IA & Architecture des données : la révolution dans la gestion des données

Dr Yaya TRAORE
Secrétaire Permanent de l'Innovation et de la Veille sur les Technologies Emergentes du Numérique
Maître de conférences en Informatique UJKZ
Responsable de l'équipe de recherche IA-FDA / LAMI
Email : yaytra@ujkz.bf ; yaya.traore@tic.gov.bf

Introduction

COMMENT TRANSFORMER DES DONNEES BRUTES EN MOTEUR DE CROISSANCE GRACE A IA?

PLAN

1. Fondamentaux sur IA et données

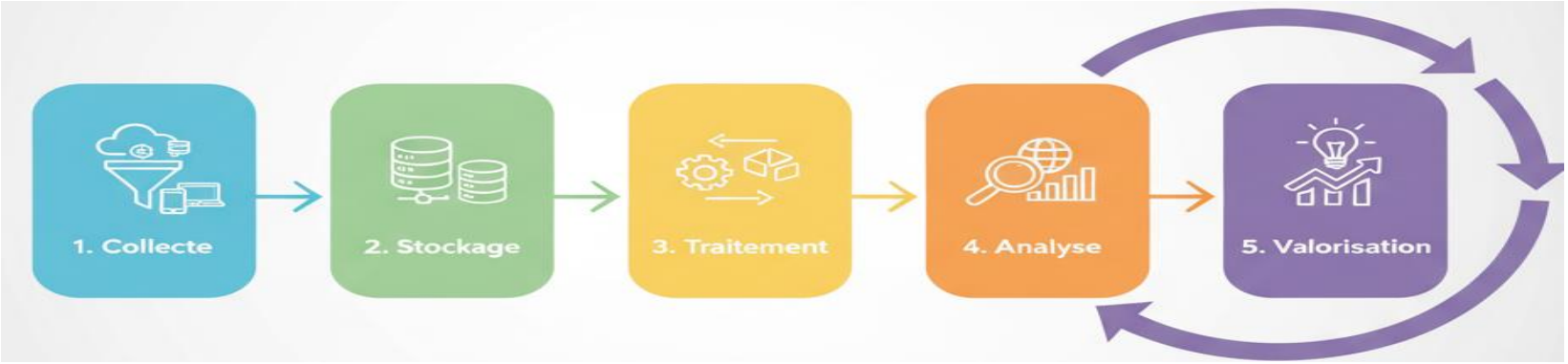
2. Principales architectures de données

3. IA et Transformation des Données

Fondamentaux sur IA et données

Qu’est-ce que la donnée?

- Les données sont des représentations brutes et non traitées de faits, d'observations, de mesures ou de descriptions
  - Elles constituent la matière première de l'information et de la connaissance dans le domaine de la data science
  - **Types de données en data science :** structurées, semi-structurées, non structurées
  - **Cycle de vie de la donnée :**



Fondamentaux : IA et Données

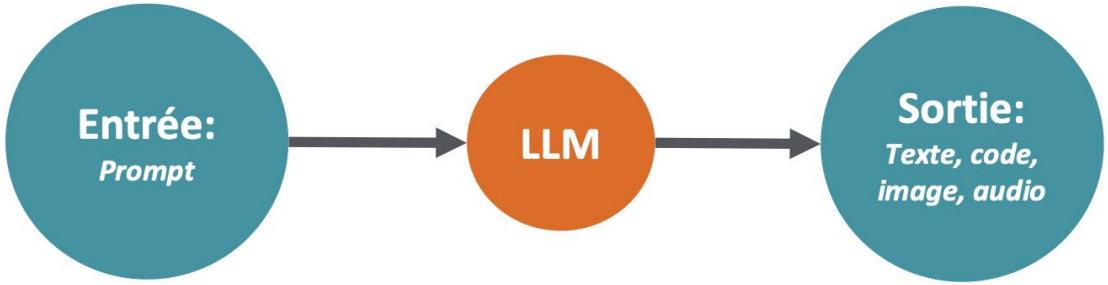
Qu'est-ce que l'intelligence artificielle ?

- Une technologie dont le but est de faire faire à une machine des tâches que l'homme accomplit en utilisant son intelligence. ***C'est un processus d'imitation de l'intelligence humaine .***
- Pour y parvenir, trois composants sont nécessaires :
    - des systèmes de calcul informatiques (infrastructure robuste d'hébergement et de traitement),
    - des données (collectées et de qualité),
    - des algorithmes (qui seront mis en œuvre dans des programmes informatiques).
- ***Données = Carburant de l'IA : sans données de qualité, pas d'IA performante***

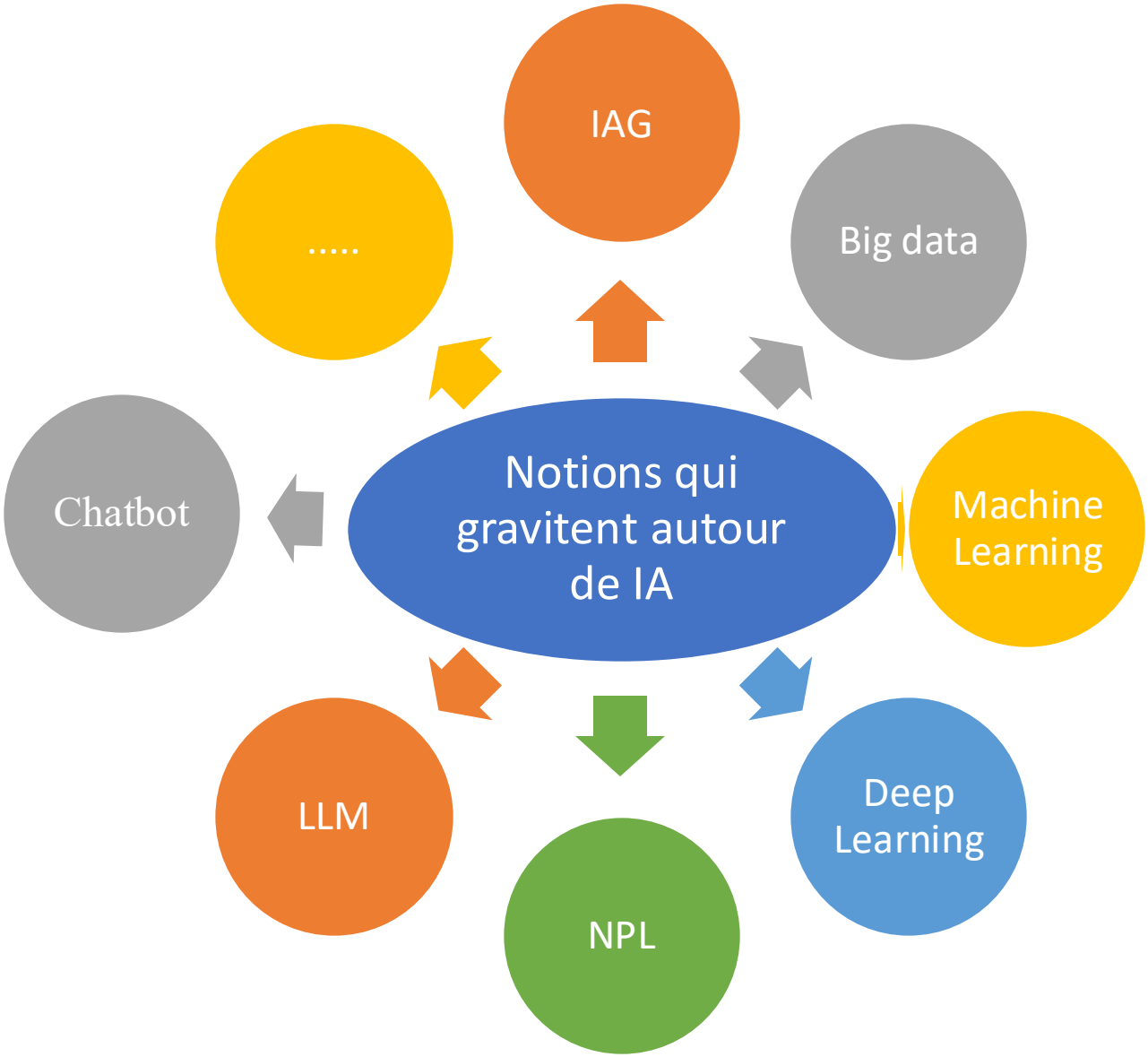
Pour se rapprocher le plus possible du comportement humain, l'intelligence artificielle a besoin d'une quantité de données et d'une capacité de traitement élevées.

Qu'est-ce que l'intelligence artificielle générative ?

- L'intelligence artificielle générative (IAG) est une IA capable de générer du contenu de manière cohérente en fonction des informations précédentes en s'associant aux modèles de large language models (LLM) – capacité prédictive



Fondamentaux sur IA et Données



Quelques outils python pour implémenter IA :

Principales architectures des données

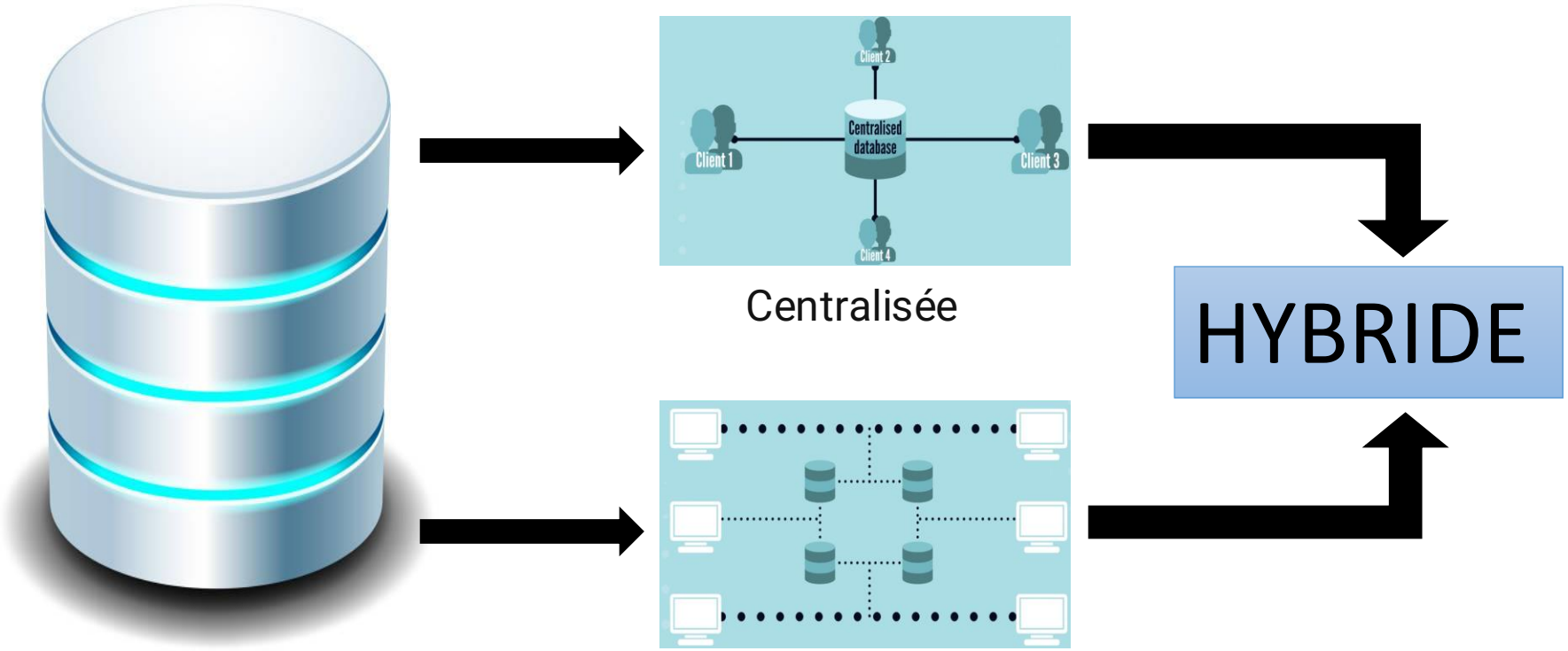
Principales architectures des données

Architecture des données :

- **Une architecture de données ?**
  - définit la structure, les processus, les technologies et les normes qui régissent la collecte, le stockage, la transformation, la distribution et la consommation des données au sein d'une organisation
  - vise à optimiser l'accès, la qualité, la sécurité et l'interopérabilité des données pour répondre aux besoins stratégiques et opérationnels
  - constitue la base des opérations de traitement des données et des applications d'intelligence artificielle (IA)

Principales architectures des données

Principaux type d'architecture :



- **Déploiement :**
  - sur site,
  - dans le cloud,
  - ou une combinaison des deux.

Principales architectures des données

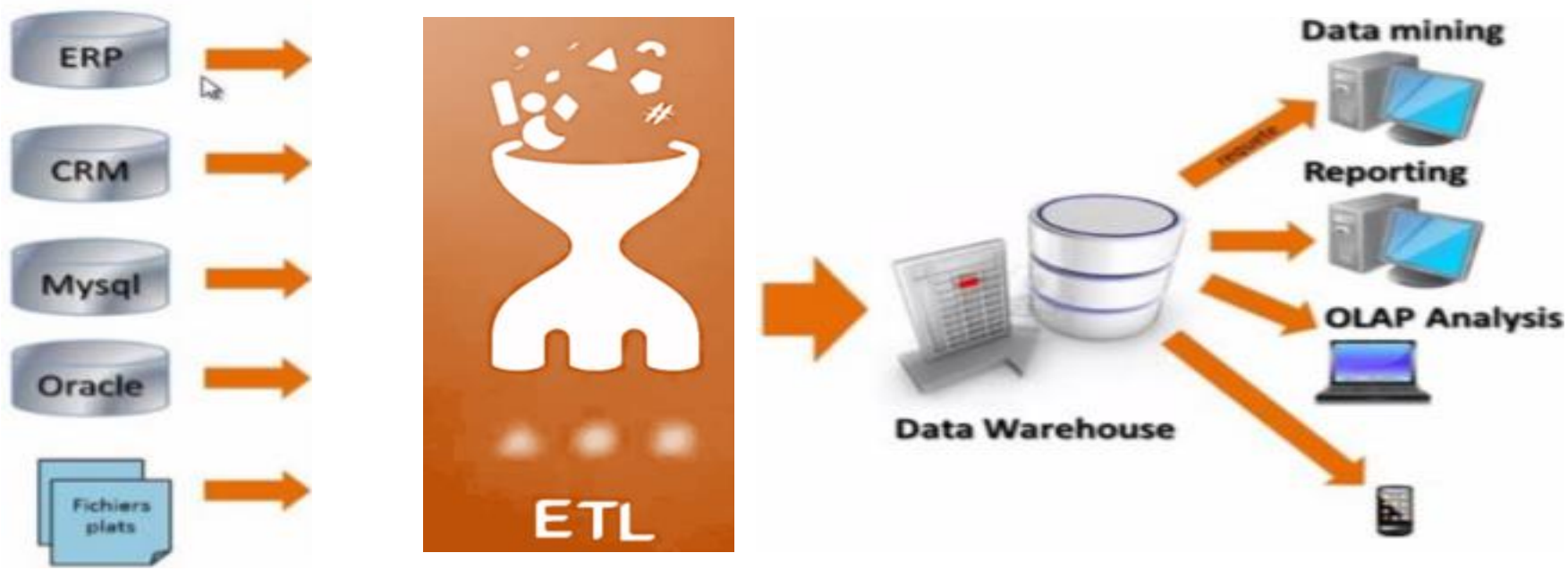
Datawarehouse ou entrepôt de données:

- Historiquement les bd soutenaient des besoins opérationnels, ce sont les systèmes OLTP
- Au fil du temps des défis d'historisation, de consolidation et d'analyste de ces données ont émergés : apparition systèmes analytiques : OLAP

OnLine Transaction Processing (OLTP)	OnLine Analytical Processing (OLAP)
Donnée Opérationlles Données courantes	Données analytiques Données historiques
Données fréquement mises à jours: optimisé pour des opérations de lecture et d'écriture	Optimisé pour des opérations de lecture
Une seule source de données Traitements impliquant peu de données à chaque opération	Plusieurs sources de données Traitements impliquant beaucoup de données

Principales architectures des données

Datawarehouse ou entrepôt de données:

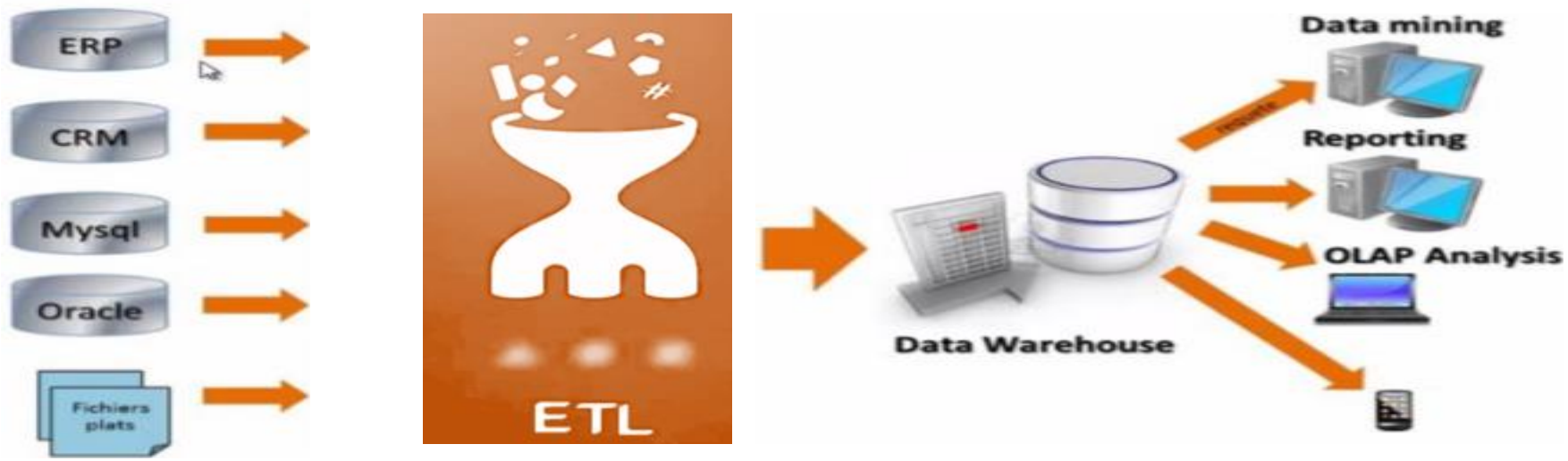


- Un datawarehouse est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décisions
- Outil open : Talend, DuckDB (https://duckdb.org/),
- Limite : utilise uniquement des données structurées

Datawarehouse ou entrepôt de données:

Exemple : A partir des données des centres de santé :

- **Besoin 1 :** produire en temps réel un tableau de bord sur les consultations (nombre de ptients par maladies, nombre de consultations par centre, nombre d'examen, ……)
- **Besoin 2 :** détecter les corrélations entres les symptômes des maladies



- **Besoins 3 : Faire une analyse prédictive pour le diagnostic d'une maladie ?**

Principales architectures des données

DataLake ou Lac de données :

- réponse au enjeux naissant liés à l'apparition du Big data caractérisé par les 5V :

- **Volume :** Quantité de données générées

- **Vélocité :** Vitesse de production et de traitement

- **Variété :** Diversité des formats et sources

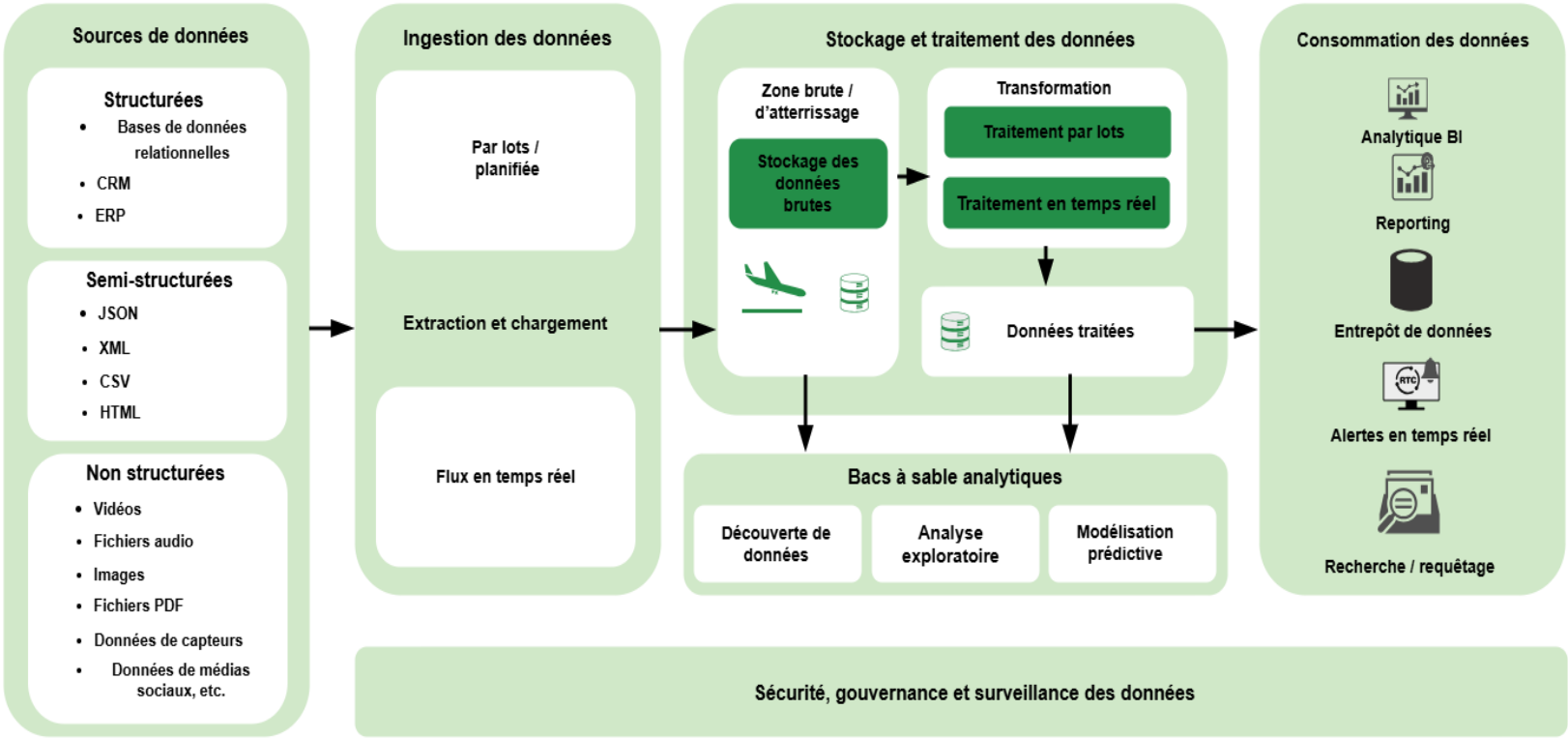
- **Véracité :** Fiabilité et qualité des données

- **Valeur :** Utilité des données pour la prise de décision

Principales architectures des données

Data Lake ou Lac de données :

- réponse au enjeux naissant liés à l'apparition du Big data caractérisé par les 5V

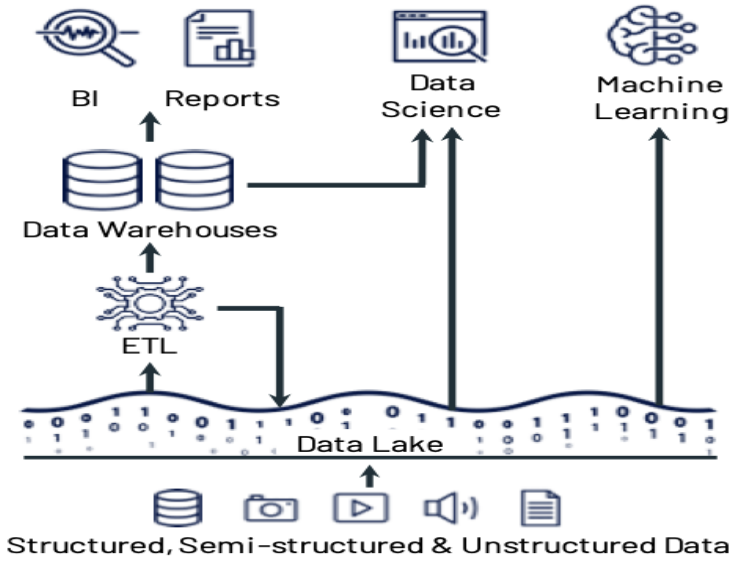


- Un lac de données (ou data lake) est une architecture de stockage centralisée conçue pour stocker de **vastes volumes de données brutes** dans leur format natif, qu'elles soient structurées, semi-structurées ou non structurées.
- **Limite : Complexité de traitement, gouvernance, Performance des requêtes**

Principales architectures des données

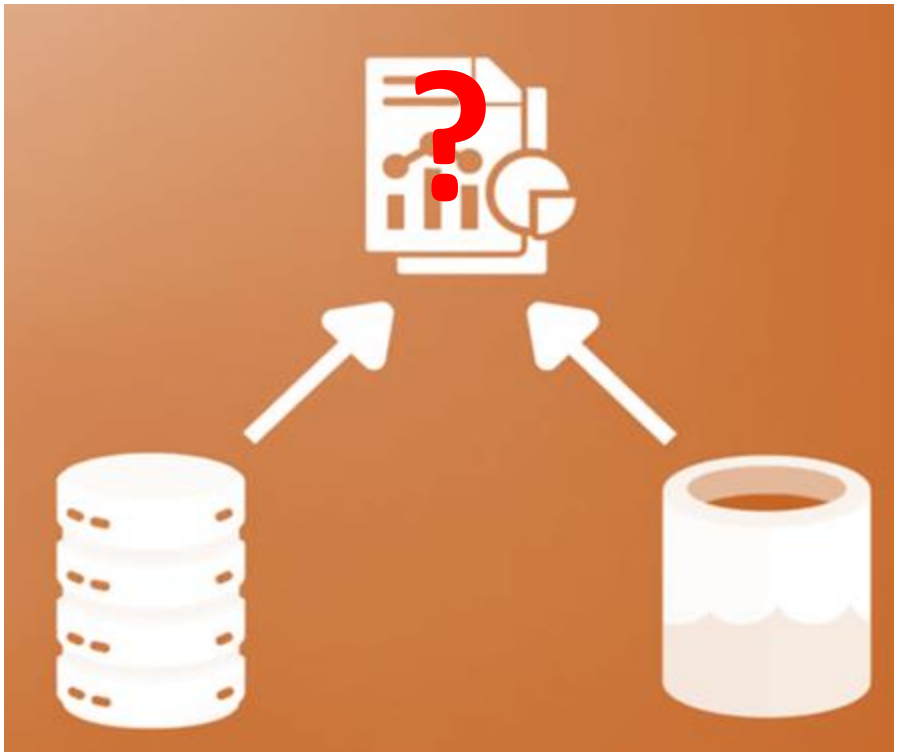
DataLake ou Lac de données :

- **Exemple d'utilisation :**
  - Un hôpital peut stocker dans un lac de données des dossiers médicaux, des images de radiologie, des données de capteurs IoT et des notes cliniques. Ces données peuvent ensuite être traitées pour des analyses prédictives ou des diagnostics assistés par l'IA.



Data LakeHouse :

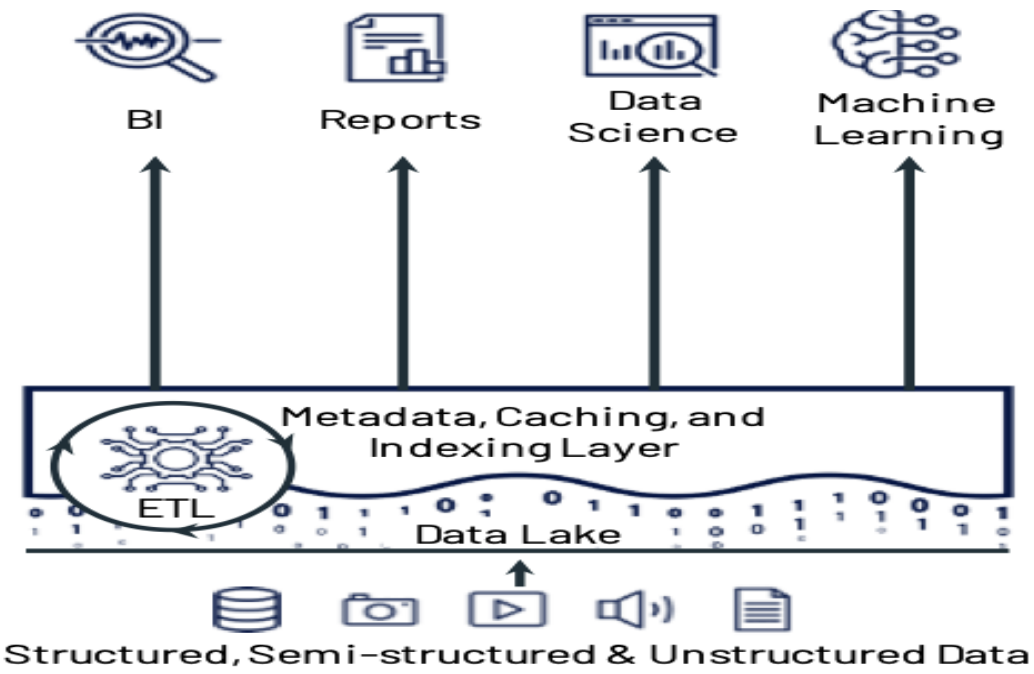
- réponse aux besoins d'analyser des rapports croisant des données issues d'un dataWarehouse et d'un data Lake



DataWarehouse Data Lake

Principales architectures des données

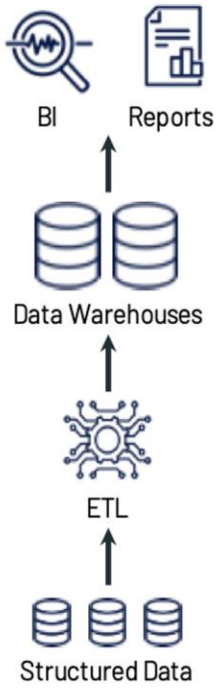
Data Lakehouse :

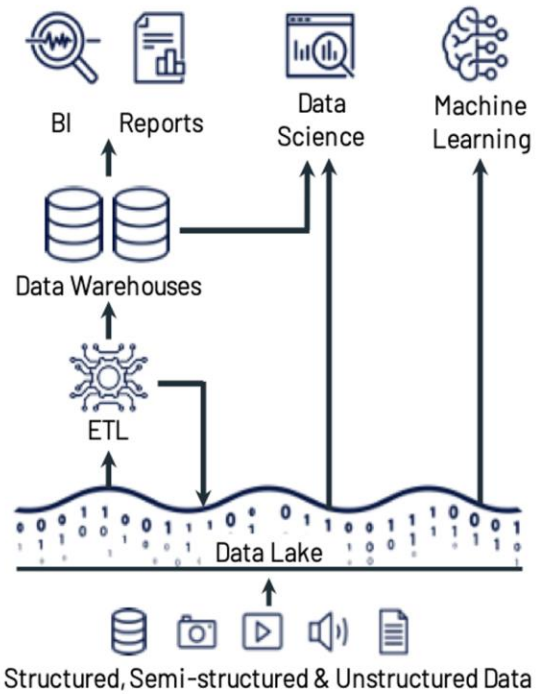


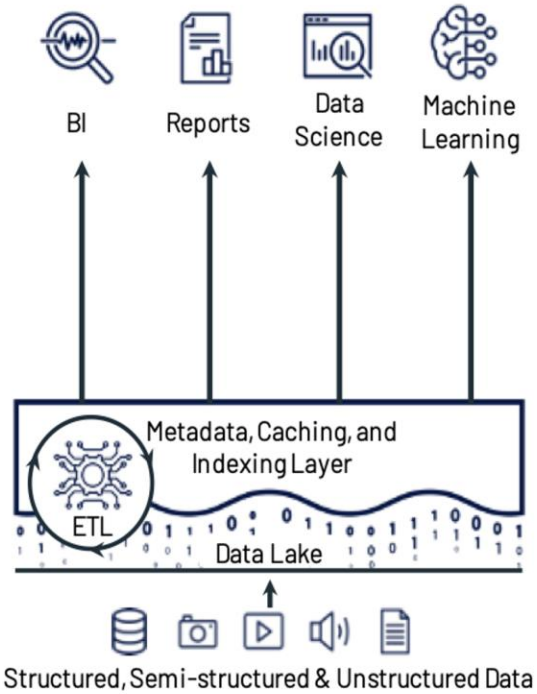
- Un data lakehouse est une architecture de gestion de données qui combine les meilleures caractéristiques d'un data lake et d'un dataWarehouse
- Il permet de stocker des données brutes (structurées, semi-structurées ou non structurées) comme un data lake, tout en offrant des capacités d'analyse et de gestion transactionnelle avancées typiques d'un data warehouse
- **Défis :** les data lakehouses sont complexes à construire de A à Z et doivent être étroitement associées aux fonctionnalités d'IA

Principales architectures des données

Différence entre DataWareHouse, DataLake et LakeHouse :







Data Warehouse :
les données sont structurées, filtrées et agrégées en amont pour répondre à des besoins d'analyse identifiés

Data Lake :
adapté à la data science et à la découverte de nouvelles informations alors qu'à l'inverse le Data Warehouse répond à des questions business connues

Data Lakehouse :
Appliquer au data Lake certains principes du Data Warehouse en matière de gouvernance, de sécurité et de gestion des métadonnées.

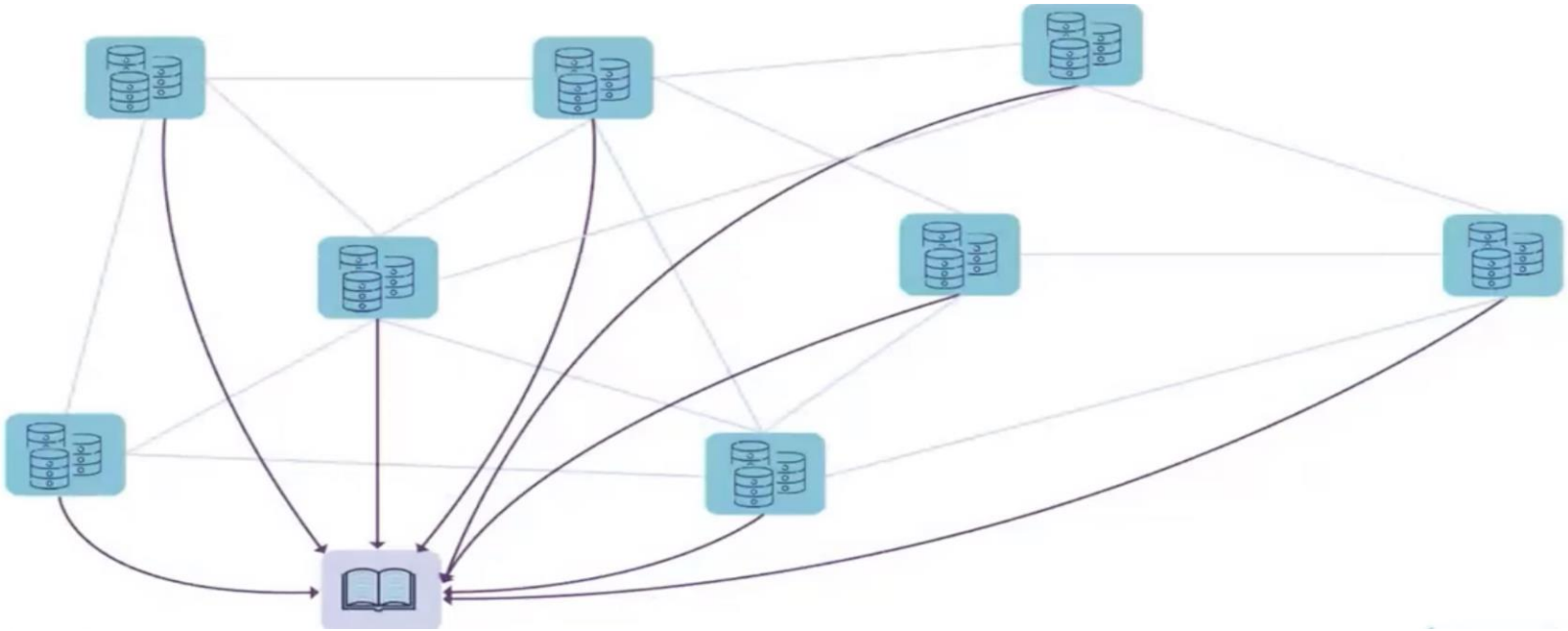
Data Lakehouse

Outils Open source pour implémenter un Data Lakehouse

- **Stockage** : Utiliser **Delta Lake**, **Apache Iceberg** ou **Apache Hudi** pour stocker les données dans un Data Lake
- **Traitement** : Utiliser **Apache Spark** ou **Apache Flink** pour le traitement batch et streaming
- **Requêtage** : Utiliser **Trino** ou **Apache Hive** pour exécuter des requêtes SQL sur les données
- **Orchestration** : Utiliser **Apache Airflow** pour planifier et gérer les workflows
- **Catalogage** : Utiliser **Apache Atlas** ou **DataHub** pour gérer les métadonnées
- **Visualisation** : Utiliser **Apache Superset** ou **Metabase** pour visualiser les données

Principales architectures des données

Data Mesh:



- data Mesh décentralise la propriété des données en alignant l'architecture sur les domaines d'activité. Il encourage les producteurs de données, ceux qui sont les plus proches de la source, à traiter les données comme un produit et à concevoir des API orientées consommateur.
- Ce modèle permet d'éliminer les goulots d'étranglement et de favoriser la démocratisation des données évolutives à l'échelle de l'entreprise.

Avantages des architectures des données :

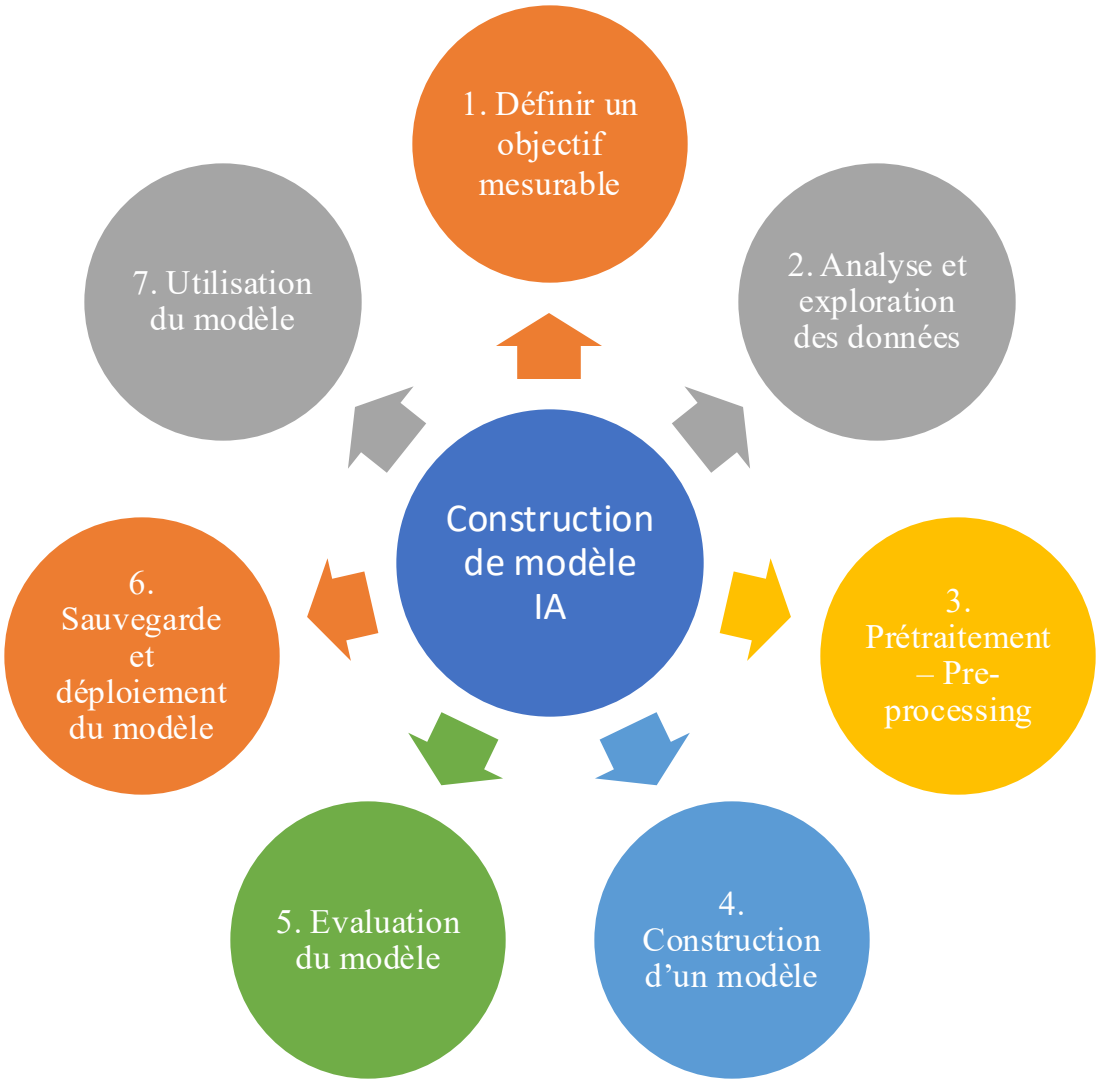
- Une architecture de données bien construite offre des avantages considérables aux entreprises, notamment :

IA et Transformation des données

IA et Transformation des données

Les différentes étapes pour construire un modèle à partir des données :

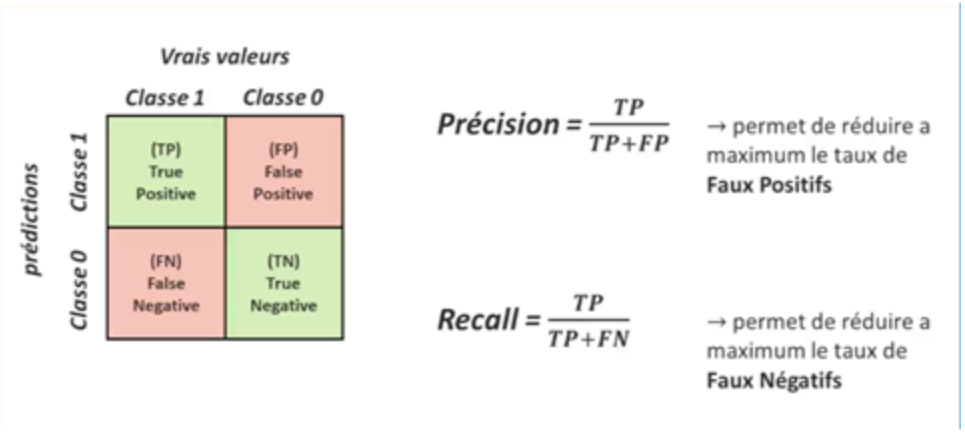
- préalable : les données sont disponibles et collectées



1. Définir un objectif mesurable

- **Objectif :** Prédire si une personne est infectée en fonction des données cliniques disponible.
- **Métrique :** Accuracy : 90%

- **Métrique :** Précision : 60%, Rappel (sensibilité) : 70%, F1 : 50%



2. Analyse et Exploration de nos Données (EDA = Exploratory Data Analysis)

- **Objectif :** Comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.
  - **Analyse de la forme :** Identification de la target, Nombre de lignes et de colonnes, Types de variables, Identification des valeurs manquantes,……..
  - **Analyse du fond :** visualisation de la target (histogramme/ Boxplot), compréhension des différentes variables, visualisation des relations features-target (histogramme/ Boxplot), identification des outliers

3. Pétraitement (Pre-processing)

- **Objectif :** Transformer le data pour le mettre dans un format propice au Machine Learning
  - Checklist de base
    - Elimination des NaN : dropna(), imputation, colonnes vides,
    - Encodage,
    - Suppression des outliers néfastes au modèle,
    - Feature Selection,
    - etc

4. Construction du modèle

- **Objectif :** Développer un modèle de machine learning qui réponde à l'objectif final.
  - Checklist de base
    - Création du Train set / Test Set
    - Choisir le bon algorithme :
      - Arbre de décision
      - Random Forest
      - Perceptron
      - Support Vector Machine (SVM)
      - CNN
      - etc

4. Evaluation du modèle

- **Objectif :** Evaluer la qualité du modèle avec les métrique de précision, rappel, accuracy, …….
  - Checklist de base
    - Entraîner et évaluer le modèle
    - Tester le modèle

4. Sauvegarder le modèle

- **Objectif :** sauvegarder le modèle avec la qualité du modèle avec les métrique de précision, rappel, accuracy, …….
  - Checklist de base
    - évaluer la performance du modèle avec un test réel
    - sauvegarder le modèle

IA et Transformation des données

- **Python** : langage de programmation en data sciences

- **Scikit-learn** est une bibliothèque Python libre et Open Source destinée à l'apprentissage automatique

- **Pandas** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données

- **Jupyter** est une application web open source pour partager du code python

- *Lien utile IA et données pour commencer : https://www.youtube.com/watch?v=x8yu8sq8mdw*

Conclusion

- **Bénéfice :** L'intégration de l'IA dans la gestion des données permettent aux entreprises de **maximiser l'utilisation de leurs données**, tout en réduisant les risques et en améliorant leur compétitivité

Catégorie	Bénéfice	Impact
Gains opérationnels	**Automatisation**	Réduction de 70-80% du temps de préparation manuelle.
	Scalabilité	Traitement de volumes massifs sans intervention humaine.
	Rapidité	Transformation en temps réel de millions d'enregistrements.
	Cohérence	Application uniforme des règles de transformation.
Amélioration de la qualité	**Précision**	Détection des anomalies et erreurs.
	Complétude	Imputation intelligente des données manquantes.
	Données à jour	Mise à jour continue via monitoring ML.
	Fiabilité	Validation continue de la qualité.
Valeur métier	**Coûts**	Réduction des coûts opérationnels de 40-60%.
	Innovation	Libération des data engineers pour des tâches à plus forte valeur.
	Conformité	Respect automatique des réglementations.

Conclusion

- **Défis** de l'intégration de l'IA dans la gestion des données

Défis	Enjeux	Impacts
Intégration des Technologies	Hétérogénéité des systèmes : une variété d'outils et de technologies (Data Lakes, Data Warehouses, bases de données SQL/NoSQL, outils d'IA, etc.). L'interopérabilité entre les outils	Difficulté à intégrer des outils variés (Data Lakes, Data Warehouses, outils d'IA) en une architecture cohérente
Qualité et Préparation des Données	Les données brutes sont souvent incomplètes, incohérentes ou biaisées. Leur nettoyage et leur normalisation sont essentiels pour garantir des résultats fiables	Des données de mauvaise qualité entraînent des analyses erronées et des décisions inefficaces
Intégration et Scalabilité	Intégrer des outils variés (Data Lakes, Data Warehouses, IA) et assurer la scalabilité pour gérer des volumes croissants de données	Une mauvaise intégration ou scalabilité limite la performance et l'efficacité des systèmes.
Sécurité et Conformité	Sécuriser les données et respecter les réglementations (RGPD, etc.).	Le non-respect des réglementations entraîne des risques juridiques et de réputation.
Compétences et Gouvernance	Manque de talents qualifiés et nécessité d'une gouvernance robuste pour gérer les métadonnées, la sécurité et la conformité.	Sans compétences adéquates et gouvernance, les projets d'IA et de données risquent l'échec

Conclusion

En résumé :

- l'IA révolutionne la gestion des données en rendant les processus plus **rapides, plus précis, plus économiques et conformes, tout en libérant du temps pour l'innovation**.
- les données alimentent les modèles d'IA, les modèles d'IA génèrent des actions/décisions, et ces actions/décisions créent de la valeur, générant à leur tour plus de données pour affiner les modèles.

Merci

pour votre aimable attention !