



ÉCOLE NATIONALE SUPÉRIEURE  
D'INFORMATIQUE ET D'ANALYSE DES SYSTÈMES  
- RABAT

Rapport du Projet de Fin d'Année

FILIÈRE : GÉNIE LOGICIEL

---

**GovernX : Moteur de  
recommandation pour la  
gouvernance des Métadonnées**

---

*Réalisé par :*  
RAJAT CHOROUK  
RAIS RIM  
ANEJARI HALIMA  
RGUIBI ASSIA

*Encadré par:*  
PR. BAINA KARIM  
MME. GASMI MANAL

# Table des figures

2.1	Diagramme DFD Niv 0 . . . . .	12
2.2	DFD Niv 1 . . . . .	13
2.3	Diagramme de dépendance fonctionnelle . . . . .	16
2.4	Modèle Conceptuel de Données (MCD) . . . . .	17
2.5	Modèle Logique de Données (MLD) . . . . .	19
2.6	MCT du processus 1 . . . . .	20
2.7	MCT du processus 2 . . . . .	21
2.8	MCT du processus 3 . . . . .	22
2.9	MCT du processus 4 . . . . .	23
2.10	MCT du processus 5 . . . . .	24
2.11	MOT du processus 1 . . . . .	25
2.12	MOT du processus 2 . . . . .	26
2.13	MOT du processus 3 . . . . .	27
2.14	MOT du processus 4 . . . . .	28
2.15	MOT du processus 5 . . . . .	29
4.1	Présentation du système . . . . .	34
4.2	Les Fonctionnalités . . . . .	34
4.3	Vue d'ensemble sur les responsabilité de chaque rôle . . . . .	35
4.4	Comment ça marche le système . . . . .	35
4.5	Interface SignUp . . . . .	36
4.6	Interface SignIn . . . . .	37
4.7	Les identifiants pour les services Atlas et Ranger . . . . .	38
4.8	Ajout d'un seul membre . . . . .	39
4.9	Ajout de plusieurs membres en utilisant un csv . . . . .	39
4.10	Affichage des membres créés et génération du pdf contenant leurs identifiants	40
4.11	Aperçu du pdf généré . . . . .	40
4.12	Activation ou Désactivation de tous les membres . . . . .	41
4.13	Création M&j Suppression des Data Domains . . . . .	41
4.14	Gestion individuelle des membres . . . . .	42
4.15	Attribution d'un Data Domain . . . . .	42
4.16	Diagrammes sur l'activité des membres . . . . .	43
4.17	Vue sur le Pré-traitement des fichiers exécuté par les membres . . . . .	43
4.18	Aperçu du pdf généré . . . . .	44
4.19	Chargement du fichier et déclenchement de l'analyse . . . . .	45
4.20	Résumé des métadonnées . . . . .	45
4.21	élimination des valeurs manquantes . . . . .	46
4.22	elimination des doublons . . . . .	46
4.23	Analyse du fichier résultant . . . . .	46

4.24	Types de donnés détectés et suggérés . . . . .	46
4.25	Normalisation des colonnes . . . . .	47
4.26	Détection des patterns . . . . .	47
4.27	Clustering sémantique . . . . .	47
4.28	Détection des valeurs abérantes et métadonnées completes . . . . .	47
4.29	Affichage des bases de données existantes . . . . .	48
4.30	Sélection de la db et affichage des tables . . . . .	48
4.31	Sélection de la table et affichage des colonnes . . . . .	48
4.32	Annotation Standard en utilisant les termes du glossaire atlas . . . . .	48
4.33	Annotation Personnelle en utilisant les termes du glossaire personnel . . . . .	48
4.34	Gestion du glossaire personnel . . . . .	49
4.35	Affichage des annotations personnelles . . . . .	49
4.36	Affichage des annotations standards . . . . .	49
4.37	Filtrage et vue globale . . . . .	50
4.38	Statistiques sur les annotations standards et personnelles . . . . .	50
4.39	Statistiques sur les actions . . . . .	50
4.40	Vue détaillée sur les actions récentes . . . . .	50
4.41	Sélection de la base de données . . . . .	51
4.42	Affichage des métadonnées sur la base de données . . . . .	51
4.43	Sélection de la table . . . . .	51
4.44	Affichage des métadonnées sur la table . . . . .	51
4.45	Sélection de la colonnes et affichage de ses métadonnées . . . . .	51
4.46	Affichage et gestion des recommandations . . . . .	52

# Table des matières

<b>Introduction générale</b>	<b>5</b>
<b>1 Contexte général du projet</b>	<b>6</b>
1.1 Gouvernance des Métadonnées : État de l'art . . . . .	6
1.1.1 Définition et enjeux . . . . .	6
1.1.2 Les outils et concepts clés . . . . .	6
1.1.3 Apache Atlas : un outil clé pour la gouvernance des métadonnées . . . . .	7
1.2 Problématique . . . . .	8
1.3 Solution . . . . .	8
1.4 Objectifs . . . . .	9
1.5 Plannification du projet . . . . .	9
<b>2 Analyse et Conception</b>	<b>10</b>
2.1 Analyse des besoins . . . . .	10
2.1.1 Besoins fonctionnels . . . . .	10
2.1.2 Besoins non fonctionnels . . . . .	11
2.2 Diagrammes de flux de données . . . . .	11
2.2.1 DFD Niveau 0 (Contexte) . . . . .	12
2.2.2 DFD Niveau 1 . . . . .	12
2.3 Dictionnaire des Données . . . . .	13
2.4 Graphe de Dépendance Fonctionnelle . . . . .	15
2.5 Modèle Conceptuel de Données (MCD) . . . . .	16
2.6 Modèle Logique de Données (MLD) . . . . .	18
2.7 Modèle Conceptuel de Traitement (MCT) . . . . .	20
2.7.1 MCT du processus 1 : Initialisation & Gestion des équipes . . . . .	20
2.7.2 MCT du processus 2 : Ingestion & Préparation des données . . . . .	21
2.7.3 MCT du processus 3 : Annotation et validation . . . . .	21
2.7.4 MCT du processus 4 : Gouvernance des métadonnées . . . . .	22
2.7.5 MCT du processus 5 : Journalisation . . . . .	23
2.8 Modèle Opérationnel de Traitement (MOT) . . . . .	24
2.8.1 MOT du processus 1 : Initialisation & Gestion des équipes . . . . .	24
2.8.2 MOT du processus 2 : Ingestion & Préparation des données . . . . .	25
2.8.3 MOT du processus 3 : Annotation et validation . . . . .	26
2.8.4 MOT du processus 4 : Gouvernance des métadonnées . . . . .	27
2.8.5 MOT du processus 5 : Journalisation & suivi des activités . . . . .	28
<b>3 Technologies Utilisées</b>	<b>30</b>
3.1 Couche Backend . . . . .	30

3.2	Couche Frontend . . . . .	31
3.3	Base de Données . . . . .	31
3.4	Sources de Métadonnées et Stockage Big Data . . . . .	31
3.5	Modèles de Recommandation . . . . .	32
<b>4</b>	<b>Réalisation</b>	<b>33</b>
4.1	Introduction . . . . .	33
4.2	Interfaces Générales : Authentification et Accueil . . . . .	33
4.2.1	Page d'accueil . . . . .	34
4.2.2	Page Registration . . . . .	36
4.2.3	Page Login . . . . .	37
4.2.4	Page Gestion des identifiants . . . . .	38
4.3	Rôle : Team Admin . . . . .	38
4.3.1	Création des membres . . . . .	39
4.3.2	Gestion des membres . . . . .	41
4.3.3	Page des statistiques . . . . .	43
4.4	Rôle : Data Analyst . . . . .	44
4.4.1	Page Analyse . . . . .	45
4.4.2	Page d'annotation . . . . .	48
4.4.3	Page statistiques . . . . .	50
4.5	Rôle : Data Steward . . . . .	51
4.5.1	Page Qualité des métadonnées . . . . .	51

# Introduction générale

Dans un contexte où les organisations collectent et exploitent des volumes massifs de données, la gouvernance des métadonnées occupe une place stratégique. Elle vise à garantir la qualité, la traçabilité, la sécurité et la valorisation des données au sein des systèmes d'information. Pourtant, la mise en place et le maintien d'une gouvernance efficace nécessitent des efforts importants, souvent réalisés manuellement, ce qui peut entraîner des oubliers, des incohérences ou des erreurs.

Face à ces défis, les entreprises s'appuient sur des catalogues de métadonnées, tels qu'Apache Atlas, pour inventorier, documenter et classer leurs données. Toutefois, ces outils restent dépendants des contributions humaines pour enrichir, corriger et maintenir les informations à jour.

Le projet présenté dans ce rapport vise à apporter une réponse à cette problématique en développant un moteur de recommandation dédié à la gouvernance des métadonnées. Ce moteur sera capable de générer automatiquement des suggestions pertinentes concernant la qualité, l'enrichissement, la conformité et la découvrabilité des métadonnées, tout en respectant les politiques de sécurité et de confidentialité.

L'objectif est d'assister les utilisateurs (data stewards, data engineers, responsables conformité) dans la gestion des métadonnées en leur fournissant des recommandations intelligentes, sans automatiser la correction, mais en leur laissant le pouvoir décisionnel final.

# Chapitre 1

## Contexte général du projet

Ce chapitre présente le contexte global du projet, en exposant l'état de l'art sur la gouvernance des métadonnées, la problématique rencontrée, la solution envisagée ainsi que les objectifs fixés.

### 1.1 Gouvernance des Métadonnées : État de l'art

#### 1.1.1 Définition et enjeux

La gouvernance des métadonnées désigne l'ensemble des pratiques, processus, outils et politiques qui permettent de gérer, organiser, contrôler et valoriser les métadonnées dans une organisation. Elle vise à assurer la qualité, la traçabilité, la sécurité et la compréhension des données par toutes les parties prenantes. Les enjeux principaux de la gouvernance des métadonnées sont la conformité réglementaire, la fiabilité des analyses, la maîtrise des risques et la valorisation des actifs de données.

#### 1.1.2 Les outils et concepts clés

- ✓ **Glossaire métier** : un référentiel qui regroupe et définit de manière standardisée les termes et concepts utilisés dans l'organisation. Il permet d'éviter les ambiguïtés et d'assurer un langage commun entre les équipes métier, technique et conformité.
- ✓ **Dictionnaire de données** : un référentiel technique décrivant chaque donnée (nom, type, format, contraintes), souvent lié directement aux bases de données et systèmes d'information. Il complète le glossaire métier et facilite la compréhension technique des jeux de données.
- ✓ **Catalogue de données** : un outil centralisé permettant d'inventorier, documenter, classer et rechercher l'ensemble des jeux de données d'une organisation. Il facilite la découverte, la gouvernance et l'utilisation des données.
- ✓ **Data lineage (traçabilité des données)** : la représentation du parcours complet d'une donnée depuis sa source jusqu'à son utilisation finale, incluant les transformations subies. Il permet de comprendre les dépendances, faciliter les audits et identifier l'origine des anomalies.

- ✓ **Tagging** : l'attribution d'étiquettes (tags) ou de classifications aux entités de données afin de faciliter leur recherche, leur catégorisation et leur gestion (par exemple : données sensibles, données personnelles).
- ✓ **Data masking (masquage des données)** : une technique qui consiste à masquer partiellement ou totalement les valeurs sensibles des données afin de limiter leur exposition, tout en conservant leur utilité pour les tests ou analyses.
- ✓ **Ontologies métier** : des structures permettant de modéliser les concepts métier et leurs relations, facilitant l'alignement entre les besoins métier et les structures techniques des données.
- ✓ **Metadata bridges** : des connecteurs ou interfaces permettant de synchroniser et partager les métadonnées entre différents outils, catalogues et plateformes pour assurer l'interopérabilité et la cohérence.

### 1.1.3 Apache Atlas : un outil clé pour la gouvernance des métadonnées

Apache Atlas est une plateforme open-source de gestion des métadonnées et de gouvernance des données. Elle fournit un cadre puissant pour cataloguer, tracer, sécuriser et enrichir les métadonnées dans des environnements big data complexes.

Les principales fonctionnalités d'Apache Atlas sont :

- **Classification des données**
  - Création dynamique de classifications telles que *PII*, *EXPIRES\_ON*, *DATA\_QUALITY*, ou *SENSITIVE*.
  - Possibilité d'ajouter des attributs personnalisés aux classifications (par exemple, une date d'expiration pour *EXPIRES\_ON*).
  - Association d'une entité à plusieurs classifications, facilitant la recherche et la mise en œuvre des politiques de sécurité.
  - Propagation automatique des classifications via le lineage, permettant aux règles de sécurité de suivre les données tout au long de leur cycle de vie.
- **Recherche et découverte**
  - Interface utilisateur intuitive pour rechercher des entités selon le type, la classification, la valeur d'un attribut ou par texte libre.
  - API REST riche permettant d'effectuer des recherches avancées par des critères complexes.
  - Langage de requête spécifique (DSL) semblable au SQL pour interroger les métadonnées.
- **Lignée centralisée (Centralized Lineage)**
  - Visualisation graphique et intuitive du parcours des données à travers différents processus et transformations.
  - API REST pour accéder et mettre à jour les informations de lineage.
- **Audit centralisé**
  - Fonctionnalités permettant de tracer toutes les actions réalisées sur les métadonnées (consultation, ajout, suppression, modification), indispensable pour les audits de conformité et la sécurité.

- **Moteur de sécurité et de politiques**

- Contrôle précis des accès aux métadonnées, avec des droits granulaires sur les entités et les opérations (ajout, mise à jour, suppression des classifications).
- Intégration avec Apache Ranger pour appliquer des politiques d'autorisation et de masquage des données en fonction des classifications.
- Par exemple, seuls certains utilisateurs peuvent accéder aux données classées *PII* ou *SENSITIVE*, et des règles de masquage peuvent être appliquées pour ne montrer que les 4 derniers chiffres d'un identifiant national.

- **Ontologies métier et « metadata bridges »**

- Atlas permet de créer et maintenir des ontologies métier, reliant des concepts métier à des éléments techniques dans le catalogue.
- Les métadonnées *bridges* facilitent l'intégration et l'interopérabilité entre différents outils et sources de données.

## 1.2 Problématique

La gouvernance des métadonnées constitue aujourd’hui un enjeu majeur pour les organisations. Assurer la qualité, la conformité, la traçabilité et la sécurité des métadonnées demande un travail considérable, souvent manuel, qui mobilise des ressources importantes et reste sujet à des oubliers ou des erreurs. La détection des incohérences, l'enrichissement des métadonnées, le suivi des politiques de gouvernance, ainsi que la détection des doublons et des incohérences nécessitent des interventions fréquentes et expertes. Ce processus, long et fastidieux, freine la valorisation des données et la capacité à réagir rapidement aux exigences réglementaires ou métiers.

## 1.3 Solution

Afin de répondre à cette problématique, la solution proposée repose sur la mise en place d'un moteur de recommandation capable de générer des suggestions pertinentes autour des métadonnées.

Ce moteur de recommandation permettra notamment de :

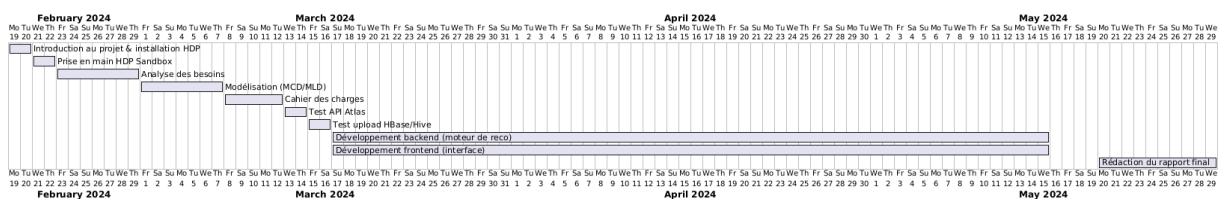
- Identifier les incohérences et manques dans les métadonnées (champs vides, formats non standard).
- Proposer des enrichissements sémantiques (ajout de tags, catégories ou relations vers des vocabulaires ou ontologies).
- Recommander des bonnes pratiques de gouvernance (règles de rétention, classification des données sensibles).
- Suggérer des améliorations pour la découverbarilité et le référencement (mots-clés, taxonomies).
- Déetecter les redondances et proposer des fusions ou des corrections manuelles.
- Renforcer la sécurité et le contrôle d'accès en recommandant des niveaux de confidentialité et des politiques d'autorisation adaptées aux classifications des données.

## 1.4 Objectifs

- Faciliter la gestion des métadonnées en allégeant le travail manuel par la génération de recommandations ciblées.
  - Aider à la détection des anomalies et incohérences pour améliorer la fiabilité des métadonnées.
  - Soutenir l'enrichissement des métadonnées en suggérant des compléments d'informations et des liens sémantiques.
  - Améliorer la gouvernance des métadonnées en proposant des recommandations conformes aux politiques internes et aux normes.
  - Renforcer la qualité et la valorisation du catalogue sans intervention corrective automatique, en laissant à l'utilisateur la décision finale.
  - Renforcer la sécurité en orientant les utilisateurs vers des pratiques d'accès contrôlé et de classification adaptée des données sensibles.

## 1.5 Plannification du projet

La planification du projet a été réalisée à l'aide d'un diagramme de Gantt, qui illustre les différentes phases et tâches du développement. Ce diagramme permet de visualiser l'échéancier du projet et de suivre l'avancement des travaux.



Ce contexte général a permis de mettre en évidence les enjeux liés à la gouvernance des métadonnées ainsi que les insuffisances des approches existantes. La problématique et les objectifs du projet étant désormais clairs, nous pouvons passer à l'analyse des besoins et à la conception de notre moteur de recommandation.

# Chapitre 2

## Analyse et Conception

Ce chapitre est consacré à l'analyse et à la conception du système. Il présente dans un premier temps l'analyse des besoins fonctionnels et non fonctionnels du projet, avant de détailler le dictionnaire de données. La deuxième partie du chapitre est dédiée à la modélisation à travers les différents modèles utilisés : le Modèle Conceptuel de Données (MCD), le Modèle Logique de Données (MLD), le Modèle Conceptuel des Traitements (MCT) et le Modèle Organisationnel des Traitements (MOT). Ces éléments permettent de structurer et de préparer la réalisation technique du moteur de recommandation pour la gouvernance des métadonnées.

### 2.1 Analyse des besoins

#### 2.1.1 Besoins fonctionnels

Les besoins fonctionnels correspondent aux fonctionnalités que le système devra offrir pour répondre à la problématique identifiée. Le moteur de recommandation pour la gouvernance des métadonnées devra permettre :

- **Analyse des métadonnées :**
  - Déetecter les champs vides ou non conformes.
  - Identifier les incohérences et les doublons.
- **Propositions d'enrichissement :**
  - Suggérer des tags, des catégories et des relations sémantiques à partir d'ontologies métier.
  - Proposer des règles de classification et des pratiques de gouvernance adaptées.
- **Suggestions de bonnes pratiques :**
  - Recommander des règles de rétention, de sécurité et de confidentialité.
  - Proposer des améliorations pour la découvrabilité (ajout de mots-clés ou taxonomies).
- **Gestion des recommandations :**
  - Afficher les suggestions de manière lisible.
  - Permettre à l'utilisateur d'accepter, de rejeter ou de modifier les recommandations.

— **Reporting et suivi :**

- Générer des rapports synthétiques sur les actions recommandées et mises en œuvre.

### 2.1.2 Besoins non fonctionnels

Les besoins non fonctionnels décrivent les contraintes de qualité et de performance du système :

— **Performance :**

- Le moteur de recommandation doit analyser et proposer des suggestions en temps raisonnable, même sur des volumes importants de métadonnées.

— **Scalabilité :**

- La solution doit pouvoir évoluer et s'adapter à l'augmentation du volume des métadonnées et à la complexité des catalogues.

— **Sécurité :**

- Les données manipulées doivent être protégées, et les recommandations doivent respecter la confidentialité des informations sensibles.

— **Interopérabilité :**

- La solution devra s'intégrer facilement avec les outils de gouvernance existants (comme Apache Atlas).

— **Ergonomie :**

- L'interface devra être simple, intuitive, et permettre une interaction fluide avec l'utilisateur.

— **Traçabilité :**

- Toutes les actions (suggestions proposées, acceptées ou rejetées) devront être journalisées pour audit et suivi.

## 2.2 Diagrammes de flux de données

Les Diagrammes de Flux de Données (DFD) sont des outils de modélisation essentiels utilisés pour représenter la circulation de l'information au sein d'un système d'information. Ils permettent de visualiser les processus, les flux de données, les entités externes et les stockages impliqués dans le traitement des données. Les DFD facilitent ainsi la compréhension du fonctionnement du système en mettant l'accent sur les échanges d'informations plutôt que sur la logique de traitement interne. Ils sont généralement structurés en plusieurs niveaux de détail, allant du diagramme de contexte (niveau 0) aux niveaux plus fins (niveau 1, niveau 2...), permettant une modélisation progressive et claire du système étudié.

### 2.2.1 DFD Niveau 0 (Contexte)

Le diagramme de flux de données (DFD) de niveau 0, également appelé diagramme de contexte, présente une vision globale et simplifiée du système de recommandation pour la gouvernance des métadonnées. Il met en évidence les flux d'information entre le système et les acteurs externes, à savoir les différents types d'utilisateurs (administrateur, data analyst, data steward, Hadoop). Ce niveau ne détaille pas les processus internes mais sert à délimiter le périmètre fonctionnel du système et à illustrer les principales entrées et sorties des données. Il constitue une base essentielle pour la compréhension globale de l'architecture du système avant son déploiement en sous-processus.

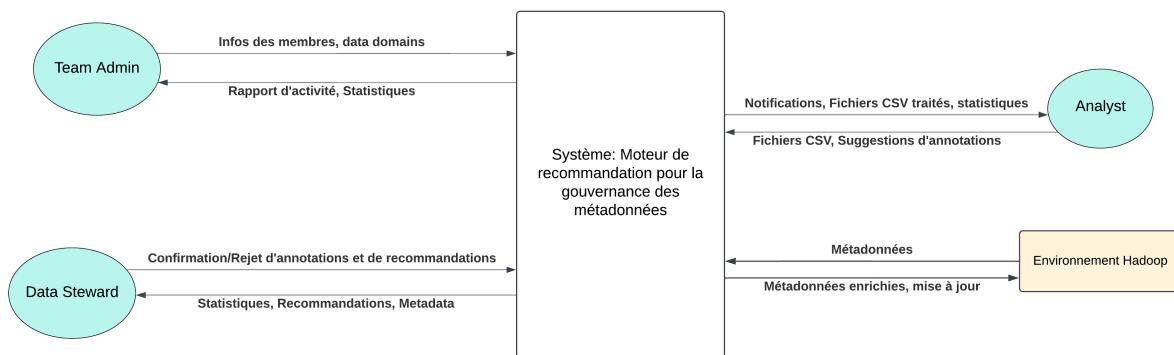


FIGURE 2.1 – Diagramme DFD Niv 0

### 2.2.2 DFD Niveau 1

Le DFD de niveau 1 permet d'approfondir la modélisation en décomposant le système en processus fonctionnels cohérents, tout en conservant une perspective globale sur les flux de données. Ce niveau décrit les interactions internes entre les entités externes, les sous-processus, les fichiers de données et les flux échangés. Chaque processus correspond à une fonctionnalité majeure du système, comme l'ingestion des données, l'annotation, la validation ou encore le suivi d'activité. Ce diagramme est crucial pour comprendre la dynamique opérationnelle du système et poser les fondations de sa conception détaillée à travers les MCT (Modèles Conceptuels de Traitement).

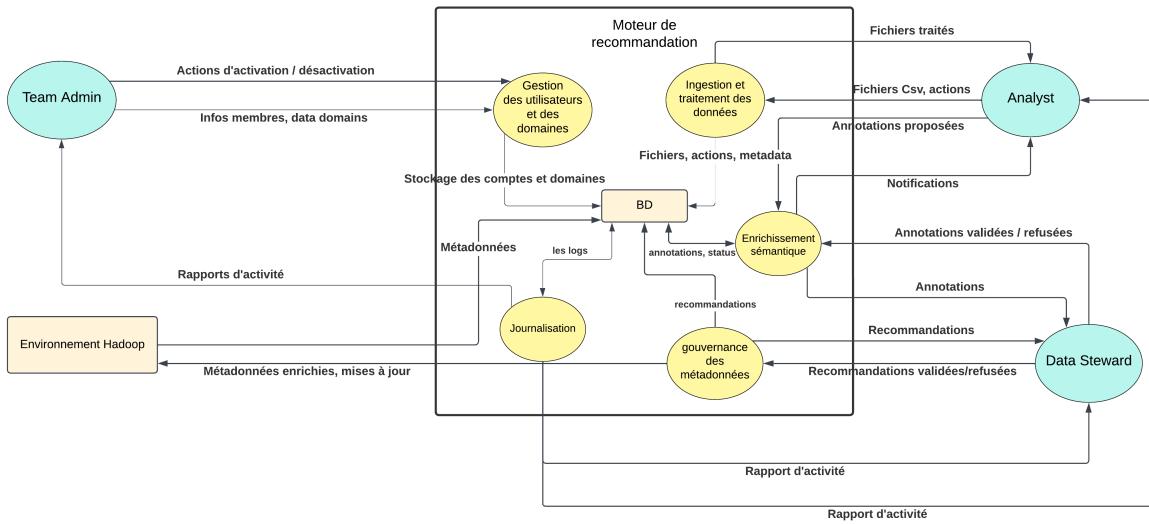


FIGURE 2.2 – DFD Niv 1

## 2.3 Dictionnaire des Données

Le dictionnaire des données accompagne le modèle conceptuel de données en détaillant les éléments qui le composent. Il décrit précisément chaque entité, attribut et valeur possible, en fournissant des informations sur les types de données, les contraintes et les relations. Ce document est essentiel pour garantir une compréhension uniforme et précise des données utilisées dans le système, et pour faciliter la transition vers les étapes suivantes de la conception.

Nom	Type	Description
email	Varchar	Adresse email de l'utilisateur
password	Varchar	Mot de passe crypté de l'utilisateur
role	Varchar	Rôle ou niveau d'accès de l'utilisateur
first_name	Varchar	Prénom de l'utilisateur
last_name	Varchar	Nom de famille de l'utilisateur
is_active	Bool	Statut d'activation du compte
last_login_date	Datetime	Dernière date de connexion
id_team	Auto_increment	Identifiant de l'équipe
team_name	Varchar	Nom de l'équipe
id_domain	Auto_increment	Identifiant du domaine
domain_name	Varchar	Nom du domaine de données
domain_description	Varchar	Description du domaine
atlas_username	Varchar	Nom d'utilisateur Atlas
atlas_key	Varchar	Clé d'accès Atlas
url	Varchar	URL d'accès ou de ressource
file_id	Auto_increment	Identifiant unique du fichier
file_name	Varchar	Nom du fichier

date_uploaded	Datetime	Date d'importation du fichier
atlas_entity_name	Varchar	Nom de l'entité dans Atlas
id_FileAction	Auto_increment	Identifiant de l'action sur fichier
date	Datetime	Date de l'action
id_user	Auto_increment	Identifiant de l'utilisateur
id_metadata	Auto_increment	Identifiant des métadonnées
column_name	Varchar	Nom de la colonne analysée
data_type	Varchar	Type de données détecté
missing_percentage	Float	Pourcentage de valeurs manquantes
is_outlier_present	Bool	Indicateur de présence de valeurs aberrantes
suggested_type	Varchar	Type suggéré pour les données
normalization	Varchar	Type de normalisation appliquée
pattern_detected	Varchar	Motif détecté dans les données
semantic_cluster	Varchar	Regroupement sémantique détecté
guid	Varchar	Identifiant global unique
qualified_name	Varchar	Nom qualifié de l'objet Atlas
location	Varchar	Emplacement du fichier ou de la ressource
owner	Varchar	Propriétaire du fichier ou entité
created_by	Varchar	Créateur de l'entité
updated_by	Varchar	Dernière personne à avoir modifié l'entité
create_time	Datetime	Date de création
update_time	Datetime	Date de la dernière mise à jour
temporary	Bool	Indique si l'objet est temporaire
table_type	Varchar	Type de table (e.g., EXTERNAL, MANAGED)
classifications	Varchar	Listes de classifications (sémantiques, métiers...)
retention_period	Int	Durée de rétention des données
position	Int	Position de la colonne dans la table
type	Varchar	Type de l'objet ou de la colonne
id_annotation	Auto_increment	Identifiant de l'annotation
entity_guid	Varchar	GUID de l'entité annotée
entity_type	Varchar	Type de l'entité annotée
entity_name	Varchar	Nom de l'entité annotée
term_guid	Varchar	GUID du terme lié à l'annotation
term_name	Varchar	Nom du terme lié
comment	Varchar	Commentaire de l'utilisateur
proposed_changes	Varchar	Modifications proposées
status	Varchar	Statut de l'annotation ou recommandation
created_at	Datetime	Date de création de l'entrée
updated_at	Datetime	Date de la dernière mise à jour
field	Varchar	Champ concerné par la recommandation
suggested_value	Varchar	Valeur recommandée
confidence	Float	Confiance en la recommandation (0-1)
json_file	Varchar	Contenu ou chemin d'un fichier JSON

last_updated	Datetime	Dernière mise à jour
is_changed	Bool	Indique si des modifications ont été faites
id_personal_glossary	Auto_increment	Identifiant du glossaire personnel
personal_glossary_name	Varchar	Nom du glossaire personnel
id_credentials	Auto_increment	Identifiant des identifiants Atlas
id_personal_annotation	Auto_increment	Identifiant d'une annotation personnelle
id_personal_term	Auto_increment	Identifiant d'un terme personnel
id_recommendation	Auto_increment	Identifiant d'une recommandation
id_glossary	Auto_increment	Identifiant du glossaire standard
db_guid	Varchar	GUID de la base Hive
db_name	Varchar	Nom de la base Hive
db_qualified_name	Varchar	Nom qualifié de la base Hive
db_owner	Varchar	Propriétaire de la base Hive
db_description	Varchar	Description de la base Hive
db_location	Varchar	Emplacement de la base Hive
table_guid	Varchar	GUID de la table Hive
table_name	Varchar	Nom de la table Hive
table_qualified_name	Varchar	Nom qualifié de la table Hive
table_owner	Varchar	Propriétaire de la table Hive
table_description	Varchar	Description de la table Hive
column_guid	Varchar	GUID de la colonne Hive
column_qualified_name	Varchar	Nom qualifié de la colonne Hive
column_owner	Varchar	Propriétaire de la colonne Hive
column_description	Varchar	Description de la colonne Hive
column_position	Int	Position de la colonne dans la table
column_type	Varchar	Type de données de la colonne
term	Varchar	Nom du terme sémantique utilisé

## 2.4 Graphe de Dépendance Fonctionnelle

Le graphe de dépendance fonctionnelle représente de manière visuelle les relations entre les données du système. Il met en évidence les dépendances entre les entités et leurs attributs, illustrant la logique et les interactions au sein du système d'information. Ce graphe est essentiel pour comprendre comment les données sont liées et influencent les processus, ce qui facilite la conception d'un modèle cohérent pour la base de données.

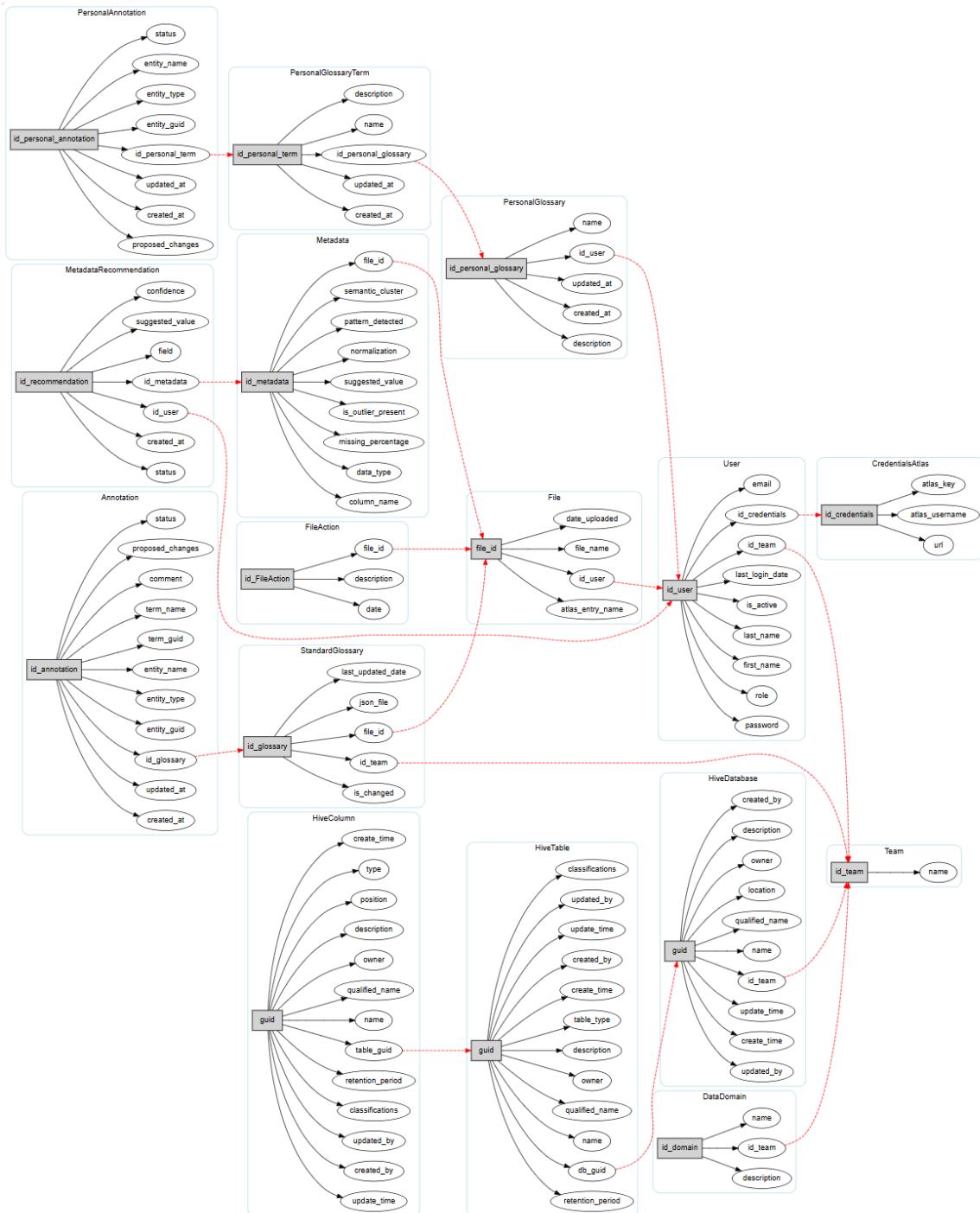


FIGURE 2.3 – Diagramme de dépendance fonctionnelle

## 2.5 Modèle Conceptuel de Données (MCD)

Le Modèle Conceptuel de Données (MCD) représente les entités principales du système ainsi que les relations qui les lient. Il constitue une abstraction permettant de visualiser la structure globale des données sans tenir compte des aspects techniques. Le MCD ci-dessous a été conçu à partir de l'analyse des besoins identifiés et servira de base à l'élaboration du modèle logique.

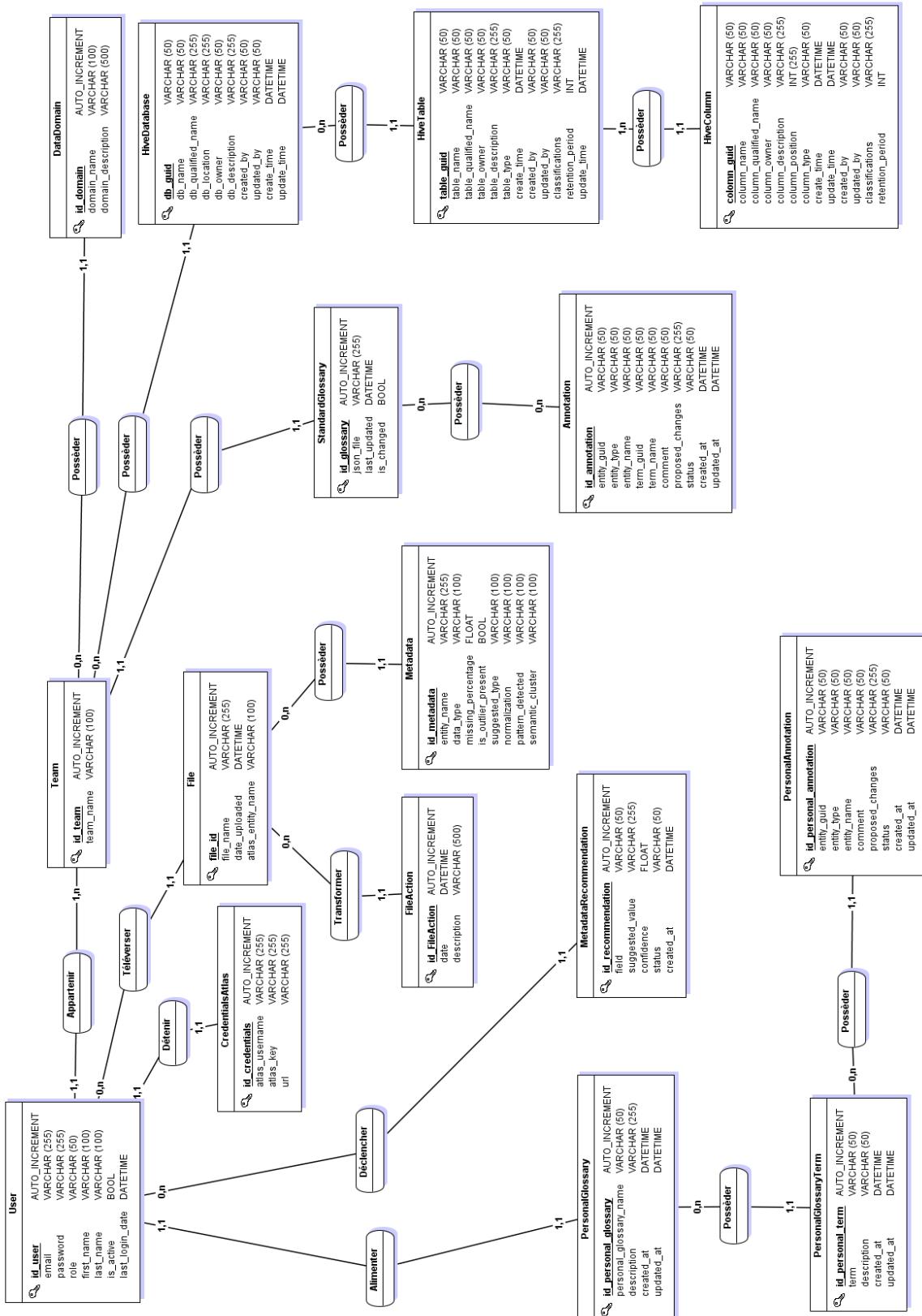


FIGURE 2.4 – Modèle Conceptuel de Données (MCD)

## 2.6 Modèle Logique de Données (MLD)

Le Modèle Logique de Données (MLD) est la traduction du MCD sous une forme plus détaillée, en intégrant les spécificités liées à la structuration des données dans un système de gestion de bases de données relationnel. Il permet de définir les tables, les clés primaires et étrangères ainsi que les contraintes d'intégrité qui seront mises en place lors du développement.

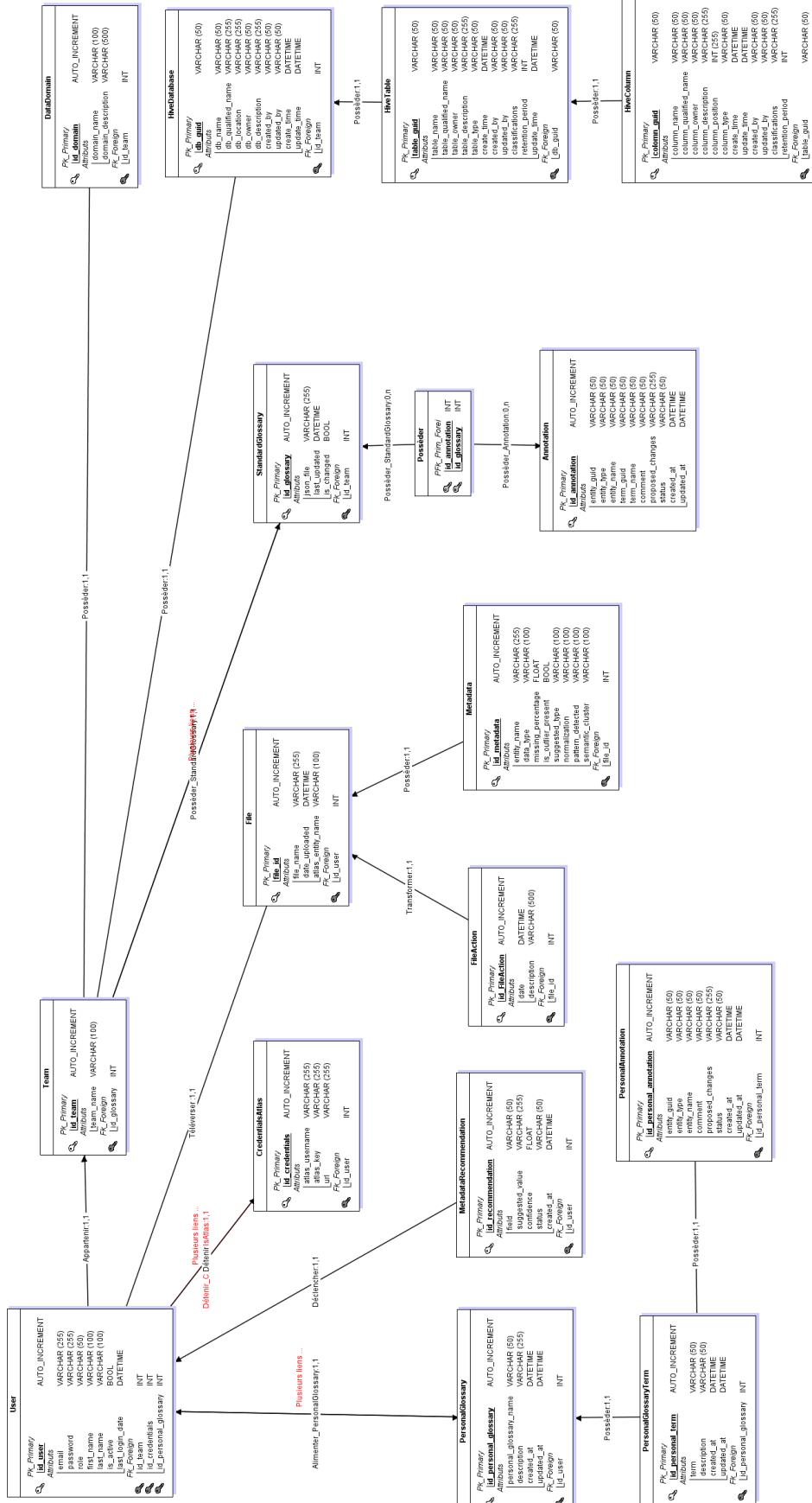


FIGURE 2.5 – Modèle Logique de Données (MLD)

## 2.7 Modèle Conceptuel de Traitement (MCT)

Le modèle conceptuel de traitement (MCT) présente les processus nécessaires au traitement des données au sein du système. Ce diagramme décrit les différentes étapes de traitement et les interactions entre les processus métiers et les données. Il offre une vue d'ensemble des mécanismes de transformation, de manipulation et de gestion des informations, permettant ainsi de comprendre comment le système traitera les données pour répondre aux besoins fonctionnels.

### 2.7.1 MCT du processus 1 : Initialisation & Gestion des équipes

Le processus Initialisation & Gestion des équipes couvre l'ensemble des tâches permettant à un administrateur de structurer son environnement de travail. Cela comprend la création de comptes utilisateurs (manuellement ou par import CSV), l'activation ou la désactivation des comptes, ainsi que la gestion des domaines de données.

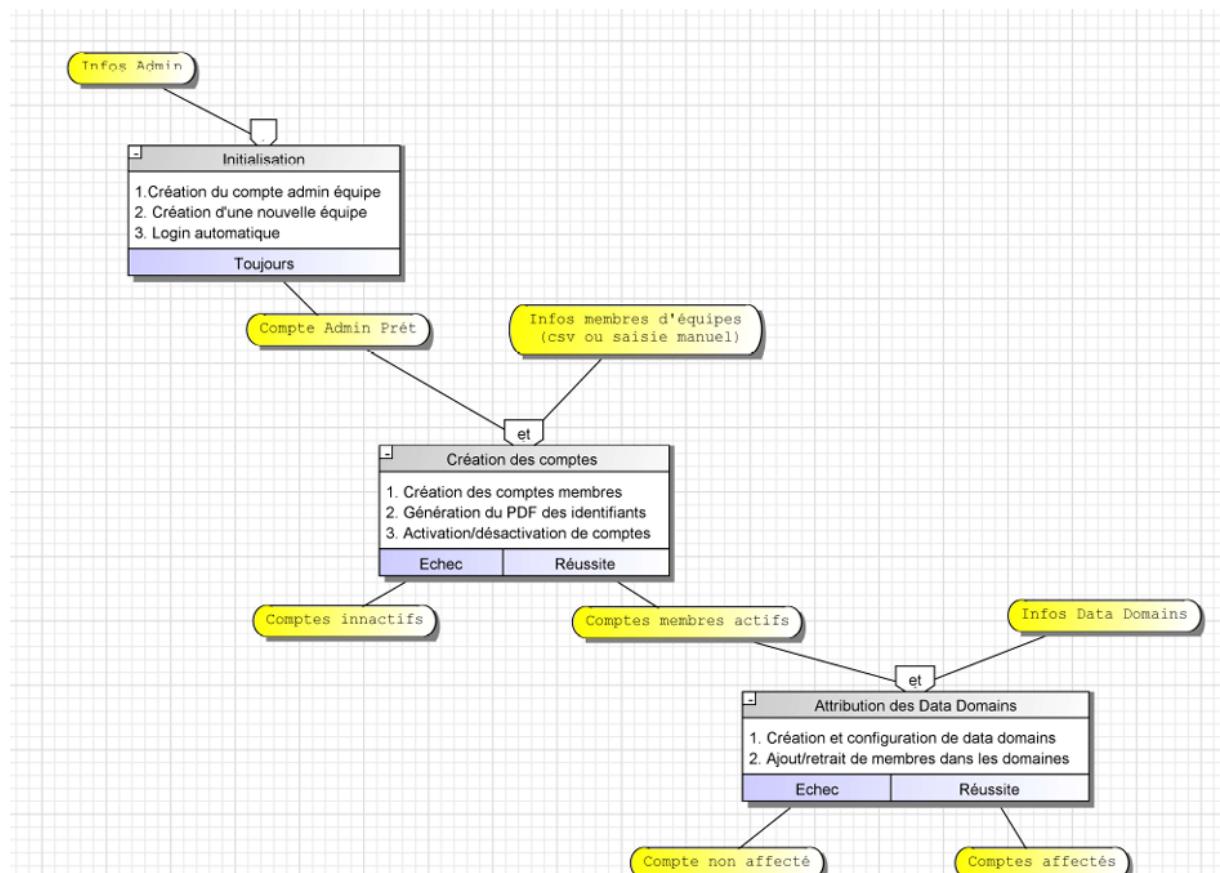


FIGURE 2.6 – MCT du processus 1

### 2.7.2 MCT du processus 2 : Ingestion & Préparation des données

Le processus Ingestion & Préparation des données représente les étapes exécutées par un Data Analyst pour intégrer des données brutes, les nettoyer, les analyser et les enrichir automatiquement. Il commence par l'import de fichiers CSV, suivi de traitements automatiques : détection et suppression des doublons et valeurs manquantes, suggestion de types de données, normalisation, reconnaissance de patterns, identification d'outliers, et clustering sémantique.

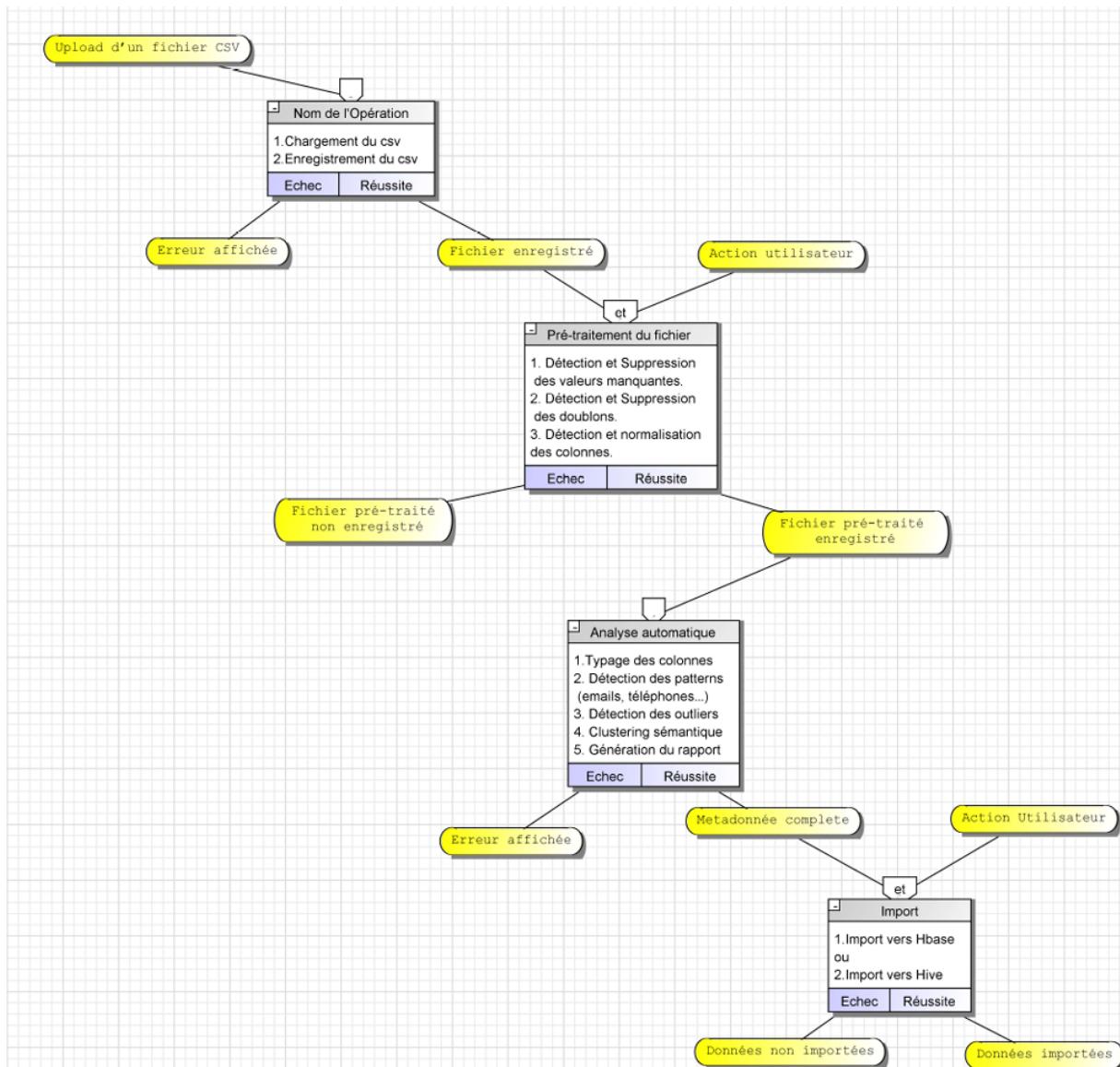


FIGURE 2.7 – MCT du processus 2

### 2.7.3 MCT du processus 3 : Annotation et validation

Ce processus couvre les activités collaboratives entre les Data Analysts et les Data Stewards dans le cadre de l'annotation des colonnes. Le Data Analyst propose des annotations sur les colonnes de données, soit à partir du glossaire métier issu d'Apache Atlas,

soit via des termes personnalisés. Ces annotations sont ensuite soumises à validation par le Data Steward, qui peut les approuver ou les rejeter.

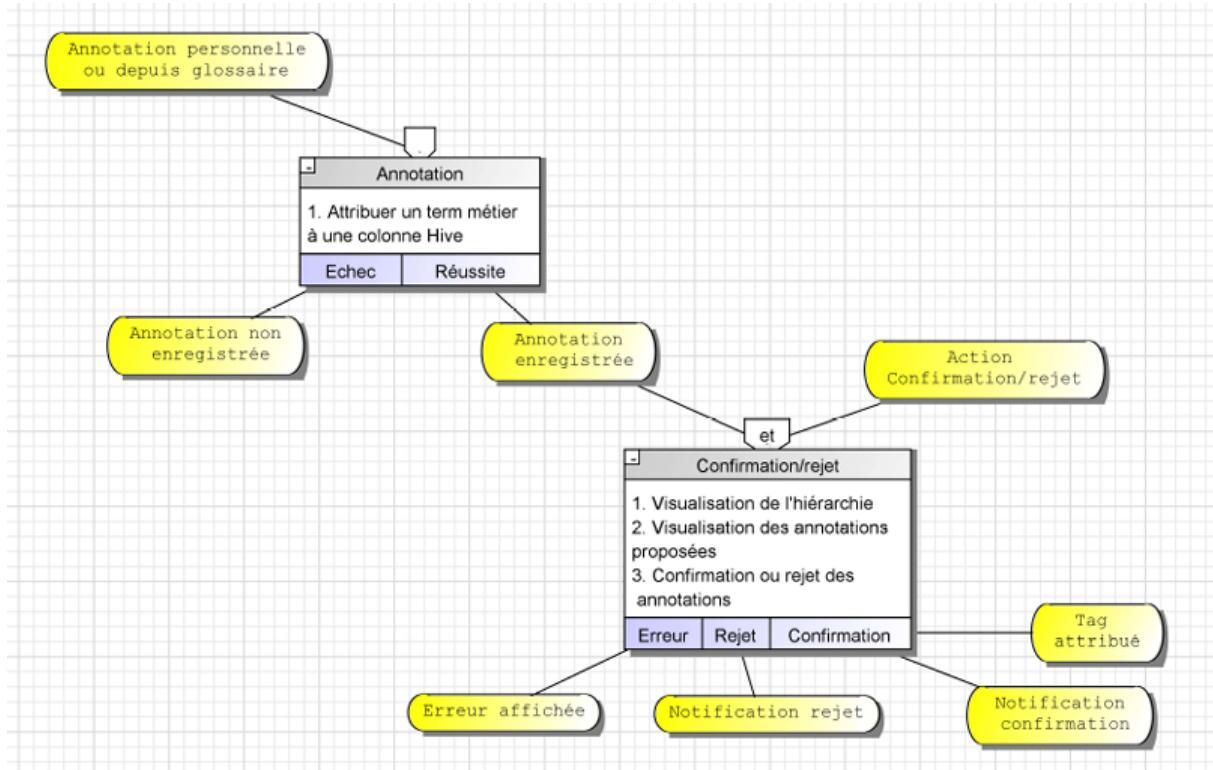


FIGURE 2.8 – MCT du processus 3

#### 2.7.4 MCT du processus 4 : Gouvernance des métadonnées

Ce processus correspond au cœur des fonctions de gouvernance exercées principalement par le Data Steward. Il intègre l'analyse et la gestion des recommandations générées automatiquement concernant la qualité des métadonnées, la classification des entités (sensibles, publiques...), la description, les relations, les périodes de rétention, et les détections de doublons ou de similitudes.

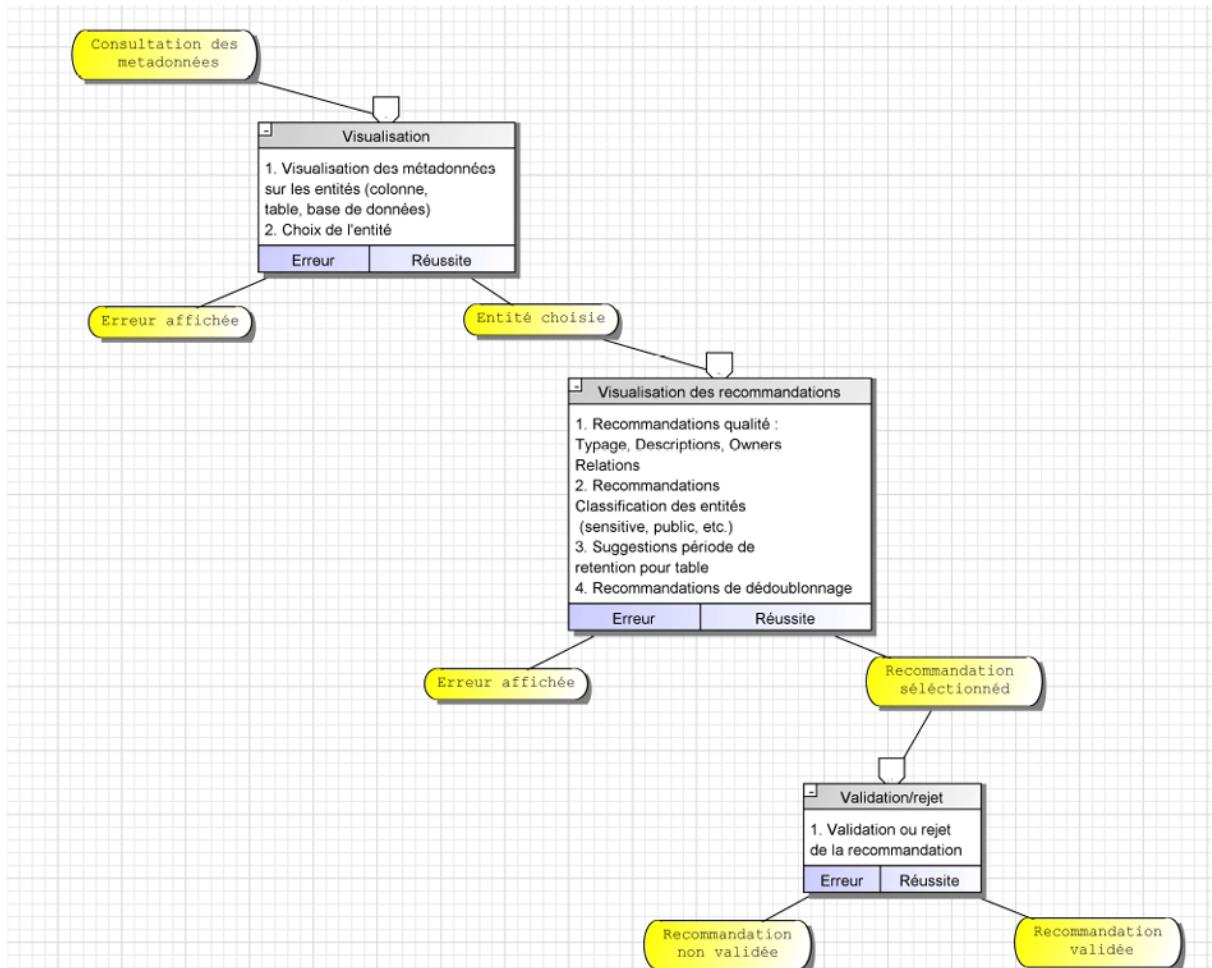


FIGURE 2.9 – MCT du processus 4

### 2.7.5 MCT du processus 5 : Journalisation

Le processus Journalisation intègre à la fois une journalisation automatique des actions des utilisateurs et une interface de visualisation/exploitation des historiques pour les utilisateurs habilités (principalement les administrateurs).

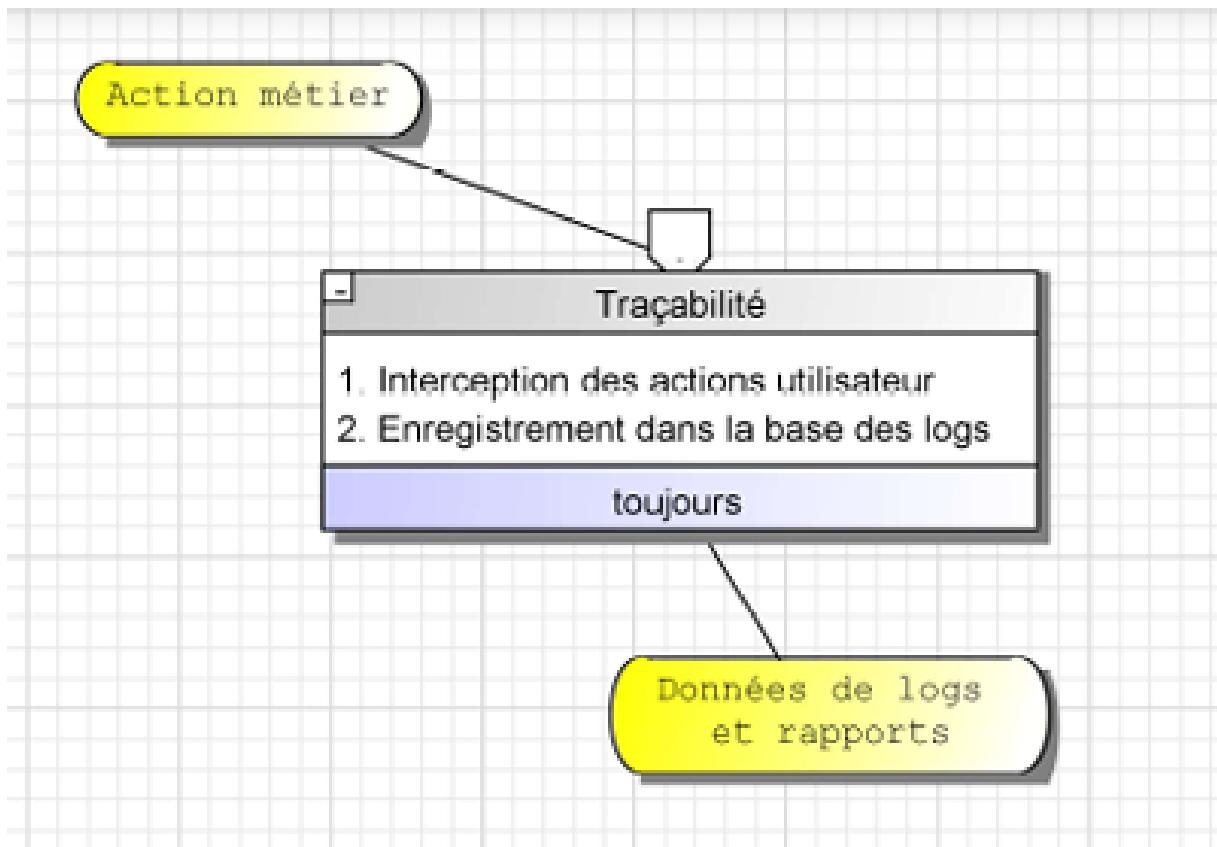


FIGURE 2.10 – MCT du processus 5

## 2.8 Modèle Opérationnel de Traitement (MOT)

Le modèle opérationnel de traitement (MOT) décrit les opérations détaillées nécessaires à l'exécution des processus métiers au sein du système. Ce diagramme met en évidence les flux de données entre les différents acteurs et entités, ainsi que les interactions spécifiques requises pour garantir le bon fonctionnement des traitements. Il constitue une représentation précise des mécanismes d'exécution des tâches dans le cadre des besoins définis.

### 2.8.1 MOT du processus 1 : Initialisation & Gestion des équipes

Le diagramme MOT associé décrit les objets manipulés dans ce processus, notamment les Utilisateurs, Comptes, Domaines de données, et Équipes. Les relations telles que l'affiliation d'un utilisateur à un domaine ou la génération de documents de connexion sont représentées, permettant de visualiser les dépendances entre les entités et les actions de l'administrateur.

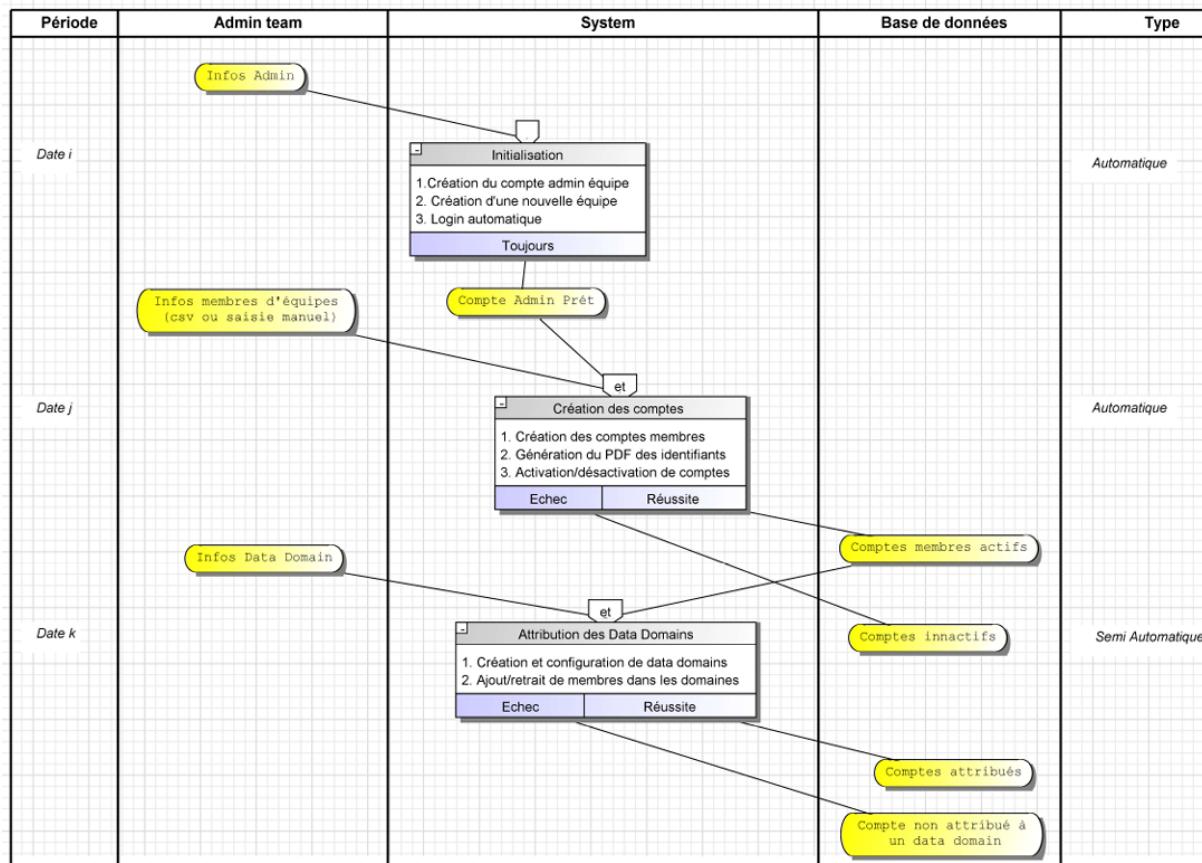


FIGURE 2.11 – MOT du processus 1

## 2.8.2 MOT du processus 2 : Ingestion & Préparation des données

Le diagramme MOT décrit ici les objets métiers essentiels tels que Fichier CSV, Jeu de données, Colonnes, Transformations, et Base de données (Hive/HBase). Les associations illustrent les transformations successives sur les données, et la manière dont elles sont liées aux sources et aux traitements appliqués.

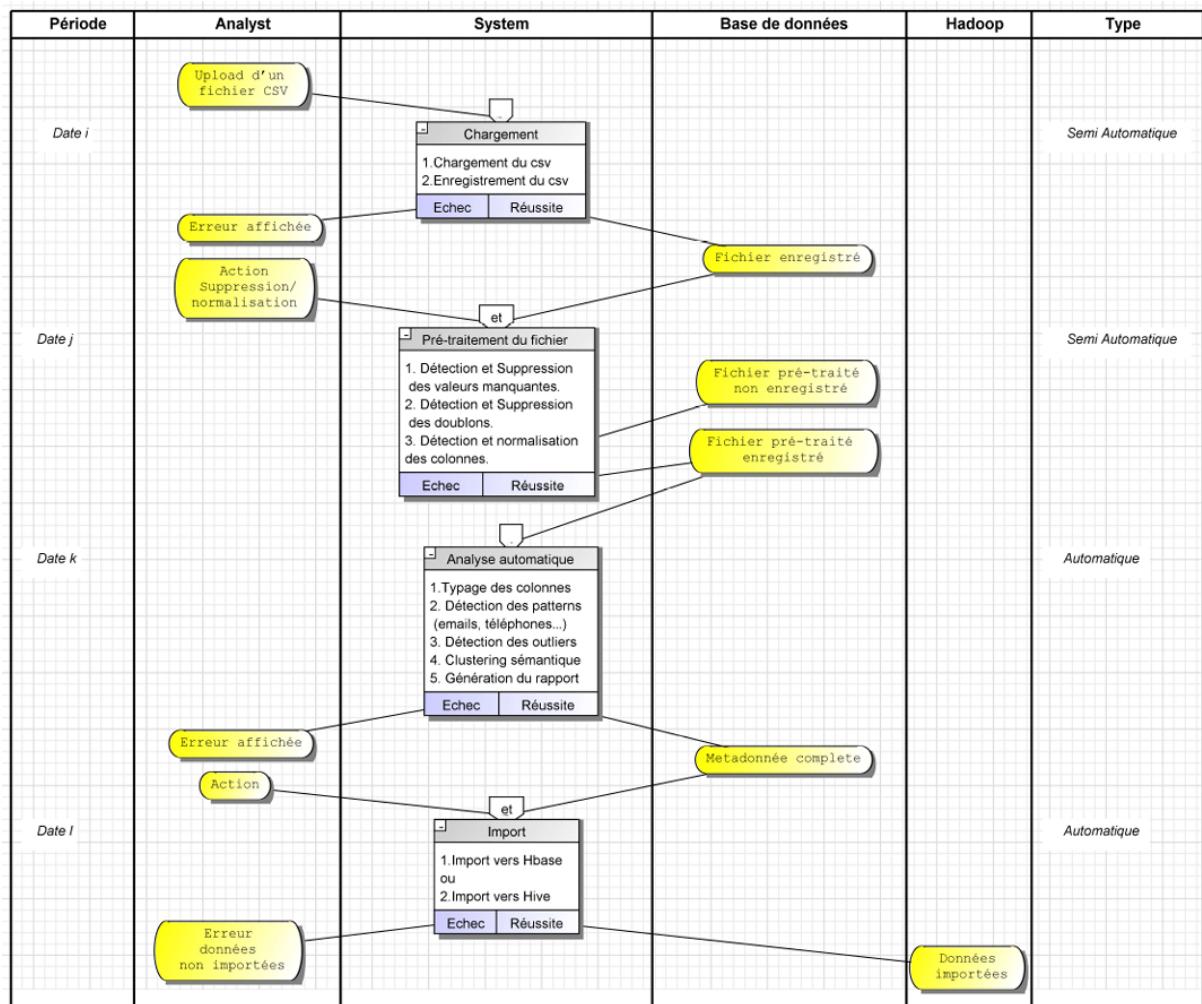


FIGURE 2.12 – MOT du processus 2

### 2.8.3 MOT du processus 3 : Annotation et validation

Le diagramme MOT associé représente les entités Annotation, Colonne, Terme métier, Utilisateur (analystes/stewards) et leur cycle de validation. Il explicite les liens entre annotations proposées et décisions prises, assurant la traçabilité et la qualité des enrichissements sémantiques.

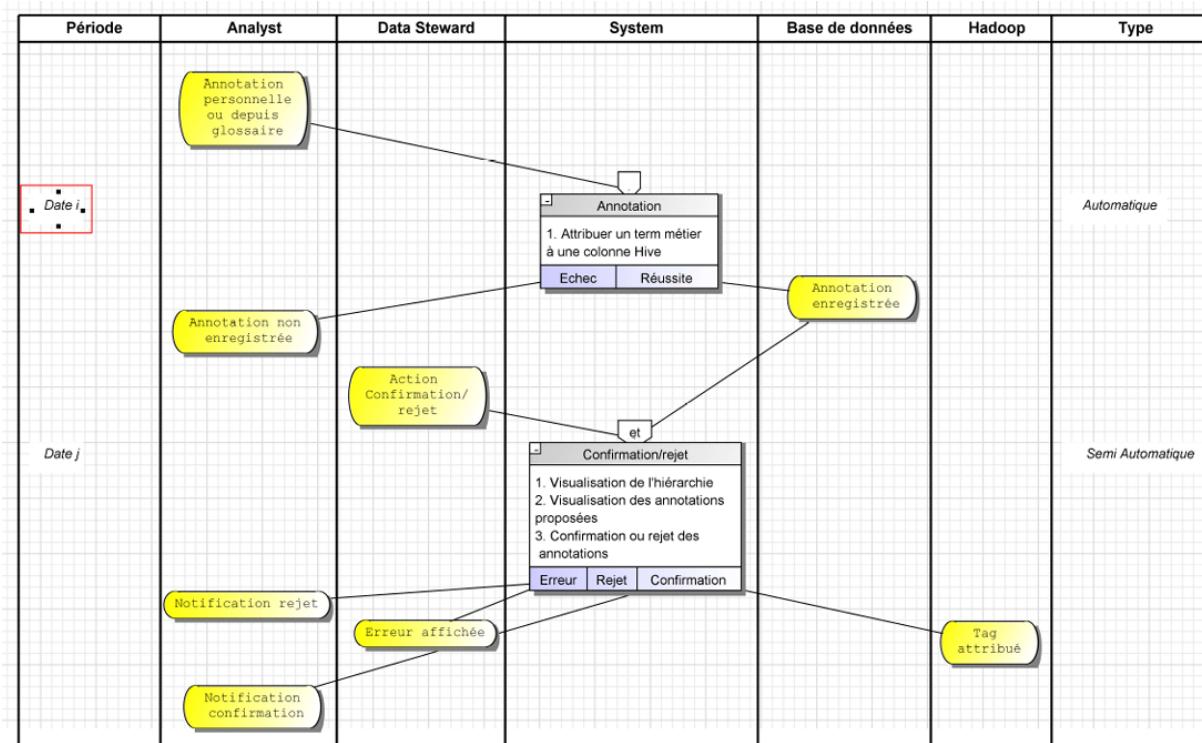


FIGURE 2.13 – MOT du processus 3

## 2.8.4 MOT du processus 4 : Gouvernance des métadonnées

Le MOT correspondant expose les objets structurants de la gouvernance : Entité, Recommandation, Classification, Période de rétention, Relation, Datalineage, etc. Il permet de visualiser comment chaque recommandation est reliée à une entité métier, et comment elle évolue selon les actions du steward.

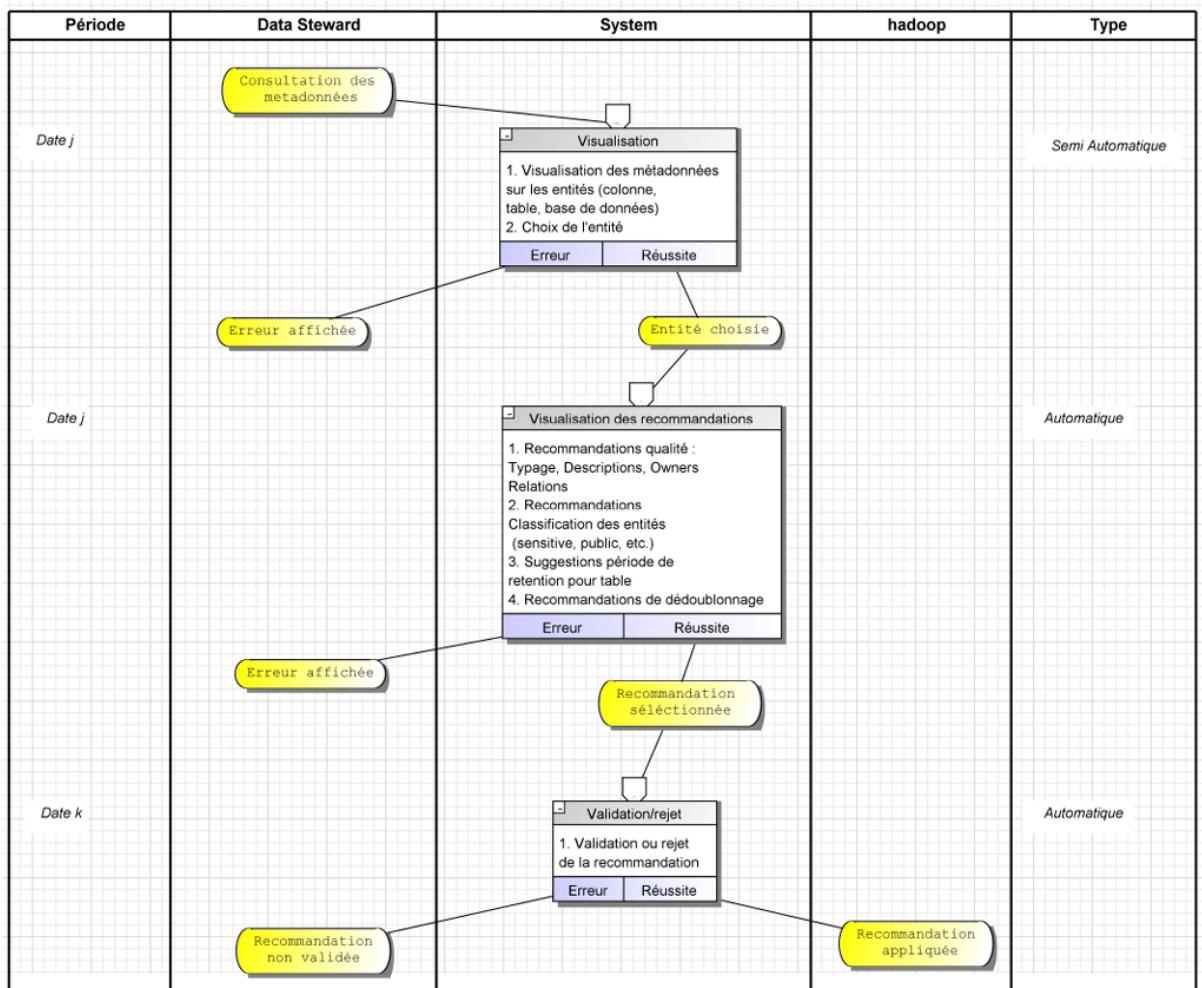


FIGURE 2.14 – MOT du processus 4

### 2.8.5 MOT du processus 5 : Journalisation & suivi des activités

Le diagramme MOT reflète la structure des Logs, Rapports d'activité, Utilisateurs, et les interactions qui les lient. Il permet de suivre l'origine des événements, leur typologie (connexion, annotation, ingestion...), et leur usage pour la génération de synthèses d'activité.

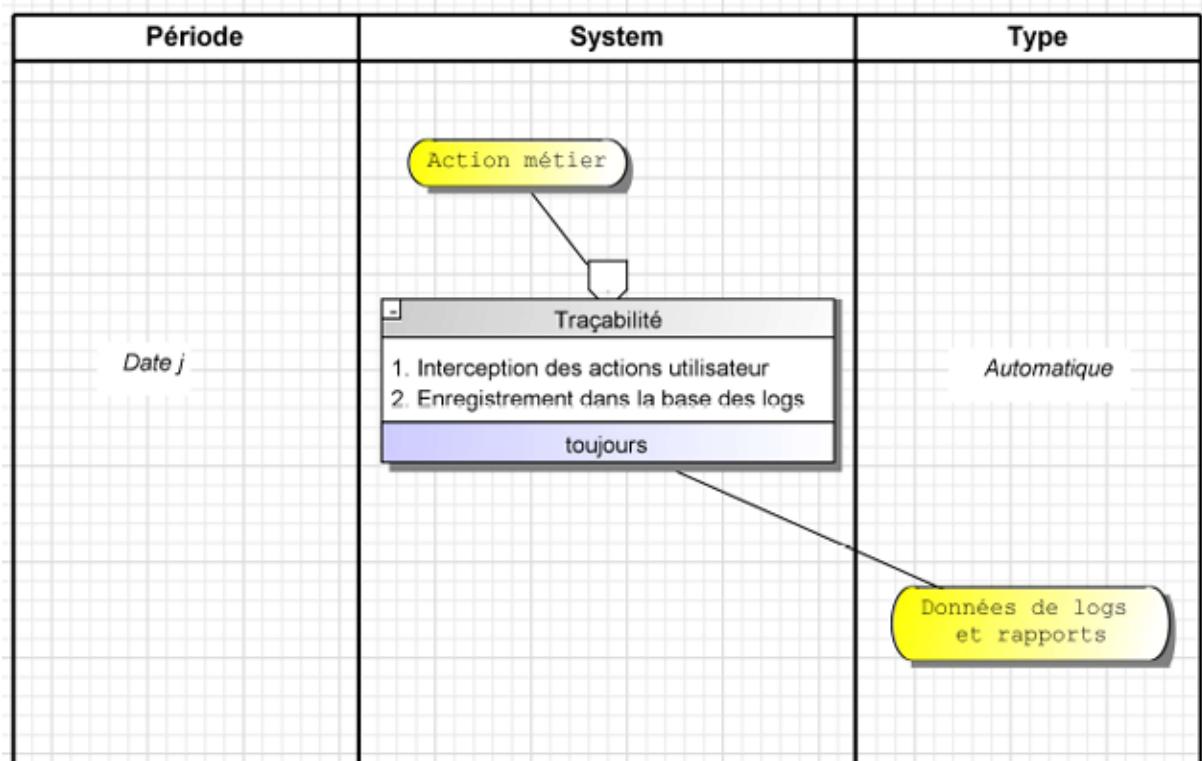


FIGURE 2.15 – MOT du processus 5

Dans ce chapitre, nous avons présenté une analyse approfondie des besoins fonctionnels et non fonctionnels du système, ce qui nous a permis d'identifier clairement les attentes et contraintes du projet.

À travers le dictionnaire de données, nous avons défini précisément l'ensemble des informations manipulées par l'application. Les modèles conceptuel (MCD) et logique (MLD) des données ont permis de structurer ces informations et de poser les bases solides pour leur organisation au sein de la base de données.

Ensuite, l'étude des processus métier via les modèles conceptuels des traitements (MCT) et les modèles organisationnels des traitements (MOT) a contribué à une compréhension détaillée du fonctionnement global du système et des interactions entre les différents acteurs.

Cette phase d'analyse et de conception constitue ainsi un socle fondamental qui guidera la phase suivante de réalisation et de mise en œuvre, en assurant cohérence, efficacité et traçabilité dans le développement de la solution proposée.

# Chapitre 3

## Technologies Utilisées

Ce projet repose sur une architecture moderne combinant des outils robustes pour le traitement, la visualisation et la gouvernance des métadonnées. Cette section présente les technologies adoptées, classées par couche fonctionnelle.

### 3.1 Couche Backend

#### Django



Django est le framework web utilisé pour développer le backend du système. Il facilite la création rapide d'applications sécurisées et maintenables en Python.

#### Python



Le langage Python est utilisé pour le prétraitement des données (nettoyage, détection de patterns, clustering) et pour interagir avec Apache Atlas à l'aide de bibliothèques spécifiques.

#### PyApacheAtlas

Cette bibliothèque Python permet d'interagir avec l'API d'Apache Atlas pour la récupération, la manipulation et l'enrichissement des métadonnées.

## 3.2 Couche Frontend

### Next.js



Next.js est un framework React utilisé pour développer l'interface utilisateur. Il offre un rendu côté serveur (SSR) performant et une navigation fluide.

## 3.3 Base de Données

### PostgreSQL



PostgreSQL est utilisé comme base de données relationnelle principale pour stocker les utilisateurs, les annotations, les historiques et les journaux d'activité.

## 3.4 Sources de Métadonnées et Stockage Big Data

### Apache Atlas



Apache Atlas est le système de gestion des métadonnées dans l'écosystème Hadoop. Il permet de centraliser les informations de gouvernance et les glossaires métiers.

### Hive et HBase



Hive et HBase sont les bases de données Big Data dans lesquelles les données préparées sont stockées. Des scripts Python ont été développés pour automatiser leur chargement à partir des fichiers CSV analysés.

## 3.5 Modèles de Recommandation

### KMeans & PCA

Des algorithmes de machine learning ont été intégrés pour générer des recommandations sur la qualité des métadonnées. Le clustering (KMeans) et la réduction de dimension (PCA) permettent d'identifier les similarités entre colonnes et d'enrichir automatiquement les annotations.

### LLM (Large Language Model)

Un modèle LLM (comme un Deepseek adapté) est utilisé pour renforcer les suggestions sémantiques et proposer des recommandations intelligentes sur les termes métier et la classification des données.

# Chapitre 4

## Réalisation

### 4.1 Introduction

Cette section présente la concrétisation du système à travers l'interface et les fonctionnalités développées. Elle illustre, à l'aide de captures d'écran et d'explications ciblées, les différentes interfaces utilisateur ainsi que les interactions spécifiques à chaque rôle du système. Cette approche permet de démontrer l'adéquation entre la conception théorique (réalisée via les modèles Merise) et la mise en œuvre effective de la solution.

Chaque sous-section est dédiée à un rôle utilisateur (Administrateur, Data Analyst, Data Steward) et met en lumière les principales tâches et fonctionnalités accessibles à ce profil, conformément aux exigences fonctionnelles identifiées. L'objectif est de montrer comment chaque rôle interagit avec le système, quelles actions il peut réaliser, et comment l'interface facilite ces interactions.

### 4.2 Interfaces Générales : Authentification et Accueil

Avant d'accéder aux interfaces spécifiques à chaque rôle, les utilisateurs interagissent avec un ensemble d'écrans communs. Ces interfaces incluent le formulaire de connexion, la page d'inscription (création de compte) et la page d'accueil (landing page) qui oriente l'utilisateur selon son profil. Elles assurent une première interaction claire et sécurisée avec le système, tout en offrant une navigation intuitive vers les espaces dédiés. Cette section présente ces écrans génériques, fondamentaux pour l'accès et l'orientation dans l'application.

### 4.2.1 Page d'accueil

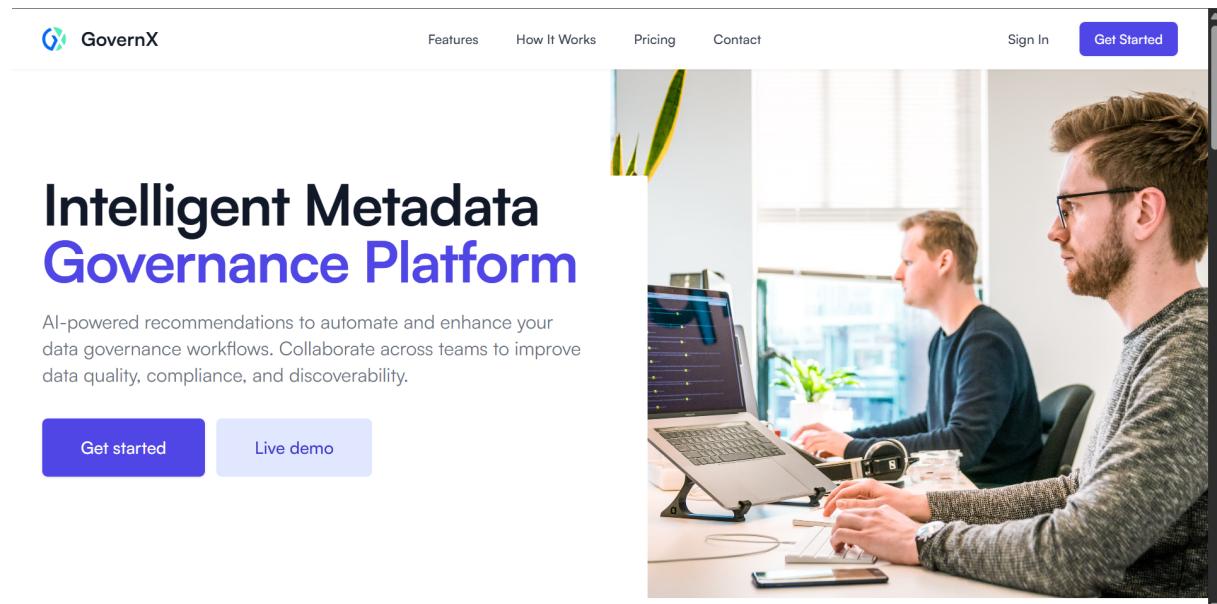


FIGURE 4.1 – Présentation du système

The screenshot shows the 'Features' section of the GovernorX website. The title 'AI-Powered Metadata Governance' is displayed in large, bold, black text. Below the title, a subtext reads: 'Automate and enhance your data governance with intelligent recommendations'. The page lists six features, each with an icon and a brief description:

- Complete Missing Fields**: AI suggests values for missing metadata like descriptions, owners, and business terms to improve data documentation.
- Standardize Naming & Types**: Automatically detect and suggest consistent naming conventions and data types across your data assets.
- Business Tagging**: Intelligent suggestions for business tags like PII, Finance, HR to improve data classification and discovery.
- Entity Relationships**: Discover and suggest relationships between data entities to build a connected data landscape.
- Retention Policies**: Automatically suggest appropriate retention periods based on data classification and regulations.
- Data Classification**: Classify data as public, sensitive, or critical with AI assistance to improve security and compliance.

FIGURE 4.2 – Les Fonctionnalités

The screenshot shows the GovernorX website's "Collaboration" section. At the top, there are navigation links: Features, How It Works, Pricing, Contact, Sign In, and a prominent blue "Get Started" button. Below this, the title "Designed for Your Team" is displayed in large, bold letters, with the subtitle "Empower every role in your data governance workflow" underneath. Three main sections are shown in boxes:

- Data Stewards**
  - Review and approve AI recommendations
  - Enforce data standards and policies
  - Manage business glossary terms
- Data Analysts**
  - Upload and analyze CSV data
  - Clean and preprocess data
  - Annotate columns with business terms
- Team Admins**
  - Manage team members and roles
  - Monitor governance activity
  - Configure governance policies

FIGURE 4.3 – Vue d’ensemble sur les responsabilités de chaque rôle

The screenshot shows the GovernorX website's "How GovernorX Works" section. At the top, there are navigation links: Features, How It Works, Pricing, Contact, Sign In, and a blue "Get Started" button. Below this, the title "How GovernorX Works" is displayed in large, bold letters, with the subtitle "Transform your metadata governance in three simple steps" underneath. Three numbered steps are listed:

- 1 Connect Your Data Sources**  
Integrate with your existing data platforms like Hive, databases, or upload CSV files. GovernorX automatically scans your metadata.
- 2 Review AI Recommendations**  
Our AI analyzes your metadata and suggests improvements for completeness, standardization, classification, and more.
- 3 Collaborate & Implement**  
Your team reviews, approves, or modifies suggestions. Changes are applied to your metadata with full audit trails.

At the bottom of the page, a large blue banner features the text "Ready to transform your data governance?" and the subtext "Start improving your metadata quality today with AI-powered recommendations."

FIGURE 4.4 – Comment ça marche le système

## 4.2.2 Page Registration

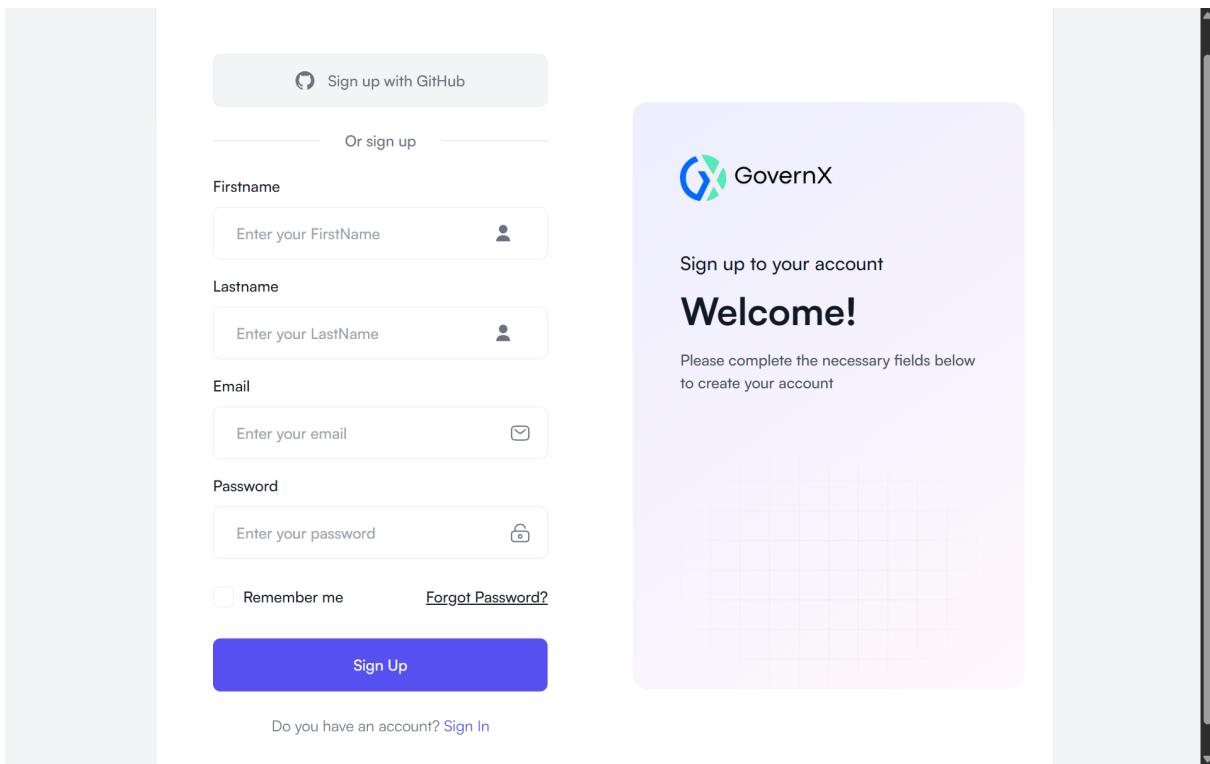


FIGURE 4.5 – Interface SignUp

### 4.2.3 Page Login

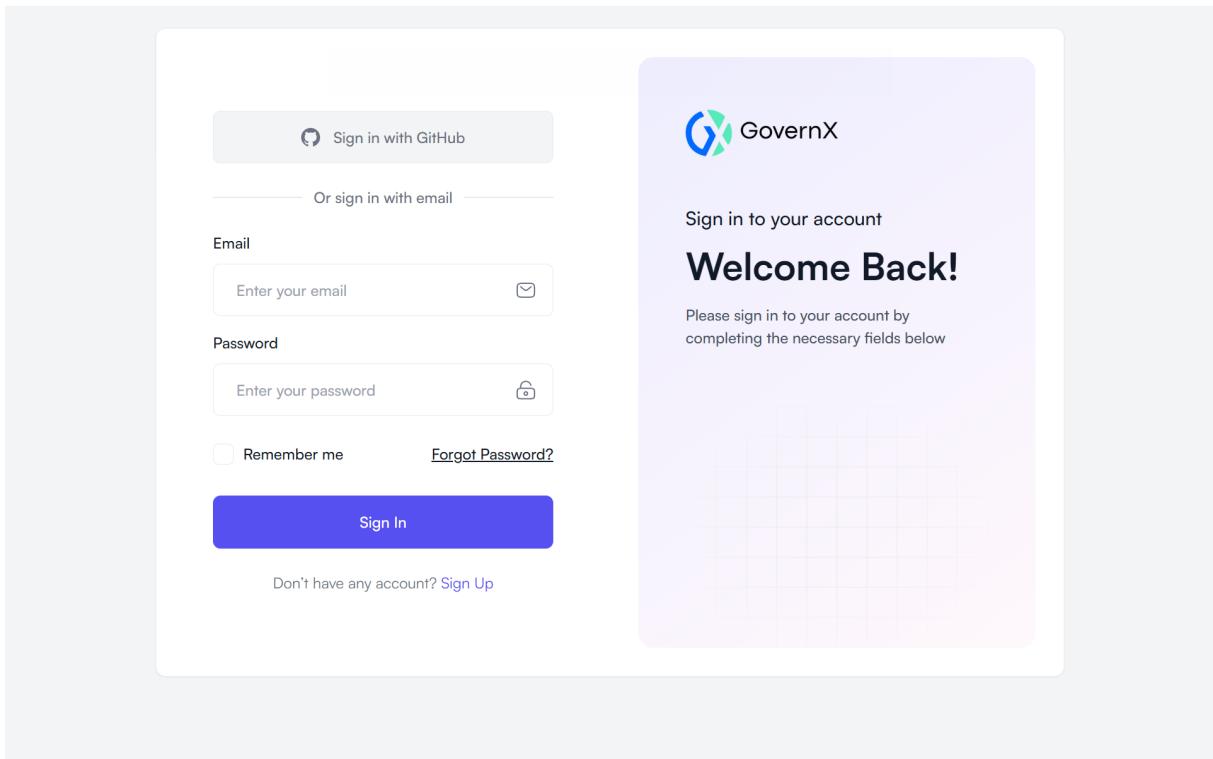


FIGURE 4.6 – Interface SignIn

#### 4.2.4 Page Gestion des identifiants

The screenshot shows a browser window with the URL `localhost:3000/role/admin/profile`. The page is titled "GovernX" and subtitle "Your Metadata Governance Solution". On the left, there is a sidebar with "Admin MENU" containing links for "Profile", "Uploads", "Ajouter des utilisateurs", and "Statistics". The main content area is titled "Manage Credentials". It has two sections: "Atlas Credentials" and "Ranger Credentials". In the "Atlas Credentials" section, the "Username" field contains "atlas". In the "Ranger Credentials" section, the "Username" field contains "ranger". There is also a "Password" field in the Ranger section, which is currently empty.

FIGURE 4.7 – Les identifiants pour les services Atlas et Ranger

### 4.3 Rôle : Team Admin

L'administrateur occupe un rôle central dans la gestion des utilisateurs et de l'organisation des équipes. Il est responsable de la création et de l'administration des comptes, de l'affectation des membres aux domaines de données, ainsi que du suivi de l'activité de l'équipe. Cette section présente l'interface dédiée à l'administrateur, illustrant les différentes fonctionnalités mises à sa disposition pour assurer une gestion fluide et sécurisée du système.

### 4.3.1 Crédation des membres

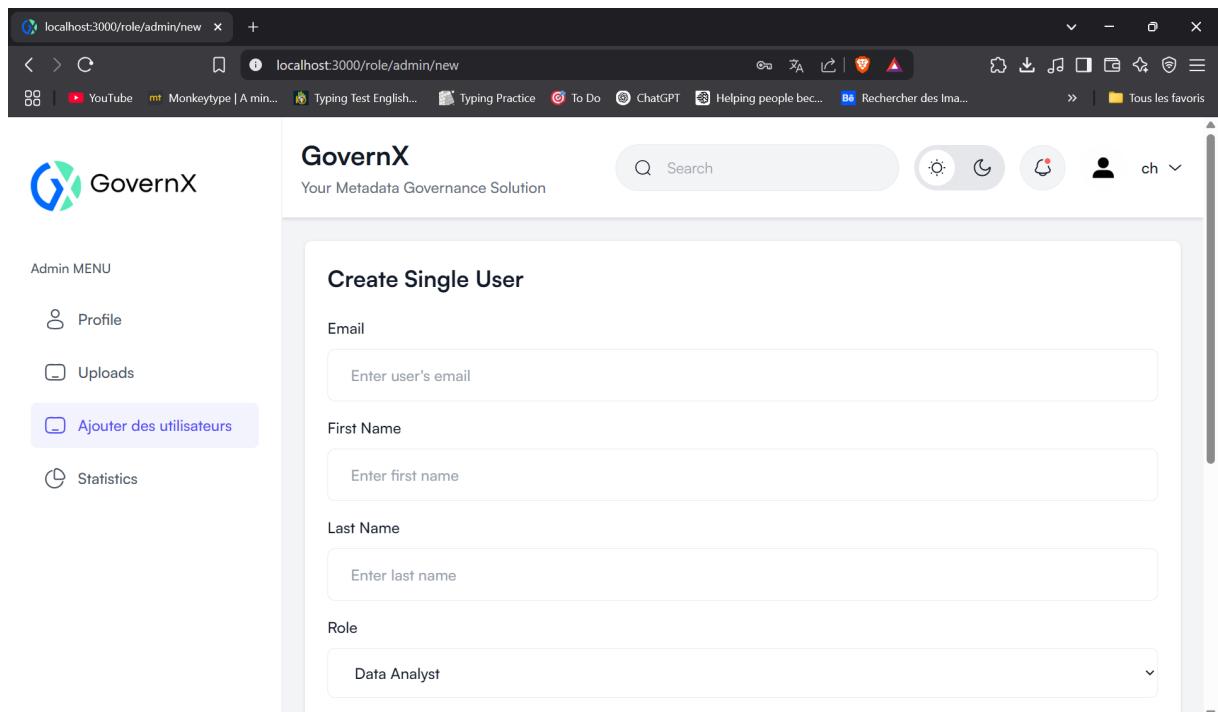


FIGURE 4.8 – Ajout d'un seul membre

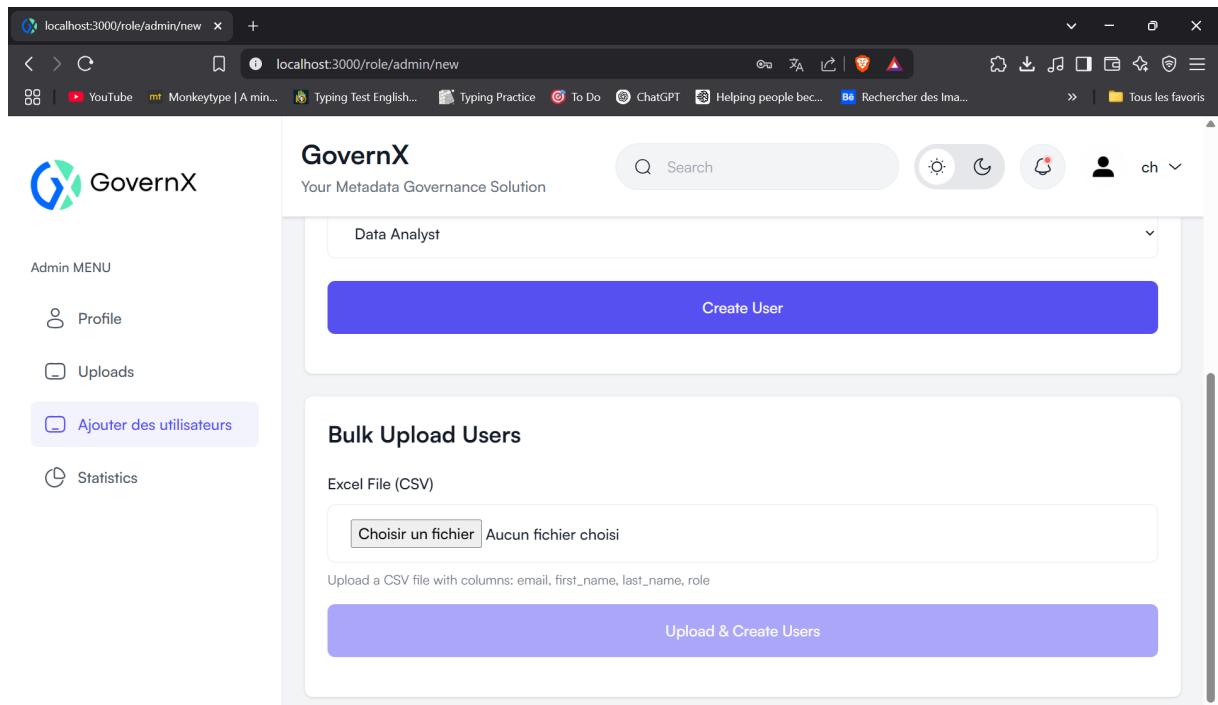


FIGURE 4.9 – Ajout de plusieurs membres en utilisant un csv

The screenshot shows the GovernX web application interface. On the left, there's a sidebar with 'Admin MENU' containing 'Profile', 'Uploads', 'Ajouter des utilisateurs' (which is highlighted in blue), and 'Statistics'. The main content area has a title 'Excel File (CSV)' with a button 'Choisir un fichier' and a message 'Aucun fichier choisi'. Below it is a note: 'Upload a CSV file with columns: email, first\_name, last\_name, role'. A large purple button says 'Upload & Create Users'. A green success message 'User created successfully!' is displayed. Below this is a section titled 'Created Users' with a table:

EMAIL	ROLE	TEAM	GENERATED PASSWORD
hello1@gmail.com	analyst	raja	yoKVlf7j68

A blue button 'Download as PDF' is located at the top right of the 'Created Users' section. A vertical scrollbar is visible on the right side of the main content area.

FIGURE 4.10 – Affichage des membres créés et génération du pdf contenant leurs identifiants

This screenshot shows a PDF document titled 'created\_users\_2025-04-24T03\_27\_50.075Z.pdf'. The PDF contains a table with the same data as the screenshot above:

EMAIL	ROLE	TEAM	GENERATED PASSWORD
jane.smith@example.com	steward	raja	eihF5V92w
mike.johnson@example.com	analyst	raja	Alkz8qe7fiz
sarah.williams@example.com	steward	raja	pQCT2A6vJB
alex.brown@example.com	analyst	raja	r2ZFqDeL1A

FIGURE 4.11 – Aperçu du pdf généré

### 4.3.2 Gestion des membres

The screenshot shows the GovernorX application interface. On the left, there is a sidebar with the title "GovernX" and "Your Metadata Governance Solution". The sidebar includes an "Admin MENU" section with links: "Profile" (selected), "Users Management" (highlighted in blue), "Ajouter des utilisateurs", and "Statistics".

The main content area has two main sections:

- Team Management**: This section displays a list titled "Your Team" with one item: "raja". Below the list are two buttons: "Activate All Users" (green) and "Deactivate All Users" (red).
- Data Domains Management**: This section has a sub-section titled "Create New Data Domain". It contains fields for "Name" (with placeholder "Enter domain name") and "Description" (with placeholder "Enter domain description"). A "Create Data Domain" button is located at the bottom of this section.

FIGURE 4.12 – Activation ou Désactivation de tous les membres

The screenshot shows the GovernorX application interface. The sidebar is identical to Figure 4.12, with "Users Management" selected.

The main content area has two main sections:

- Create New Data Domain**: This section contains fields for "Name" (placeholder "Enter domain name") and "Description" (placeholder "Enter domain description"). A "Create Data Domain" button is located at the bottom of this section.
- Existing Data Domains**: This section displays a table with one row of data:

NAME	DESCRIPTION	ACTIONS
Product Data	test	<a href="#">Edit</a> <a href="#">Delete</a>

FIGURE 4.13 – Crédation M&aj Suppression des Data Domains

The screenshot shows the 'Team Members' section of the GovernX interface. On the left, there's a sidebar with 'Admin MENU' containing links for Profile, Users Management (which is highlighted), Ajouter des utilisateurs, and Statistics. The main area has a header 'GovernX Your Metadata Governance Solution' with a search bar and user icons. Below is a table titled 'Team Members' with columns: NAME, EMAIL, ROLE, STATUS, and ACTIONS. The table lists several users with their status (e.g., Inactive, Active) and actions (e.g., Activate, Deactivate, Manage Domains).

NAME	EMAIL	ROLE	STATUS	ACTIONS
Alex Brown	alex.brown@example.com	analyst	Inactive	<button>Activate</button> <button>Manage Domains</button>
Alex Brown	alex.brwn@example.com	analyst	Active	<button>Deactivate</button> <button>Manage Domains</button>
etudiant3 3	etudiant3@gmail.com	analyst	Active	<button>Deactivate</button> <button>Manage Domains</button>
hel hello	hel	analyst	Active	<button>Deactivate</button> <button>Manage Domains</button>
hello hell	hello	steward	Active	<button>Deactivate</button> <button>Manage Domains</button>
hello 1	hello1@gmail.com	analyst	Active	<button>Deactivate</button> <button>Manage Domains</button>
Jane Smith	jane.sith@example.com	steward	Active	<button>Deactivate</button> <button>Manage Domains</button>

FIGURE 4.14 – Gestion individuelle des membres

The screenshot shows a modal dialog titled 'Manage Data Domains for Alex Brown'. It has sections for 'Current Data Domains' (No data domains assigned), 'Available Data Domains' (listing 'Product Data' and 'Customer Data' with 'Add' buttons), and a 'Bulk Update' section with a list of domains ('Product Data', 'Customer Data'). A note at the bottom says 'Hold Ctrl/Cmd to select multiple domains'.

FIGURE 4.15 – Attribution d'un Data Domain

### 4.3.3 Page des statistiques

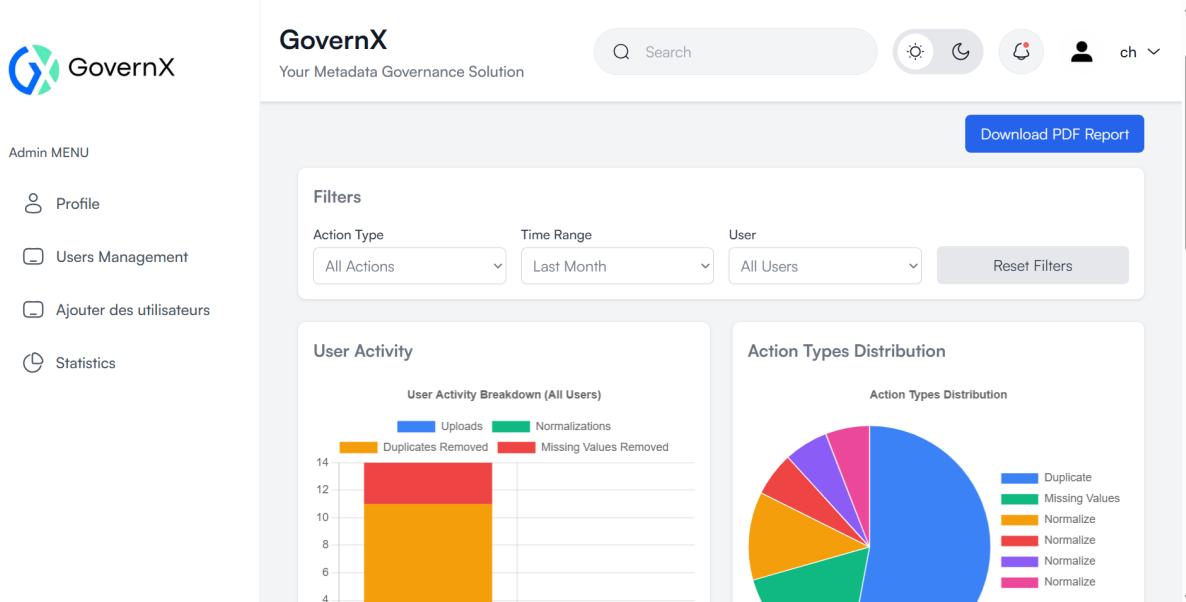


FIGURE 4.16 – Diagrammes sur l'activité des membres

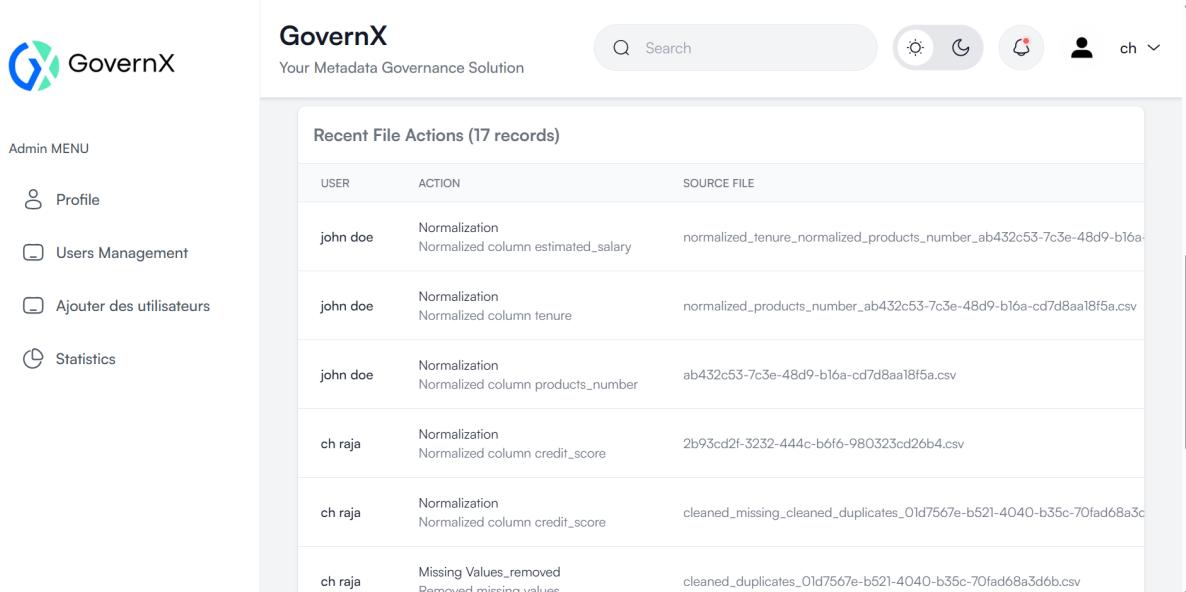


FIGURE 4.17 – Vue sur le Pré-traitement des fichiers exécuté par les membres

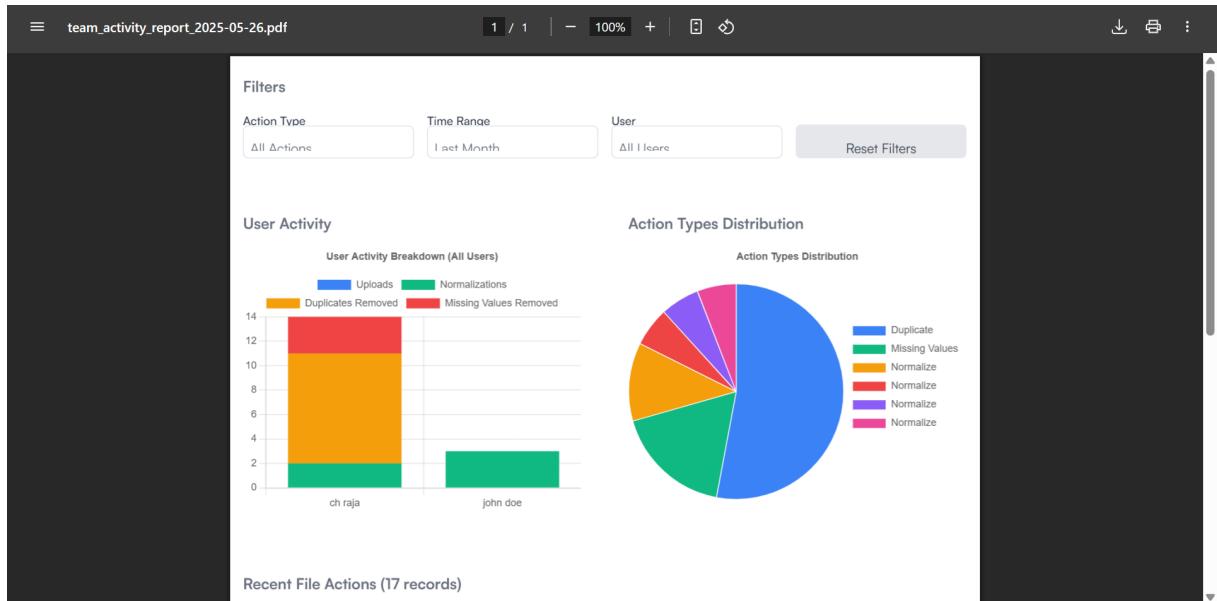


FIGURE 4.18 – Aperçu du pdf généré

## 4.4 Rôle : Data Analyst

Le data analyst est en charge de l'ingestion, de la préparation et de l'annotation des données. À travers son interface, il peut charger des fichiers, appliquer des traitements automatiques (nettoyage, typage, normalisation, détection de patterns), annoter les colonnes avec des termes métier et suivre l'évolution de ses actions. Cette section met en avant les écrans développés pour accompagner efficacement l'analyste dans l'enrichissement des métadonnées.

#### 4.4.1 Page Analyse

The screenshot shows the GovernorX interface for CSV analysis. On the left, a sidebar titled "DATA ANALYST MENU" includes links for Profile, Upload (which is highlighted in blue), Annotation, and Statistics. The main area is titled "CSV Analysis" and shows a file named "Bank Customer Churn Prediction - Copie.csv" uploaded. It features two buttons: "Analyse CSV" (blue) and "Download PDF Report" (green). Below this is a "Metadata Summary" section. The "Columns" list includes: customer\_id, credit\_score, country, gender, age, tenure, balance, products\_number, credit\_card, active\_member, estimated\_salary, and churn. The "Issues" section indicates 1 duplicate and 2 missing values across 2 columns.

FIGURE 4.19 – Chargement du fichier et déclenchement de l'analyse

This screenshot shows the GovernorX interface after the analysis has been run. The sidebar and top navigation are identical to Figure 4.19. The main content is the "Metadata Summary" section, which now displays a more comprehensive list of columns: customer\_id, credit\_score, country, gender, age, tenure, balance, products\_number, credit\_card, active\_member, estimated\_salary, and churn. The "Issues" section remains the same, showing 1 duplicate and 2 missing values across 2 columns. At the bottom of the summary section are three buttons: "Download Original File" (blue), "Remove Duplicates" (red), and "Remove Missing Values" (yellow).

FIGURE 4.20 – Résumé des métadonnées

The figure consists of two side-by-side screenshots of the GovernX web application. Both screenshots show the 'Metadata Summary' page with a list of columns on the left and an 'Issues' section on the right.

**Left Screenshot (Figure 4.21):** A modal dialog titled 'Confirm Remove Duplicates' is displayed. It asks, 'Are you sure you want to remove 1 duplicate records?' with 'Cancel' and 'Remove' buttons. Below the modal, there are three buttons: 'Download Original File', 'Remove Duplicates', and 'Remove Missing Values'.

**Right Screenshot (Figure 4.22):** A modal dialog titled 'Confirm Remove Missing Values' is displayed. It asks, 'Are you sure you want to remove records with missing values?' with 'Cancel' and 'Remove' buttons. Below the modal, there are two buttons: 'Download Processed File' and 'Remove Missing Values'.

FIGURE 4.21 – élimination des valeurs manquantes

FIGURE 4.22 – élimination des doublons

A screenshot of the GovernX interface showing the 'Metadata Summary' page. The left sidebar shows the 'DATA ANALYST MENU' with 'Upload' selected. The main area displays the 'Metadata Summary' with a list of columns and an 'Issues' section indicating 0 duplicates and 0 missing values.

**Columns:**

- customer\_id
- credit\_score
- country
- gender
- age
- tenure
- balance
- products\_number
- credit\_card
- active\_member
- estimated\_salary
- churn

**Issues:**

Duplicates: 0  
Missing values: 0 columns

At the bottom is a blue button labeled 'Download Processed File'.

FIGURE 4.23 – Analyse du fichier résultant

A screenshot of the GovernX interface showing the 'Data Types Comparison' page. The left sidebar shows the 'DATA ANALYST MENU' with 'Upload' selected. The main area displays a table comparing current and suggested data types for various columns.

COLUMN	CURRENT TYPE	SUGGESTED TYPE
customer_id	int64	int64
credit_score	int64	int64
country	object	string
gender	object	string
age	int64	int64
tenure	int64	int64
balance	float64	float64
products_number	int64	int64

FIGURE 4.24 – Types de données détectés et suggérés

The screenshot shows the GovernX interface with the title "GovernX Your Metadata Governance Solution". On the left, there is a "DATA ANALYST MENU" with options: Profile, Upload (highlighted in blue), Annotation, and Statistics. The main content area is titled "Normalization Suggestions". It contains a table with columns "COLUMN", "SUGGESTION", and "ACTION". The table rows are:

COLUMN	SUGGESTION	ACTION
customer_id	no normalization needed	
credit_score	z-score normalization suggested	Normalized
age	z-score normalization suggested	Normalize
tenure	z-score normalization suggested	Normalize
balance	z-score normalization suggested	Normalize
products_number	z-score normalization suggested	Normalize
credit_card	z-score normalization suggested	Normalize
active_member	z-score normalization suggested	Normalize

FIGURE 4.25 – Normalisation des colonnes

The screenshot shows the GovernX interface with the title "GovernX Your Metadata Governance Solution". On the left, there is a "DATA ANALYST MENU" with options: Profile, Upload (highlighted in blue), Annotation, and Statistics. The main content area is titled "Pattern Detection". It contains a table with columns "COLUMN" and "DETECTED PATTERNS". The table rows are:

COLUMN	DETECTED PATTERNS
customer_id	phone
credit_score	none
country	none
gender	none
age	none
tenure	none
balance	none
products_number	none

FIGURE 4.26 – Détection des patterns

The screenshot shows the GovernX interface with the title "GovernX Your Metadata Governance Solution". On the left, there is a "DATA ANALYST MENU" with options: Profile, Upload (highlighted in blue), Annotation, and Statistics. The main content area has two sections: "Semantic Column Clusters" and "Outliers Count".

**Semantic Column Clusters:**

- Cluster: france\_germany\_spain
  - country
- Cluster: male\_female\_france
  - gender

**Outliers Count:**

A bar chart titled "Outliers Count" showing the number of outliers for different columns. The y-axis ranges from 0 to 2500. The x-axis categories are credit\_score, age, products\_number, and churn. The bars show values approximately 300, 100, 100, and 2000 respectively.

FIGURE 4.27 – Clustering sémantique

FIGURE 4.28 – Détection des valeurs abérantes et métadonnées complètes

#### 4.4.2 Page d'annotation

The screenshot shows the Data Annotation Tool interface. On the left, a sidebar titled "DATA ANALYST MENU" contains links for Profile, Upload, Annotation (which is highlighted in blue), and Statistics. The main area is titled "Data Annotation Tool" with the subtitle "Annotate your data assets with business glossary terms". It features two sections: "Databases" and "Personal Glossaries". The "Databases" section lists "pfa" with the email "pfa@Sandbox". The "Personal Glossaries" section has a "Create New Glossary" form with fields for "Name" and "Description", and a "Create Glossary" button.

FIGURE 4.29 – Affichage des bases de données existantes

This screenshot shows the "Data Annotation Tool" interface after selecting the database "pfa". The sidebar remains the same. The main area now displays "Tables in pfa" with entries for "test2" and "test". The "test" table is selected, indicated by a blue border around its row.

FIGURE 4.30 – Sélection de la db et affichage des tables

This screenshot shows the "Data Annotation Tool" interface after selecting the "test" table. The sidebar remains the same. The main area displays "Columns in test" with columns for "name", "email", and "description", each with their respective details from the "pfa" database.

FIGURE 4.31 – Sélection de la table et affichage des colonnes

This screenshot shows the "Standard Annotation" form for the "name" column of the "test" table. It includes fields for "name" (pfa.test.name@Sandbox), "email" (pfa.test@email@Sandbox), and "description" (pfa.test.description@Sandbox). Below these are sections for "Column Metadata" (with "name", "owner", and "qualifiedName" fields) and "Glossary Term" (with a dropdown menu for selecting a glossary term). A comment field "Annotation Comment" is also present at the bottom.

FIGURE 4.32 – Annotation Standard en utilisant les termes du glossaire atlas

This screenshot shows the "Personal Annotation" form for the "name" column of the "test" table. It includes fields for "name" (pfa.test.name@Sandbox), "email" (pfa.test@email@Sandbox), and "description" (pfa.test.description@Sandbox). Below these are sections for "Column Metadata" (with "name", "owner", and "qualifiedName" fields) and "Personal Glossary Term" (with a dropdown menu for selecting a personal term). A comment field "Annotation Comment" is also present at the bottom.

FIGURE 4.33 – Annotation Personnelle en utilisant les termes du glossaire personnel

The screenshot shows the 'Personal Glossaries' section of a Data Analyst menu. On the left, a sidebar lists 'DATA ANALYST MENU' items: Profile, Upload, **Annotation** (which is selected), and Statistics. The main area is titled 'Personal Glossaries' and contains a 'Create New Glossary' form with fields for 'Name' and 'Description', and a 'Create Glossary' button. Below this is a section titled 'Your Glossaries' containing two entries:

- finance**: nothing special just a test. Includes a 'View Terms' link.
- personal data**: it groups personal data terms. Includes a 'Hide Terms' link.

Each glossary entry has an 'ADD NEW TERM' section with 'Term Name' and 'Description' fields, and a 'Add Term' button.

FIGURE 4.34 – Gestion du glossaire personnel

This screenshot shows the 'Personal Annotations' interface. It features tabs for 'Standard Annotations' and 'Personal Annotations', with 'Personal Annotations' being active. A 'Refresh' button is at the top right. The main area displays a list of annotations:

- pds.test.name(@Sandbox (hive\_column))**: No term associated. Status: PENDING. View button.
- from postman again**: Date: 04/05/2025 19:46:15. View button.

FIGURE 4.35 – Affichage des annotations personnelles

This screenshot shows the 'Standard Annotations' interface. It features tabs for 'Standard Annotations' and 'Personal Annotations', with 'Standard Annotations' being active. A 'Refresh' button is at the top right. The main area displays a list of annotations:

- description (hive\_column)**: Term: orders. Date: 04/05/2025 19:47:02. Status: PENDING. View button.
- name (hive\_column)**: Term: orders. Date: 04/05/2025 19:30:42. Status: PENDING. View button.

FIGURE 4.36 – Affichage des annotations standards

### 4.4.3 Page statistiques

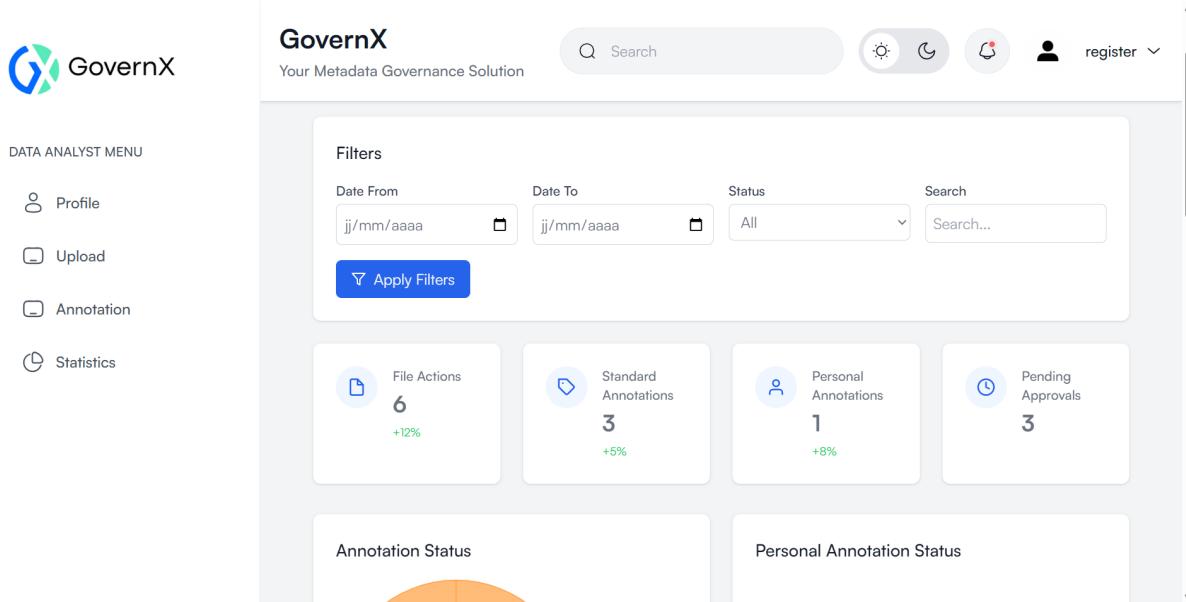


FIGURE 4.37 – Filtrage et vue globale

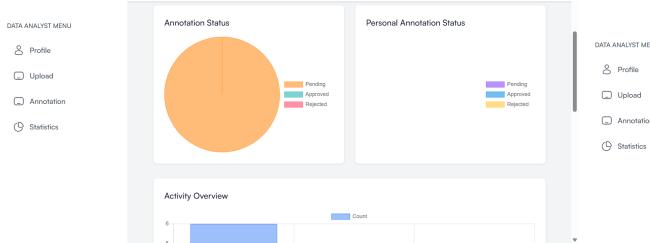


FIGURE 4.38 – Statistiques sur les annotations standards et personnelles



FIGURE 4.39 – Statistiques sur les actions

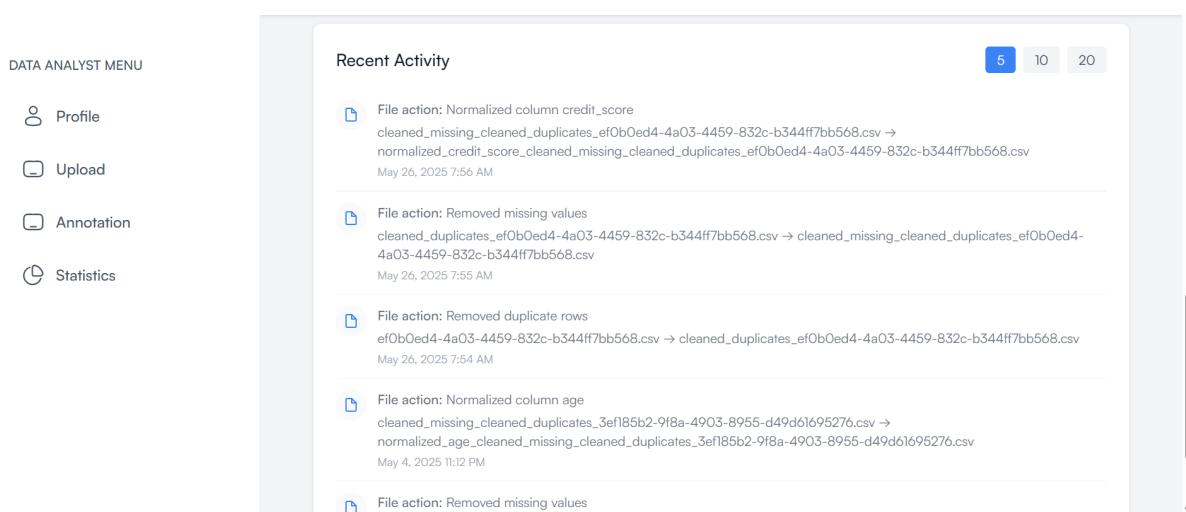


FIGURE 4.40 – Vue détaillée sur les actions récentes

## 4.5 Rôle : Data Steward

Le data steward joue un rôle de contrôle et de validation dans le processus de gouvernance des métadonnées. Il examine les annotations proposées, évalue la qualité des données, confirme ou rejette les suggestions, et supervise la classification et la traçabilité des données. Cette section décrit les fonctionnalités qui lui sont dédiées, en illustrant comment le système soutient son travail d'arbitrage et de gouvernance.

### 4.5.1 Page Qualité des métadonnées

The screenshot shows the 'Metadata Quality Assistant' interface. On the left, the 'DATA STEWARD MENU' includes 'Profile', 'MetaData Quality' (which is selected and highlighted in blue), 'Business Tags', and 'Statistics'. The main area displays 'Databases' with a single entry: 'pfa' under 'pfaSandbox'. A small info icon is visible next to the database name.

FIGURE 4.41 – Sélection de la base de données

The screenshot shows the 'Database: pfa' details page. The 'Basic Information' section includes the qualified name 'pfa@Sandbox', type 'MANAGED\_TABLE', owner 'hive', created by 'hive', updated by 'admin', and creation date '2020-05-02 13:02:14'. The 'Full Members' section contains a JSON object representing the database's metadata, including fields like 'qualifiedName', 'status', 'type', 'owner', 'createdBy', 'updatedBy', 'description', and 'name'.

FIGURE 4.42 – Affichage des métadonnées sur la base de données

The screenshot shows the 'Tables in pfa' page. It lists two tables: 'sales\_data' and 'test2'. 'sales\_data' is described as a 'Table comment des données de vente'. 'test2' is described as 'This is the new description.' Below this, the 'Table: test2' details are shown, including basic information like qualified name 'pfa.test2@Sandbox', type 'MANAGED\_TABLE', and owner 'hive'.

FIGURE 4.43 – Sélection de la table

The screenshot shows the 'Table: test2' details page. It includes sections for 'Basic Information', 'Full Members', and 'Description'. The 'Full Members' section contains a JSON object representing the table's metadata, similar to the database object in Figure 4.42.

FIGURE 4.44 – Affichage des métadonnées sur la table

The screenshot shows the 'Columns in test2' page. It lists three columns: 'email', 'description' (which is selected and highlighted in blue), and 'name'. The 'Recommendations for description' section provides 'Column Details' for the 'description' column, including its name, type (string), position (2), and description ('No description'). A 'Refresh' button is also present.

FIGURE 4.45 – Sélection de la colonnes et affichage de ses métadonnées

The screenshot shows the 'DATA STEWARD MENU' on the left with options: Profile, MetaData Quality (selected), Business Tags, and Statistics. The main area is titled 'AI Recommendations' and lists four items:

- description**: Description of the column 'description' in the table 'test2' within the 'pfa' database. Status: accepted. Buttons: Accept (green) and Reject (red).
- name**: description. Status: pending. Buttons: Accept (green) and Reject (red).
- type**: string. Status: pending. Buttons: Accept (green) and Reject (red).
- owner**: admin. Status: pending. Buttons: Accept (green) and Reject (red).

FIGURE 4.46 – Affichage et gestion des recommandations