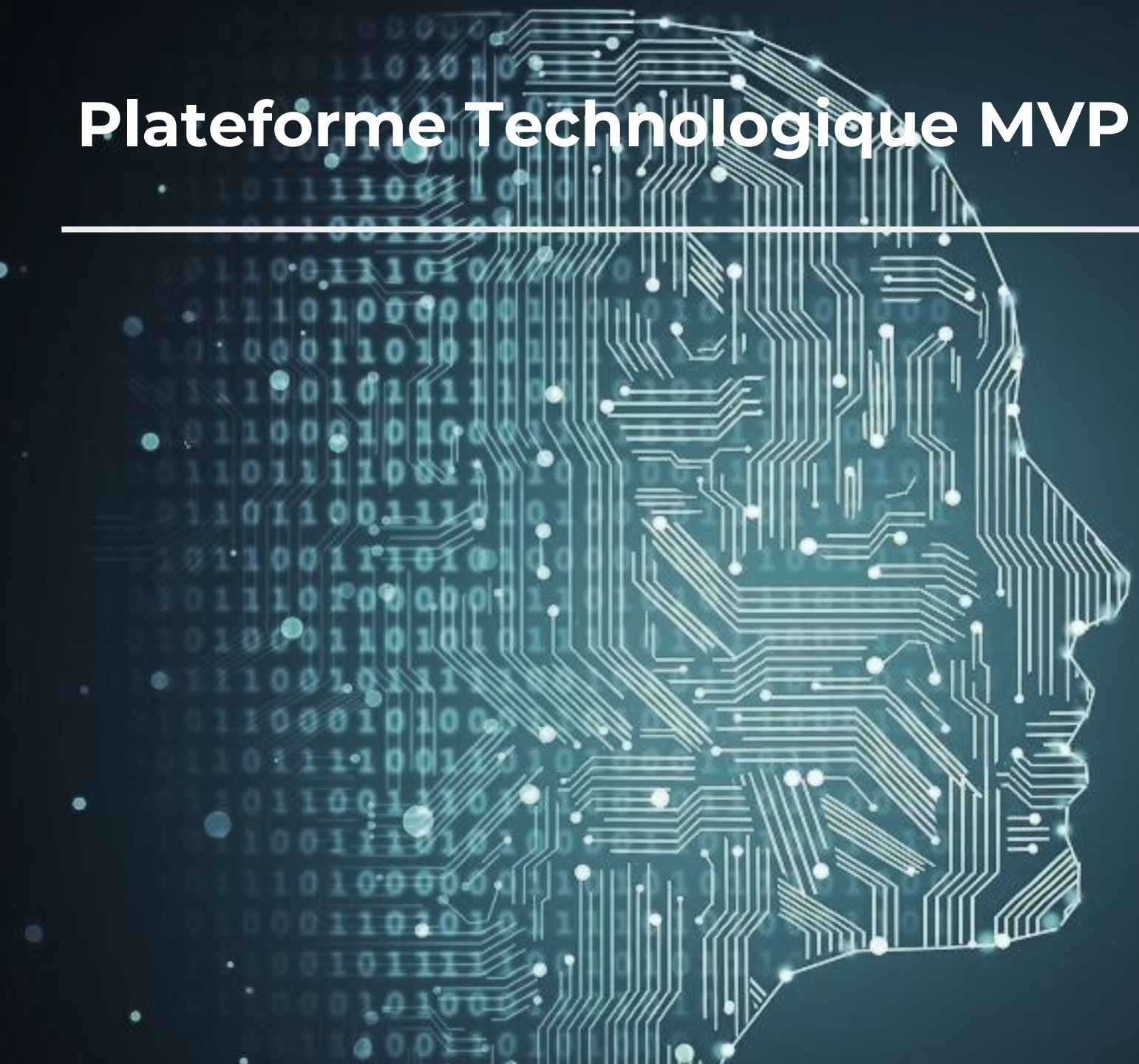
HEALTH DATA HUB

Besoins fonctionnels, exigences techniques et de sécurité

Plateforme Technologique MVP



- **Cible fonctionnelle MVP**
- Cas d’usage
- Exigences technique et de sécurité

Vue d’ensemble des objectifs métiers du MVP

Mettre à disposition des porteurs de projet un environnement technologique sécurisé permettant de réaliser les projets pilotes

Mettre à disposition d’un panel d’utilisateurs test un premier catalogue de données accessibles à la demande selon la gouvernance Hub

Tester les fonctions d’administration et de suivi de l’offre de service du Hub

Profils utilisateurs

Porteurs de projets pilotes

Utilisateurs cible MVP : ~20

- Niveau de compétences techniques :
  - Data : Statistiques, code, création d'algorithmes
- Usages :
  - Traitement de la donnée
  - Tests et exécution d'algorithmes
- Principales attentes :
  - Accès à la donnée
  - Outils de traitement, analyse et visualisation à l'état de l'art
  - Puissance de calcul

Utilisateurs de l'offre de service « Catalogue de données à la demande »

Utilisateurs cible MVP : ~50

- Niveau de compétences techniques :
  - Data : Statistiques, code, création d'algorithmes
  - Data : création de clés d'appariement
- Usages :
  - Consultation des données
  - Traitement de la donnée
  - Tests et exécution d'algorithmes
- Principales attentes :
  - Accès à la donnée
  - Outils de traitement, analyse et visualisation à l'état de l'art
  - Puissance de calcul

Administrateurs du « Hub »

Utilisateurs cible : ~5

- Niveau de compétences techniques :
  - Data : appariement (matching probabiliste ou déterministe)
  - Hub : Suivi technique des composants
- Usages :
  - Appariement
  - Gestion technique, administrative et financière du Hub
- Principales attentes :
  - Interface de gestion complète (technique, financière et admin)
  - Interface d'appariement des bases de données

User Stories
Porteurs de projet pilote (1/2)

Usage

- Je dispose d'un « **espace projet** » dédié préparé par l'équipe du Hub présentant en lecture les jeux de données « source » de mon projet, pour lesquels j'ai reçu une habilitation, et selon le(s) moyen(s) de stockage de mon choix (fichier plat, base relationnelle, clef valeur, colonne, document, graphe) :
  - Vues sur les bases de données présentes au catalogue du Hub ;
  - Jeux de données appariés réalisés spécifiquement pour le projet par l'équipe Hub ;
  - Jeux de données propriétaires complémentaires spécifiques au projet, fournis par le porteur de projet à l'équipe Hub.

- Je dispose sur mon « espace projet » **d'outils de requête, d'analyse, de visualisation et de développement** permettant à partir de mes jeux de données « sources » :
  - De créer de nouvelles tables ou jeux de données enrichis sur le moyen de stockage de mon choix, qui viendront automatiquement s'ajouter à mon espace projet ;
  - De réaliser des traitements statistiques ou d'entrainer des modèles de machine learning, en utilisant des librairies de références ;
  - De visualiser un jeu de donnée au moyen d'un outil de data visualisation ;
  - De définir, d'enregistrer et d'exécuter une séquence automatisée de traitements (« pipeline », « workflow ») ;
  - D'importer, d'enregistrer et de gérer différentes versions des algorithmes que je souhaite exécuter.

- Je dispose sur mon « espace projet » d'une **capacité de stockage** (standard, rapide) **et de calcul** (CPU, GPU) garantie dimensionnée à mon besoin, définie en lien avec un expert de l'équipe du Hub et allouée par cette dernière.

User Stories
Porteurs de projet pilote (2/2)

Usage

- J'ai la garantie que l'ensemble des jeux de donnés crées au sein de mon « espace projet » et les algorithmes développés ne sont **accessibles** qu'aux **administrateurs** du Hub et **utilisateurs autorisés** pour le projet.

- J'ai la possibilité depuis mon « espace projet » **d'exporter les commandes et les résultats obtenus** sur l'interface de requête. Cet export peut être **limité en volume** (X Mo) et **en fréquence** (à définir) et est sujet à une approbation explicite de l'administrateur du hub, il est **conditionné par une confirmation** me rappelant :
  - Les conditions d'usage des données du Hub, et en particulier l'interdiction d'exporter des données non anonymes non agrégées ;
  - Mes responsabilités personnelles concernant la protection de ces données individuelles et les conséquences pénales et légales encourues en cas de défaut ;
  - La traçabilité complète de mes opérations et la réalisation d'audit réguliers sur les exports par les équipes du Hub.

- J'ai à ma disposition des API, connectées via un VPN Ipsec, me permettant de :
  - Déclencher des « pipelines »
  - D'exposer des données anonymisées

User Stories
Utilisateurs de l’offre de service (1/3)

Usage

- J’ai accès à un **espace « utilisateurs »** présentant les différents **« espaces projets »** auxquels je suis rattaché et un rappel de leur finalité

- J’ai accès à un **catalogue des jeux de données « publics »** disponibles sur le Hub. Ce catalogue présente :
  - Les jeux de données « publics » disponibles ;
  - Une information sur la structure et le contenu des jeux de données disponibles : champs, couverture, profondeur d’historique, niveau de qualité, fréquence de mise à jour ;
  - Un accès en téléchargement à une documentation publique ;
  - Un accès en téléchargement à un échantillon anonymisés et/ou synthétique des données.

- J’ai accès à un **formulaire mail prérempli** me permettant de formuler ma **demande d’habilitation**. Ce formulaire m’indique les pièces à fournir et les procédures à suivre.

User Stories
Utilisateurs de l'offre de service (2/3)

Usage

- Je dispose d'un ou plusieurs « espace projets » préparés par l'équipe du Hub présentant une vue en lecture des jeux de données « source » du catalogue sur le périmètre sur lequel j'ai obtenu une habilitation pour le projet et la finalité déclarée.

- Je dispose sur mon « espace projet » **d'outils de requête, d'analyse, de visualisation et de développement** permettant à partir de mes jeux de données « sources » :
  - De créer de nouvelles tables ou jeux de données enrichis sur le moyen de stockage de mon choix, qui viendront automatiquement s'ajouter à mon espace projet ;
  - De réaliser des traitements statistiques ou d'entrainer des modèles de machine learning, en utilisant des librairies de références ;
  - De visualiser un jeu de donnée au moyen d'un outil de data visualisation ;
  - De définir, d'enregistrer et d'exécuter une séquence automatisée de traitements (« pipeline », « workflow ») ;
  - D'importer, d'enregistrer et de gérer différentes versions des algorithmes que je souhaite exécuter.

- Je dispose sur mon « espace projet » d'une **capacité de stockage** (standard, rapide) **et de calcul** (CPU, GPU) garantie dimensionnée à mon besoin, définie en lien avec un expert de l'équipe du Hub et allouée par cette dernière.

User Stories
Utilisateurs de l’offre de service (3/3)

Usage

- J’ai la garantie que l’ensemble des jeux de donnés crées au sein de mon « espace projet » et les algorithmes développés ne sont **accessibles** qu’aux **administrateurs** du Hub et **utilisateurs autorisés** pour le projet.

- J’ai la possibilité depuis mon « espace projet » **d’exporter les commandes et les résultats obtenus** sur l’interface de requête. Cet export est **limité en volume** (X Mo) et **en fréquence** (à définir) et est **conditionné par une confirmation** me rappelant :
    - Les conditions d’usage des données du Hub, et en particulier l’interdiction d’exporter des données non anonymes non agrégées ;
    - Mes responsabilités personnelles concernant la protection de ces données individuelles et les conséquences pénales et légales encourues en cas de défaut ;
    - La traçabilité complète de mes opérations et la réalisation d’audit réguliers sur les exports par les équipes du Hub.

User Stories
Administrateurs du « Hub » (1/2)

Usage

- Je peux **ajouter, retirer et modifier des utilisateurs** et leur **attribuer un accès** à la plateforme.

- Je peux **ajouter, retirer et modifier des « espaces projets »** pour lesquels je peux spécifier :
  - La finalité déclarée du projet ;
  - Les utilisateurs habilités à accéder à l'espace projet ;
  - Les outils mis à disposition sur l'espace projet ;
  - La capacité de calcul (CPU, GPU) et de stockage maximum affectée au projet ;
  - Les fichiers, jeux de données ou champs visibles en lecture en tant que jeux de données « source » du projet.

- J'ai accès à un **catalogue** présentant l'ensemble des **jeux de données** présents sur la plateforme, leurs **métadonnées** et les espaces projets y ayant accès. Je peux apposer des marques (« **tags** ») au niveau des fichiers, des métadonnées et des champs, par exemple pour spécifier **des restrictions particulières** concernant l'accès à la donnée. En particulier, je peux apposer une marque pour indiquer les jeux de données qui seront présentés au catalogue de jeux de données « publics ».

User Stories
Administrateurs du « Hub » (2/2)

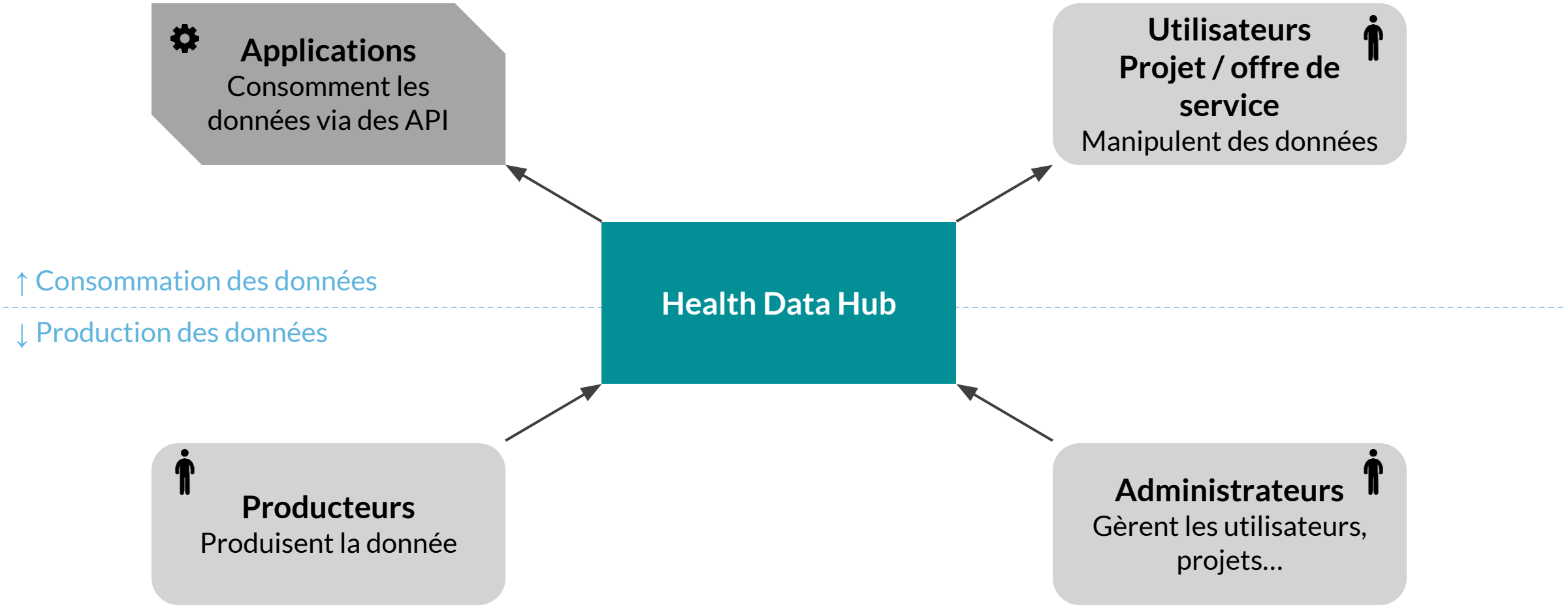
Usage

- Je dispose **d'une interface de supervision unique** couvrant l'ensemble des composants de la plateforme. Elle permet :
  - De suivre la consommation technique (ressources de calcul et de stockage) et financière d'ensemble et par projet ;
  - D'accéder aux interfaces d'administration de l'ensemble des composants (infrastructure, socle applicatif) ;
  - De suivre l'état de la plateforme et de disposer d'un historique des pannes et des opérations de maintenance prévues.

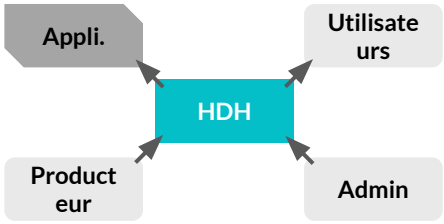
- Je dispose d'un **outil de collecte et d'analyse des journaux** qui me permet de trier, d'ordonnancer et de filtrer en fonction de différents critères l'ensemble des événements du système (authentification, gestion des comptes et des droits, accès aux ressources, modification des stratégies de sécurité, activité des processus, activité des systèmes). En particulier je peux :
  - Disposer d'un visibilité complète sur tous les exports réalisés dans la journée pour réaliser un audit quotidien ;
  - Déterminer pour une période donnée les actions menées par un utilisateur donné ou les utilisateurs ayant procédé à une action données ;
  - Lister tous les événements associés à chacun des objets visualisés et générer des alarmes selon le paramétrage de mon choix.

- Je dispose d'un **espace, de capacité de calcul et de stockage et d'outils de traitement de la donnée cloisonnés physiquement du reste de l'infrastructure** me permettant de **réaliser les opérations d'appariements**.

Cible fonctionnelle
Schéma d'ensemble



Cible fonctionnelle
Focus : Health Data Hub



Health Data Hub

Applications

Utilisateurs

Exposition des données

Espace projet

Ingestion des données

Stockage des données

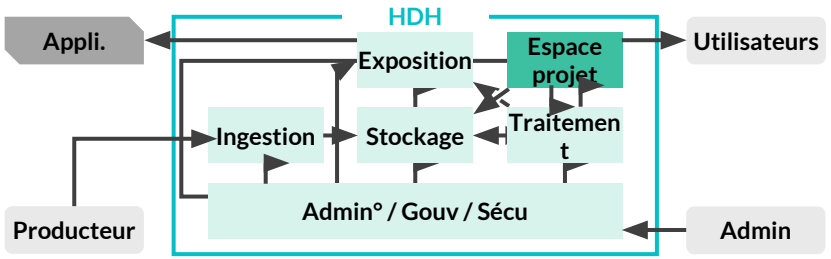
Traitement des données

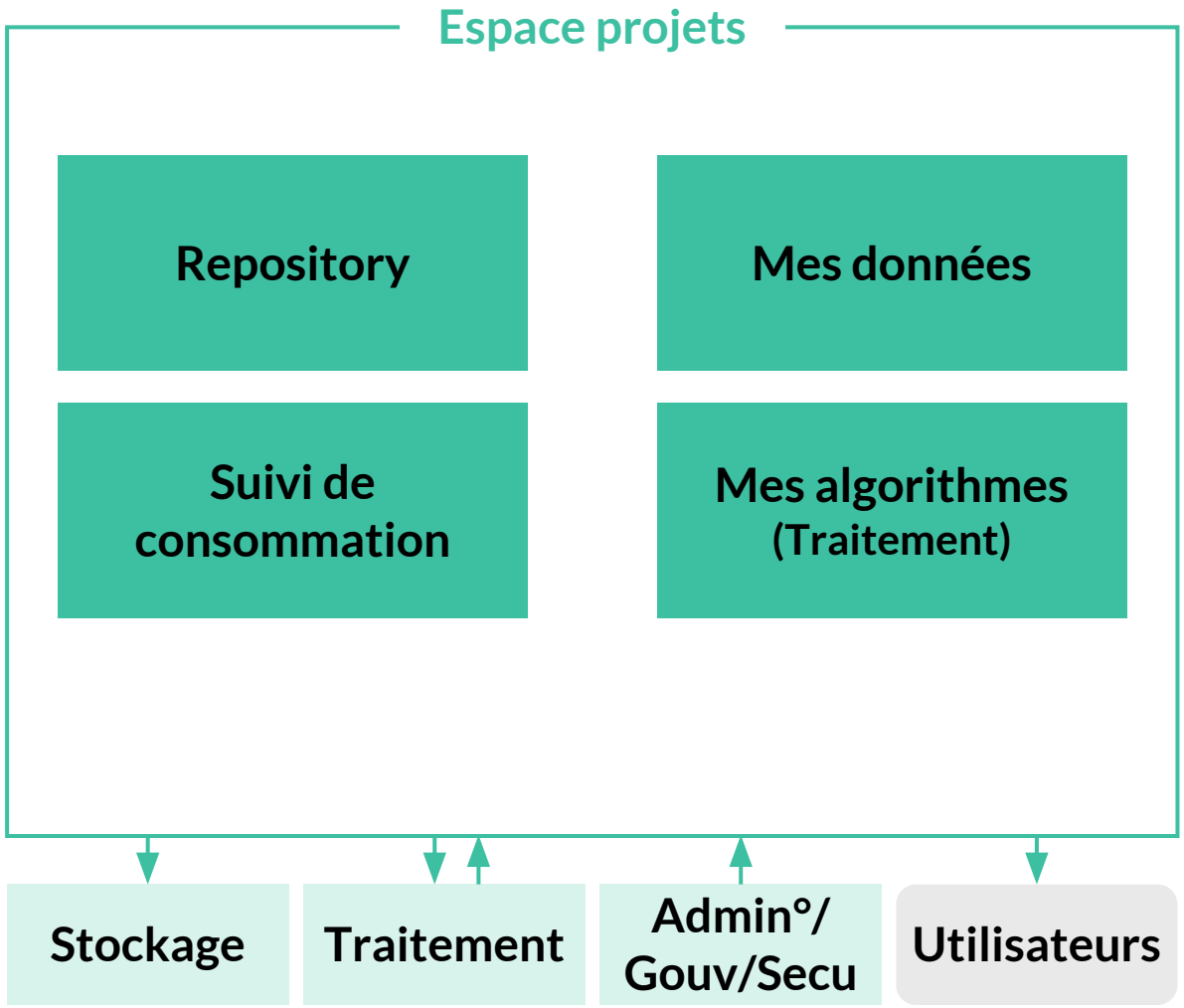
Administration / Gouvernance / Sécurité

Producteur

Administrateurs

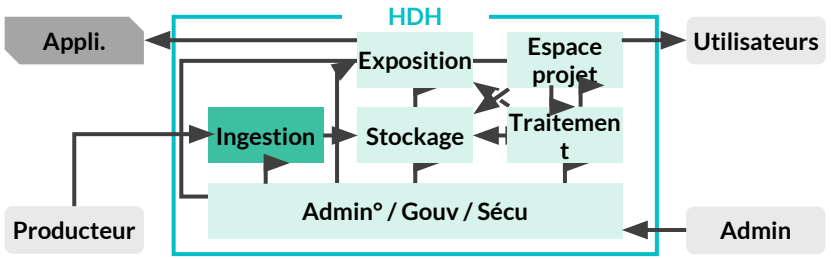
Cible fonctionnelle
Focus : Espace projets

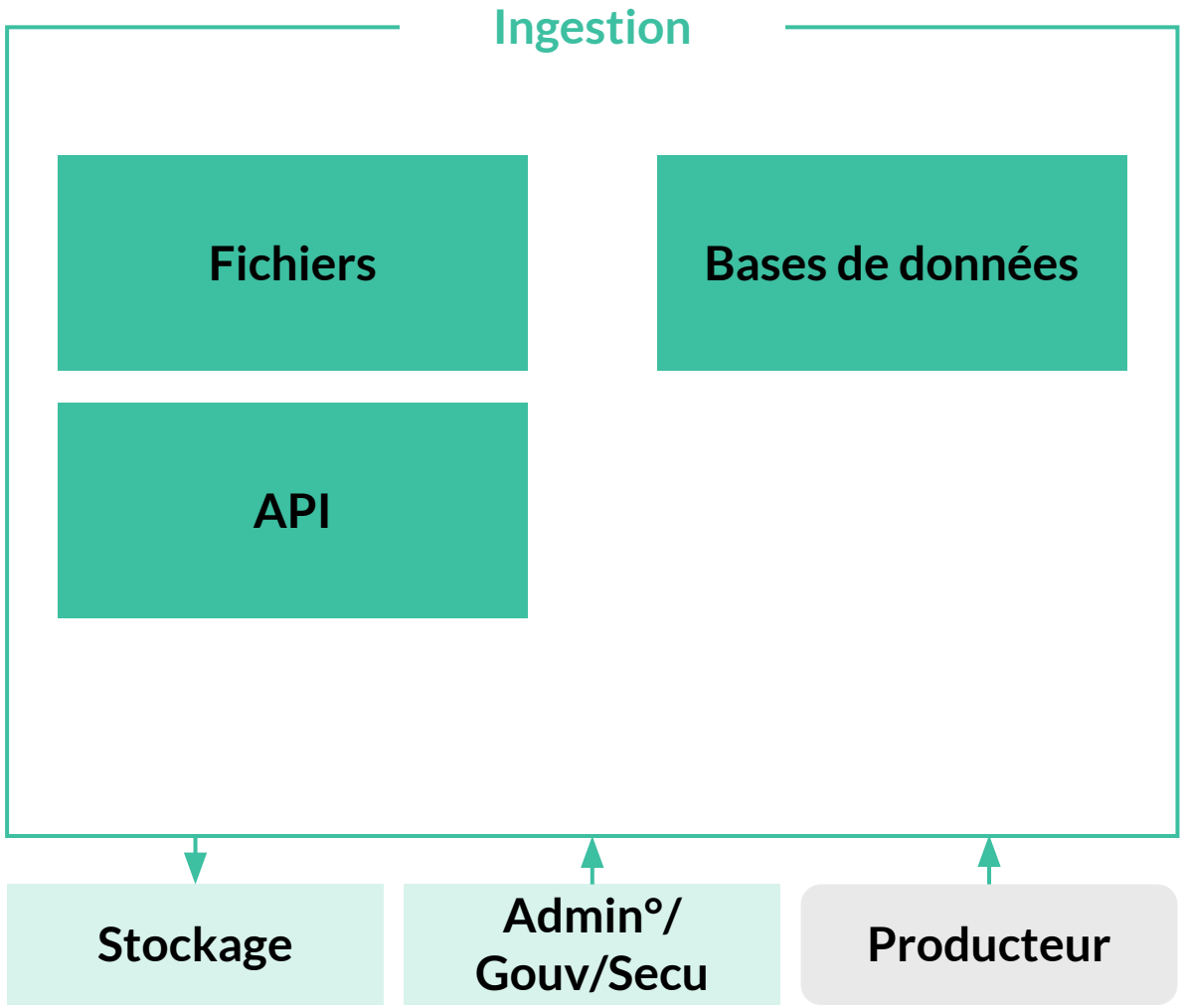




- **Repository**
  - Données propres aux utilisateurs : codes, fichiers, librairies…
  - Historique des exports de données
- **Mes données**
  - Vues sur les données calculées et catalogues de données disponibles pour l'utilisateur
- **Suivi de consommation**
  - Suivi des consommations par projet en volume et temps de calcul
- **Mes algorithmes**
  - Notebook
  - Codes sources

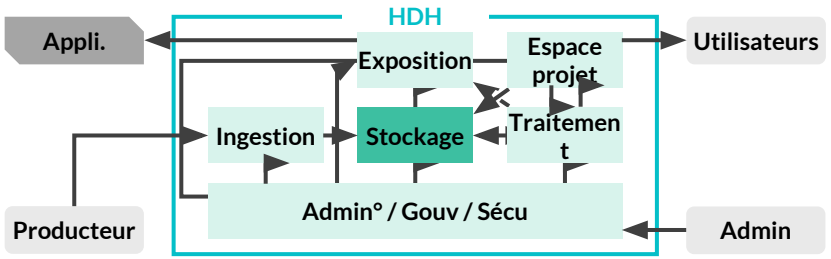
Cible fonctionnelle
Focus : Ingestion des données

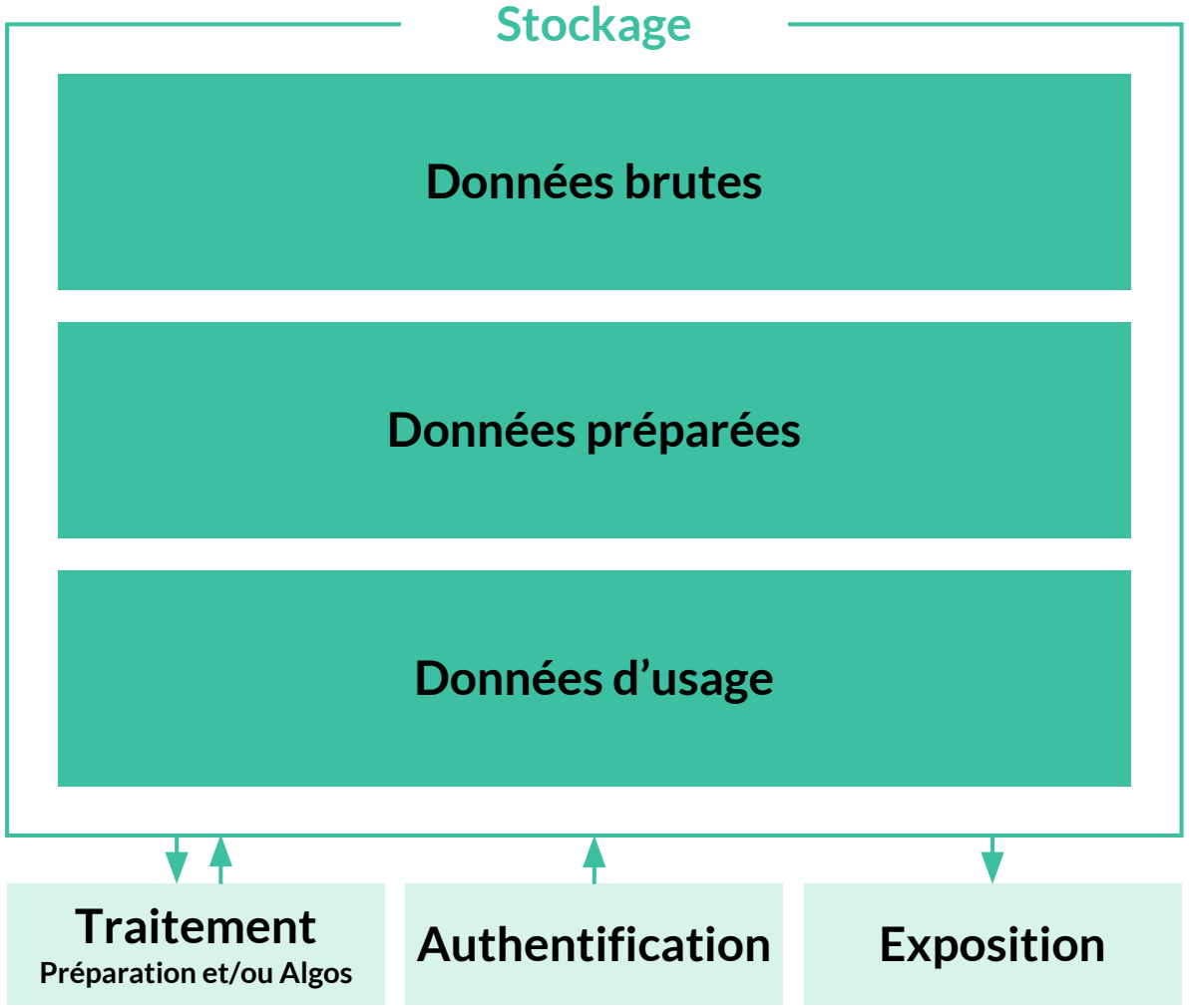




- **Fichiers**
  - Tout type de fichiers (Images, texte, audio, vidéo…)
- **Bases de données**
  - Bases relationnelles
- **API**
  - Mise à disposition d'une API pour permettre aux producteurs de déposer de la donnée

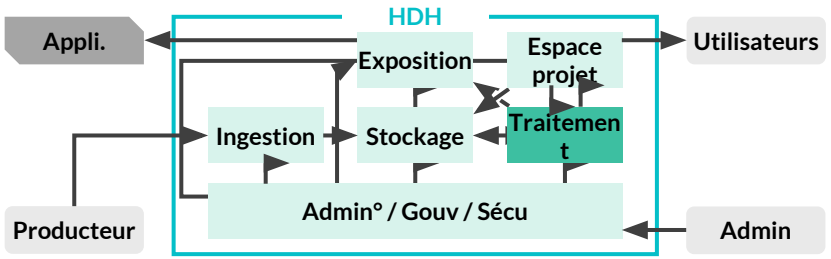
Cible fonctionnelle
Focus : Stockage des données

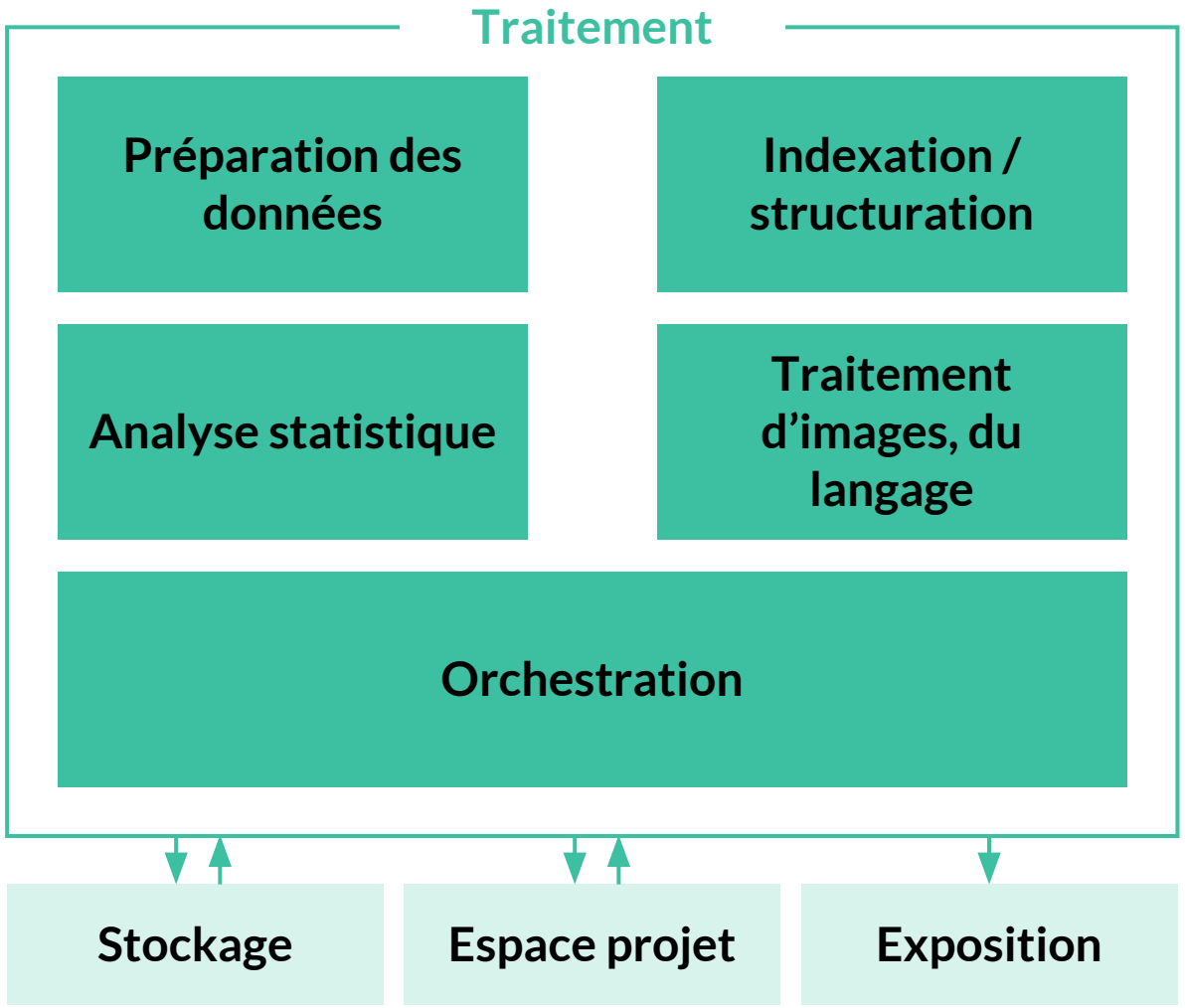




- **Données brutes**
  - Données issues d'une source non modifiées pour rejeu si besoin
- **Données préparées**
  - Données nettoyées, ajout de colonnes afin de répondre a des cas d'usages
- **Données d'usage**
  - Données à destination de l'utilisateur pour faire ses agrégations, statistique ou recherche

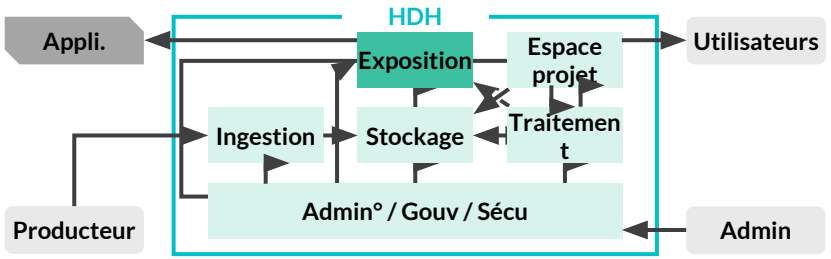
Cible fonctionnelle
Focus : Traitement des données

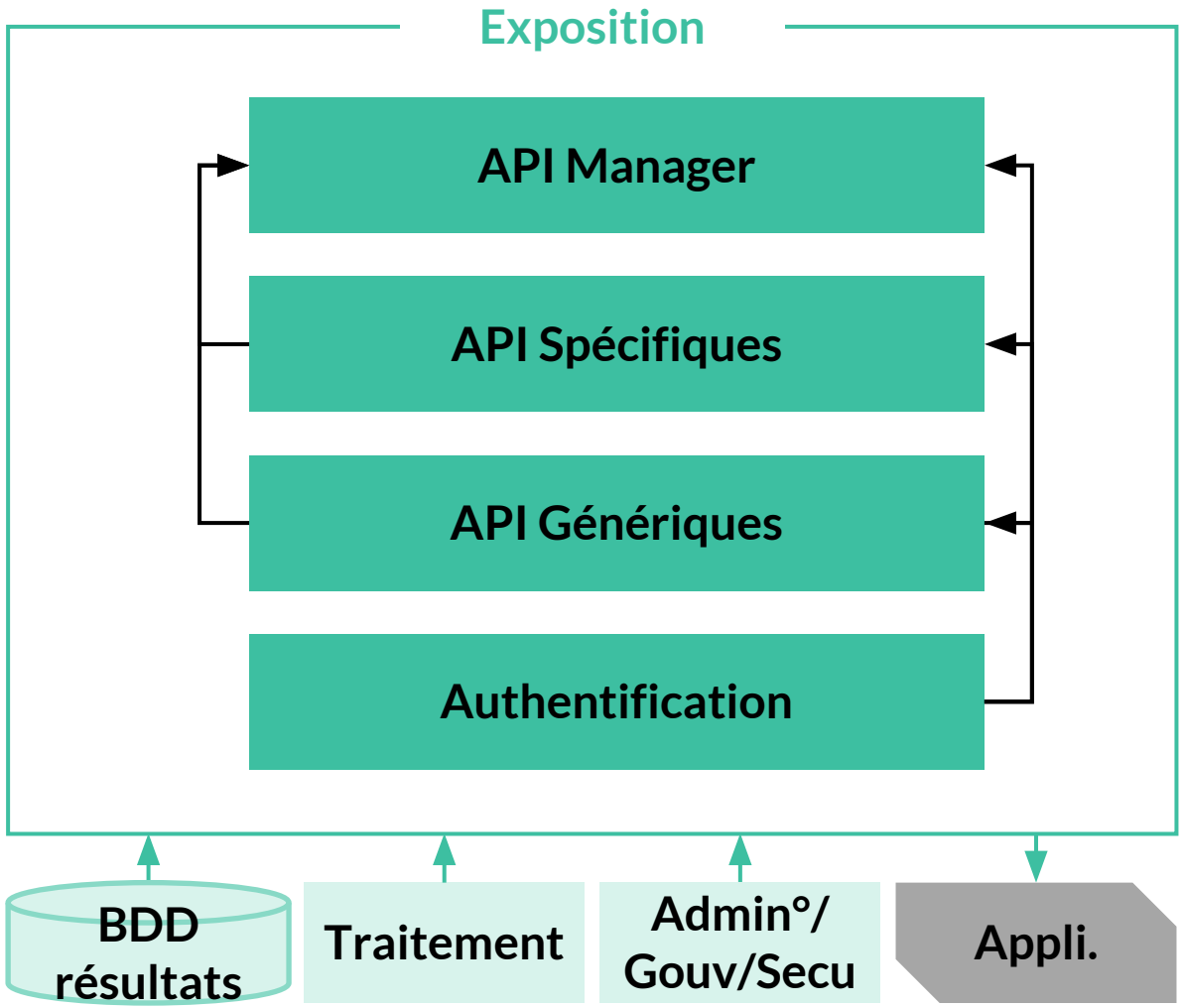




- **Préparation des données**
  - Nettoyage
  - Feature engineering
- **Indexation / structuration**
- **Analyse statistique**
  - Calculs distribués
  - Machine learning
  - Requêtes statistiques
- **Traitement d'images, du langage**
  - Deep learning
- **Orchestration**
  - Gestion du pipeline des traitements

Cible fonctionnelle
Focus : Exposition des données

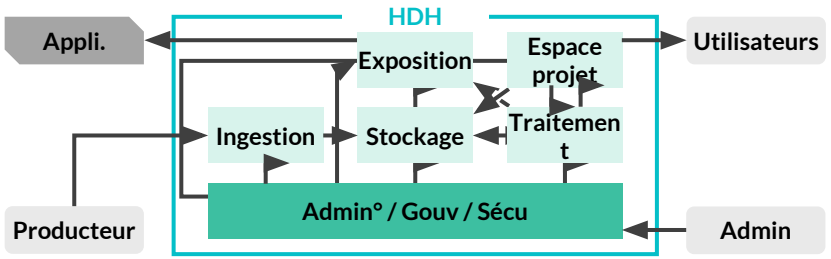


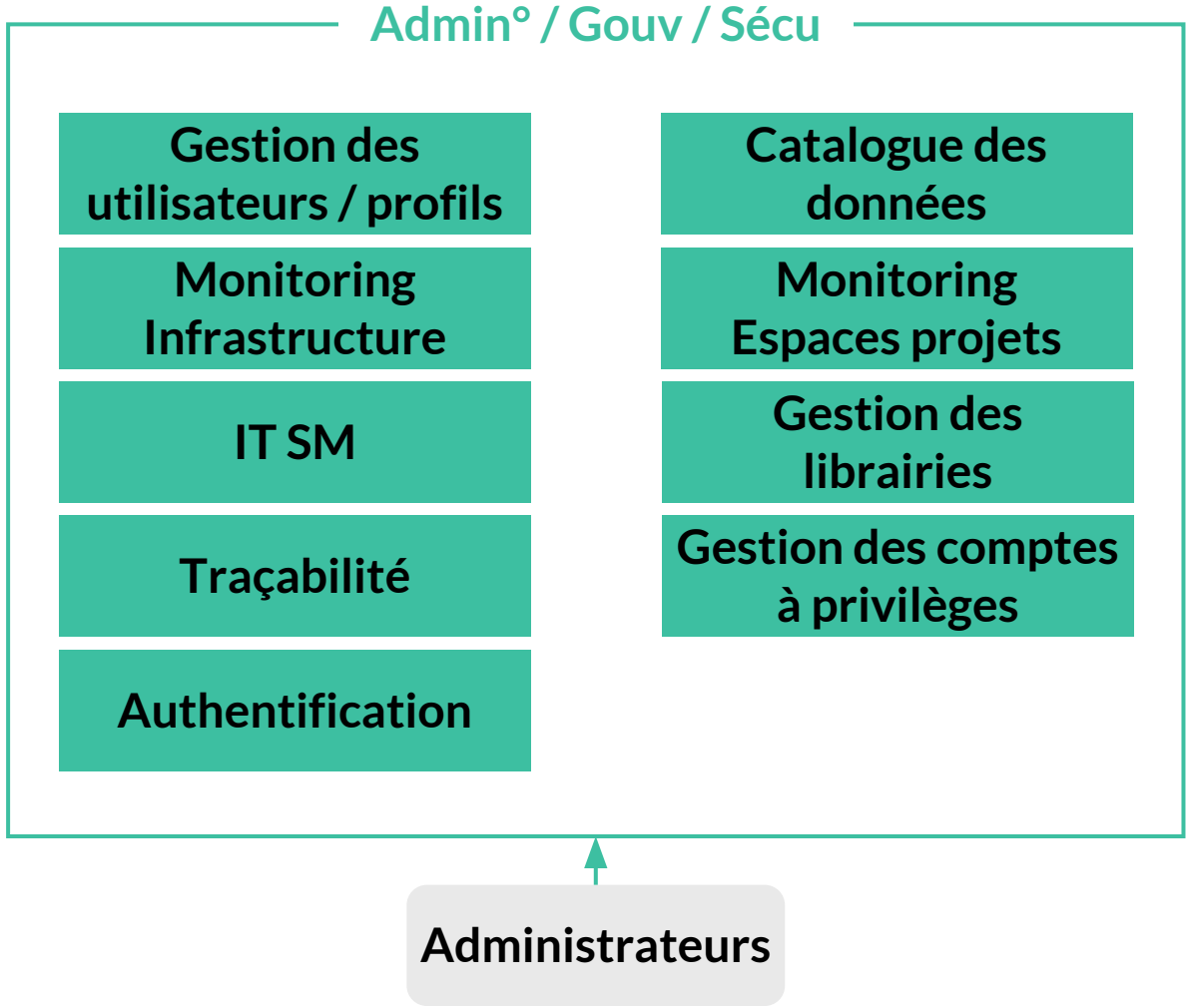


- **API Manager**
  - Exposition des API de manière sécurisée
- **API Spécifiques**
  - API développées sur mesure pour/par les projets
- **API Génériques**
  - API standards sur étagère
- **Authentification**
  - Propagation d'authentification de l'usager auprès des API
  - Validation des droits

⚠ **On interdit l'exécution de requêtes directement sur le cluster**

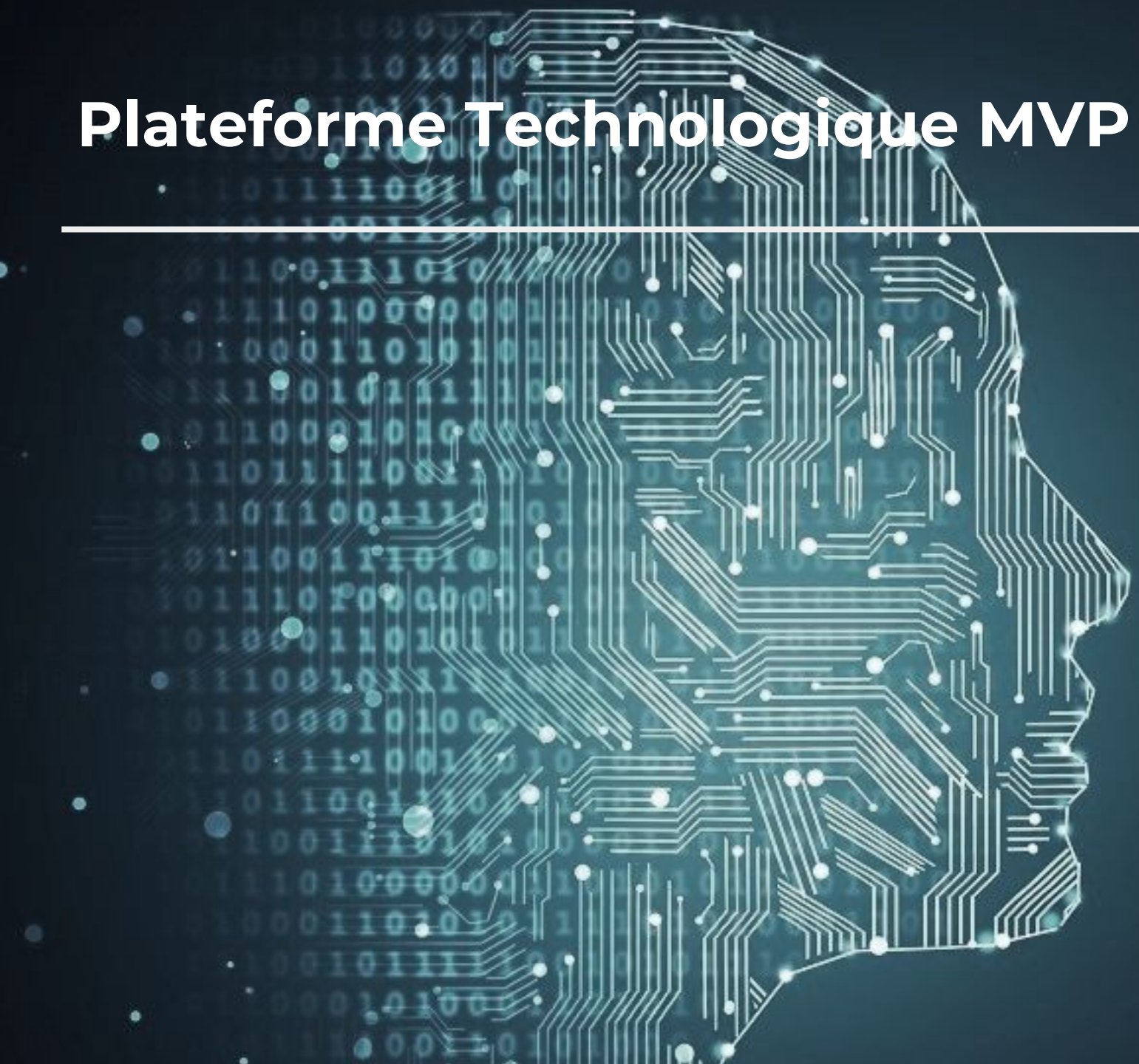
Cible fonctionnelle
Focus : Admin°/Gouvernance/Sécurité





- **Gestion des utilisateurs / profils**
  - Ajouts / retraits d'utilisateurs
  - Gestion des droits
- **Catalogue de données**
  - Gouvernance de la donnée
  - Administration des catalogues de données : niveaux d'accès
- **Monitoring Infrastructure**
  - Tableau de bord de suivi et consommation des composants
- **Monitoring des espaces projets**
  - Tableau de bord de suivi et consommation des projets
- **IT System Management**
  - Suivi, traçabilité des demandes et incidents
- **Gestion des librairies**
  - Gestion des versions et catalogues de librairies (jar, …)
- **Traçabilité**
  - Consultation des logs
  - Génération d'audit
- **Gestion des comptes à privilèges**
  - Gestion des super utilisateurs
- **Authentification**
  - Gestion des protocoles d'identification des utilisateurs

Plateforme Technologique MVP



- Cible fonctionnelle MVP
- **Cas d’usage**
- Exigences techniques et de sécurité

Exemple de cas d’usage : cas d’usage n°1

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Démontrer le lien entre l’exposition aux antibiotiques et la survenue d’une pathologie définie : • Décrire l’épidémiologie détaillée • Analyser le lien à l’échelle de l’individu et temporo-spatial à l’échelle des populations	**Base résultant de l’appariement des bases suivantes :** • **Entrepôt de données de santés codé avec la pathologie identifiée :** ○ Volume 2To ○ Description : données administratives, sociales et médicales, recueillies lors des consultations et hospitalisations, des patients soignés dans les hôpitaux • **SNDS** ○ Volume : non défini ○ Données extraites du SNDS, selon les critères d’appariement **> Besoin en stockage final** : non défini	• **Traitement** ○ Analyses descriptives, modélisations temporelles, modèles mathématiques. • **Langage** ○ R ○ Python • **Outils** ○ Données scannées : OCR, NLP

Exemple de cas d'usage : cas d'usage n°2

Finalité du cas d'usage	Données à traiter	Traitement requis et outils
Identifier des interactions médicamenteuses bénéfiques ou délétères chez des patients atteints d'une pathologie définie : • Identifier les combinaisons médicamenteuses délétères nécessitant la mise en place d'alertes de pharmacovigilance • Identifier les combinaisons médicamenteuses bénéfiques justifiant de la réalisation d'essais cliniques de «repositionnement» de médicaments	**Base résultant de l'appariement des bases suivantes :** • **SNDS** : o Volume : 1 à 2 To o Données extraites du SNDS, selon certains critères, représentant environ 750 tables par patient. Le nombre approximatif de patients attendu est de 500 000. • **Données patients porteurs** : o Volume : 5 Go o Description : Dossier médicaux o Données stockées sous forme d'une matrice texte. o Possibilité éventuelle d'intégrations à faire avec données étrangères **> Besoin en stockage final** : 8 To	• **Traitement** o Analyse statistique o Approches machine learning pour structurer données – et traitement du langage naturel • **Outils** o Bureau virtuel, o Serveur R studio o Outils de machine learning

Exemple de cas d’usage : cas d’usage n°3

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Créer une base de données rassemblant un set de données minimum pour l’étude de pathologies rares définies : • Déterminer une série d’indicateur de santé publique: prévalence, incidence, typologie des maladies, éléments sur l’histoire des maladies et des parcours des patients • Déterminer autant que possible l’errance diagnostique pour certaines pathologies Les données du SNDS permettront de retracer la totalité du parcours médical des patients et de réaliser des études médico-économiques	**Base résultant de l’appariement des bases suivantes :** • **SNDS** ○ Volume : non défini ○ Données extraites du SNDS, selon les critères d’appariement • **Base de données sur les pathologies étudiées** ○ Volume : 0,2 To ○ Description : environ 30 tables provenant de grands hôpitaux/CHU **> Besoin en stockage final** : 1 To	• **Traitement** ○ Analyse statistique ○ Création d’indicateurs de santé publique • **Langage** ○ R ○ Python • **Outils** ○ Jupyter ○ R studio

Exemple de cas d’usage : cas d’usage n°4

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Etudier l’impact d’une pathologie sur la génèse d’une autre : • Chainer des cohortes rassemblant les données cliniques, biologiques radiologiques et génétiques de ces patients • Analyser l’évolution du parcours de soins à partir d’hypothèses établies sur les données cliniques ou biologiques • Faire émerger des liens de causalité (inférence causale) voire de nouvelles hypothèses de prise en charge, grâce aux approches de machine learning	**Base résultant de l’appariement des bases suivantes :** • **SNDS** ○ Volume : 10To ○ Description : Extraction d’une cohorte • **Données de deux cohortes** ○ Volume : 0.1 To ○ Données épidémiologiques, cliniques, anatomo-pathologiques structurées ○ Données génétiques structurées ○ 40 tables, 900 millions d’entrées **> Besoin en stockage final** : 10 To	• **Ingestion** ○ Ingestion de la base depuis un disque externe • **Traitement** ○ Analyse statistique ○ Machine Learning • **Langage** ○ Python • **Outils** ○ Spark-Scala-Python ○ Jupyter

Exemple de cas d’usage : cas d’usage n°5

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Analyser des images enrichies et créer un modèle prédictif de réponse aux thérapies • Identifier les marqueurs de réponse ou non réponse aux traitements • Identifier les marqueurs de toxicité médicamenteuse	**SNDS** o Volume : Non défini o Description : Extraction d’une cohorte **Entrepôt de données** o Volume : 10 To o Description : Données structurées code diagnostics, Traitements, Informations patient, images et compte rendus textuels, tumeurothèque **> Besoin en stockage final** : > 30 To	• **Ingestion** o ETL pour les données structurées (cliniques, annotation) • **Traitement** o Analyse statistique o Machine Learning • **Langage** o Python • **Outils** o Serveur R studio o Jupyter Python, TensorFlow, Keras o Module de text mining pour l’annotation automatique à partir des compte rendus textuels

Exemple de cas d’usage : cas d’usage n°6

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Proposer un framework pour l’intégration et l’enrichissement de données hétérogènes • Développer une méthodologie et un algorithme de TAL (traitement automatique du langage) pour traiter les comptes rendus médicaux et extraire les concepts médicaux pertinents • Lancer des cas d’usages médicaux concrets Les données du SNDS permettront de consolider une vision sur l’ensemble du parcours du patient, de limiter les efforts d’annotation et d’évaluer la validité des données extraites.	**Base résultant de l’appariement des bases suivantes :** • **SNDS** ○ Volume : Non défini ○ Description : Données extraites du SNDS, selon les critères d’appariement • **Entrepôt de données** ○ Volume : 0,5 To ○ Description : données textuelles de rapports médicaux partiellement structurées, totalité des Comptes Rendus disponibles : entre 20-30 millions **> Besoin en stockage final** : 1 To	• **Traitement** ○ Analyse statistique ○ NLP • **Langage** ○ Python • **Outils** ○ Serveur Jupyter

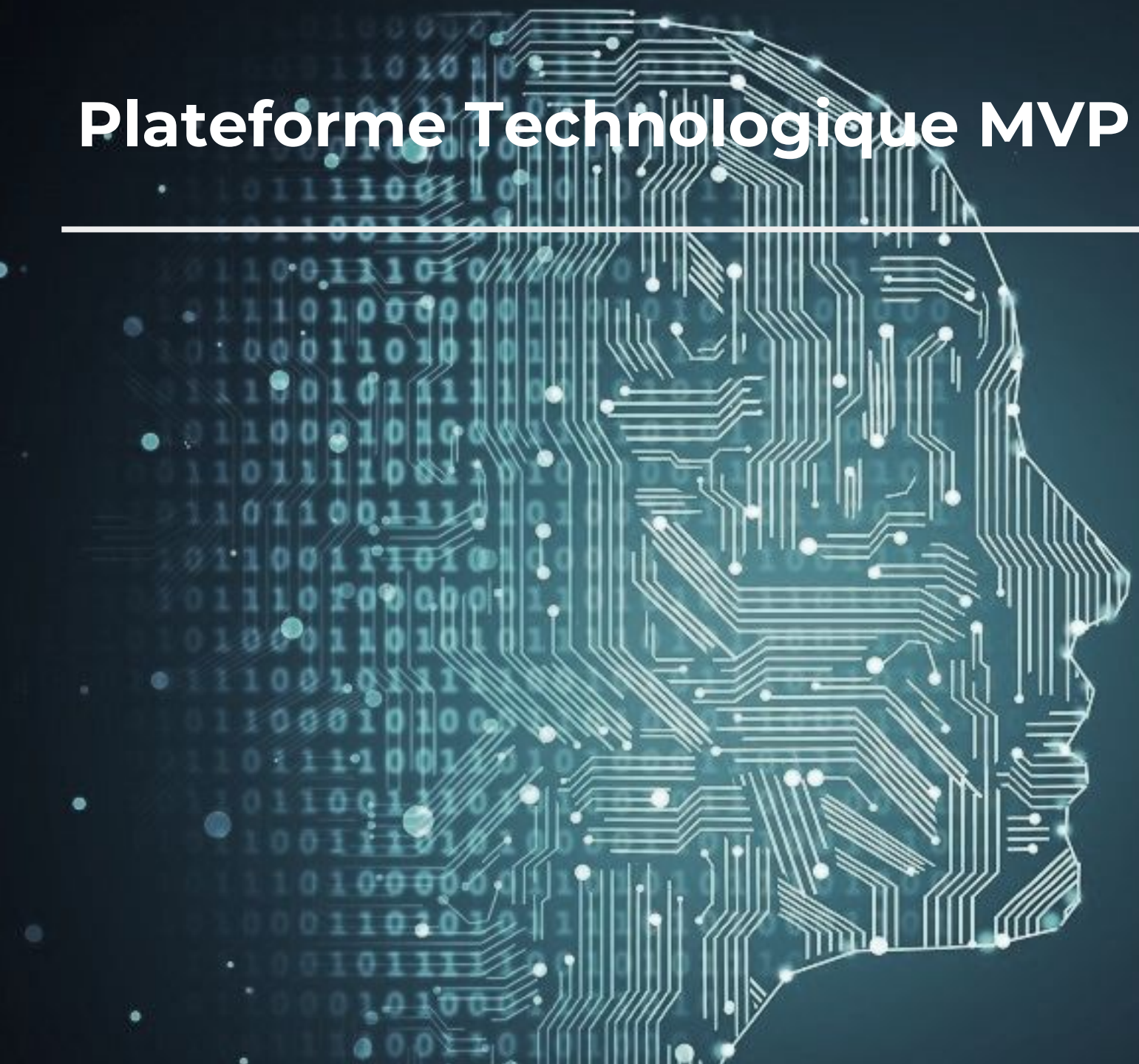
Exemple de cas d’usage : cas d’usage n°7

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Etudier les facteurs influençant les choix entre deux types d’allocations pour personnes handicapées en termes de coûts, de contenus des plans d’aide, de restes à charge, etc. **Mettre en œuvre des techniques d’appariement innovantes**, sur la base d’informations identifiantes : nom, prénom(s), date de naissance, adresse postale, mais non du NIR	**Base résultant de l’appariement des bases suivantes :** • **Base 1** ○ Volume : 10 Go ○ Description : données structurées au format SAS, nombre d’observations : 1 300 000 • **Base 2** ○ Volume : 10 Go ○ Description : données structurées au format SAS, nombre d’observations : 120 000 **> Besoin en stockage final** : 1 To	• **Traitement** ○ Analyse statistique • **Langage** ○ Python • **Outils** ○ Serveur R studio

Exemple de cas d’usage : cas d’usage n°8

Finalité du cas d’usage	Données à traiter	Traitement requis et outils
Concevoir des algorithmes d’Intelligence Artificielle (IA) pouvant aider des équipes médicales dans leur pratique quotidienne d’imagerie, et ainsi contribuer à encore davantage populariser la technique en facilitant son utilisation afin d’améliorer les soins**. L’enjeu médical est de développer et évaluer des algorithmes d’IA** détectant différents organes et leurs lésions	• **Données d’imagerie d’hôpitaux :** ○ Volume : Non défini ○ Description : images (nombre : 1 440 000) et comptes rendus (format DICOM et TXT) **> Besoin en stockage final** : 4 To	• **Ingestion** ○ Copie des données par disques USB cryptés • **Traitement** ○ Extraction automatique de contenu dans les comptes rendus: NLP ○ Machine Learning : apprentissage supervisé, semi-supervisé et non supervisé : détection et segmentation d’organes et de pathologies dans des images 2D, génération d’images de synthèse • **Langage** ○ Python • **Outils** ○ Image processing et visualisation ○ API permettant de se connecter aux bases de données

Plateforme Technologique MVP



- Cible fonctionnelle MVP
- Cas d’usage
- **Exigences techniques et de sécurité**

Exigences techniques et de sécurité
Infrastructure et maintenance

Brique	Principales exigences
Infrastructure	Hébergeur agrée ou certifié "Hébergeur de données de santé" > Quelle couverture sur l'offre de service ?
	[SNDS] Territorialité : France
	Possibilité d'utiliser des CPU et des GPU comme capacité de calcul (y compris sur domaine HDS)
	Accès à des supports de stockages rapides (SSD ou mémoire flash) ou standards
	Elasticité : capacité d'accueil d'une centaine nœuds sans modification autre qu'une configuration et sans impact sur la performance du système
	Capacité de tag des objets cloud en vue d'une intégration dans un référentiel d'entreprise (ITSM)
	Le format des données en fin de contrat doit être exploitables et permettre une récupération des données (préciser média et coût)
	Taux de disponibilité minimum : 99,95%

Exigences techniques et de sécurité

Bureau Virtuel, Plateforme Data : sujets transverses, ingestion

Brique	Principales exigences
Bureau Virtuel	Solution VDI avec capacité à customiser les bureaux virtuels
	Capacité à capturer le keylogger et de la capture vidéo sur sessions VDI
	Partage de l'annuaire d'authentification entre le hub et la plateforme VDI
Plateforme Data - Transverse	Système d'exploitation avec un support valide de la part d’un éditeur ou d’un prestataire de service.
	Possibilité d'exposer des données par le développement d'API spécifiques
Plateforme Data – Ingestion / Echanges	Broker de messagerie asynchrone
	Ingestion Batch / Micro-batch

Exigences techniques et de sécurité
Plateforme Data : moyens de stockage

Brique	Principales exigences
Plateforme Data – Moyens de stockage	Moyen de stockage en cache
	Moyen de stockage objet
	Moyen de stockage distribué
	Moyen de stockage structuré : base relationnelle
	Moyen de stockage structuré : base clef-valeur
	Moyen de stockage structuré : base orientée colonne
	Moyen de stockage structuré : base orientée document
	Moyen de stockage structuré : base orientée graphe
	Moteur de recherche

Exigences techniques et de sécurité
Plateforme Data : Traitement de la donnée

Brique	Principales exigences
Plateforme Data – Traitement	Moteur de calcul distribué pour les traitements en batch
	Moteur de calcul distribué pour les traitements en flux
	Moteur d'exécution d'applications conteneurisées
	Moteur de calcul distribué pour l'apprentissage automatique supportant le calcul sur GPU
	Calcul non distribué dans un container pour les data set de faible volume

Exigences techniques et de sécurité
Plateforme Data : Usages

Brique	Principales exigences
Plateforme Data – Usages	IDE Python
	IDE R
	L'import de librairies extérieures (Datascience / Machine Learning de référence (Tensorflow, Keras, Sci-kit, Pandas, Theano…)) peut se faire online via un nexus synchronisé avec l'extérieur, miroir de cran ou pip, ou en en offline en important les librairies dans un repository Git
	Outil de data visualisation : quelles solutions sur étagère vs solutions dans la market place (SAS, R Studio…) ?
	Système de contrôle des versions, gestion de code source, jalons, anomalies
	Outils Devops (SonarQube, Jenkins)

Exigences techniques et de sécurité
Sécurité

Brique	Principales exigences
Sécurité	Authentification forte (selon les exigences du palier 2 - PGSSI-S) cascadée sur toutes les couches de la solution
	Certificat management ayant la capacité d'importer des PKI d'entreprises externes et certifié critères communs > Quelle étendue de l'application des critères communs sur l'ensemble de l'offre de service ?
	Gestion des accès à la donnée selon les droits attribués à l'utilisateur et les politiques définies au niveau fichier, méta-données, champs
	Journalisation complète des activités (voir palier 3 de la PGSSI-S ci après), y compris celles des administrateurs, sur tous les niveaux de profondeur de l'infrastructure et des applicatifs
	Possibilité de transmettre quotidiennement des traces vers une infrastructure extérieure à celle du hub
	Scellement quotidien des traces et transmission de la preuve de scellement à un tiers de confiance
	Solution d'exploitation et d'analyse des traces
	Dispositif de détection d'intrusion et d'attaques
	Possibilité de génération d'audit
	Fonctionnalité de chiffrement des données sensibles
	Liste des certifications de sécurité et conformité

Exigences techniques et de sécurité
Gouvernance de la donnée, administration / supervision

Brique	Principales exigences
Gouvernance de la donnée	Catalogue de données, tagging des fichiers, des méta-données, des champs
	Gestion du cycle de vie de la donnée : planification de l'alimentation, des traitements, du stockage, de l'archivage, et durée de rétention
Administration / Supervision	Ajout, retrait et modifications des utilisateurs, définition des droits et profils qui conditionnent les accès, définition d'un role base access
	Allocation et suivi des ressources (CPU, RAM, GPU) affectées à un utilisateur ou un traitement
	Administration et suivi de l'intégralité des composants de la plateforme depuis une interface de supervision unique

Exigences techniques et de sécurité
Prestations : Conception et intégration, Sécurité

Brique	Principales exigences
Conception et intégration	Définition de l'architecture technique et fonctionnelle cible du MVP
	Définition de la stratégie et du plan de test
	Installation, configuration, sécurisation et test de la plateforme
	Transfert de connaissances et accompagnement
Sécurité	Maintien en condition de sécurité et application des correctifs > Couverture dans le cas de la consommation d'un service IAAS
	Maintien en condition de sécurité et application des correctifs > Couverture dans le cas de la consommation d'un service PAAS

Focus : Authentification forte et PKI
Exigences

- En matière d'authentification, le SNDS impose d'être conforme aux exigences du palier 2 du Référentiel d'identification de la PGSSI-S : soit une authentification forte à multi facteurs.

- De plus, la PGSSI-S, indique que pour ce dispositif, l'utilisateur voulant accéder au système, utilise une bi-clé d'authentification (couple clé privée, clé publique) en provenance d'une PKI.

- La solution devra être un service sécurisé et résilient qui utilise des modules de sécurité matérielle validés **critères communs** pour protéger les clés.

- Les clés privées seront gérées et stockées par le ministère, le certificat management doit donc avoir la capacité d'importer des PKI d'entreprises externes

- Les journaux de toutes les utilisations clés devront être mis à disposition afin de répondre aux besoins en matière de réglementation et de conformité.
- Ces journaux devront pouvoir être envoyés vers un service (SIEM: Security Information and Event Management) d'analyse et de détection des menaces.

- Le MVP doit donc disposer de capacités :
    - D'authentification forte
    - De gestion des clés avec importation de PKI d'entreprises externes

Focus : Traçabilité
Exigences

- Les exigences de traçabilité (typologie, profondeur, journalisation) sont fixées par le SNDS et doivent être conformes au **palier 3** d'imputabilité décrit dans la PGSSI-S (détail ci après)

- Des solutions techniques doivent être mises en regard de chacune de ces exigences pour assurer la conformité de plateforme et son homologation. En particulier : Chaque composant applicatif doit journaliser avec l'utilisateur et par conséquent doit permettre l'authentification ou la cascade d'authentification afin de tracer les actions alignée sur les mêmes time stamp. Les logs sont tous remontés dans une console centrale. Il faut prévoir une application pour reconstruire le parcours utilisateurs
  Ex: Le notebook trace la connexion de l'utilisateur et l'exécution de code qui fait une requête à la base de données. La base de données reçoit une requête authentifiée et trace la requête et potentiellement les données (interdiction de réexécuter par la suite).

- Les traces sont transmises quotidiennement vers une infrastructure extérieure au hub

Focus : Traçabilité
Exigences de traçabilité du palier 3 de la PGSSI-S

Prérequis	Génération de piste d'audit	Conservation des traces	Restitution de la piste d'audit	Documentation spécifique
• Palier 1 du référentiel d'identification et d'authentification • Gestion dans le temps des identités, des rôles et des habilitations • Heure partagée par l'ensemble des composants du SIS	• Traces fonctionnelles : o type d'action o horodatage o Identité utilisateur o résultat (succès, erreur, refus) o données métiers concernées et traces embarquées (ex: identifiant du document) o version o contexte de réalisation, informations fournies (ex : message d'avertissement) o paramétrage technique de l'application • Traces techniques d'au moins un type de composant : o type d'action o horodatage o identité (utilisateur, machine , programme)	• Possibilité d'extraction des traces pour conservation dans des endroits multiples pour réduire le risque de modifications systémiques • Archives journalières regroupant l'ensemble des traces • Scellement quotidien des traces • Conservation des traces sur une durée glissante de 6 mois	• Outil de gestion permettant : o la restitution ergonomique des traces utilisable par des non spécialistes de la sécurité o la réconciliation des traces autant que de besoin o la gestion d'un format pivot ou gère de nombreux formats de traces • Guide didactique d'utilisation de l'outil de gestion de la preuve	• Documentation des dispositifs d'authentification, de gestion des identités, des rôles, des habilitations et des traces • Description des sources des traces et des processus mis en œuvre de la génération à la constitution de l'archive journalière • Description des processus mis en œuvre de la génération à la réconciliation