Paris School of Business

M2 Data Management
Groupe 7

SPRINT 1

Thomas Dahrouj - Nadir Boudjerida - Jean louis Hu

I. Importance de la data gouvernance

Avant tout, il faut savoir pourquoi la gouvernance des données est importante ?

La gestion des données est cruciale pour assurer la précision et la sécurité des informations exploitées au sein d'une entreprise. Son but principal est d'accroître la fiabilité des données, en veillant à ce qu'elles soient exactes, actualisées et utilisables. Cela permet également de repérer les éventuels problèmes liés aux données, comme les incohérences ou les erreurs, afin de les résoudre rapidement. Une gestion efficace des données contribue à améliorer la performance opérationnelle en optimisant l'utilisation des informations pour les processus métiers, tout en respectant les réglementations pour éviter les sanctions et préserver l'intégrité des données.

1. Améliorer la fiabilité des données

La gestion des données veille à ce que les informations utilisées par les entreprises soient précises, complètes et actualisées. En supprimant les inexactitudes, les duplicatas ou les données dépassées, elle garantit que les décisions stratégiques prises s'appuient sur des données fiables. Cette fiabilité renforce la confiance dans les analyses, réduisant ainsi les risques d'erreurs dans les processus opérationnels et les rapports. Une gestion efficace des données garantit également que chaque collaborateur dispose d'informations cohérentes pour travailler de manière optimale.

2. Identifier les problèmes potentiels liés aux données

La gestion des données permet de repérer rapidement les anomalies susceptibles de compromettre la qualité des informations. Cela comprend l'identification de données manquantes, incohérentes ou erronées. Par exemple, des mécanismes automatisés ou des contrôles réguliers facilitent la détection d'incohérences dans les bases de données avant qu'elles n'affectent les prises de décision. Cette surveillance proactive contribue à anticiper les problèmes en amont et assure que les données restent fiables et exploitables pour les utilisateurs finaux.

3. Augmenter l'efficacité opérationnelle

Une gestion optimisée des données permet aux équipes de travailler plus vite et de manière plus intelligente. Lorsque les informations sont propres, facilement accessibles et bien structurées, les employés peuvent éviter les tâches répétitives et le temps perdu à chercher des informations. Une bonne gouvernance des données améliore les processus de travail, rend l'automatisation plus fiable et aide à réduire les coûts opérationnels. Cela favorise une augmentation de la productivité en libérant du temps pour des activités à haute valeur ajoutée.

4. Assurer le respect des normes réglementaires

Dans un cadre réglementaire de plus en plus rigoureux, se conformer aux lois et aux régulations relatives à la gestion des données devient essentiel. Des législations comme le RGPD en Europe ou le CCPA aux États-Unis imposent des exigences strictes quant aux méthodes de collecte, de stockage et d'utilisation des informations personnelles. La gouvernance des données permet à l'entreprise de se conformer à ces obligations, évitant ainsi des sanctions financières et légales, tout en assurant la protection de la confidentialité et la sécurité des données sensibles.

Après avoir souligné l'importance de la gouvernance des données, il est crucial de se pencher sur les bonnes pratiques à adopter dans ce domaine. Nous allons en mentionner et en approfondir quelques-unes.

II. Gestion de la qualité des données

1. Référentiel de données

Tout d'abord, l'élaboration d' un glossaire métier et un dictionnaire de données commun semble pertinent. Cependant il est nécessaire de définir ces notions afin de mieux comprendre ce qui les différencie :

Glossaire métier : Un glossaire métier regroupe l'ensemble des termes, définitions et abréviations spécifiques à un projet ou à une organisation. Il permet à tous les membres de l'équipe de partager une vision unifiée des concepts, réduisant ainsi les risques de malentendus et d'interprétations erronées.

Dictionnaire de données partagé: Ce répertoire contient les métadonnées relatives à chaque élément de données exploité dans les systèmes d'information. Il détaille des aspects tels que la nature des données, les valeurs possibles, les liens avec d'autres données, ainsi que les règles de validation. Un dictionnaire de données bien organisé favorise une gestion efficace et une intégration harmonieuse des données entre les différentes équipes.

2. Objectifs clairs et mesurables

Il est important de définir des indicateurs de performance clés (KPIs) de manière précise pour évaluer la qualité des données, notamment en tenant compte de la complétude, de l'exactitude et de la cohérence. La **complétude** correspond à la proportion de champs renseignés par rapport au nombre total de champs obligatoires dans un enregistrement. L'**exactitude** se mesure par le pourcentage d'enregistrements qui reflètent fidèlement la réalité, comme la validation des adresses, par exemple. Quant à la **cohérence**, elle se traduit par le pourcentage d'enregistrements conformes à des règles de validation spécifiques, telles que le respect d'un format prédéfini pour les dates.

3. Approche itérative (Agile)

Les sprints sont focalisés sur l'amélioration continue de la qualité des données.

En effet, dans une méthodologie Agile, les projets sont segmentés en courtes phases de développement appelées sprints. Chacune de ces périodes doit être orientée vers une avancée particulière en matière de qualité des données, qu'il s'agisse de corriger un lot d'erreurs identifiées ou d'introduire de nouveaux mécanismes de contrôle de qualité.

De plus, toujours dans une optique d'amélioration continue, à l'issue de chaque sprint, une analyse approfondie est effectuée afin de déterminer les aspects réussis et ceux nécessitant des ajustements. Ce bilan permet de réajuster les méthodes et de maintenir une attention constante sur l'amélioration progressive de la qualité des données.

4. Rôles et responsabilités

Il est essentiel de veiller à la surveillance et à l'amélioration de la qualité des données. Chaque membre de l'équipe devrait avoir des attributions clairement établies en matière de qualité des données. Par exemple, un responsable des données pourrait s'occuper de la gouvernance, tandis qu'un analyste de données se concentrerait sur l'analyse et l'identification des anomalies.

Il est également important de favoriser une culture axée sur la qualité, où chaque membre de l'équipe comprend l'importance cruciale de la qualité des données et est incité à signaler les éventuels problèmes rencontrés.

5. Contrôles automatisés

Optimiser la détection et la correction des erreurs grâce à des outils spécialisés est essentiel. Pour cela, il convient d'exploiter des solutions dédiées à la qualité des données, permettant d'identifier les anomalies de manière automatisée. Cela peut se traduire par des validations en temps réel lors de l'entrée des données, ainsi que par des procédures de nettoyage des données préétablies. Par ailleurs, il est possible d'intégrer des scripts ou des workflows capables d'effectuer des corrections automatiques pour certaines erreurs, telles que l'harmonisation des formats ou l'élimination des doublons. Cette approche permet ainsi de diminuer considérablement le besoin d'interventions manuelles.

6. Traçabilité et transparence

La cartographie des données consiste à élaborer une représentation des flux d'informations à travers les différents systèmes. Cette démarche permet de comprendre la circulation des données, leur emplacement de stockage et les utilisateurs impliqués. Elle joue un rôle crucial dans l'identification des éléments essentiels pour garantir la qualité des données.

Par ailleurs, il est essentiel de réaliser des vérifications régulières de la qualité des données. Ces audits visent à s'assurer que les normes de qualité sont respectées et à repérer les domaines susceptibles d'être améliorés.

7. Collaboration inter-équipe

Pour favoriser une meilleure communication entre nous, nous pouvons recourir à des plateformes collaboratives telles que Discord.

Nous pouvons également tirer parti d'outils de coopération comme Slack ou Microsoft Teams, qui facilitent des échanges fluides entre les équipes. Ces plateformes nous permettront de partager des mises à jour, d'aborder les enjeux liés à la qualité des données et de collaborer pour trouver des solutions adéquates.

Il serait également judicieux d'organiser des réunions régulières entre les différentes équipes afin d'évaluer la qualité des données, de partager nos succès et nos difficultés, et d'encourager des retours d’information constructifs.

8. Amélioration continue

Il est essentiel d'organiser des rétrospectives, enrichies de retours d'expérience, à l'issue de chaque sprint afin de perfectionner les processus en place. À la fin de chaque cycle, il est important de mener une évaluation pour identifier les points forts et les aspects à améliorer. Cela doit encourager un échange constructif de feedbacks, favorisant ainsi un climat d'apprentissage.

Suite aux échanges lors de ces rétrospectives, il convient d'adapter les processus de gestion de la qualité des données. L'objectif est de mieux satisfaire les besoins de l'équipe tout en assurant une amélioration continue.

III. Users Story

En nous basant sur les objectifs globaux du projet, nous allons définir et concevoir des users stories afin de répondre à des aspects et des besoins spécifiques qui seront répondus par le logiciel choisi, OpenMetaData. Nous allons d'abord définir les acteurs et leurs user stories.

Acteurs	User Story
Client	En tant que client interne, je veux pouvoir prioriser les besoins dans un tableau clair, afin de m'assurer que les nouvelles fonctionnalités développées répondent d'abord à mes attentes essentielles.
Collaborateur métier	En tant que collaborateur métier, je veux accéder à un catalogue répertoriant toutes les sources de données disponibles afin de trouver rapidement les informations dont j'ai besoin.
Equipe Projet	En tant que membre d'équipe projet, je veux pouvoir ajouter des commentaires et annotations aux jeux de données afin de partager mes retours avec mes collègues en temps réel.
Administrateur	En tant qu'Administrateur, je veux gérer les droits d'accès aux données de la plateforme en fonction des rôles et des responsabilités de chaque utilisateur, tout en garantissant une traçabilité complète des actions, afin de protéger les données sensibles.

IV. Fonctionnalité OpenMetaData

Les différentes fonctionnalités d'OpenMetadata permettent de répondre aux besoins variés liés à la gestion des métadonnées au sein des organisations.

Tout d'abord, OpenMetadata se distingue par sa capacité à centraliser l'ensemble des métadonnées. Elle regroupe des informations cruciales telles que la provenance, la qualité et les types de données, offrant ainsi une vue globale et consolidée des données disponibles dans l'organisation.

Ensuite, la plateforme facilite la création d'un catalogue complet répertoriant toutes les sources de données. Ce catalogue rend la recherche et la découverte de jeux de données pertinents plus simples et efficaces pour les utilisateurs.

OpenMetadata se révèle également précieux en matière de gouvernance des données. Ses outils permettent de définir des règles de sécurité strictes, de gérer les accès aux données et d'assurer que celles-ci sont utilisées de manière conforme aux politiques internes.

La traçabilité des données, ou Data Lineage, est un autre aspect essentiel. Elle permet de suivre l'historique des données, en identifiant leur origine et les modifications qu'elles ont subies au fil du temps. Cette fonctionnalité aide à comprendre l'impact des transformations au sein des pipelines de données.

La collaboration entre les équipes est encouragée grâce aux options d'annotations et de commentaires. Les utilisateurs peuvent échanger des idées, signaler des erreurs, poser des questions ou initier des discussions autour des jeux de données, renforçant ainsi le travail d'équipe et l'amélioration continue des données.

Par ailleurs, OpenMetadata s’intègre facilement à divers outils de l’écosystème, comme Apache Airflow et Snowflake. Cela permet d'automatiser la collecte des métadonnées et de les synchroniser avec d'autres systèmes, rendant les processus plus fluides.

L'automatisation est encore renforcée par la présence d'API extensibles, permettant la gestion automatisée de tâches telles que la classification des données et le suivi des performances.

Enfin, en tant que solution open-source, OpenMetadata bénéficie du soutien d'une communauté active. Cette dernière contribue à son évolution et offre aux entreprises la possibilité de l’adapter à leurs besoins spécifiques, sans être contraintes par un fournisseur propriétaire.

V. Backlogs

Fonctionnalités	USER STORY	CRITÈRES D'ACCEPTANCE	EFFORT	PRIORITÉ
Gouvernance des données	En tant que product owner, je veux pouvoir définir des règles de qualité des données pour garantir la conformité des données	Étant donné que je suis Product Owner ,quand je suis sur la page de configuration des règles, alors je veux pouvoir définir des politiques de sécurité, gérer les accès, et surveiller les flux de données pour assurer une utilisation conforme et sécurisée	8 points	5
Gestion du backlog des fonctionnalités	En tant que Product Owner, je veux pouvoir créer et gérer le backlog des fonctionnalités, afin que je puisse prioriser les besoins des utilisateurs et garantir que l'équipe de développement se concentre sur les tâches les plus importantes.	Étant donné que je suis Product Owner, quand je consulte le backlog des fonctionnalités, alors je peux ajouter, prioriser et gérer les éléments de backlog, avec des critères d'acceptance clairs et une visibilité sur leur avancement.	6 points	4
Gestion des rôles et des permissions	En tant que scrum master/ administrateur, je veux pouvoir assigner des rôles et des responsabilités aux utilisateurs.	Étant donné que je suis connecté en tant que scrum master, quand je consulte la page de gestion des rôles, alors je peux attribuer des droits spécifiques à chaque utilisateur en fonction de son rôle mais également vérifier la progression de nos travaux.	9 points	5

Intégration avec des outils	En tant que data analyst, je souhaite intégrer la solution avec des outils d'automatisation pour déclencher des contrôles.	Étant donné que je suis connecté en tant que data analyst, quand je consulte la page des intégrations, alors je peux intégrer divers outils de traitement, analyse, et gouvernance pour automatiser la collecte et synchroniser les métadonnées avec les systèmes existants	7 points	2
Centralisation des métadonnées	En tant que data steward, je veux pouvoir centrer les informations liées aux données pour avoir une vue d'ensemble	Étant donné que je suis data steward, quand je suis connecté à l'interface de gestion des données, alors je peux accéder à une vue d'ensemble centralisée des informations, avec des mises à jour en temps réel et la possibilité de générer des rapports d'analyse.	7 points	4
Traçabilité des données (Lineage)	En tant que data analyst, je veux suivre l'origine et les transformations des données, afin que je puisse comprendre leur évolution et anticiper les impacts des modifications.	Étant donné que je suis data analyst, quand je consulte les données, alors je peux suivre l'origine et les transformations des données pour comprendre leur évolution et anticiper les impacts des modifications	8 points	4

Catalogue des données	En tant qu'utilisateur je veux créer un catalogue de données répertoriant les sources disponibles, afin que je puisse rechercher et découvrir les données rapidement.	Étant donné que je suis utilisateur, quand je consulte le catalogue de données, alors je peux voir une liste complète des sources de données disponibles, effectuer des recherches par mots-clés ou filtres et accéder à des informations détaillées sur chaque source	6 points	3
Collaboration et annotations	En tant que Scrum Master, je veux permettre la collaboration en autorisant les commentaires et annotations sur les jeux de données, afin que l'équipe puisse suivre les discussions et partager des retours en temps réel.	Étant donné que je suis Scrum Master, quand je consulte un jeu de données, alors je peux voir et ajouter des commentaires et annotations, permettant ainsi à l'équipe de suivre les discussions et de partager des retours en temps réel.	5 points	2

VI. Réalisation passés

L'équipe a réalisé plusieurs avancées notables. Les logiciels ont été sélectionnés avec soin après une analyse approfondie des besoins et des critères de performance, garantissant ainsi leur adéquation avec les objectifs fixés. La répartition des tâches s'est faite en tenant compte des compétences de chaque membre, permettant à chacun de jouer un rôle essentiel dans le projet.

Des objectifs clairs ont été définis dès le départ afin d'assurer une vision commune et une parfaite cohésion dans l'exécution. Chaque membre a su trouver sa voie et contribuer efficacement à l'avancement du projet. Ce parcours témoigne d'une organisation minutieuse et d'une collaboration efficace, des éléments clés ayant permis de mener à bien l'ensemble des initiatives.

VII. Réalisation futures

L'équipe a de grandes ambitions pour l'avenir. Il s'agit tout d'abord de créer un référentiel regroupant les principaux mots clés liés à l'activité de la plateforme, dans le but d'harmoniser le vocabulaire utilisé par les différentes parties prenantes et ainsi faciliter la collaboration. Un autre objectif est de développer un catalogue complet des données présentes sur la plateforme, détaillant chaque source de données, son origine, son type et son usage dans les processus métier.

En complément, il sera nécessaire de documenter chaque table de la base de données, en fournissant des informations précises sur leur structure, comme les colonnes, les types de données ou encore les relations entre les tables, afin de garantir une meilleure compréhension technique. La mise en place d'une cartographie des flux de données et de leur localisation dans les différents systèmes de la plateforme sera également cruciale pour visualiser les relations entre les sources et leur utilisation.
Grâce à ce référentiel unifié, l'équipe pourra collaborer de manière plus efficace en partageant une compréhension commune des données, ce qui améliorera la transparence et facilitera la communication.

VIII. Les difficultés rencontrées

L'équipe a rencontré plusieurs difficultés au cours du projet. Un manque de clarté sur les priorités a compliqué l'harmonisation des attentes des parties prenantes, notamment concernant les fonctionnalités essentielles à développer. Cette situation a créé des divergences dans la priorisation des tâches, rendant plus difficile l'avancée fluide du projet.

La complexité des données a également posé un défi important. Il a été difficile d'identifier des indicateurs de qualité adaptés, en raison de la diversité des sources de données et de leur structure. Cela a impacté la capacité à garantir une exploitation optimale des données dans les processus métier.

Enfin, des ressources limitées ont constitué un autre obstacle, notamment en ce qui concerne les compétences spécialisées nécessaires pour mener à bien certaines tâches techniques. Cette insuffisance a ralenti certaines étapes du projet, nécessitant des ajustements pour compenser le manque de personnel qualifié.

IX. Notre avancé sur le projet

Voici un tableau extrait de la plateforme Trello qui récapitule l'ensemble des tâches que nous avons traités ensemble :