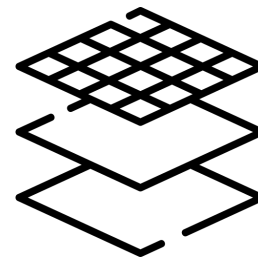
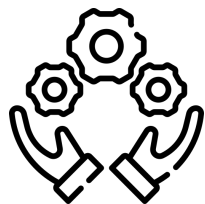


# Architecture et Gouvernance des données

*Niveau Master 1*



# Objectifs pédagogiques ([cf. IA School](#))

- Comprendre les **concepts fondamentaux** de l'architecture et de la gouvernance des données (type de données, data warehouse...)
- Maîtriser les techniques de **modélisation des données** (MCD, MLD).
- Connaître les différentes **architectures de données** (centralisée, décentralisée, distribuée).
- Identifier les **enjeux et les défis** de la gestion des données dans les organisations (RGPD, cloud, big data...)

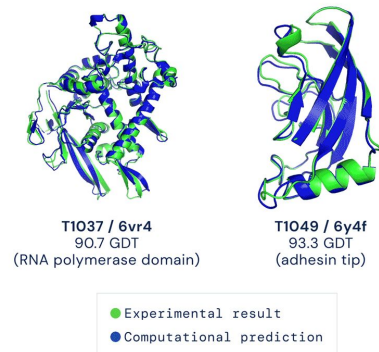
# Plan séance 1

- Introduction et définitions
- Modélisation des données
  - MCD
  - MLD
  - Tutoriel et exercices
- Architectures de données classiques
  - Centralisée
  - Décentralisée
  - Distribuée
  - Étude de cas

# Votre formateur

Mehdi MUNIM

- Background data appliquée à la biologie (**AlphaFold**)
- 2 ans XP en enseignement intro IA / data science
  - Universités (UPCité, UPEC)
  - Cité des sciences
  - École de commerces
  - ESIEE-IT
  - **IA School (machine learning, probabilités, python...)**
  - IA générative / ChatGPT pour l'entreprise



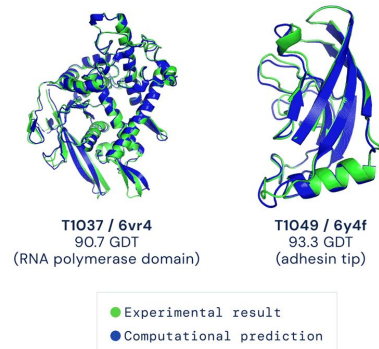
# Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

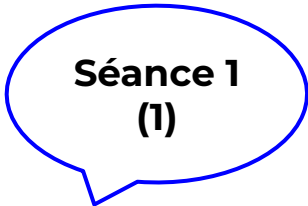
**This year's prize celebrates computational tools that have transformed biology and have the potential to revolutionize drug discovery.**

By [Ewen Callaway](#)



For the first time — and probably not the last — a scientific breakthrough enabled by artificial intelligence (AI) has been recognized with a Nobel prize. The 2024 chemistry Nobel was awarded to John Jumper and Demis Hassabis at Google DeepMind in London, for developing a game-changing [AI tool for predicting protein structures called AlphaFold](#), and David Baker, at the University of Washington in Seattle, for his work on computational protein design, which





**Séance 1  
(1)**

# Introduction

## **Un exemple pour commencer...**

# Le CHU de Nantes



## Un établissement de santé de référence :

- Plus de 13 000 professionnels
- Un acteur majeur de la santé en France et à l'international

## Objectifs :

- Améliorer la prise en charge des patients
- Accélérer la recherche médicale et l'innovation
- Promouvoir une utilisation éthique et responsable des données de santé





# Le Défi

En 2019, le CHU de Nantes a mis en place une politique de Gouvernance des Données pour améliorer la prise en charge des patients.

## Le défi:

- **Données dispersées** : Les informations sur les patients (imagerie médicale, analyses, traitements...) étaient stockées dans différents systèmes, ce qui rendait difficile l'accès et le partage des données entre les équipes médicales (30 millions de documents en 2019)
- **Manque d'harmonisation** : Les données n'étaient pas toujours structurées de la même manière, ce qui compliquait leur analyse et leur exploitation (code spécifique, noms des variables...)

Source :

<https://actus.nantes-saintnazaire.fr/article/sante-bigdata-clinique-donnees-interesse-hopitaux-francais>

# La Solution

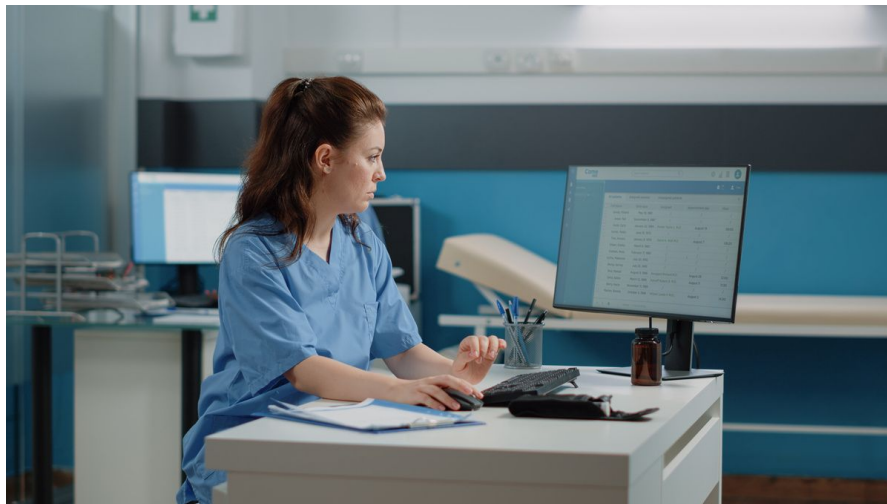
Création d'une **“clinique des données”** en plusieurs étapes:

- Inventaire des données disponibles, avec des définitions claires et des métadonnées.
- Centralisation des données patients dans un système unique (*data warehouse*)
- Assurer la fiabilité et la cohérence des données (*data quality*)
- Garantir la confidentialité des données patients et le respect des réglementations (RGPD).



# Le cas du CHU de Nantes - Les Résultats

**Amélioration de la prise en charge des patients :** Accès rapide et facile aux informations pertinentes pour les médecins, permettant une prise de décision plus éclairée.



# Le cas du CHU de Nantes - Les Résultats

**Amélioration de la prise en charge des patients :** Accès rapide et facile aux informations pertinentes pour les médecins, permettant une prise de décision plus éclairée.

**Recherche médicale facilitée :** Analyse des données pour identifier de nouveaux traitements et améliorer les protocoles de soins.



# Le cas du CHU de Nantes - Les Résultats

**Amélioration de la prise en charge des patients :** Accès rapide et facile aux informations pertinentes pour les médecins, permettant une prise de décision plus éclairée.

**Recherche médicale facilitée :** Analyse des données pour identifier de nouveaux traitements et améliorer les protocoles de soins.

**Optimisation des ressources :** Meilleure gestion des lits, des équipements et du personnel grâce à une meilleure connaissance des besoins.



# Extension “Ouest DataHub” (2020)

Source :

<https://actus.nantes-saintnazaire.fr/article/sante-bigdata-clinique-donnees-interesse-hopitaux-francais>



## NOTRE CLINIQUE DES DONNÉES INTÉRESSE BEAUCOUP LES AUTRES HÔPITAUX FRANÇAIS

Publié le 03/05/2021  
Santé / Biotech, Numérique, COVID-19



- **5** millions de patients
- **130** millions de documents
- **1,3** milliard de données structurées

# Un peu de théorie...

# Définitions

## **Donnée :**

Représentation brute d'un fait, d'une observation ou d'une mesure.

- 25 °C
- Paris
- [mehdi.munim.int@groupe-gema.com](mailto:mehdi.munim.int@groupe-gema.com)
- ...



# Définitions

## **Donnée :**

Représentation brute d'un fait, d'une observation ou d'une mesure.

## **Information :**

Donnée structurée et contextualisée, porteuse de sens.

⇒ Il fait 25°C à Paris

# Définitions

## **Donnée :**

Représentation brute d'un fait, d'une observation ou d'une mesure.

## **Information :**

Donnée structurée et contextualisée, porteuse de sens.

## **Système d'information :**

Ensemble organisé de ressources pour collecter, stocker, traiter et diffuser l'information.

# Définitions

**Exemples de SI** (Système d'information) :

- Une “clinique des données”
- Système E-commerce
- CRM (Salesforce...)
- Gestion des ressources humaines (SIRH)
- Réservation de billets d'avions

**⇒ Tout notre cours concerne les SI !**

# Architecture des données

## **Définition :**

Structure, organisation et intégration des données au sein d'un SI.

## **Permet :**

- la prise de décision éclairée
- la réduction des coûts
- la collaboration,
- l'innovation.

# Architecture des données

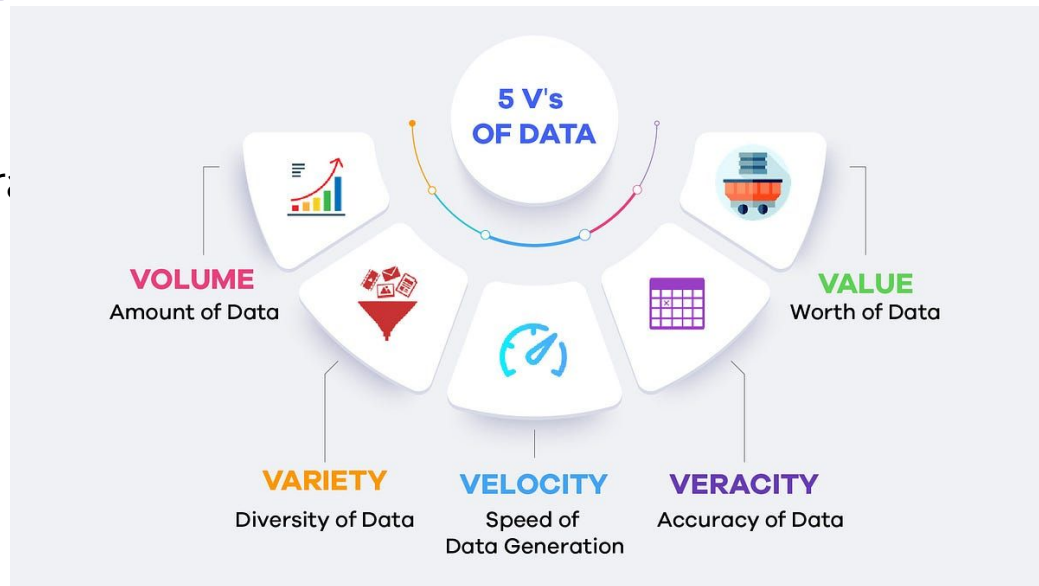
## Définition :

Structure, organisation et intégration

## Permet :

- la prise de décision éclairée
- la réduction des coûts
- la collaboration,
- l'innovation.

**Enjeux :** Volume, variété, véracité, sécurité des données



# Gouvernance des données

## **Définition :**

Ensemble de processus pour gérer les données de manière efficace et sécurisée.

## **Principes :**

Disponibilité, intégrité, confidentialité, traçabilité, conformité.

## **Objectifs :**

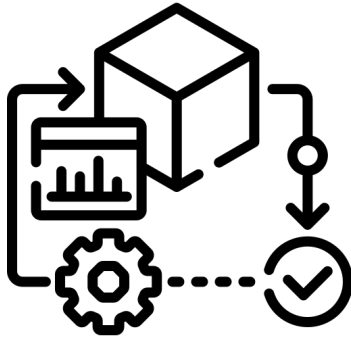
Qualité des données, réduction des risques, optimisation de l'utilisation, conformité réglementaire.



**Pour mieux comprendre les enjeux,  
rendez-vous ci dessous :**

**<https://bit.ly/4hpGWqV>**

# Modélisation des données





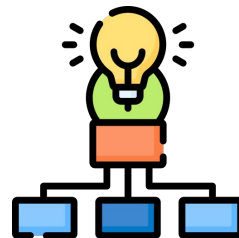
# Introduction à la modélisation des données

## Objectifs :

- Comprendre et représenter la structure des données.
- Faciliter la communication entre les acteurs d'un projet.
- Guider la conception de la base de données.
- Assurer la cohérence et l'intégrité des données.

## Types de modèles :

- **Conceptuel (MCD)** : Vue abstraite et globale des données.
- **Logique (MLD)** : Traduction du MCD en un modèle relationnel.
- **Physique (MPD)** : Adaptation du MLD à un SGBD spécifique.

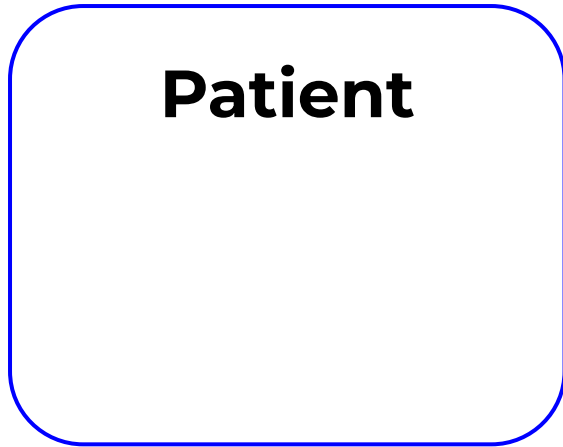


# Modèle conceptuel

*Entités, attributs, relations, cardinalités...*

# Entité

- **Entité** : Objet, concept ou événement du monde réel. (ex: Patient, Médecin)
  - Représentée par un rectangle.



*On se place dans le cadre d'un SI, dans notre cas : un système de suivi des patients*

# Attribut

- **Attribut** : Caractéristique d'une entité. (ex: Prénom, Nom, Âge...)
  - Représenté par une liste d'items

## Patient

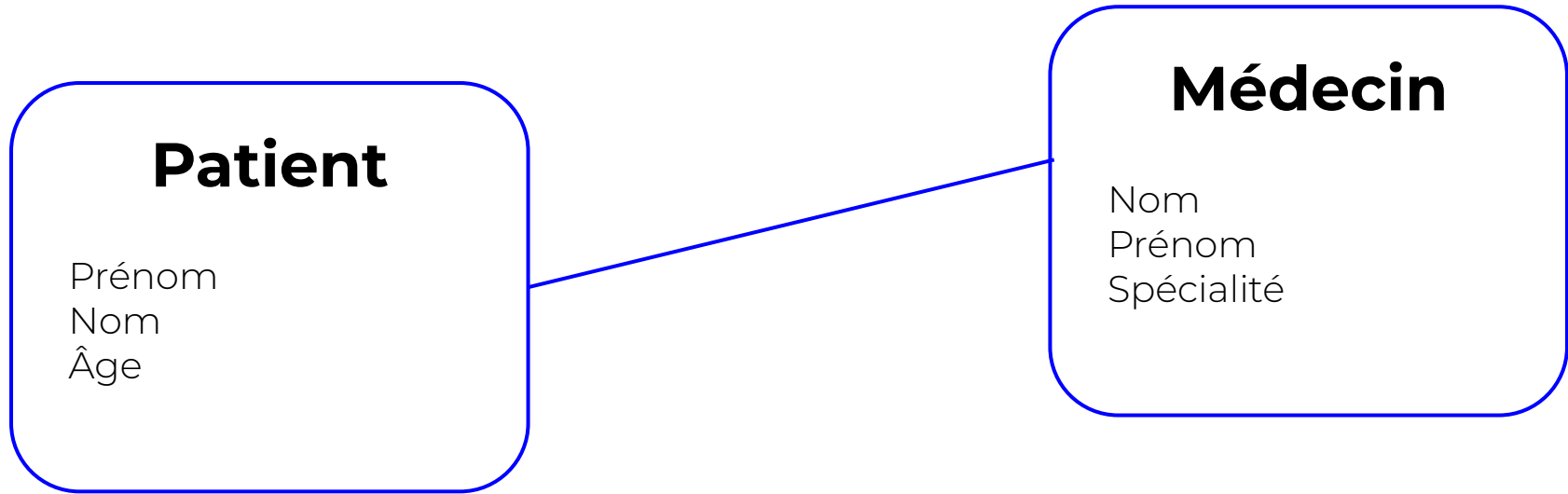
Prénom

Nom

Âge

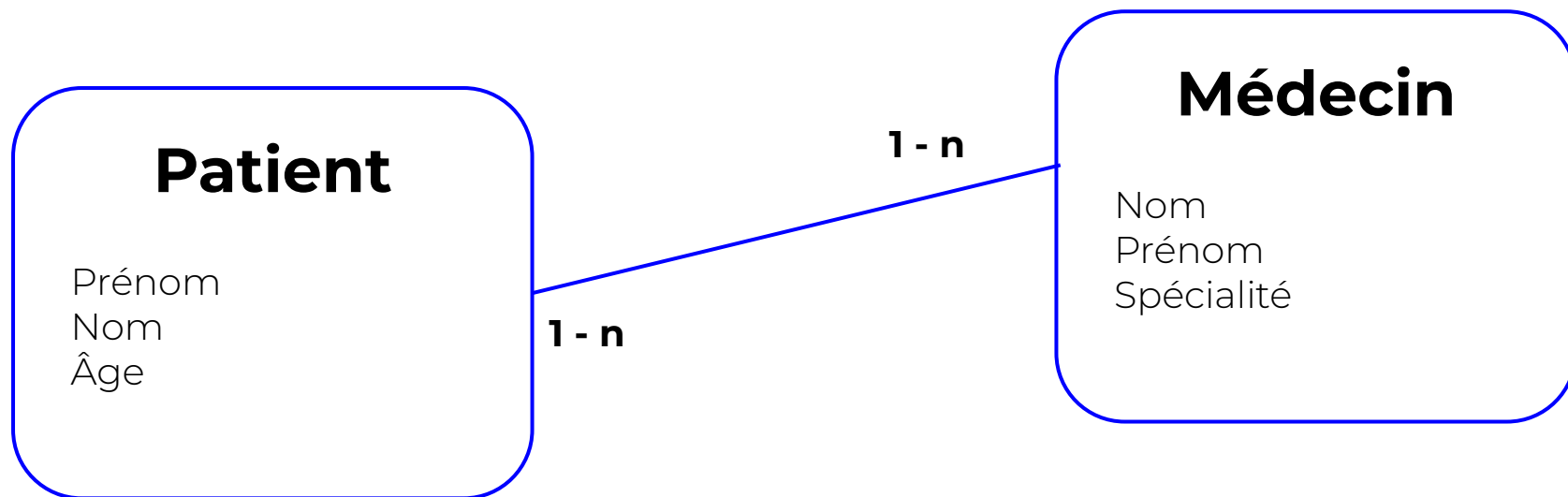
# Relation

- **Relation** : Lien entre deux entités. (ex: "un patient est suivi par un médecin")
  - Représentée par un trait.



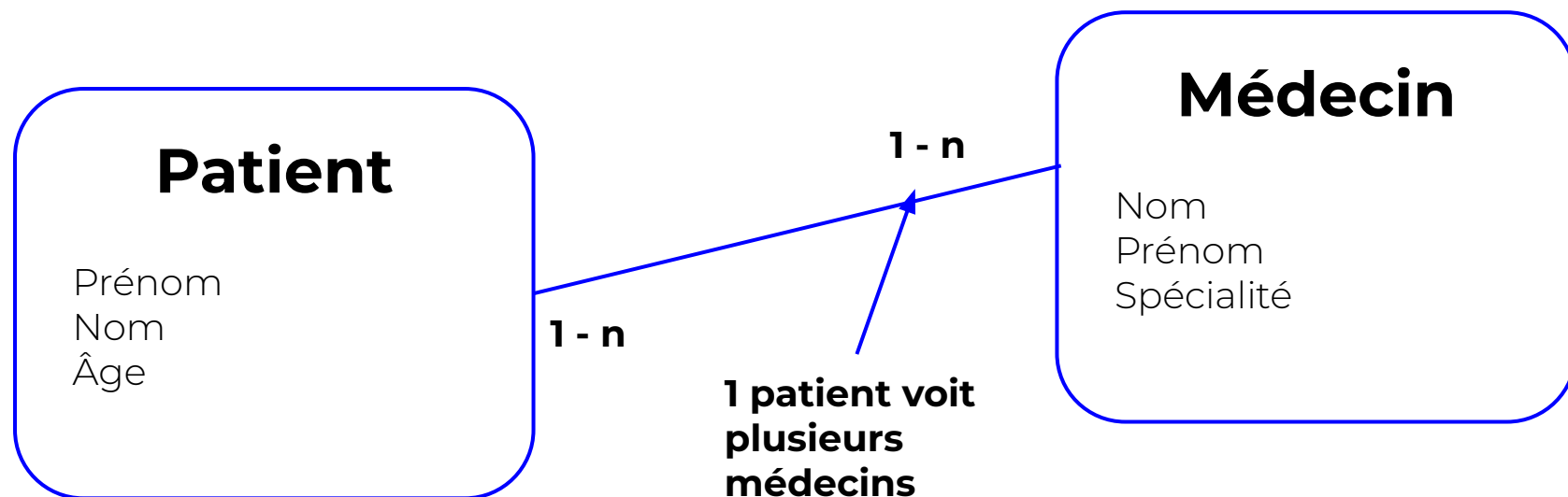
# Cardinalité

- **Cardinalités** : Nombre d'instances d'une entité liées à une autre.



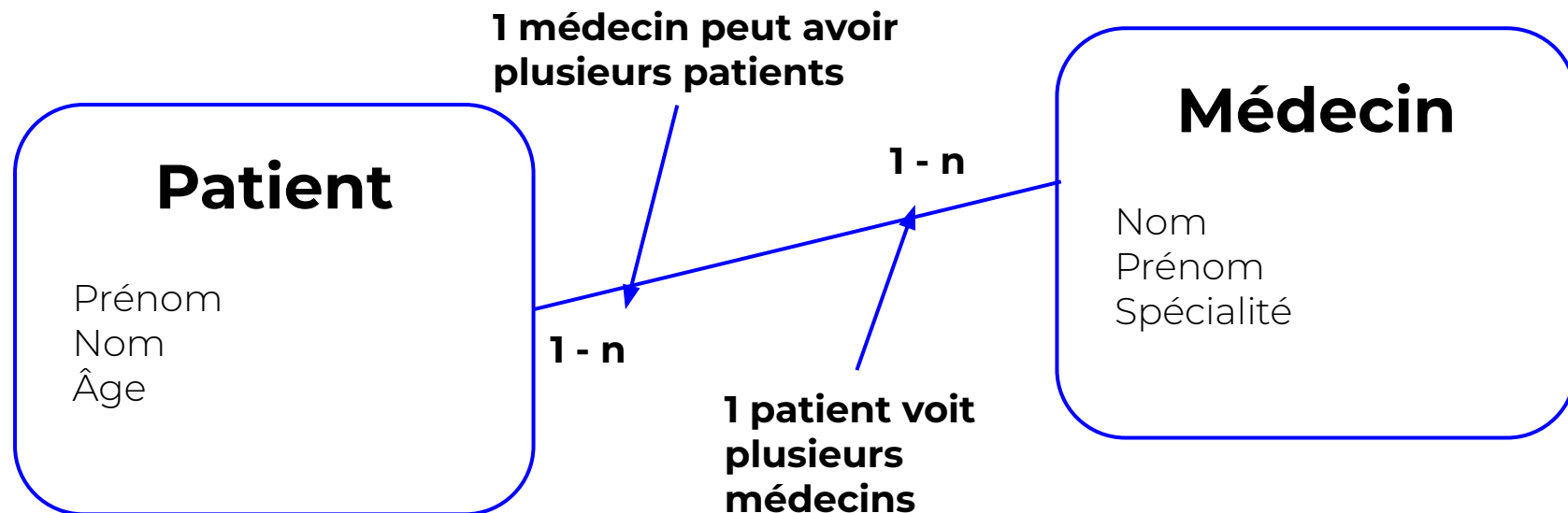
# Cardinalité

- **Cardinalités** : Nombre d'instances d'une entité liées à une autre.

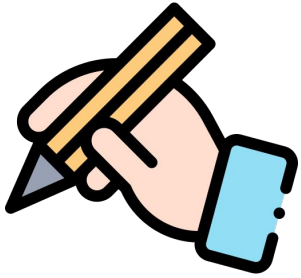


# Cardinalité

- **Cardinalités** : Nombre d'instances d'une entité liées à une autre.







## Activité : Modélisation d'une base de données

<https://bit.ly/4fvmqU7>

# Modèle logique

*Clés primaires, étrangère, jointure...*

# Du modèle conceptuel au modèle logique

## Rappel du MCD :

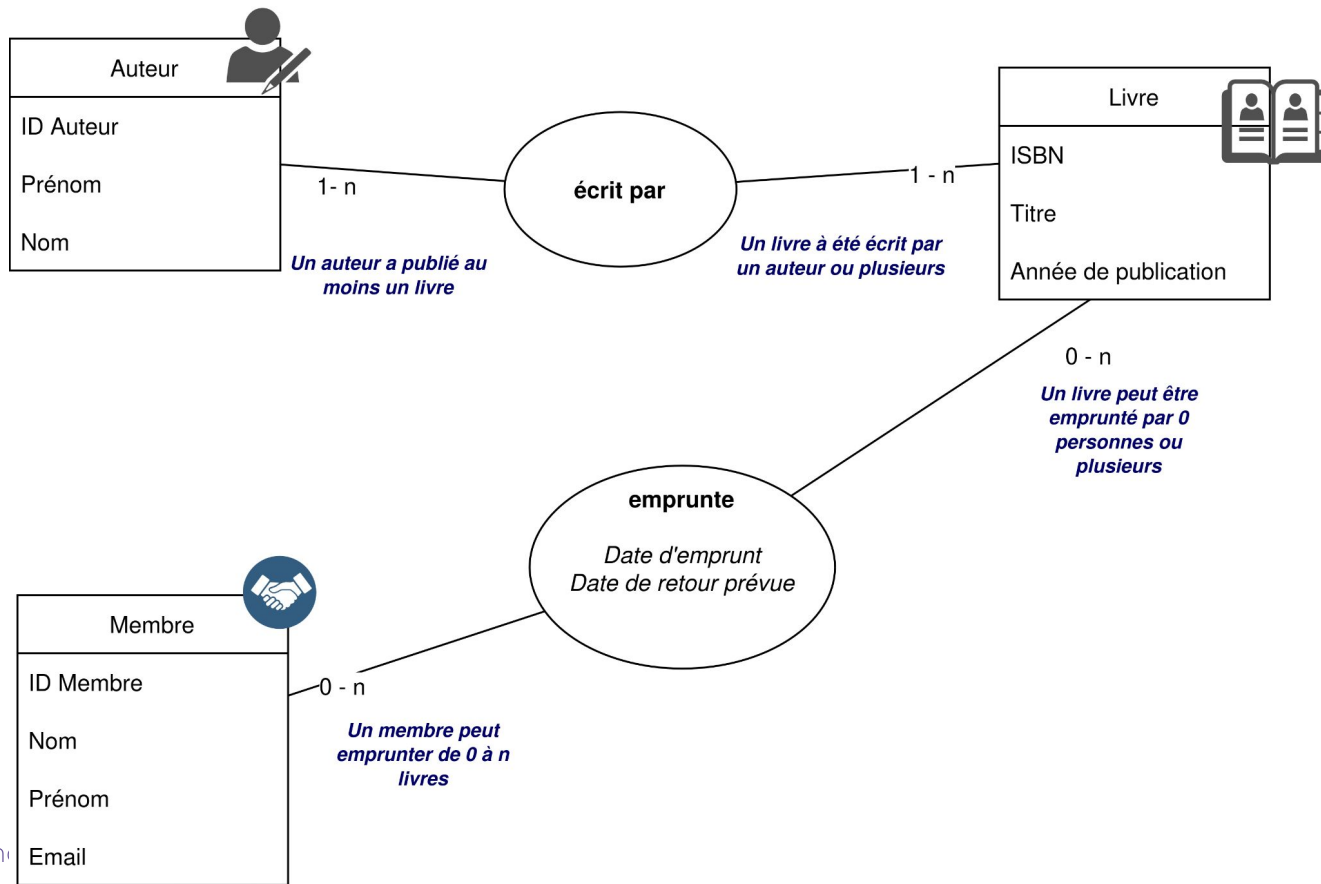
- Le MCD décrit les entités, les attributs et les relations de manière abstraite.
- Il est indépendant de toute considération technique (SGBD).

## Objectif du MLD :

- Traduire le MCD en un modèle relationnel, compréhensible par un SGBD.
- Définir les tables, les colonnes, les clés primaires et étrangères.
- Préparer la mise en œuvre physique de la base de données.

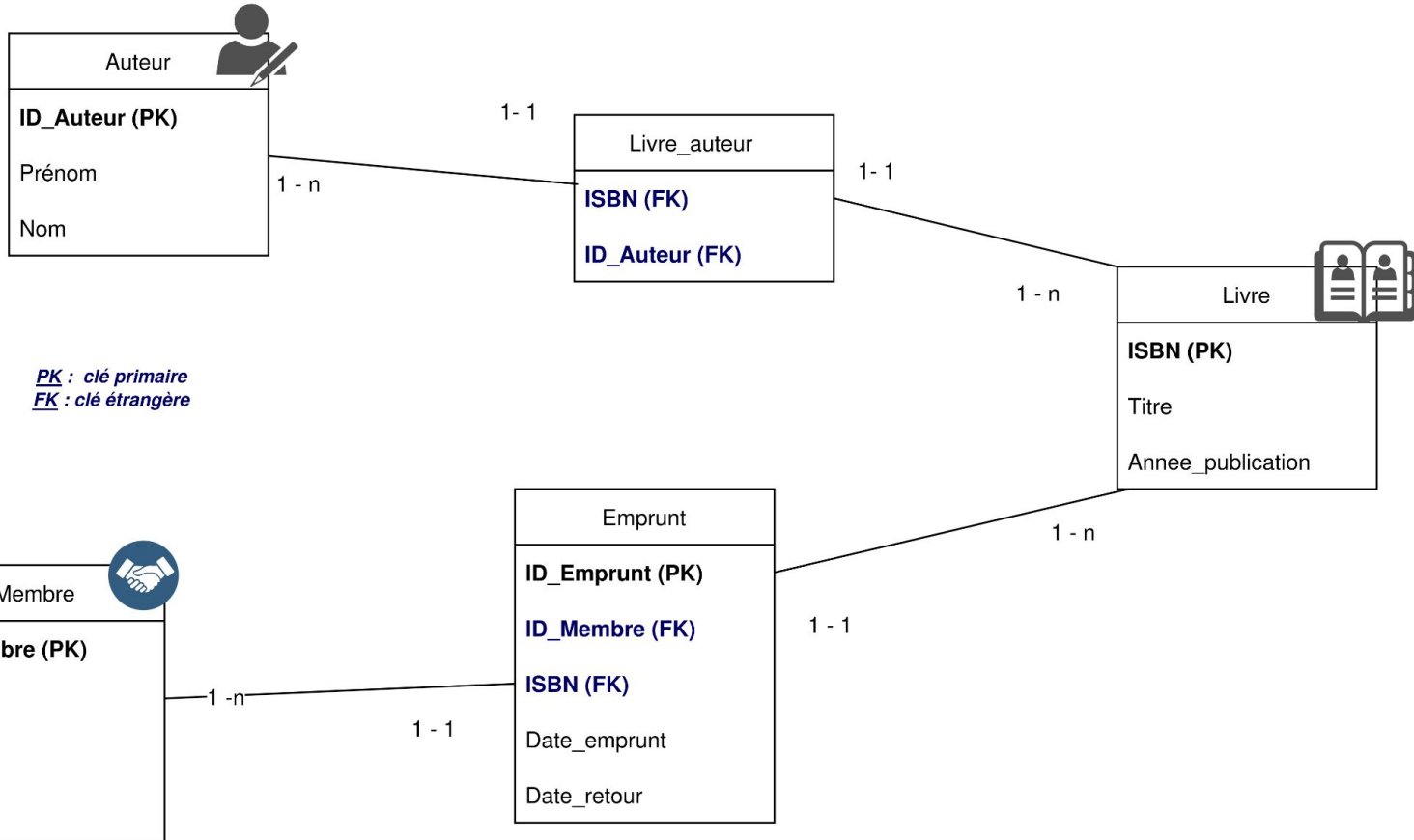
# Modèle conceptuel (MCD)

## BIBLIOTHEQUE



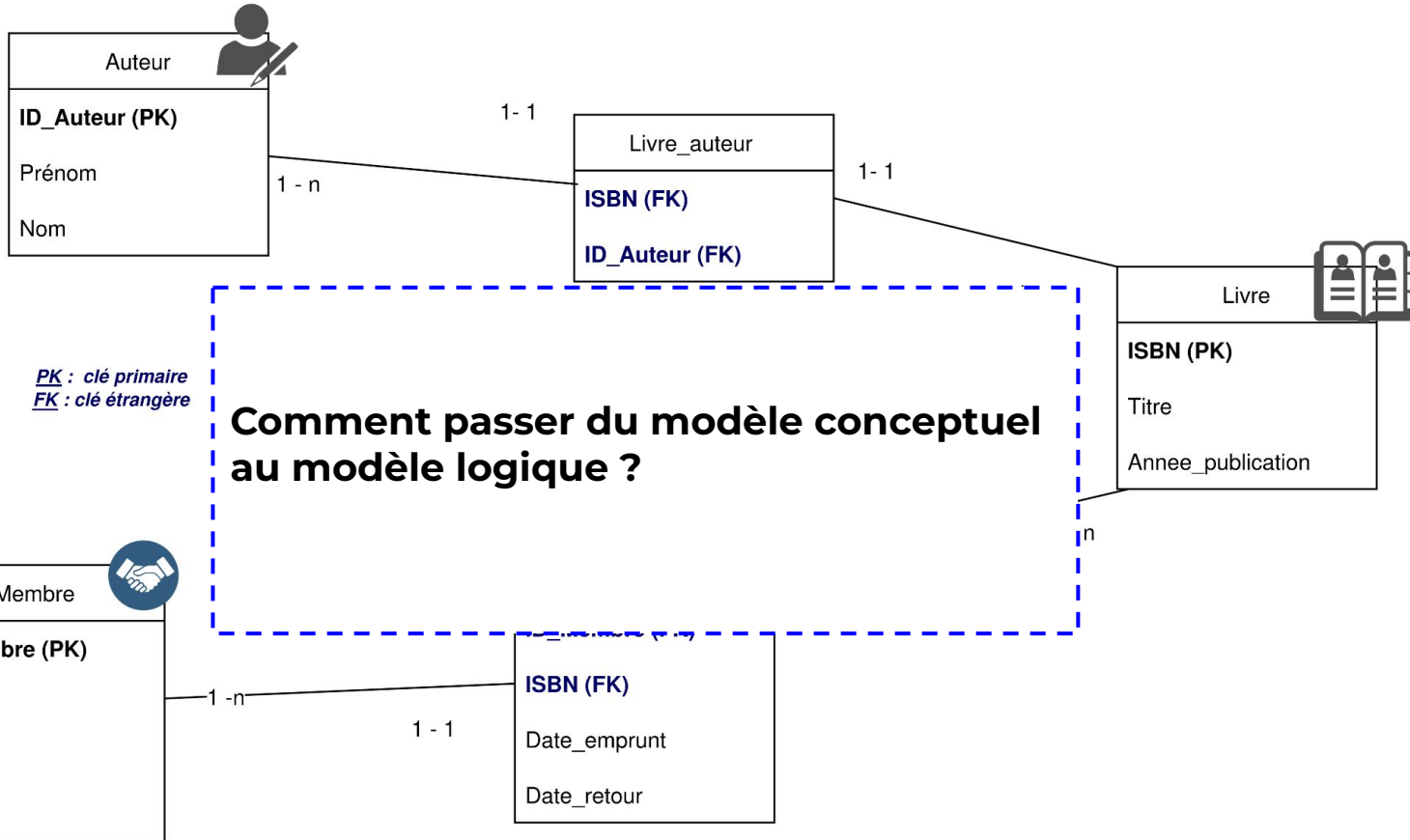
# Modèle logique (MLD)

## BIBLIOTHEQUE



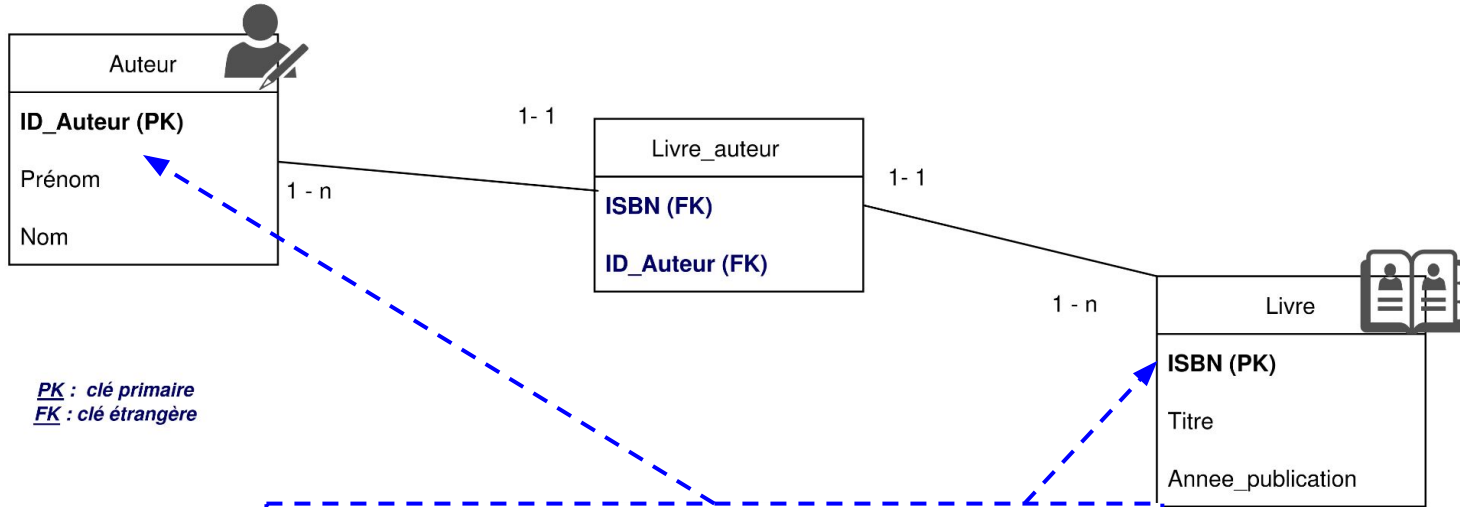
# Modèle logique (MLD)

## BIBLIOTHEQUE



# Modèle logique (MLD)

## BIBLIOTHEQUE



PK : clé primaire  
FK : clé étrangère

### Étape 1 :

On identifie et on définit les **clés primaires** de chaque table.

⇒ Qu'est-ce qu'une clé primaire...

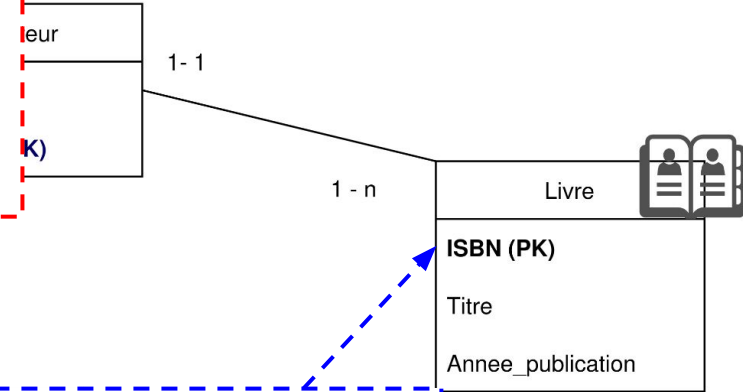
# Modèle logique (MLD)

## BIBLIOTHEQUE

### Clé primaire (PK) :

Ensemble d'attributs qui identifie de manière unique chaque enregistrement (ligne) dans une table de base de données.

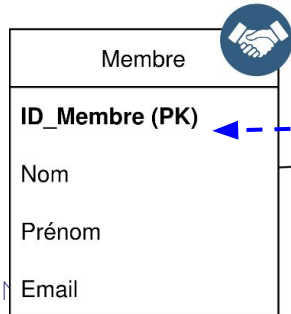
PK : clé primaire  
FK : clé étrangère



### Étape 1 :

On identifie et on définit les **clés primaires** de chaque table.

⇒ Qu'est-ce qu'une clé primaire...



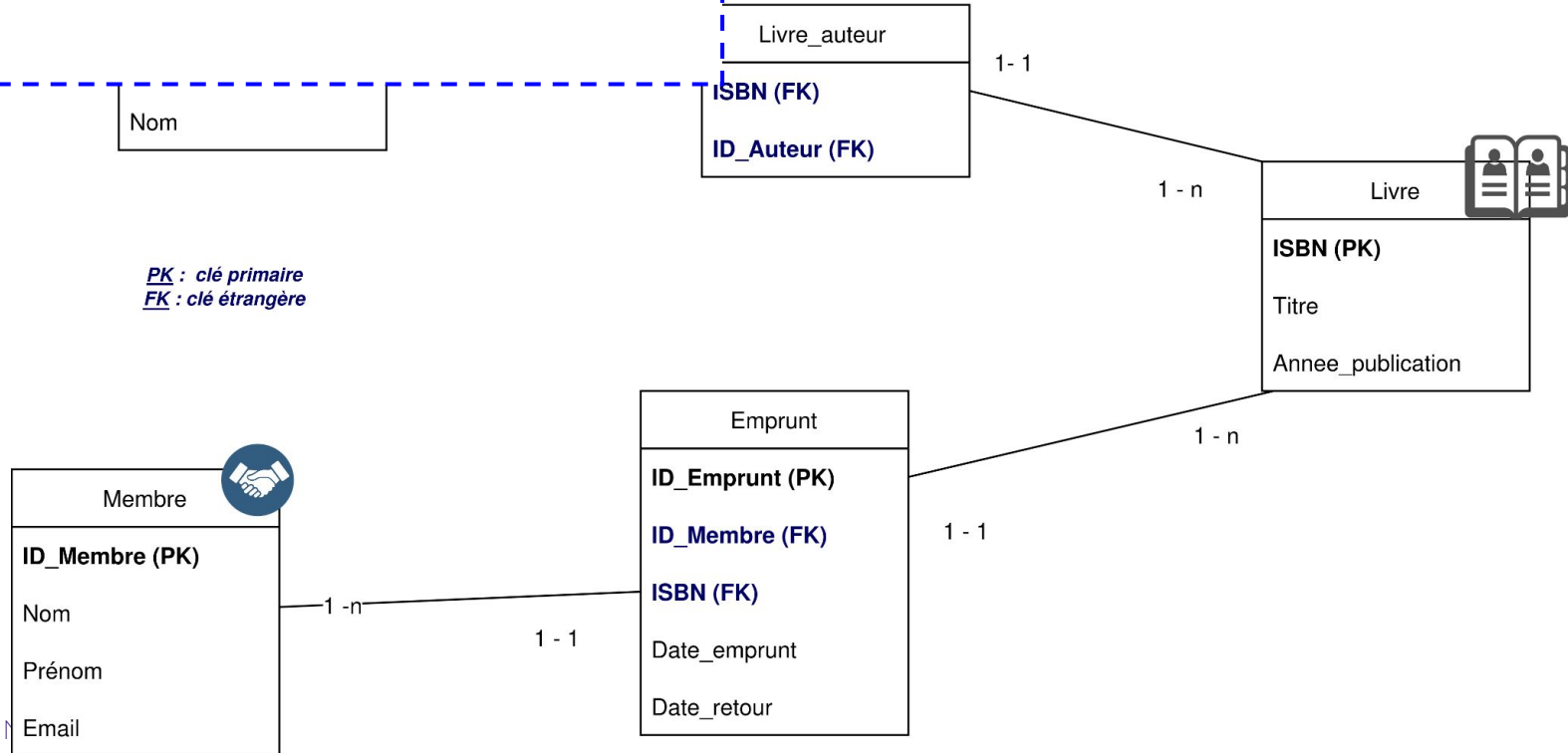


# Modèle logique (MLD)

## BIBLIOTHEQUE

### Étape 2 :

Traduire les **associations en entités**



# Modèle logique (MLD)

## BIBLIOTHEQUE

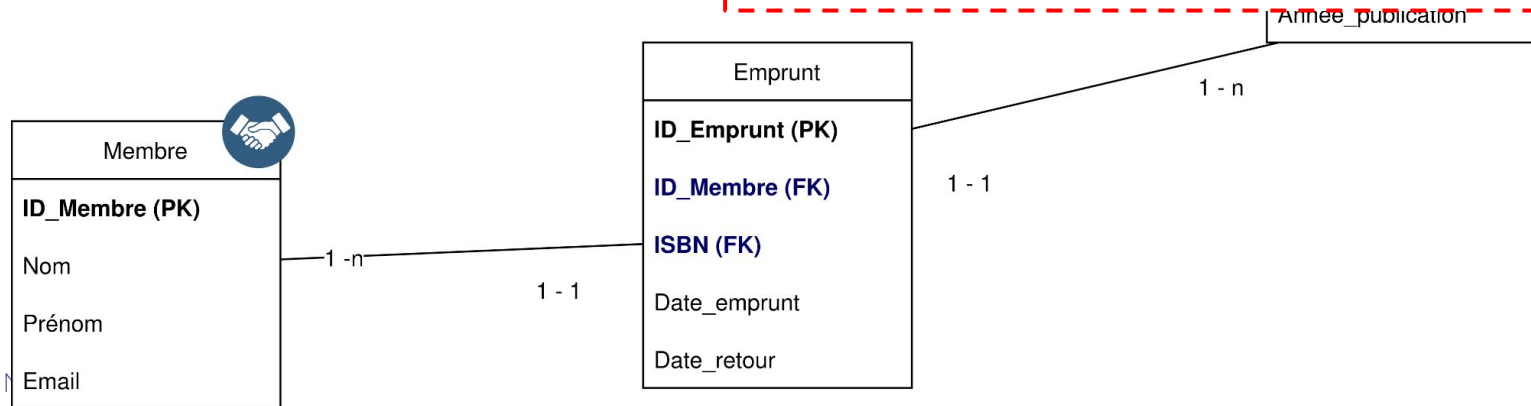
### Étape 2 :

Traduire les **associations en entités**

Pour les **associations 1:N**, ajouter une clé étrangère dans la table du côté "plusieurs".

Pour les **associations N:N**, créer une table de jonction avec les clés étrangères des deux entités.

PK : clé primaire  
FK : clé étrangère



# Modèle logique (MLD)

## BIBLIOTHEQUE

### Étape 2 :

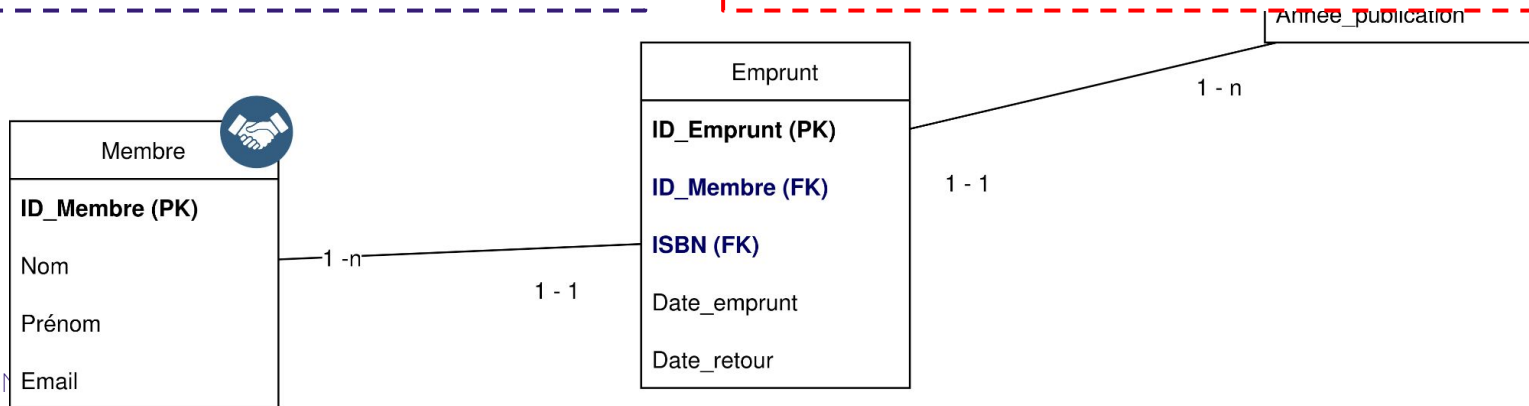
Traduire les **associations en entités**

#### Clé étrangère (FK) :

Ensemble d'attributs dans une table qui fait référence à la clé primaire d'une autre table. Elle permet d'établir un lien logique entre les deux tables.

Pour les **associations 1:N**, ajouter une clé étrangère dans la table du côté "plusieurs".

Pour les **associations N:N**, créer une table de jonction avec les clés étrangères des deux entités.



# Modèle logique (MLD)

## BIBLIOTHEQUE

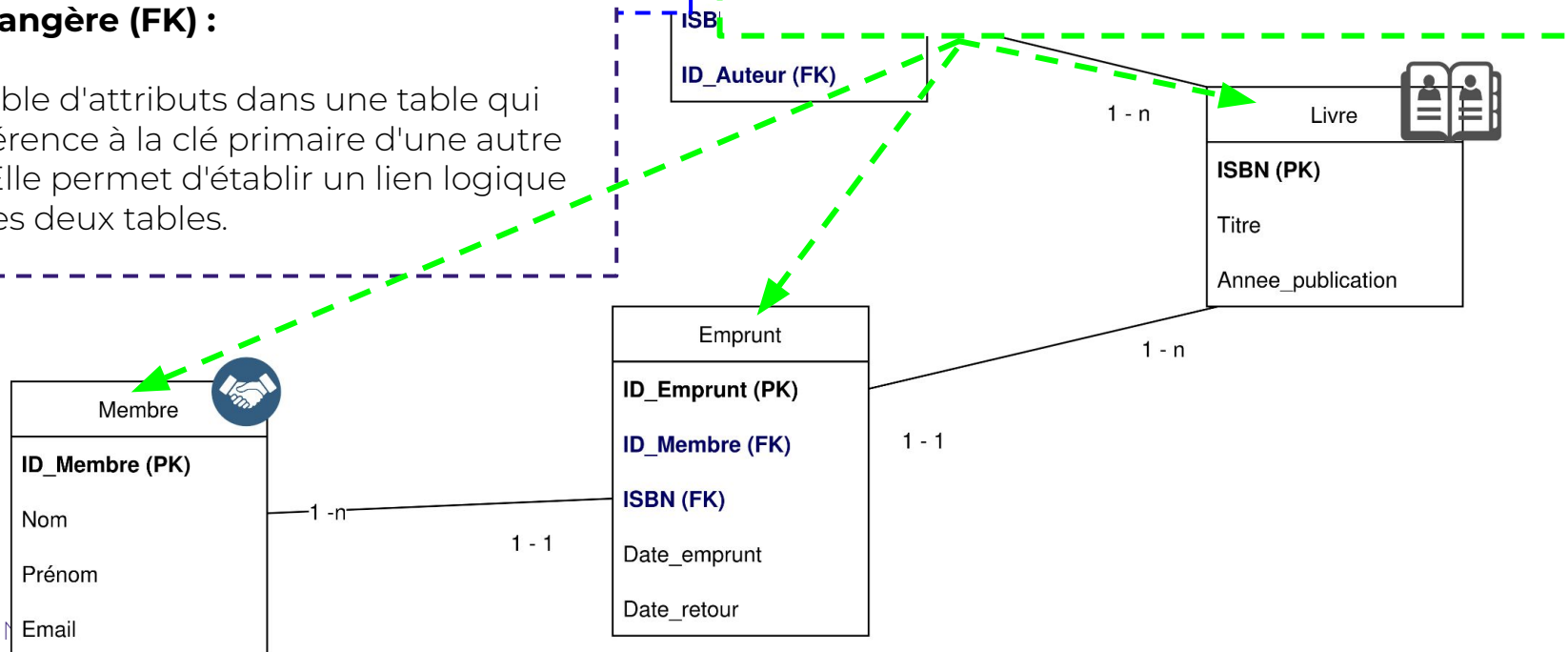
### Étape 2 :

Traduire les **associations en entités**

### Clé étrangère (FK) :

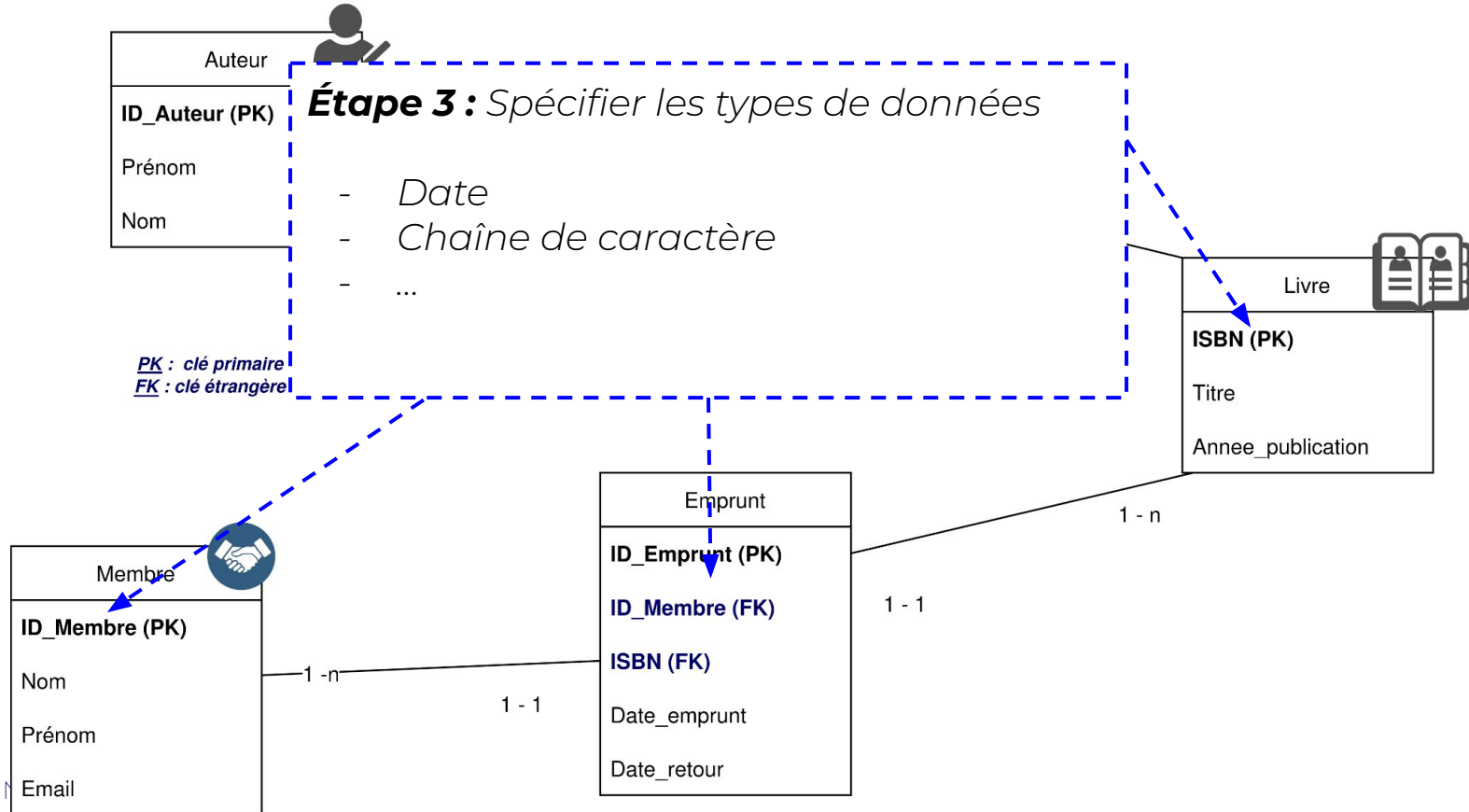
Ensemble d'attributs dans une table qui fait référence à la clé primaire d'une autre table. Elle permet d'établir un lien logique entre les deux tables.

⇒ **Emprunt** est une table de jonction entre "Livre" et "Membre"



# Modèle logique (MLD)

## BIBLIOTHEQUE

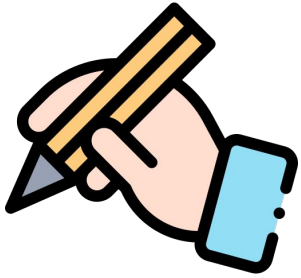


# MPD : Adaptation au SGBD, types de données

**Choix du SGBD :** MySQL, PostgreSQL, Oracle, etc.

**Types de données :** VARCHAR, INT, DATE, BOOLEAN, etc.

```
CREATE TABLE Livre (  
    ISBN VARCHAR(13) PRIMARY KEY,  
    Titre VARCHAR(100) NOT NULL,  
    Année_publication INT  
);  
  
CREATE TABLE Auteur (  
    ID_Auteur INT AUTO_INCREMENT PRIMARY KEY,  
    Nom VARCHAR(50) NOT NULL,  
    Prénom VARCHAR(50) NOT NULL  
);  
  
CREATE TABLE Membre (  
    ID_Membre INT AUTO_INCREMENT PRIMARY KEY,  
    Nom VARCHAR(50) NOT NULL,
```



**Dans le cas du CHU de Nantes, convertissez votre MCD en MLD en suivant les trois étapes :**

- 1. Identifiez les clés primaires**
- 2. Traduire les associations en entités**
- 3. Spécifiez les types de données**

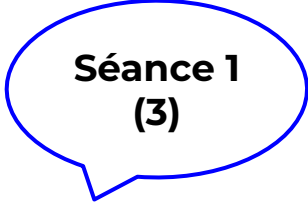


**Maintenant à vous !**

**Modéliser un système de votre choix (par groupe)  
réaliser un MCD / MLD**

**<https://bit.ly/3YLS2Uq>**





**Séance 1  
(3)**

# **Architectures de données classiques**

*centralisée, décentralisée, distribuée*

# Architecture centralisée

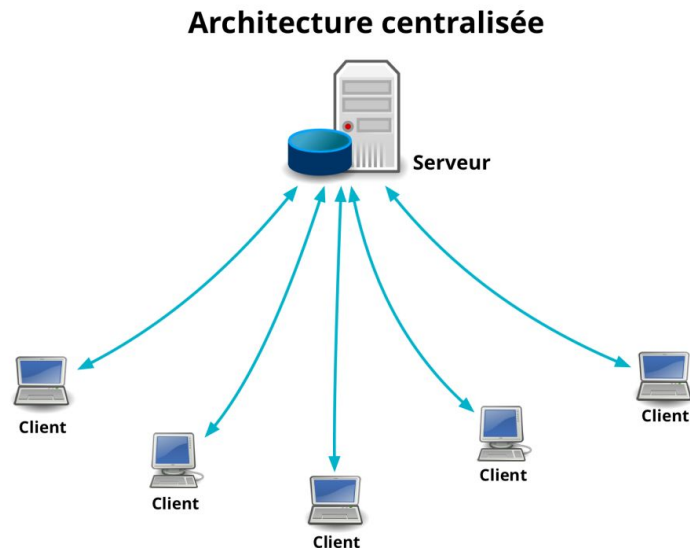
**Définition :** Toutes les données sont stockées et gérées à partir d'un emplacement central

## Caractéristiques :

- Un seul point de contrôle et de gestion
- Facilité de maintenance et de mise à jour
- Sécurité et cohérence des données simplifiées

Un **serveur central** puissant héberge les données, les applications et les ressources du système

Des **clients** (ordinateurs, tablettes, smartphones) se connectent au serveur pour accéder à ces services.



# Architecture centralisée

**Définition :** Toutes les données sont stockées et gérées à partir d'un emplacement central

## Caractéristiques :

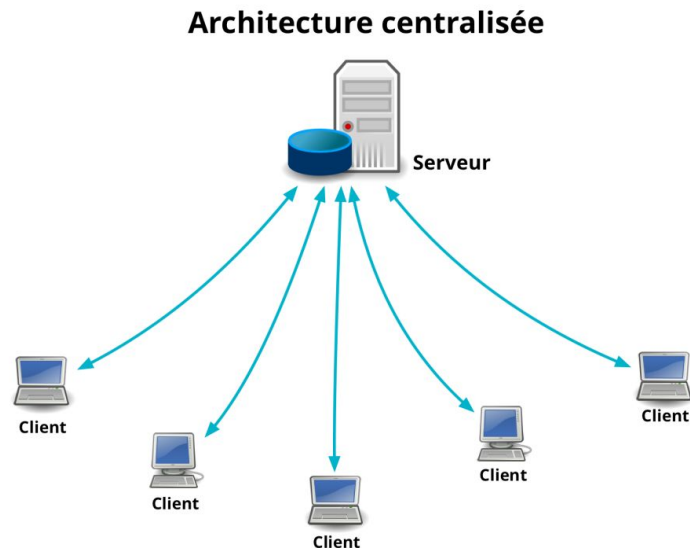
- Un seul point de contrôle et de gestion
- Facilité de maintenance et de mise à jour
- Sécurité et cohérence des données simplifiées

## Avantages :

Contrôle, cohérence, sécurité

## Inconvénients :

Point unique de défaillance, évolutivité limitée

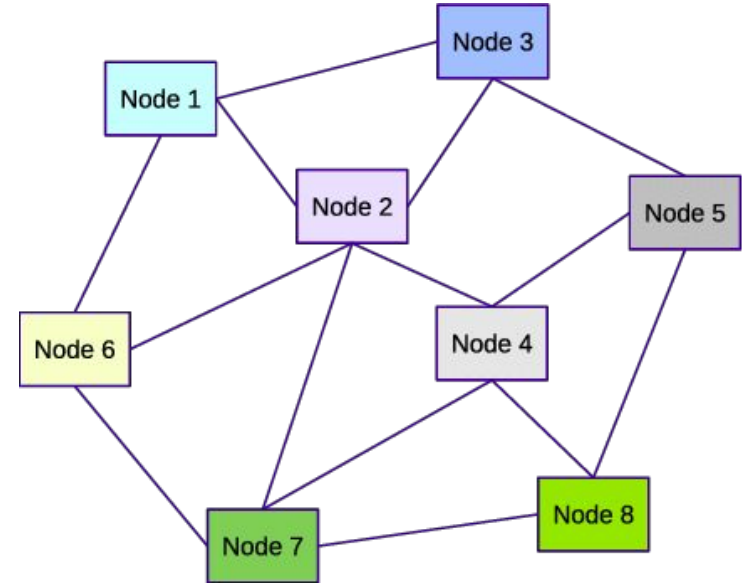


# Architecture décentralisée

**Définition :** Les données sont réparties sur plusieurs sites ou unités autonomes

## Caractéristiques :

- Chaque unité gère ses propres données
- Autonomie locale avec coordination centrale
- Flexibilité et adaptation aux besoins locaux



# Architecture décentralisée

**Définition :** Les données sont réparties sur plusieurs sites ou unités autonomes

## Caractéristiques :

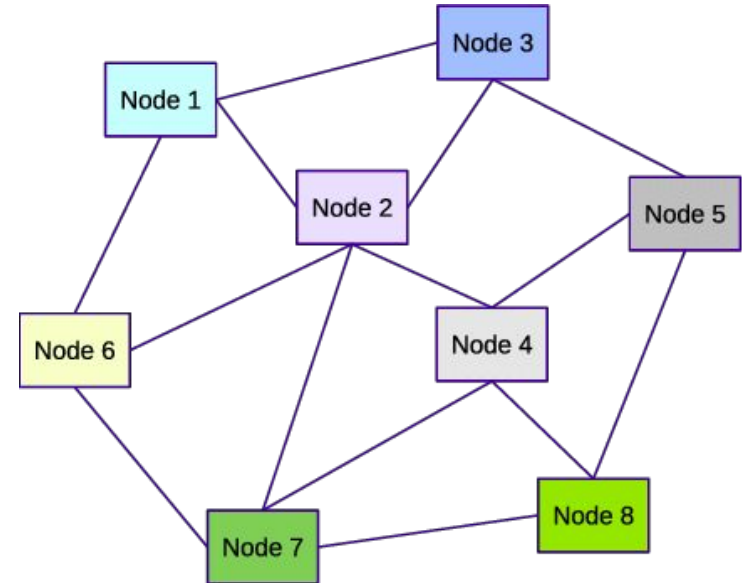
- Chaque unité gère ses propres données
- Autonomie locale avec coordination centrale
- Flexibilité et adaptation aux besoins locaux

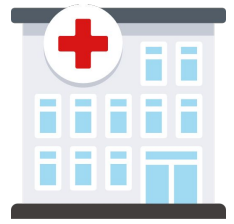
## Avantages :

Flexibilité, résilience, adaptation locale

## Inconvénients :

Complexité de gestion, risque d'incohérence

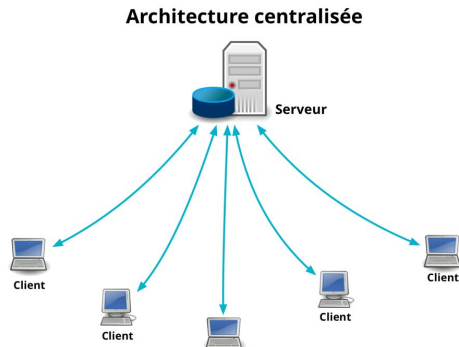




# Exemples

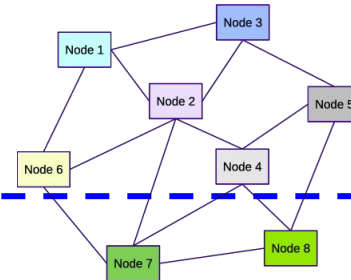
## Architecture centralisée

- Le CHU de Nantes stocke toutes les données médicales des patients (DPI, résultats d'examens, etc.) sur un **serveur central unique**, situé dans l'hôpital.
- Les médecins et le personnel médical accèdent à ces données via des terminaux connectés au réseau interne de l'hôpital.



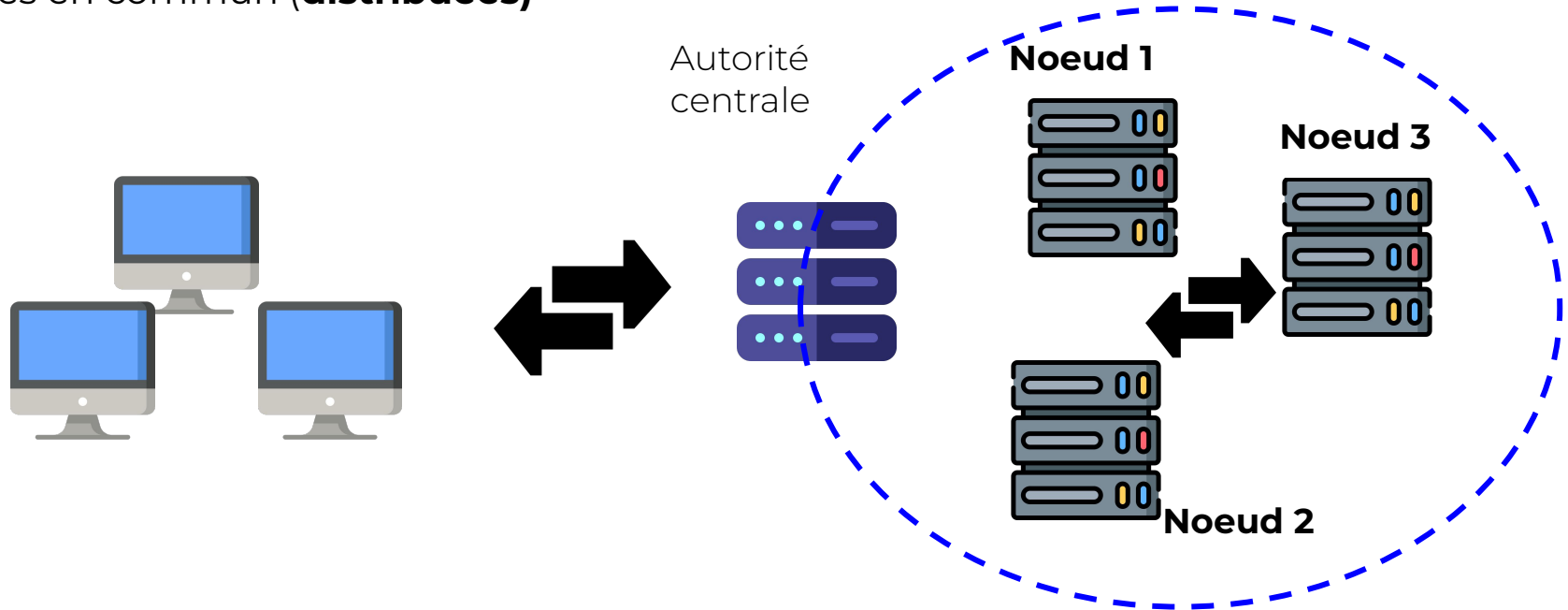
## Architecture décentralisée

- Le CHU de Nantes dispose de plusieurs sites hospitaliers (Hôtel-Dieu, site de laennec, etc.).
- Chaque site possède son propre serveur** qui stocke les données des patients pris en charge sur ce site.
- Chaque **serveur est autonome** et gère ses propres données et applications. Il n'y a pas de coordination centralisée.



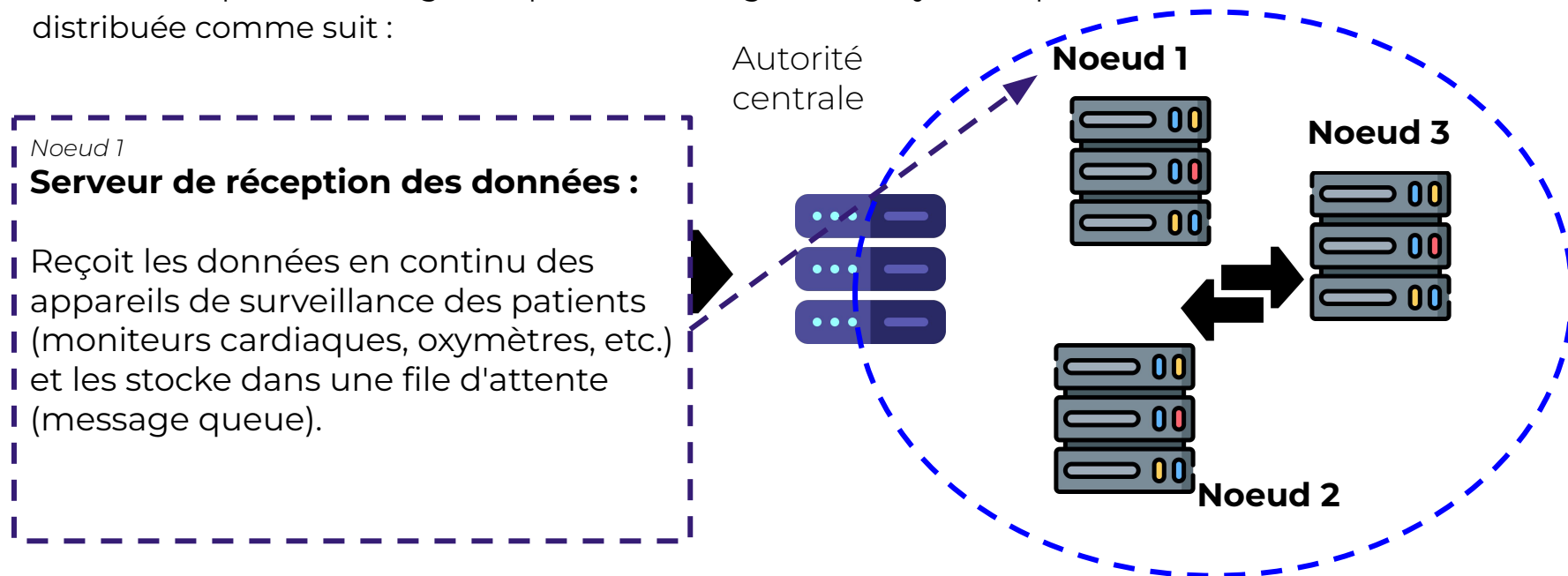
# Architecture distribuée

Les données sont réparties sur plusieurs nœuds **interconnectés** avec des ressources mises en commun (**distribuées**)



# Exemple d'architecture distribuée

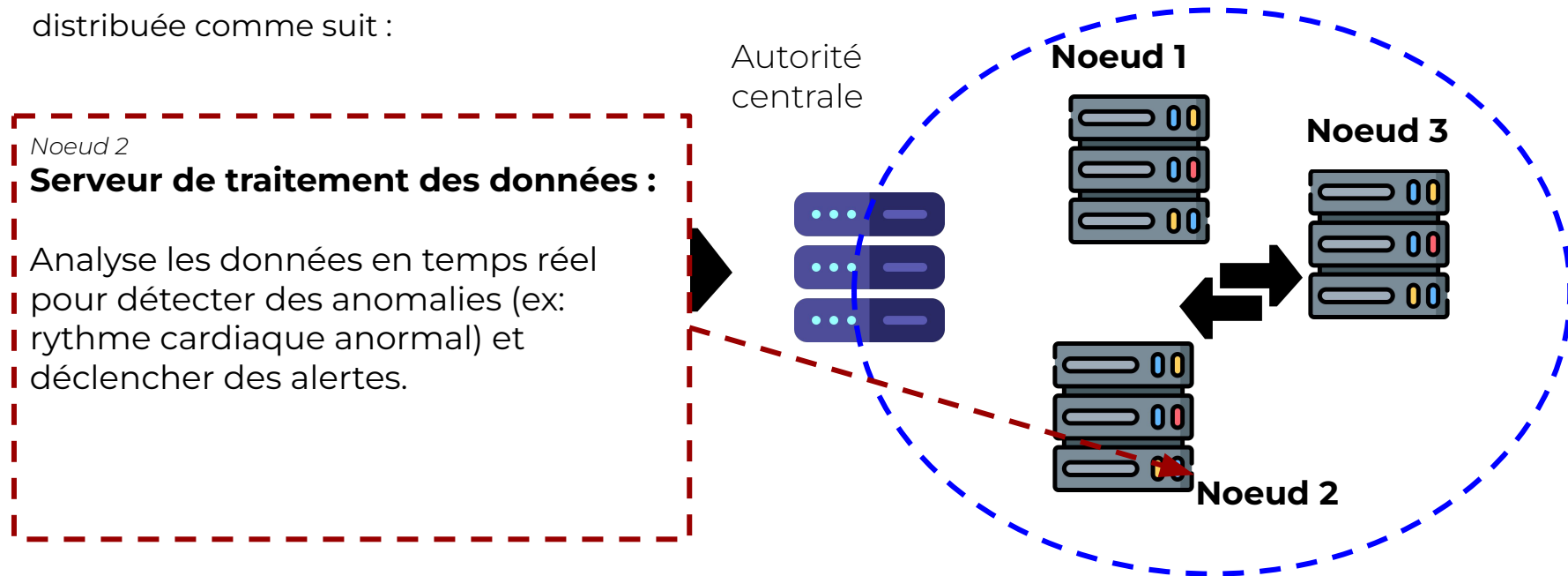
Le CHU de Nantes souhaite mettre en place un système d'analyse des données en temps réel pour améliorer la prise en charge des patients en urgence. Ce système pourrait utiliser une architecture distribuée comme suit :





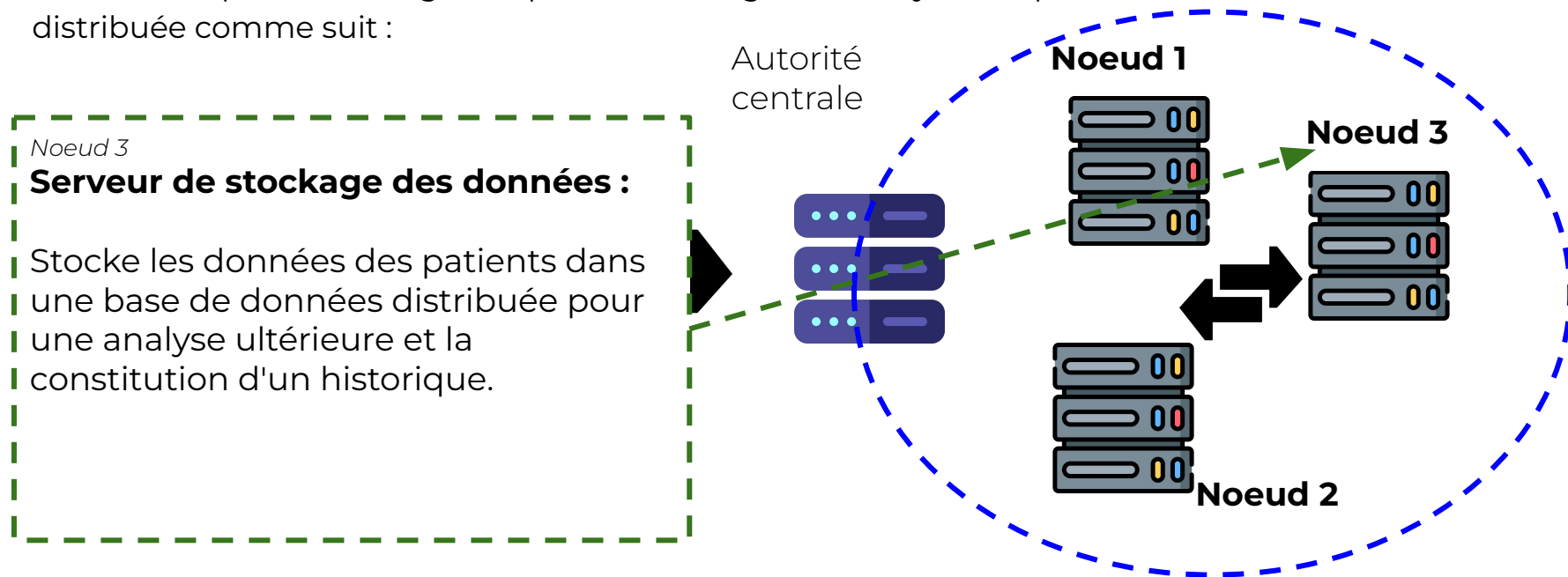
# Exemple d'architecture distribuée

Le CHU de Nantes souhaite mettre en place un système d'analyse des données en temps réel pour améliorer la prise en charge des patients en urgence. Ce système pourrait utiliser une architecture distribuée comme suit :



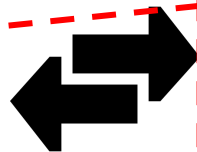
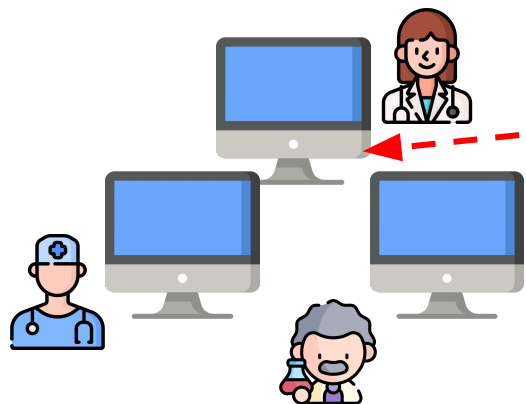
# Exemple d'architecture distribuée

Le CHU de Nantes souhaite mettre en place un système d'analyse des données en temps réel pour améliorer la prise en charge des patients en urgence. Ce système pourrait utiliser une architecture distribuée comme suit :



# Exemple d'architecture distribuée

Le CHU de Nantes souhaite mettre en place un système d'analyse des données en temps réel pour améliorer la prise en charge des patients en urgence. Ce système pourrait utiliser une architecture distribuée comme suit :



## C'est une architecture distribuée !

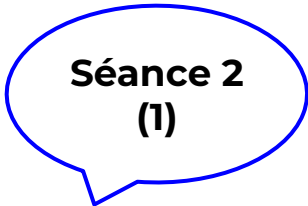
Les serveurs ne sont pas simplement indépendants, ils travaillent ensemble de **manière coordonnée** pour analyser les données en temps réel et fournir des informations aux soignants.



Noeud 2

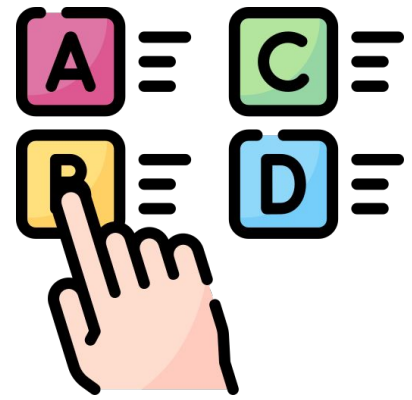


**À vous !**  
**Une étude de cas à choisir (par équipe) :**  
**<https://bit.ly/4edf8TG>**



**Séance 2  
(1)**

*Séance 2*  
**Architectures modernes, gouvernance et  
réglementations**



**Révisons les acquis de la séance 1**

***Rendez-vous sur Kahoot !***

# Plan de la séance 2

- Introduction et révision
  - Data Warehousing et Data Lake
  - Définitions et concepts clés
  - Études de cas
  - Activité
  
- Gouvernance des données
  - Principes, rôles, mise en place
  - Référentiels TOGAF / Zachman
  - Activité
  
- Aspects réglementaires
  
- Conclusion



# Data Warehousing & Data Lake





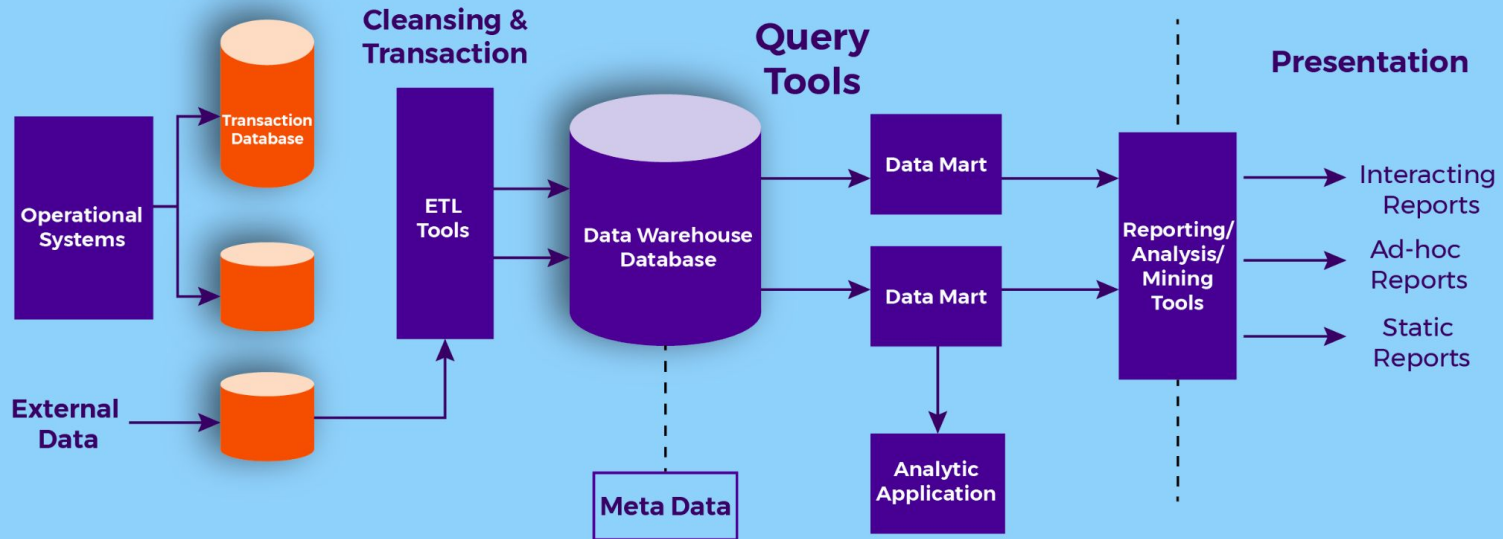
# Data Warehouse

Un **Data Warehouse** est un système centralisé qui stocke des données historiques provenant de différentes sources opérationnelles.

**Objectif principal** : Faciliter l'analyse et le reporting décisionnel.

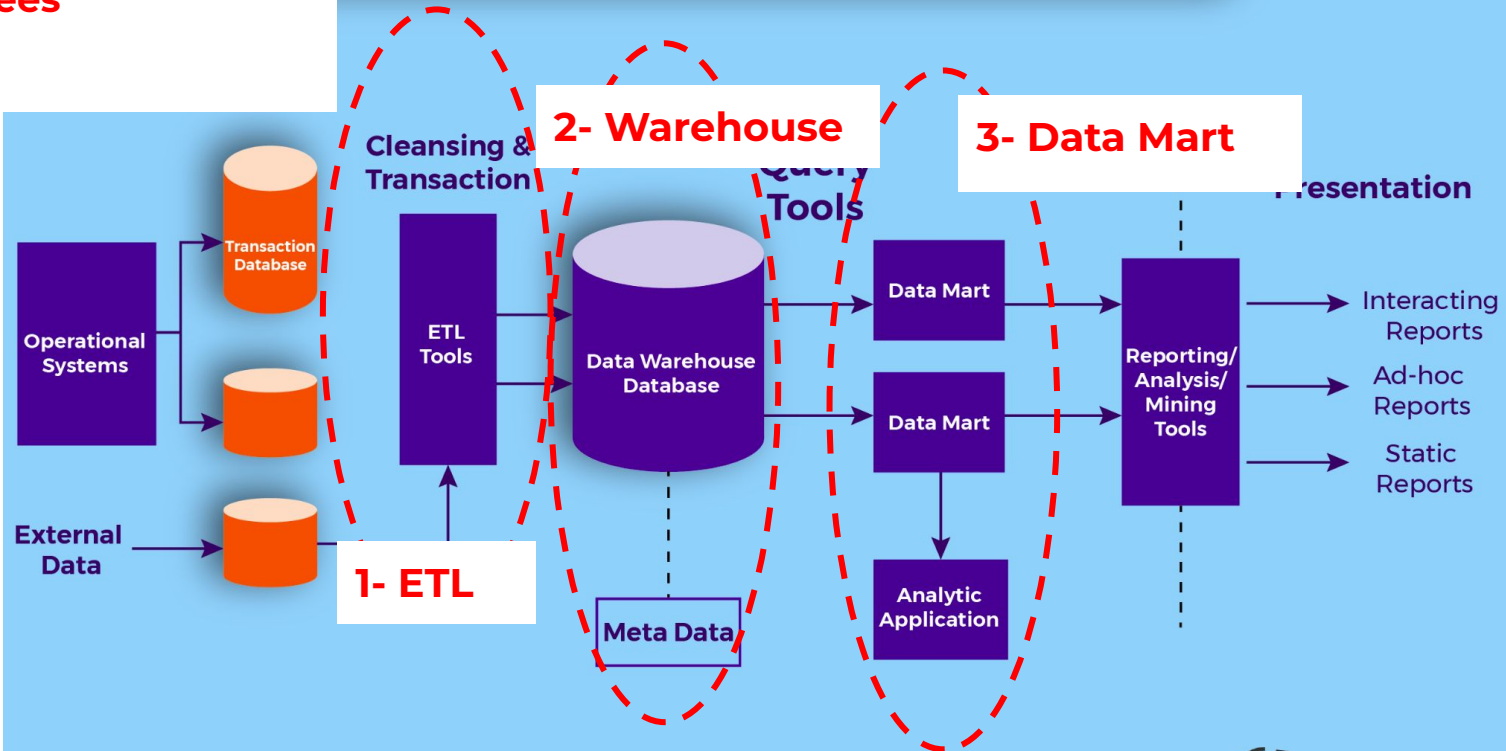


# Data Warehouse Architecture



**3 points importants  
dans le cadre d'un  
entrepôt des  
données**

## Data Warehouse Architecture

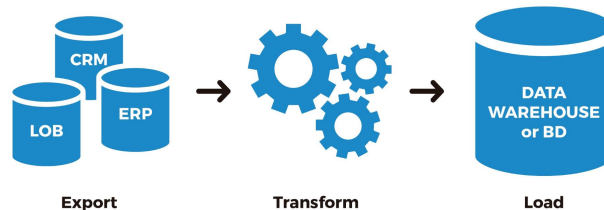


# 1- ETL (*Extract, Transform, Load*)

**Définition :** Processus qui extrait les données des sources, les transforme et les charge dans le Data Warehouse.

## Etapes :

- **Extraction :** Collecte des données depuis les systèmes sources.
- **Transformation :** Nettoyage, transformation et harmonisation des données.
- **Chargement :** Insertion des données dans le Data Warehouse.



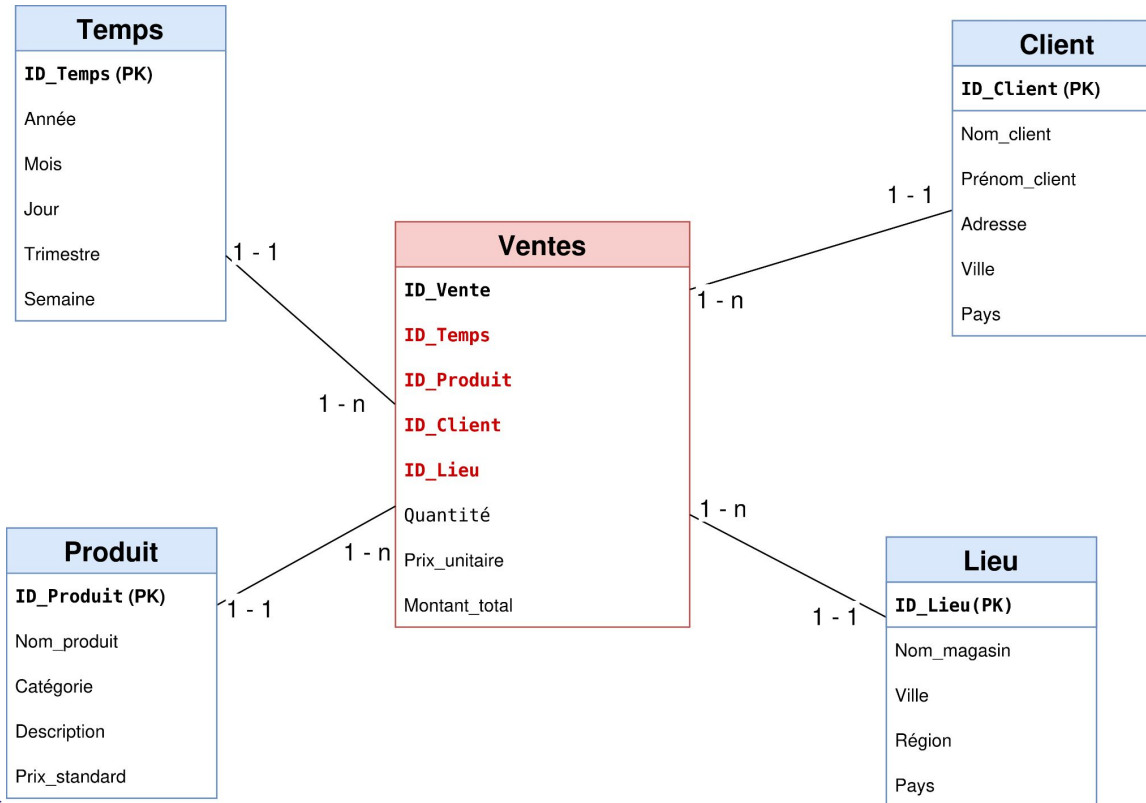
# The ETL Process Explained



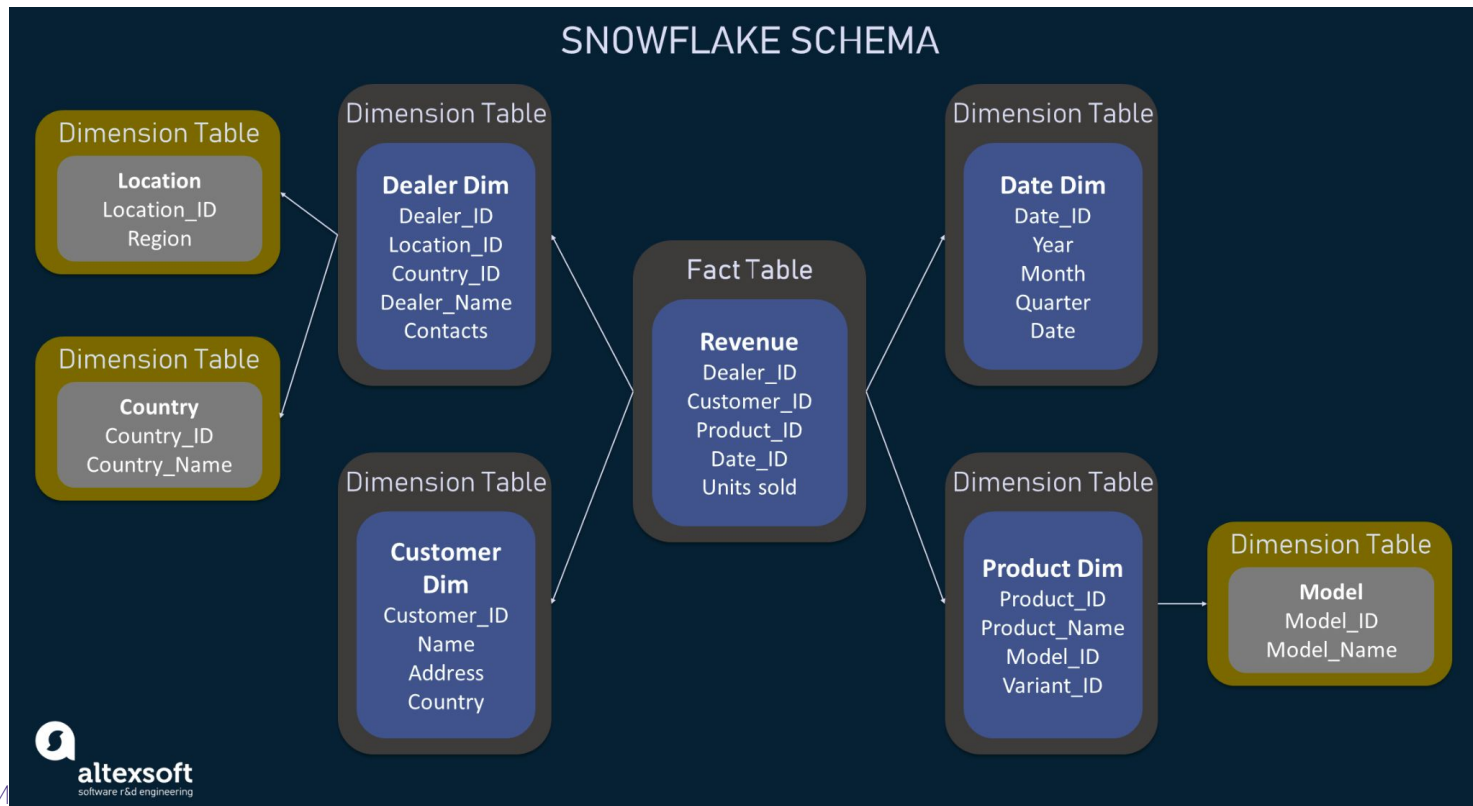
## 2- Le Warehouse : Schéma en étoile

- Modèle de données dimensionnel.
- Les données sont organisées autour d'une **table de faits** centrale connectée à plusieurs **tables de dimensions**.
- **Table de faits :**  
Contient les mesures ou indicateurs clés de performance (ex: nombre de ventes, montant des ventes).
- **Tables de dimensions :**  
Fournissent le contexte des mesures (ex: temps, produit, client...)

# Schéma en étoile



## 2- Le Warehouse : Schéma Snowflake





## 3- Data Mart

### Définition :

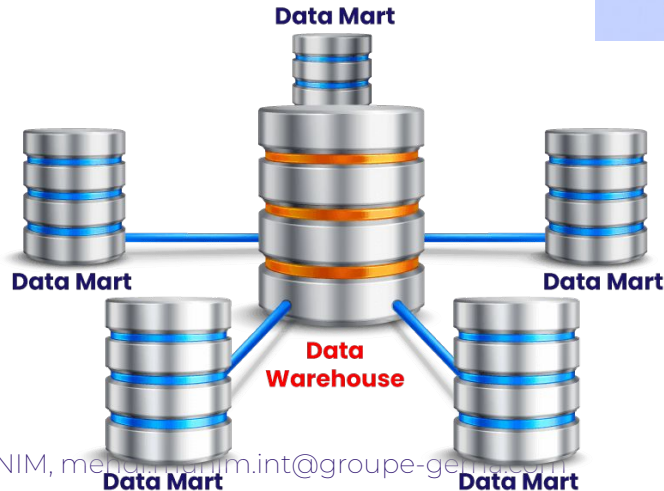
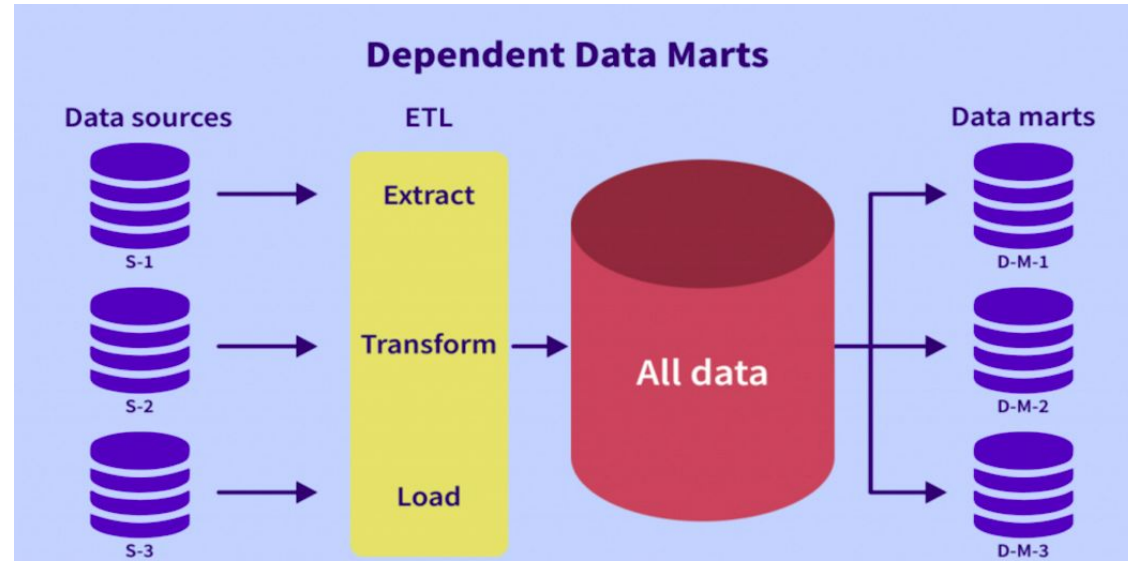
- Un **Data Mart** est un sous-ensemble d'un Data Warehouse qui se concentre sur un domaine spécifique ou un département particulier.
- Il est conçu pour répondre aux besoins d'un groupe d'utilisateurs précis.
- Contient des données extraites du Data Warehouse et organisées pour faciliter l'analyse d'un domaine spécifique.

### Exemple :

Une entreprise possède un Data Warehouse centralisant les données de tous ses départements (ventes, marketing, finance, RH).

⇒ Le département marketing crée un Data Mart contenant **uniquement les données marketing** (campagnes, clients, prospects, etc.).

Un **Data Mart** est un sous-ensemble d'un Data Warehouse qui se concentre sur un domaine spécifique ou un département particulier.



# Data Lake

**Définition :** Référentiel centralisé pour stocker toutes les données d'une organisation, structurées ou non structurées.

## Objectifs :

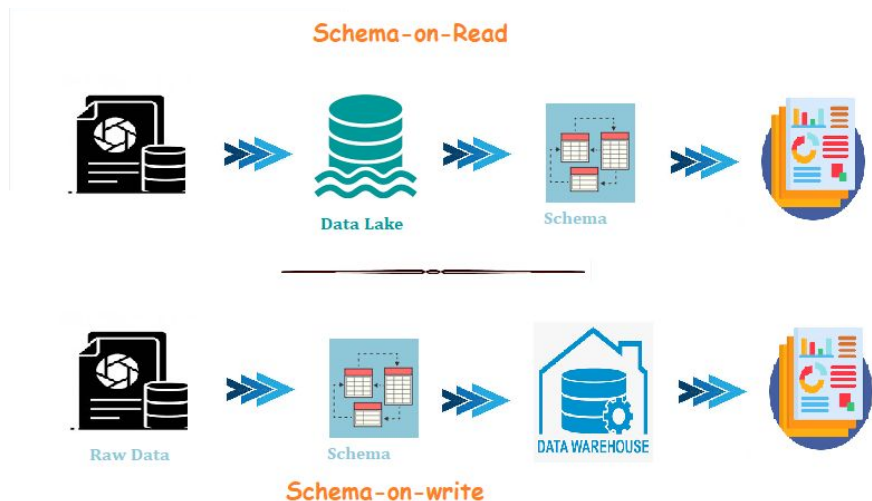
- Analyse exploratoire et flexible des données
- Données brutes pour différents cas d'utilisation
- Analyse Big Data et Machine Learning

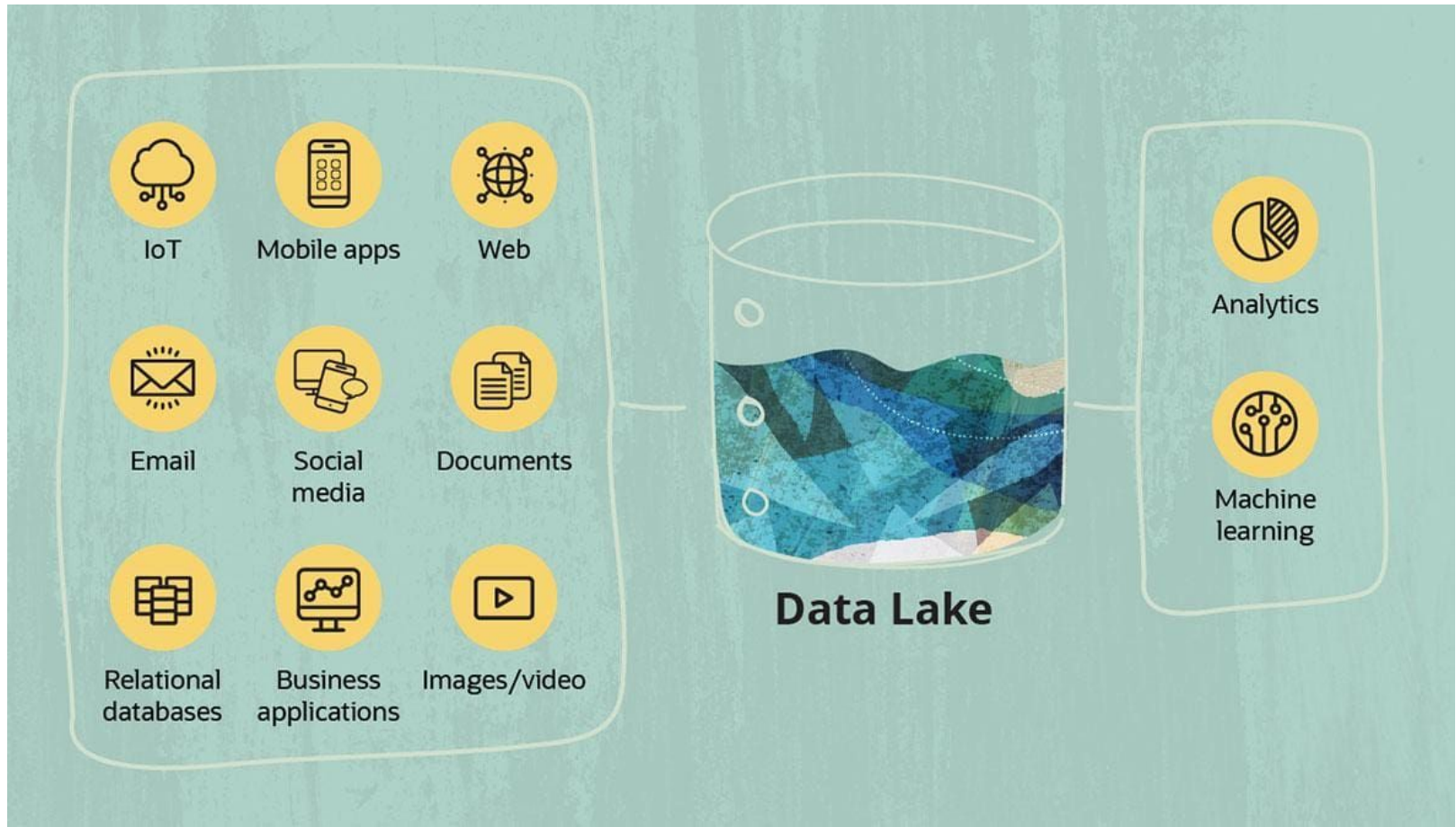


# Caractéristiques d'un Data Lake

**Schéma flexible :** "Schema-on-read" - le schéma (structure des données) est appliqué lors de la lecture des données.

**Variété des données :** Tous types de données (texte, images, vidéos, logs, etc.).





<https://www.netsuite.com/portal/resource/articles/data-warehouse/data-lake.shtml>

# Avantages d'un Data Lake

**Flexibilité :** Permet d'explorer et d'analyser les données de différentes manières.

**Évolutivité :** Conçu pour gérer des volumes de données importants.

**Coût-efficacité :** Le stockage de données brutes est généralement moins coûteux.



# Data Lake - Exemples

## **Analyse du Big Data :**

Un Data Lake peut stocker et traiter des données massives provenant de sources variées, telles que les médias sociaux, les capteurs IoT et les logs web.

## **Machine Learning :**

Les Data Scientists peuvent utiliser les données brutes d'un Data Lake pour entraîner des modèles de Machine Learning.

## **Archivage de données :**

Un Data Lake peut servir à archiver des données historiques qui ne sont plus activement utilisées.



**Discutons de ces concepts !  
Quiz et étude de cas ci-dessous :**

<https://bit.ly/4f71Jhp>

**NETFLIX**



# Gouvernance des données

# Pourquoi la gouvernance des données ?

## **Gouvernance des Données :**

Ensemble de processus, de rôles, de politiques et de métriques qui permettent de gérer les données d'une organisation de manière efficace et sécurisée.

## **Objectif :**

Assurer que les données sont utilisées de manière appropriée, fiable et conforme aux réglementations.

# Principes clés de la Gouvernance des Données

## Disponibilité :

Accès facile aux données pour les utilisateurs autorisés.

## Intégrité :

Exactitude, cohérence et fiabilité des données.

## Confidentialité :

Protection des données sensibles.

## Traçabilité :

Suivi de l'origine et des modifications des données.

## Conformité :

Respect des lois et réglementations (ex: RGPD).





**Un petit exposé à réaliser par équipe.**  
**Lien ci-dessous**  
**[bit.ly/3YObhbc](https://bit.ly/3YObhbc)**

