Intégration de données :

Introduction aux ETL

Jeudi 27 octobre 2022

Module Interopérabilité et Standards de Données Médicales (ISDM) - Majeure Santé EPITA

Dr Damien Leprovost – AP-HP , Limics

Licence Creative Commons BY-SA 4.0

Ce support de cours est distribué sous licence
Attribution - Partage dans les Mêmes Conditions 4.0 International (CC BY-SA 4.0)

Vous êtes autorisé à :

- **Partager** — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats
- **Adapter** — remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.

Selon les conditions suivantes :

Attribution — Vous devez créditer l'œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son œuvre.

Partage dans les Mêmes Conditions — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'œuvre originale, vous devez diffuser l'œuvre modifiée dans les même conditions, c'est à dire avec la même licence avec laquelle l'œuvre originale a été diffusée.

Objectifs du cours

- Savoir ce qu'est un **ETL**
- Se familiariser avec les **Data Warehouse** et **Data Mart**
- Comprendre les **enjeux** de l'intégration de données
- Saisir l'intérêt de **spécifier** un ETL
- Se sensibiliser à la complexité d'une **implémentation réelle**

Plan du cours

I. Introduction

Contexte

II. ETL

Définitions et design

III. DataWarehouse et DataMart

Structures et organisation

IV. En pratique : intégration de données

Au sein de l'EDS de l'AP-HP

V. En pratique : Implémentation

A l'AP-HP

VI. Conclusion

I. Introduction

- Contexte
- Besoins
- Obstacles

Contexte

- Les données sont stockées dans des systèmes hétérogènes qui collectent de l'information en temps réel, avec des natures, volumétries, fiabilité et criticité propres.
- Exemples:
  - Entrées du dossier patient
  - Rapports d'analyses
  - Données issues de capteurs
  - Livraison de données externes (INSEE)
  - Etc.

Besoins

- Collecter et centraliser l'information pour répondre à une demande
  - Prise de décision
  - Génération d'informations
  - Exécution de recherche
  - Amélioration des performances
  - Anticipations des évolutions
  - …

Obstacles

- Homogénéité des systèmes d'informations
  - plateformes, protocoles, référentiels
- Incohérences et conflits possibles des contenus entre les sources
- Absence fréquente d'historisation des sources
- Dynamisme imprévisible des sources
  - contenus et/ou protocoles

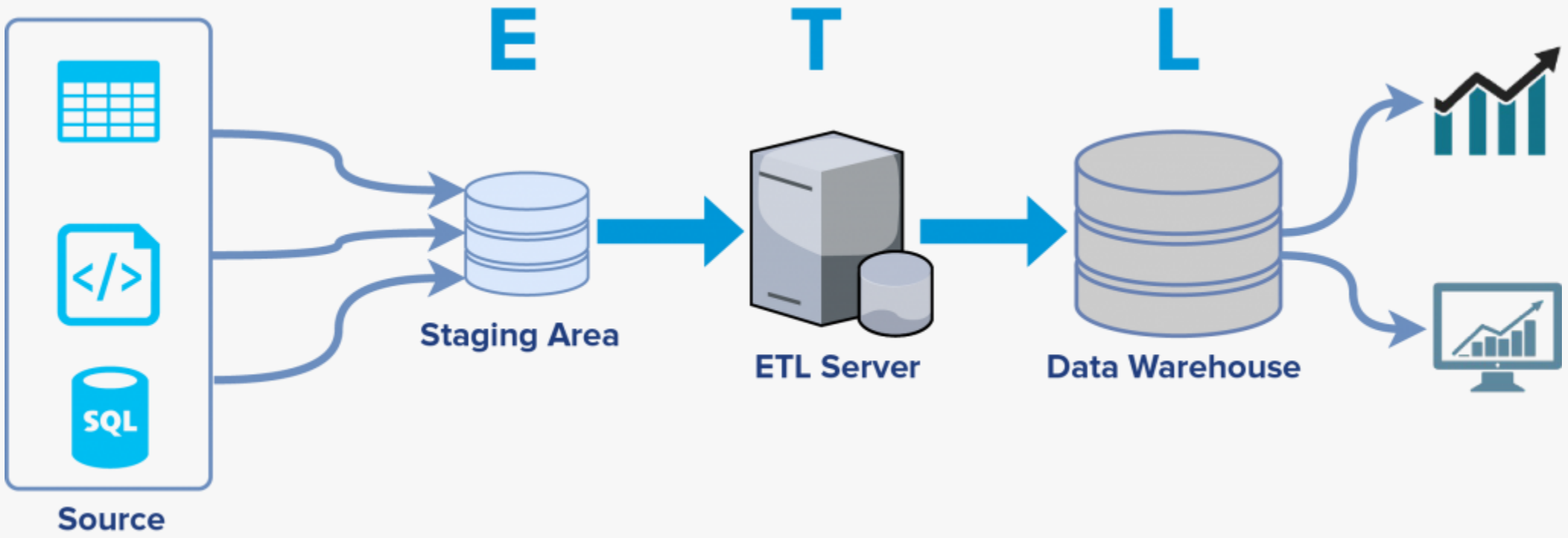
II. ETL

- Qu’est-ce qu’un ETL ?
- Force et défis du paradigme
- Design d’ETL

Définition de l'ETL

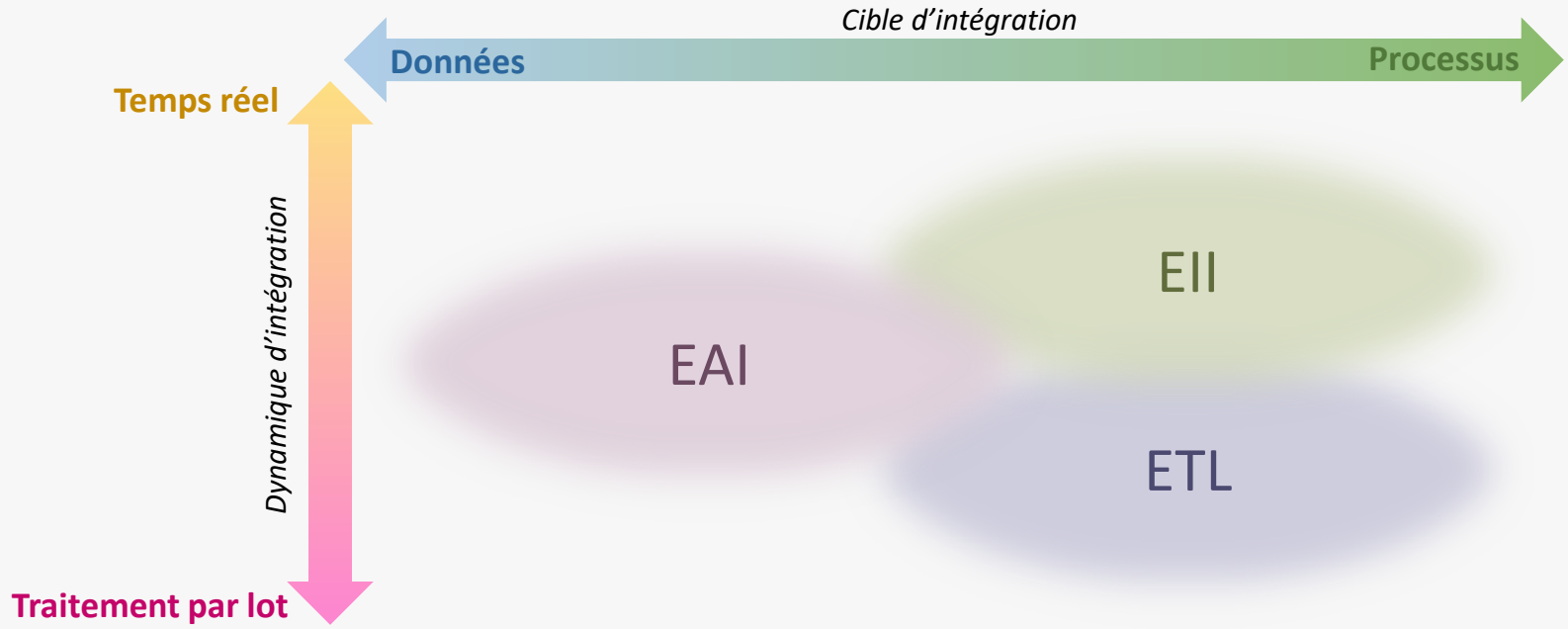
- Un processus en trois phases dans lequel les données sont **extraites**, **transformées** et **chargées** dans un conteneur de données de sortie.
- Les données peuvent être:
  - agrégées à partir d'une ou plusieurs sources
  - exportées vers une ou plusieurs destinations
- Automatise généralement l'ensemble du processus
- Peut être exécuté
  - manuellement
  - selon des calendriers récurrents (cron)

Définition de l'ETL



Définition de l'ETL

- Il existe plusieurs système d'intégration de données :
  - L'intégration de données d'entreprise (**EII**).
  - L'intégration des applications d'entreprise (**EAI**).
  - L'**ETL** (Extract - Transform - Load)



Forces

- Optimisé pour les structures de données
- Périodique, orienté batch
  - non destiné au temps réel
- Peut déplacer de gros volumes de données en une seule étape
- Permet des transformations de données complexes
  - nécessitant des calculs, des agrégations ou plusieurs étapes
- Planification contrôlée par l'administrateur
- Haut niveau de réutilisation des objets et des transformations

Principaux défis

- Délai de mise en place (« time to market »)
- Gestion du changement
- Données déplacées indépendamment du besoin réel
- Forte consommation de stockage
- Données non synchronisées avec la source d'origine
- Grandes exigences en ressources pour le staging
- Unidirectionnel

Design d'ETL

- Le rapatriement des données peut se faire de 3 façons différentes :
  - **Push :** la logique de chargement est dans le système de production, il pousse les données vers le Staging quand il en a l'occasion.
  - **Pull :** le Pull tire les données de la source vers le Staging.
  - **Push-Pull :** La source prépare les données à envoyer et prévient le Staging qu'elle est prête. Le Staging va récupérer les données. Si la source est occupée, le Staging fera une autre demande plus tard.

III. DataWarehouse et DataMart

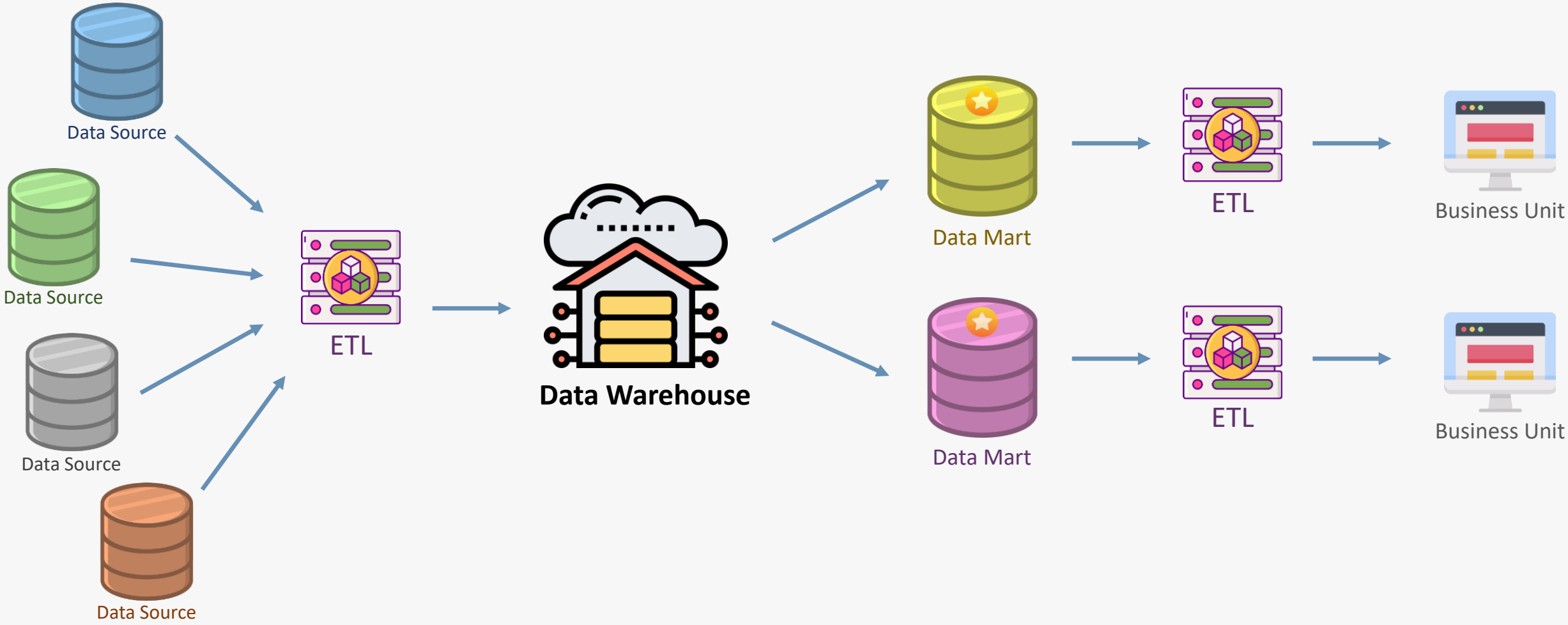
- Qu’est-ce qu’un DataWarehouse ?
- Qu’est-ce qu’un DataMart ?
- Articulation DW – DM
- Modèles d’architecture

Qu'est-ce qu'un DataWarehouse ?

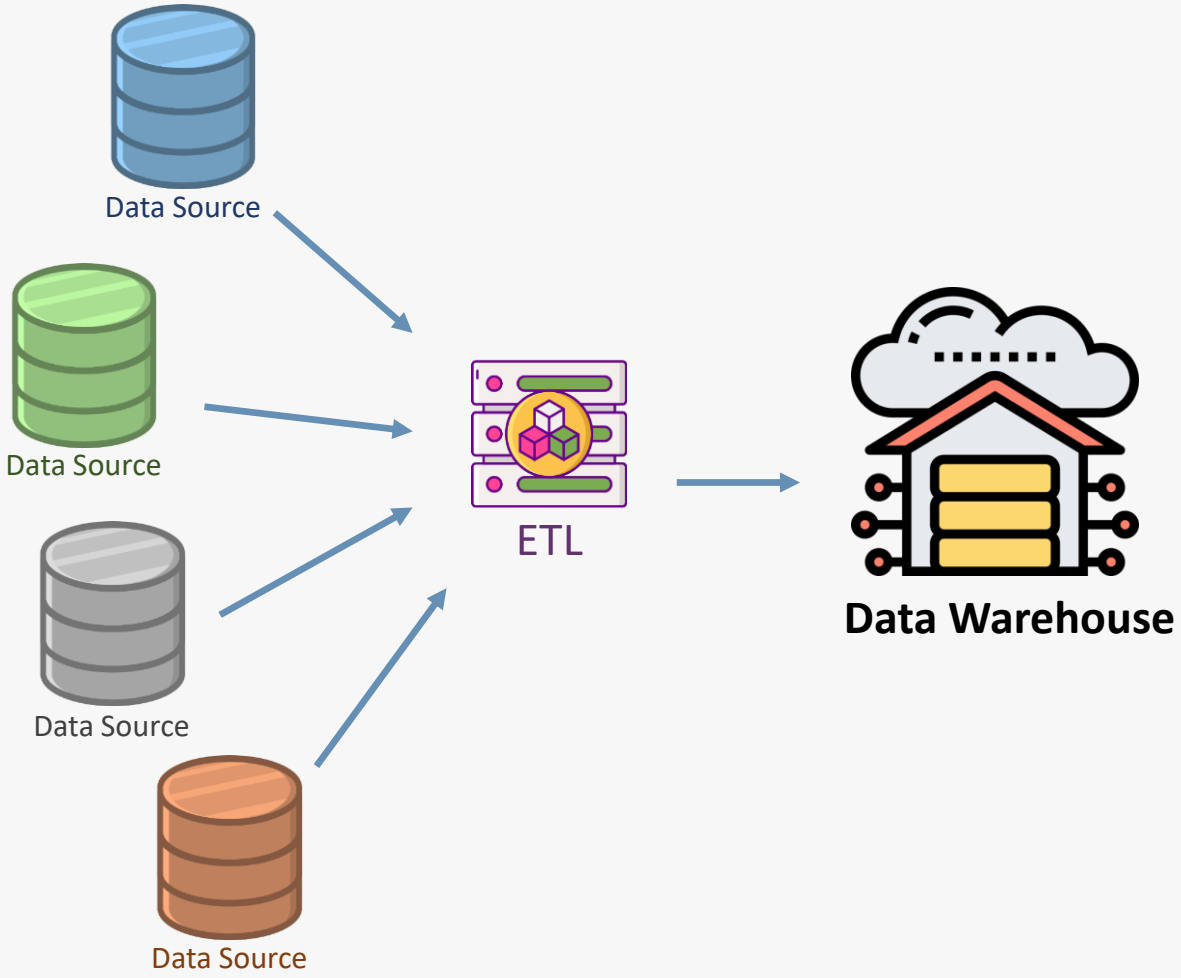
- **Wikipédia**: Le terme entrepôt de données (en anglais, data warehouse ou DWH) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise.

- ***Collecter :*** récupérer l'information utile
- ***Ordonner :*** structurer l'information en vue de son exploitation
- ***Journaliser :*** organiser l'historisation des données

Qu'est-ce qu'un Data Warehouse ?



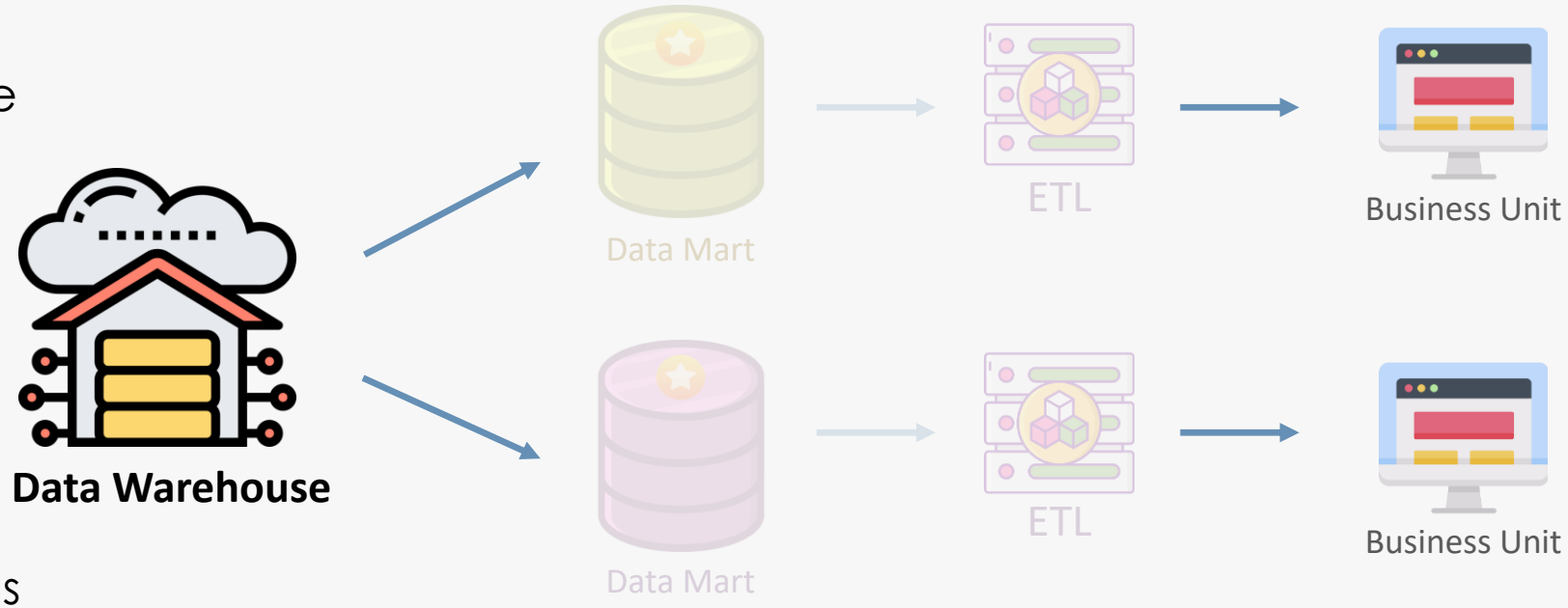
Qu'est-ce qu'un Data Warehouse ?



- Le Data Warehouse conserve une copie des informations des systèmes sources.
- Il permet de :
  - Rassembler des données provenant de sources multiples en une seule base de données
  - Exécuter des requêtes longues, bloquantes, sur des données opérationnelles
  - Maintenir l'historique des données, même si les systèmes de transaction source ne le font pas

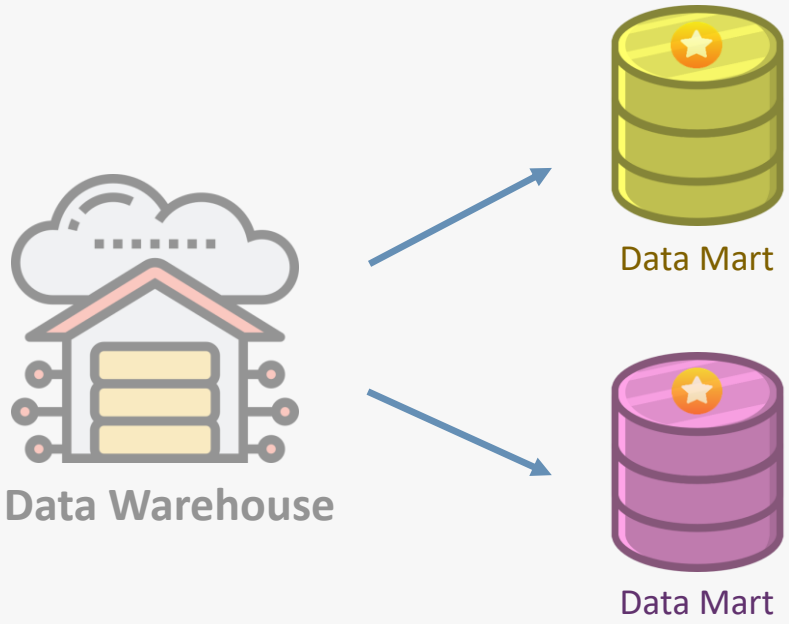
Qu'est-ce qu'un Data Warehouse ?

- Le Data Warehouse permet au niveau supérieur :
  - Présenter l'information de l'organisation en une vue centrale
  - Fournir un modèle de données commun pour toutes les données d'intérêt
  - Restructurer les données de sorte qu'elles prennent sens (décisionnel)
  - Ajouter de la valeur aux applications métiers opérationnels
  - Faire des requêtes d'aide à la décision plus faciles à écrire
  - Auditer et améliorer la qualité des données



Qu'est-ce qu'un Data Mart ?

- Le Data Mart :
  - *magasin de données*
  - sous-ensemble de Data Warehouse destiné à fournir des données aux utilisateurs
  - souvent spécialisé vers un groupe ou un type d'affaire



- Au niveau technique :
  - une base de données relationnelle
  - utilisée en informatique décisionnelle
  - exploitée pour restituer des informations ciblées sur un métier spécifique
  - constituant un ensemble d'indicateurs, utilisés pour :
    - le pilotage
    - l'aide à la décision

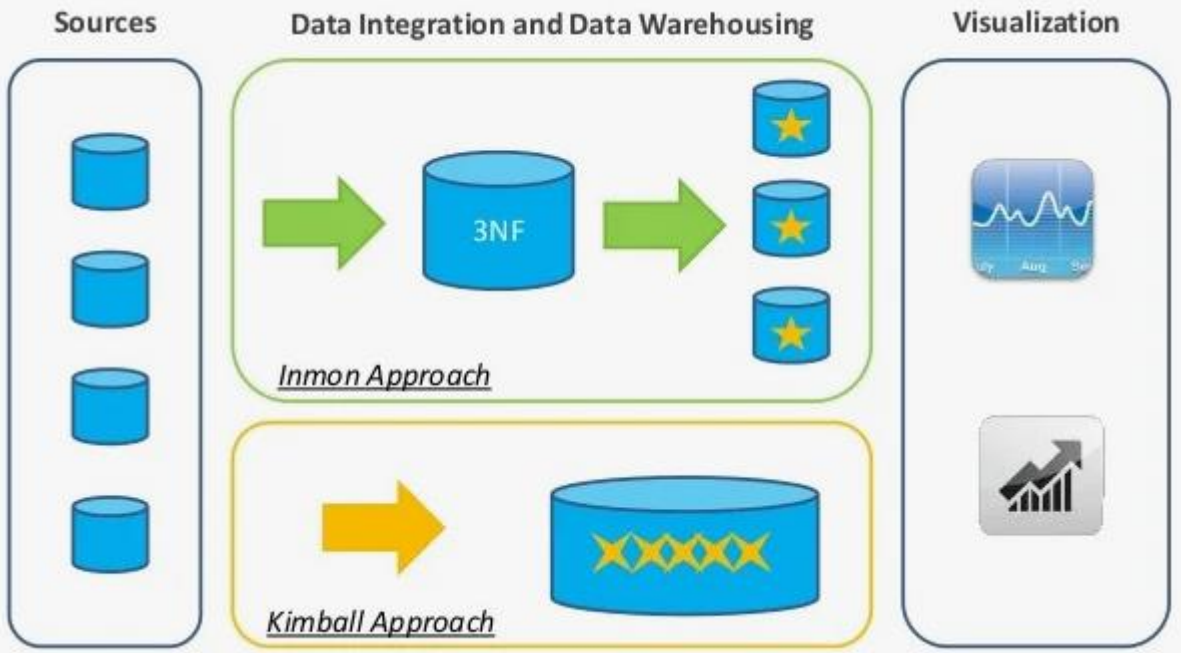
Le datawarehouse est **général**, le datamart est **spécifique à un métier**

DataWarehouse vs DataMart

Data Warehouse	Data Mart
Cadre • Indépendant de l'applicatif • Centralisé • Planifié	*Cadre* • Lié à un besoin • Décentralisé • Organique
Données • Historisées • Légèrement transformés	*Données* • Possiblement agrégées • Hautement transformées
Couverture • Multiples	*Couverture* • Dépendante d'un ou plusieurs cas d'usages
Sources • Nombreuses • Internes et externes	*Sources* • Limitées et sélectionnées • Via Data Warehouse
Dynamique • Flexibilité • Orientée données • Très longue durée de vie • Massif	*Dynamique* • Restrictif • Orientée projet • Durée de vie courte à moyenne • Volume modéré (à surveiller)

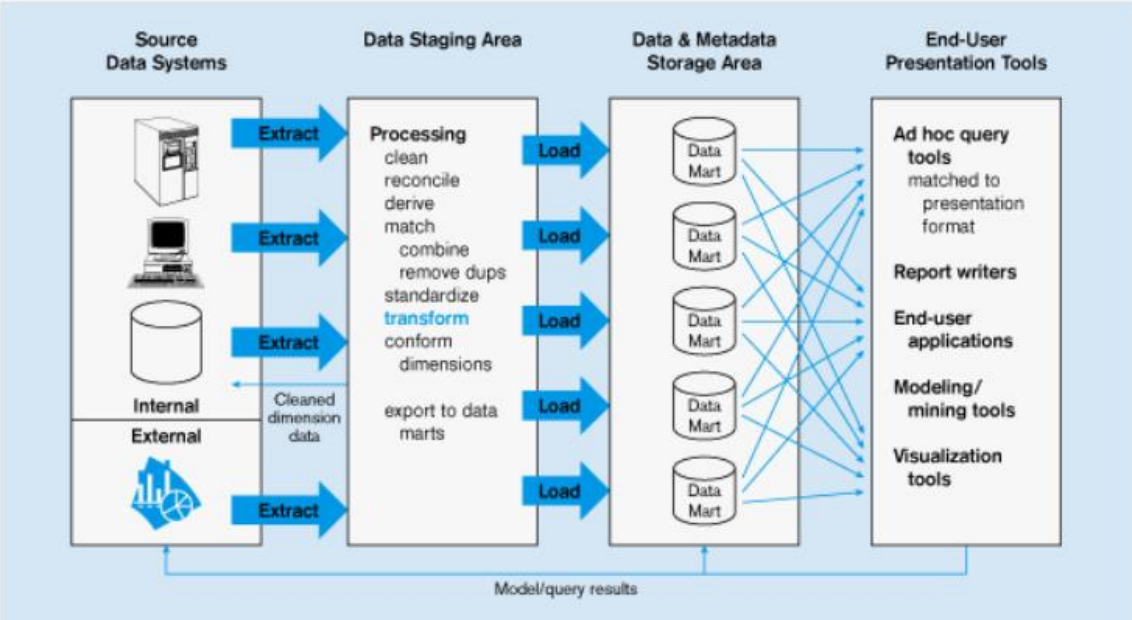
DataWarehouse vs DataMart

- Deux conceptions existantes :
  - **Définition d'Inmon :**
    - Le DataMart est issu d'un flux de données provenant du DataWarehouse
    - Il présente la donnée de manière spécialisée, agrégée et regroupée fonctionnellement.
  - **Définition de Kimball :**
    - Le DataMart est un sous-ensemble du DataWarehouse
    - Il est constitué de tables à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier.
    - L'ensemble des DataMarts constitue le DataWarehouse.



Articulation DataWarehouse - DataMart

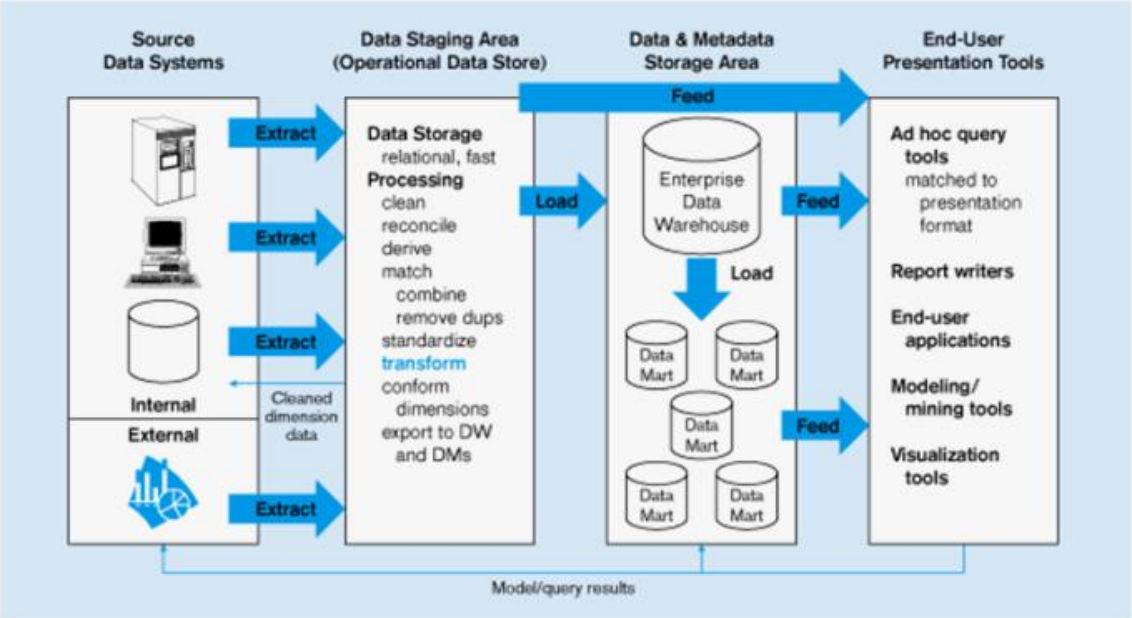
- **Datamarts indépendants** *(Inmon)*



- Extraction et chargement tubulaires
- Accès à la donnée complexe

Articulation DataWarehouse - DataMart

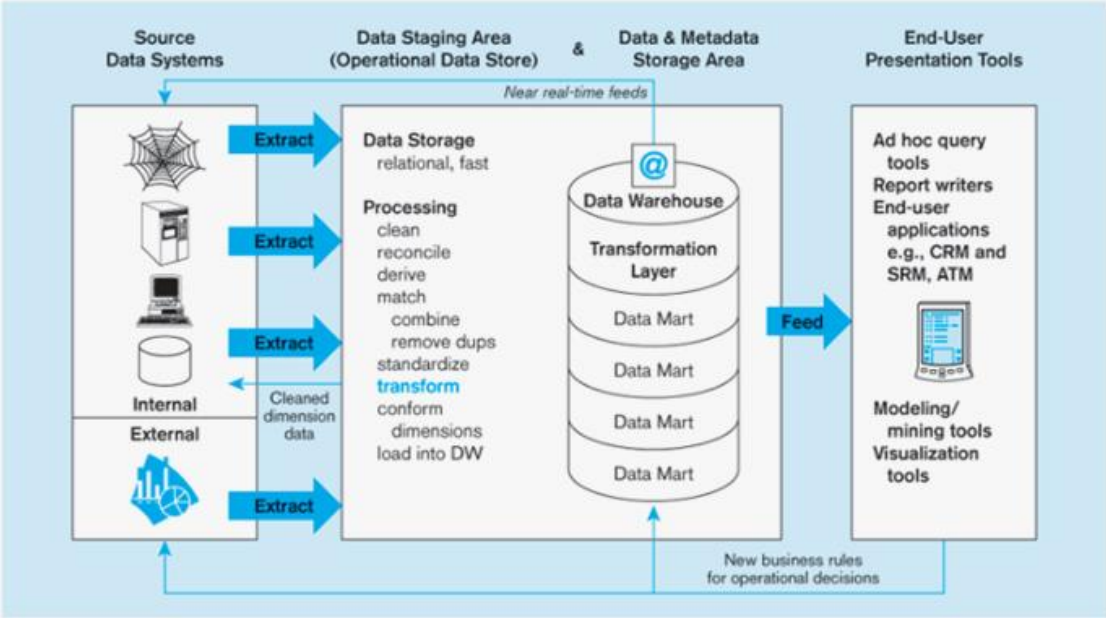
- **Datamarts dépendants** *(hybride)*



- Stockage opérationnel
- Dépendances fonctionnelles

Articulation DataWarehouse - DataMart

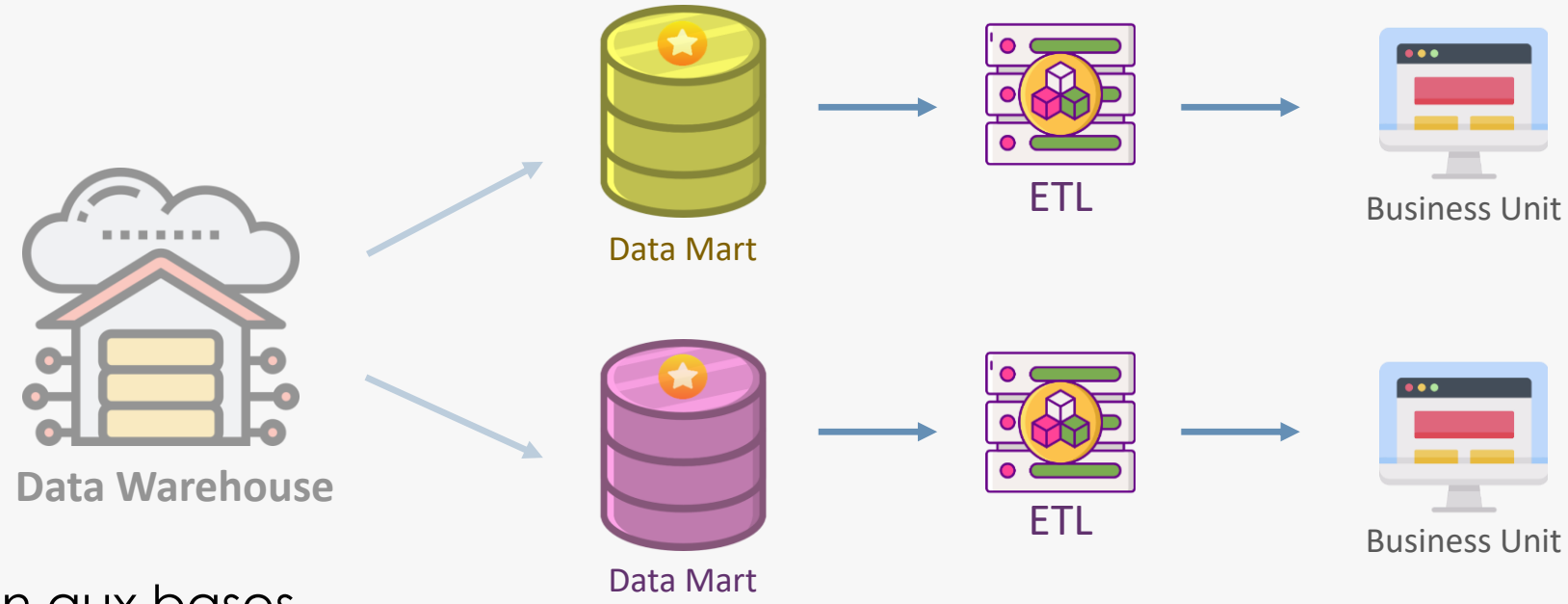
- **Datamarts logiques** *(Kimball)*



- Les datamarts sont des vues logiques du DWH
- Dépendances techniques

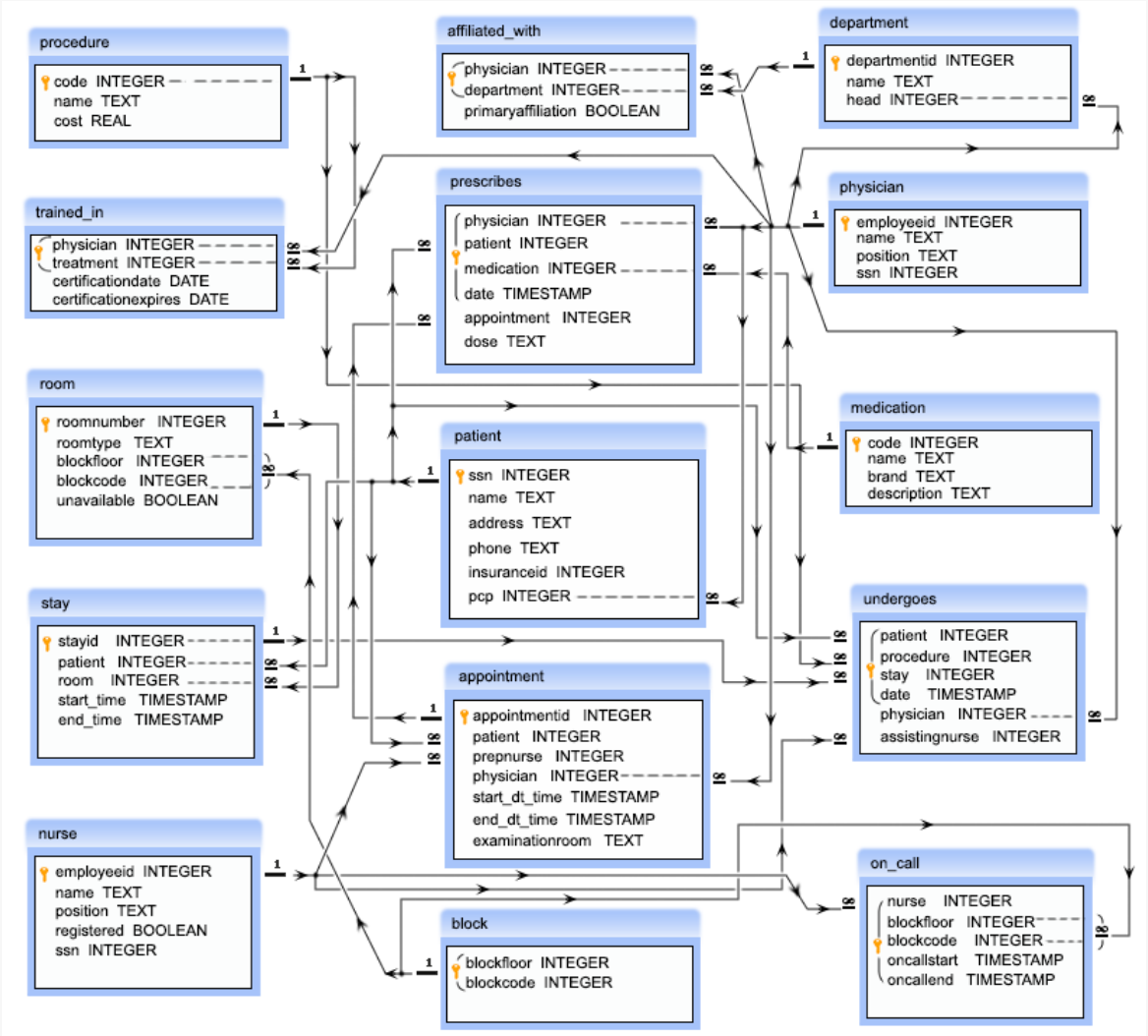
Utilité pour les Business Units

- Données orientées sujets
  - En production : données organisées par **processus fonctionnels**
  - Données structurées par thème, potentiellement **transverses** par rapport aux domaines fonctionnels et organisationnelles
- Exemples :
  - Actes, Séjours en opposition aux bases de données par services ou applicatif



Modèle relationnel

- Normalisé (3eFN)
- Répond aux besoins transactionnels
- Avantages :
  - Réduction de l'entrée de données
  - Réduction du nombre d'index
  - Ajouts/destructions/modifications plus rapides
- Désavantages :
  - Peu efficace pour l'extraction de données analytiques
  - Beaucoup de relations
  - Trop complexe pour l'utilisateur BI



peu approprié pour les DWH

Modèle dimensionnel

- Part du besoin « client »
- Définit des faits et des dimensions
  - **Fait :** un sujet d'analyse. Un fait est caractérisée par plusieurs informations
  - **Dimension :** un critère selon lesquels on souhaite faire de l'analyse

Aussi connu sous le nom de modèle en étoile

IV. En pratique : intégration de données

au sein de l'EDS de l'AP-HP

- Processus d'intégration
- Méthodologie de cadrage
- Cartographie simplifiée

Processus d'intégration d'un flux dans l'EDS

Définir les besoins

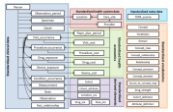
- Comprendre les enjeux métiers
- Identifier les besoins métier pour la recherche et le pilotage
- Définir les objectifs en termes de résultats attendus

Collecter & Stocker

- Comprendre les données sources (rétro-engenniering)
- Définir le périmètre des données à intégrer dans l'EDS
- Collecter et stocker les données dans l'EDS

Modéliser & Standardiser

- Aligner les structures et les concepts suivant le besoin
  Modèles : FHIR, OMOP, ontologies i2b2, BI …
- Standardiser par rapport à des référentiels (terminologies)



FHIR

Qualifier

- Evaluer la qualité des données

Exploiter & visualiser

- Mettre en œuvre l'analyse des données
- Définir les méthodes de visualisation des données

COHORT 360

i2b2

jupyter

COGNOS

Méthodologie de cadrage

Quoi ?

Compréhension du contexte
- Périmètre et vision globale du système source et cible
- Compréhension des besoins (Recherche, Pilotage)

Analyse fonctionnelle
- Système source, fonctionnalités, processus
- Organisation et utilisateurs métier
- Définition des domaines et glossaire des concepts métier

Analyse technique
- Architecture applicative et flux d'informations
- Analyse applicative (volumétrie, schémas, …)

Périmètre des données
- Inventaire des données métier et de référence sources (dictionnaire de données)

Comment ?

Stratégie d'extraction des données sources

Stratégie d'intégration des données dans l'EDS
- Exigences fonctionnelles (règles de gestion de transformation, contrôles, ...) avec la prise en compte des contraintes fonctionnelles et applicatives
- Exigences non fonctionnelles (initialisation, profondeur d'extraction, volumétrie, fréquence d'extraction, mode d'alimentation, historisation, rejeux, ordonnancement, calendrier, contrôles techniques, …)

Collecte des données
- Intégration des données en mode silo
- Dictionnaire des données intégrées

Modélisation et transformation
- Transformation et restructuration des données
- Enrichissement des données
- Chaînage des données entre les différentes sources et par domaine

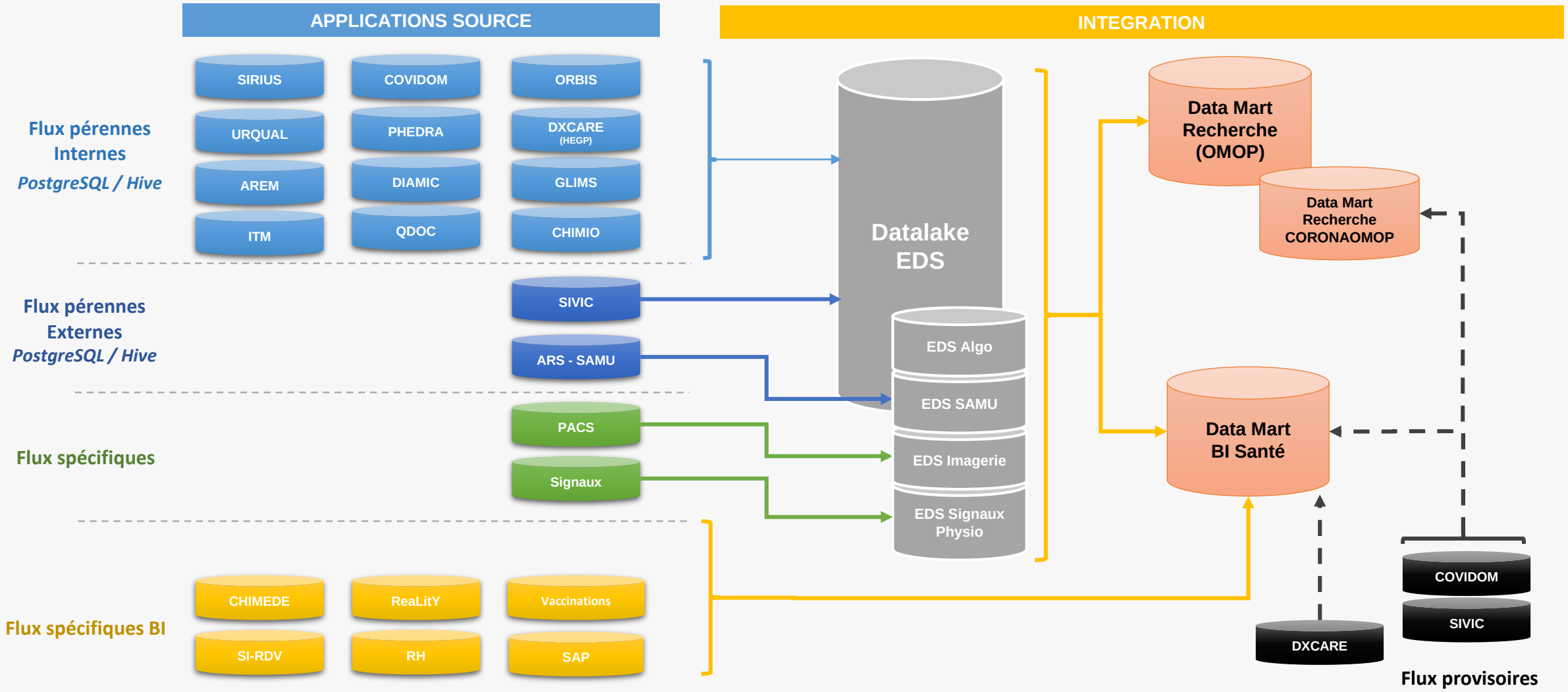
Recette et stratégie de qualité des données
- Validation des données intégrées
- Identification du cadre d'évaluation de la qualité des données

Stratégie de standardisation
- Standardisation par rapport à des référentiels (terminologies)
- Modèle de standardisation (modèles FHIR, OMOP)

Gestion des risques
- Identification des risques
- Contraintes inter-applications

Sécurité
- Sécurité liée aux données intégrées (CNIL, anonymisation, droits d'accès)

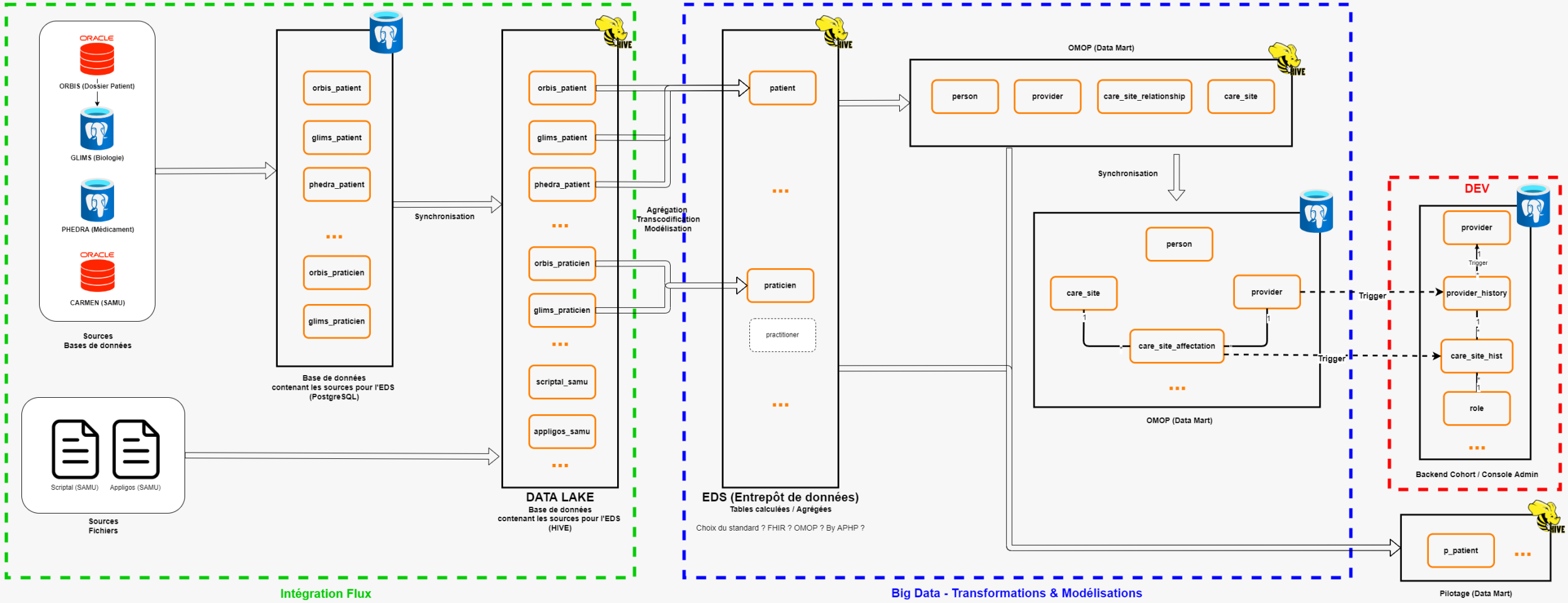
Cartographie simplifiée



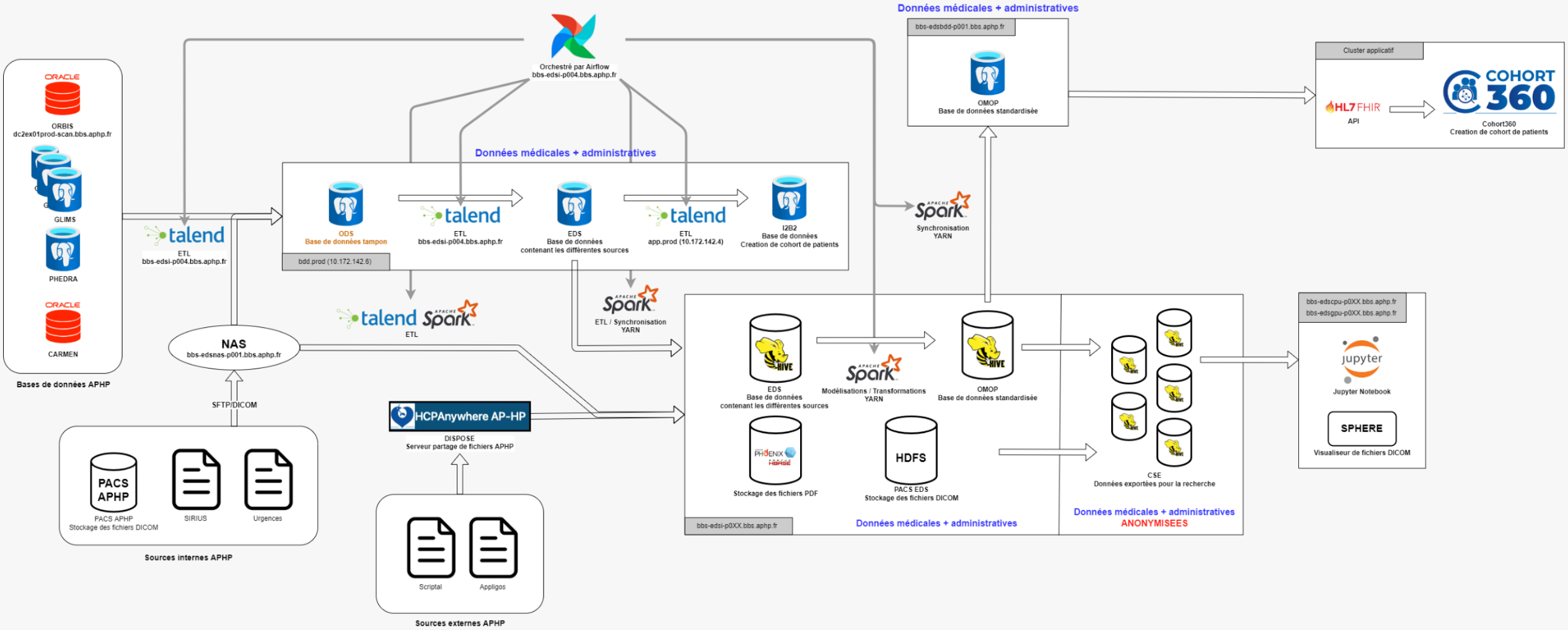
V. En pratique : implémentation

- Cartographie d'implémentation
- Technologies impliquées

Cartographie d'implémentation (simplifiée)



Technologies impliquées (simplifié)



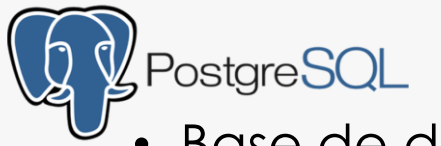
Technologies impliquées (exemples)

- Gestion de flux de travail
- Workflows en scripts Python

- Intégration des données
- Orchestrateur graphique

- Framework de calcul distribué
- Scripts en Scala, Java ou Python

- Infrastructure Hadoop
- HiveQL (SQL-like) et conversion Spark



- Base de données relationnelle
- Compatibilité applicative utilisateur

VI. Conclusion

Points d'attention à souligner

- Processus critique de gestion de la **donnée stratégique**
- Chantier long terme, consommation importante de ressources
- Structure dépendante des **besoins** et des **usages**

→ Enjeu fondamental de la **planification** et de la **gouvernance**

URL du cours et licence

Permalien : https://www.damien-leprovost.fr/enseignements/ETL.2022.pdf