvernance des données au sein de l'IT et habilite les entreprises à relever les challenges liés à la réglementation, aux mises en conformité, à la collaboration entre les équipes, et à élever le concept de qualité des données au rang de culture d'entreprise.

Opérationnellement, la solution est No-code by design. Elle permet un travail pluridisciplinaire entre les équipes responsables de la qualité des données et les équipes métiers, pour les rendre autonomes.
De son traitement à son utilisation, Tale of Data ouvre la boîte noire de la donnée en entreprise, garantissant sa traçabilité, son traitement, son contrôle et son automatisation dans un cadre collaboratif et sécurisé.
Notre mission chez Tale of Data se concentre sur trois axes principaux : Aligner, Fiabiliser et Industrialiser les processus de qualité des données. En collaborant avec de grandes entreprises à travers le monde, nous avons acquis une profonde compréhension des défis auxquels elles sont confrontées.

Tale of Data

Glossaire

A 3
Algorithme de matching / fuzzy matching 3
API Call 3
API ou Interface de programmation d'applications 3
B 3
BAN - Base Adresse Nationale 3
Base de données relationnelle 3
BCBS 239 3
C 4
Code IRIS 4
Connecteurs 4
Crowd sourcing 4
D 4
Data Catalog 4
Data Compliance 4
Data Discovery 4
Data driven 5
Data Gouvernance 5
Data Lake 5
Data Lineage 5
Data Mesh 5
Data Observability 5
Data Product 6
Data Quality 6
Data scientist 6
Data steward 6
Data Stories 6
Data Warehouse 6
Databases 6
Data Visualisation (dataviz) 7
Dédoublonnage 7
Distance de Levenshtein 7
Données à enrichir 7
Données d'enrichissement 7
Données PI (Plant Information) 7
E 7
Enregistrement 7
Enrichissement des données 8
F 8
Flow 8
Flow Designer 8
J 8
Jeux de données 8
Jointure 8
Jointure full-text 8
Jointures floues 8
Langage naturel 9
M 9
Machine Learning 9
Mass Data Discovery 9
Master Data 9
Métadonnées 9
Mode Cluster 9
N 10
N-gramme ou N-Gram 10
Nature 10
Nœud d'aggrégation 10
Nœud diffusion 10
Noeud filtre 10
O 10
Open Data 10
P 10
Pattern 10
Phonétique / Algorithme phonétique 11
Préparation de données (ou Data Preparation) 11
R 11
Random forest 11
Réconciliation des données 11
Record Lineage 11
Redressement 11
Référentiel 12
Règles de gestion 12
Règles métier 12
Remédiation 12
Runtime 12
Runtime Environment 12
S 12
Script 12
Séries temporelles 13
Shadow IT 13
Système Legacy 13
T 13
Traitement Fenêtre 13
Type 13
U 13
Union 13

Algorithme de matching / fuzzy matching

Procédé algorithmique basé sur une correspondance approximative de deux entrées, plutôt que sur une correspondance exacte. En pratique, différents algorithmes sont mis à disposition dans Tale of Data pour prendre appui, par exemple, sur les spécificités de la phonétique française ou anglaise. D'autres approches sont proposées comme de donner davantage de poids aux consonnes ou d'utiliser des procédés mathématiques éprouvés comme la distance de *Levenshtein**.

API Call

Demande de service faite à une API pour déclencher des traitements ou récupérer ou envoyer des données entre différentes applications.

API ou Interface de programmation d'applications

Interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

BAN - Base Adresse Nationale

La Base Adresse Nationale est la base regroupant les adresses officielles du territoire français. Cette base est dite « ouverte » : son accès et l'usage sont laissés libres aux usagers, qui peuvent être d'origine privée ou publique.

Base de données relationnelle

En informatique, une base de données relationnelle est une base de données où l'information est organisée dans des tableaux à deux dimensions appelés des relations ou tables. Selon ce modèle relationnel, une base de données consiste en une ou plusieurs relations.

BCBS 239

Norme bancaire visant à augmenter les capacités des banques en matière d'agrégation de données de risques financiers ; à produire des reportings et à améliorer la qualité de ces données risques.

Code IRIS

Les « Ilots Regroupés pour l'Information Statistique » sont des briques de découpage du territoire créées par l'INSEE de taille homogène. Chaque maille élémentaire regroupe 2 000 habitants.

Connecteurs

Moyen pour se connecter à une source de données d'un type particulier (par exemple une base de données SQL Server, ou un serveur de fichiers de type Azure Blob Storage, etc) -> cf section Architecture.

Crowd sourcing

Mode d'organisation faisant appel à des contributions d'un grand nombre de personnes pour enrichir et améliorer un contenu. Par exemple, Wikipédia est une encyclopédie dont le contenu est enrichi à l'aide d'un très grand nombre de contributeurs.

Data Catalog

Référentiel centralisé de métadonnées, qui permet de gérer, rechercher et documenter les données disponibles dans une organisation, facilitant ainsi leur découverte et leur utilisation.

Data Compliance

Approche visant à garantir la conformité de l'entreprise vis-à-vis des lois, règlements et normes qui régissent la collecte, le traitement, le stockage et le partage des données. Suivant les domaines d'activité de l'entreprise, celle-ci peut inclure des réglementations sectorielles, en particulier dans le cas des banques, des assurances et des établissements traitant la donnée médicale.

Data Discovery

Procédé permettant d'explorer les données disponibles dans un système informatique pour en découvrir la structure, le contenu et les interrelations, facilitant ainsi la compréhension et l'analyse des données.

Data driven

Adjectif anglais qui peut se traduire par « pilotée par les données ». Autrement dit, il s'agit d'une entreprise qui s'appuie sur l'analyse de ses données pour prendre des décisions et orienter son évolution plutôt que sur l'intuition.

Data Gouvernance

Ensemble des pratiques visant à organiser et gérer les données d'une entreprise. Une gouvernance efficace maximise la valeur des données tout en réduisant les risques de non-conformité. Elle clarifie les rôles et responsabilités liés à l'accès, à la qualité et au cycle de vie des données, tout en assurant l'alignement de l'organisation autour de normes et méthodes communes. La gouvernance des données se doit de mettre en œuvre les pratiques et les outils nécessaires pour contrôler et améliorer la qualité des données. En effet, sans données fiables, la gouvernance des données est totalement inopérante.

Data Lake

Espace de stockage centralisé qui permet de conserver des données structurées, semi-structurées et non structurées à grande échelle, facilitant leur analyse et traitement ultérieurs.

Data Lineage

Représentation permettant de tracer l'origine et le parcours des données à travers différents systèmes, garantissant transparence, conformité ainsi que l'analyse des impacts.

Data Mesh

Approche de l'architecture des données visant à exploiter les bénéfices d'une organisation décentralisée mais maîtrisée. La donnée est organisée autour de *Data Products**, appartenant à différents domaines métiers spécifiques, chacun alignés avec le business. Soutenue par des technologies permettant aux métiers de travailler avec la data en libre service, une approche Data Mesh permet de fédérer toutes les équipes autour d'une *Data Gouvernance** commune, en exploitant les talents et les compétences disponibles à travers toute l'entreprise.

Data Observability

Capacité à surveiller et comprendre l'état des données dans un système, en utilisant des métriques et des visualisations pour assurer leur qualité, intégrité et performance.

Data Product

Ensemble de données organisées et prêtes à être consommées, souvent associées à des outils et interfaces permettant de les exploiter efficacement pour répondre à des besoins spécifiques. Considérer la donnée comme un produit sur étagère permet de lui associer des caractéristiques intéressantes pour le consommateur, telles que la facilité d'accès, une description détaillée, la qualité, la fraîcheur et même des conseils d'utilisation.

Data Quality

Ensemble des processus et techniques visant à assurer que les données sont précises, complètes, fiables et pertinentes pour leur utilisation prévue.

Data scientist

Spécialiste de la donnée, il recueille, traite, analyse et fait parler les données pour améliorer les performances de l'entreprise.

Data steward

Personne responsable de la gestion et de la supervision de certaines données appartenant à l'entreprise. À ce titre, il est chargé de garantir la qualité des données et leur adéquation aux objectifs l'organisation.

Data Stories

Narrations basées sur les données, utilisant des visualisations et des analyses pour communiquer des informations et des insights de manière claire et engageante.

Data Warehouse

Système central de stockage de données, conçu pour faciliter la prise de décision. Il intègre et structure des données provenant de sources variées. Contrairement au *Data Lake**, les données y sont nettoyées, transformées et prêtes à l'exploitation, nécessitant un traitement en amont.

Databases

Collections organisées de données, stockées et accessibles électroniquement à partir d'un système informatique, permettant de gérer, manipuler et interroger les données efficacement.

Data Visualisation (dataviz)

Méthode qui consiste à communiquer des chiffres ou des informations brutes en les transformant en objets visuels faciles à lire : points, barres, courbes, cartographies. Tale of Data inclut un module de DataViz, accessible à tous les utilisateurs de la solution, ainsi qu'à ceux souhaitant utiliser uniquement ce module.

Dédoublonnage

Méthode qui permet d'éliminer les doublons.

Distance de Levenshtein

Méthode qui mesure à quel point deux mots ou phrases se ressemblent. Elle compte combien de changements (comme supprimer, ajouter ou remplacer des lettres) sont nécessaires pour transformer un mot en un autre.

Données à enrichir

Jeu de données existant (par exemple, une liste de prospects dans un CRM) sur lequel on souhaite ajouter des informations supplémentaires sous forme de nouvelles colonnes (par exemple, l'effectif de chaque société).

Données d'enrichissement

Il s'agit d'un jeu de données de référence, interne (ex : liste des clients publiée dans le *Data Warehouse**) ou externe (ex : la base SIRENE) qui contient des informations supplémentaires dont vous avez besoin pour augmenter votre capacité d'analyse.

Données PI (Plant Information)

Ces données, produites sur des sites industriels, sont issues de capteurs installés sur des sites de production et envoyés dans un système de stockage.

Enregistrement

Lignes dans une base ou un fichier (par opposition aux colonnes).

Enrichissement des données

Consiste à compléter les données, à les améliorer et à les structurer via l'utilisation d'une autre source (référentiel, fichier base ...).

Flow

Traitement construit interactivement par drag and drop dans Tale of Data, permettant d'effectuer des tâches de remédiation, de préparation et de monitoring de données. Un flow est conçu spécifiquement pour être utilisé en production.

Flow Designer

Environnement dans le logiciel Tale of Data pour mettre au point des Flows* dans le but de concevoir des transformations sur les données.

Jeux de données

Ensemble de données structurées sous forme de table.

Jointure

Opération reliant deux ou plusieurs tables sur la base de conditions communes. Une opération de jointure est analogue à une recherche verticale dans un tableur, en associant les données des tables lorsque des correspondances existent entre leurs lignes.

Jointure full-text

Assemblage de plusieurs sources en recherchant dans l'ensemble des données textuelles spécifiées. Cela permet par exemple de repérer des correspondances entre des enregistrements, même si les différences résident dans l'ordre des mots. Contrairement aux algorithmes classiques, une jointure full-text détecte ces correspondances subtiles, souvent évidentes pour un humain.

Jointures floues

Assemblage de plusieurs sources en faisant des correspondances entre elles à l'aide d'algorithmes de fuzzy matching.

Langage naturel

Signifie que l'utilisateur n'a pas besoin de connaître de langages informatiques pour utiliser la solution. Les fonctions sont toutes utilisables via des menus explicites.

Machine Learning

Processus permettant de créer automatiquement, à partir d'exemples, un programme ayant la capacité de résoudre des problèmes ou faire des prédictions sur un domaine similaire à celui des exemples utilisés.

Mass Data Discovery

Procédé d'exploration du système informatique permettant de découvrir et cartographier toutes les données présentes dans le-dit système. Ceci permet notamment d'établir un atlas des données sensibles stockées (comme les données personnelles). Il permet aussi la génération d'un rapport analysant la qualité des données stockées.

Master Data

Donnée de référence de l'entreprise.

Métadonnées

Informations qui décrivent les caractéristiques, l'origine, la structure (ex : noms et types des colonnes) et l'usage d'un jeu de données. Elles servent de "données sur le jeu de données" et facilitent sa gestion, compréhension, et utilisation, en permettant notamment de tracer sa provenance, d'assurer sa qualité, et d'en réguler l'accès.

Mode Cluster

Mode de fonctionnement distribué sur plusieurs serveurs, qui permet de traiter en parallèle un grand nombre de données.

N-gramme ou N-Gram

Méthode utilisée dans Tale of Data pour évaluer la similarité entre plusieurs mots ou plusieurs phrases. Plus généralement, il s'agit de la succession de N éléments de même type extraits d'un texte, d'une séquence ou d'un signal ; les éléments pouvant notamment être des mots ou des lettres.

Langage naturel

Signifie que l'utilisateur n'a pas besoin de connaître de langages informatiques pour utiliser la solution. Les fonctions sont toutes utilisables via des menus explicites.

Nœud d'aggrégation

Outil de traitement qui permet regrouper et de combiner des données provenant de différentes sources ou niveaux de détail pour produire des résumés, des statistiques ou des vues consolidées. Le nœud d'agrégation facilite l'analyse en offrant une vue d'ensemble simplifiée, par exemple en calculant des totaux, des moyennes ou des tendances à partir de données brutes.

Nœud diffusion

Outil de traitement permettant de diffuser les mêmes données à plusieurs autres outils de traitement.

Noeud filtre

Outil de traitement permettant de choisir le passage de certaines données aux outils de traitement en aval de celui-ci.

Open Data

Littéralement, « données ouvertes » , se dit des données auxquelles l'accès est totalement public et libre de droit, au même titre que l'exploitation et la réutilisation. La Base des Adresses Nationales ou encore la base des SIRET sont des illustrations d'informations consultables en Open Data.

Pattern

Motif défini par l'utilisateur et qui peut être recherché dans les données, ou utilisé dans le cadre de leur transformation.

Phonétique / Algorithme phonétique / Analyse phonétique

Rapprochement de termes selon une identité de son. Exemple : recherche de similarité entre des noms de famille avec le son [o], pouvant s'orthographier o, ô, au, eau.

Préparation de données (ou Data Preparation)

Étape préalable à l'analyse, comprenant des tâches comme le nettoyage et l'enrichissement des données. La préparation des données est l'étape clé pour une analyse des données. Elle transforme les données brutes en informations fiables et exploitables, garantissant ainsi une analyse valide et maîtrisée.

Random forest

Méthode d'apprentissage automatique utilisée pour la classification et la prévision lors de l'étude des données. Elle fonctionne en construisant une multitude d'arbres de décision lors de l'entraînement. Leur combinaison permet l'obtention de résultats plus précis et robustes.

Réconciliation des données

Processus visant à mettre en correspondance des ensembles de données, en utilisant des règles spécifiques ou des outils adaptés pour assurer leur cohérence et leur homogénéité.

Record Lineage

Représentation dans Tale of Data montrant l'enchaînement des flux de données qui alimentent un jeu de données (flux amont) et ceux qui en dépendent (flux aval). Cette visualisation permet d'opérer une navigation aisée des traitements et des jeux de données activement utilisés.

Redressement

Phase pendant laquelle les données « brutes » sont analysées pour être corrigées. C'est une des actions de la préparation de données.

Référentiel

Un jeu de données utilisé comme référence. Par exemple, un référentiel produits est une liste des produits d'une entreprise, contenant des informations officielles et précises sur leurs caractéristiques. Ces données sont mises à jour selon des règles définies et peuvent être utilisées pour des contrôles, des enrichissements ou des corrections.

Règles de gestion

Directives qui régissent les activités d'une organisation ou d'un système. Elles visent à assurer la cohérence et la conformité des opérations, minimiser les risques d'erreurs ou de fraudes et améliorer la qualité des produits ou des services.

Règles métier

Ensemble d'opérations de transformation sur des données, qui est défini par l'utilisateur de Tale of Data sans écriture de code, pour transformer les données selon des conditions spécifiques. Ces règles sont lisibles, réutilisables et peuvent être appliquées à d'autres processus de transformation (Flows).

Remédiation

Résolution des problèmes de qualité présents dans les données.

Runtime

Environnement dans le logiciel Tale of Data pour exécuter des Flows* dans le but d'opérer des transformations sur les données. L'exécution des Flows* peut être déclenchée directement par l'utilisateur, ou être planifiée de manière extrêmement flexible.

Runtime Environment

Environnement logiciel dans lequel les programmes s'exécutent. Cela inclut le système d'exploitation, les bibliothèques et les outils nécessaires pour exécuter les applications.

Script

Programme informatique qui, en s'exécutant, permet de réaliser une action ou afficher une page Web.

Séries temporelles

Série de données indexées par le temps. Le PIB d'un pays ou l'évolution de la population sont des séries temporelles.

Shadow IT

Données et logiciels utilisés en dehors de la DSI, comme des bases MS Access ou des fichiers Excel avec macros. Ces pratiques présentent des risques de sécurité et de non-conformité, notamment vis-à-vis du RGPD, car elles échappent au contrôle de la DSI.

Système Legacy

Un système legacy, appelé encore système "hérité" est un système informatique (comme un ERP) répondant toujours aux besoins mais il ne pouvant plus évoluer. L'organisation s'appuie toujours sur ce système, mais pourra être limitée car il ne peut pas interagir pas avec les outils analytiques les plus récents, comme ceux hébergés sur le cloud.

Traitement fenêtre

Outil dans Tale of Data qui permet d'effectuer des calculs sur le voisinage local des données, à l'aide de "fenêtres", pour analyser des tendances locales tout en gardant chaque valeur individuelle visible. Par exemple, pour calculer une moyenne mobile, une fenêtre peut inclure les données des 7 derniers jours pour produire un résultat pertinent à chaque point. Cette méthode est très utile pour examiner des données liées au temps (comme des transactions ou des séries chronologiques) ou organisées en catégories (comme des catalogues ou des listes). Elle permet d'étudier à la fois des détails locaux et des tendances globales.

Type

Caractéristique des données représentant leur caractère fondamental de donnée textuelle, numérique, temporelle, etc.

Union

Opération dans Tale of Data permettant d'empiler verticalement plusieurs tables pour les regrouper en une seule table.

Écrivez avec nous le langage de la Data Quality !

Merci pour votre lecture et votre intérêt pour ce document. Si vous avez des définitions manquantes ou des suggestions à nous apporter, n'hésitez pas !

Ce document est à vous et pour vous : faites-nous part de vos idées pour enrichir et améliorer la prochaine édition du **"Langage de la Data Quality".**

👉 **Contactez-nous à l'adresse suivante :** contact@taleofdata.com

Mes nouvelles définitions

Document rédigé par :
Jean-Christophe Bouramoué – CTO & Founder at Tale of Data
Robbie Jameson – CEO at Tale of Data
Valérie Varela – Sales Director at Tale of Data
Adnan Joudeh – Digital Marketing Officer at Tale of Data
Carole Ellouk – CMO and COO at Tale of Data

Tale of Data

01 89 16 61 75

contact@taleofdata.com

taleofdata.com

2024