Pr. Schiappa

Entrepôts des données

ECUE 5 - SANTÉ NUMÉRIQUE

Solèmostase

(IDR/CDW) :

Plateformes utilisées pour l'intégration de sources de données au travers d'outils d'analyses spécialisés.

BIG DATA :

Gros volumes qui alimentent la vie quotidienne d'un hôpital

ENTRÊPOT DES DONNÉES CLINIQUES

3 DIMENSIONS :

- **Volume** : Les données proviennent de diverses sources (poids différents)
- **Vitesse** : Les données sont produites à un rythme de plus en plus soutenu et doivent être traitées rapidement
- **Variété** : Les données sont sous des formats différents

Objectif : Faciliter le traitement + analyse de données massives
(meilleure prise en charge meilleures recherches)

Solèmostase

PROBLÉMATIQUE BIG DATA :

90% du volume total des données ont été produites ces 2 dernières années MAIS + de **80%** ne sont pas exploitées.

- **Base de la sécurité sociale :**

- Feuilles de soins : 8,9 milliards + 2,3 Go (Sniiram)

- **Centre Antoine Lacassagne (CAL)**

Données structurées : format prédéfini (20% des données)
CAL : 80% des données sont non structurées (texte libre)

RÉFÉRENCE COURS DONNÉES ET QUALITÉS DE DONNÉES :

❖ **Données structurées** = informations (mots, signes, chiffres...) contrôlées par des référentiels et présentées dans des cases (les champs d'une base de données) qui permettent leur interprétation et leur traitement par des machines.

❖ **Données non-structurées** = le reste, tout ce qui n'est pas organisé en base de données, c'est-à-dire la bureautique, la messagerie, les images, les vidéos, etc.

ENTREPÔT DE DONNÉES CLINIQUES :

Les centres hospitaliers :

- La pratique de la médecine
- Les Files active
- Traitements/répartition
- Questions cliniques et/ou fondamentales

DÉFINITIONS :

« Un entrepôt de données va **recueillir** et **regrouper** les données importantes et les **associer** aux patients. Les propriétés des variables, des champs, leurs noms, les règles sont définies, idéalement utilisent un standard international. Les données sont solides et ne changeront pas à chaque mise à jour, elles retraceront le parcours du patient et seront à jour »

ETL “EXTRACT – TRANSFORM – LOAD “ :

- ***Extract :*** connecter les différentes sources de données et d’extraire les données nécessaires

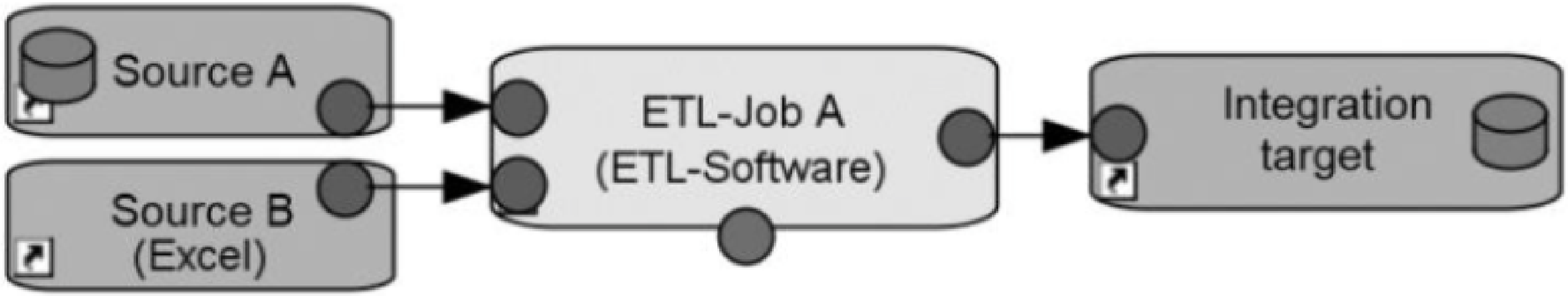
Problématique : hétérogénéité des sources de données qui nécessiteront de multiples approches pour la connexion et l’extraction des données.

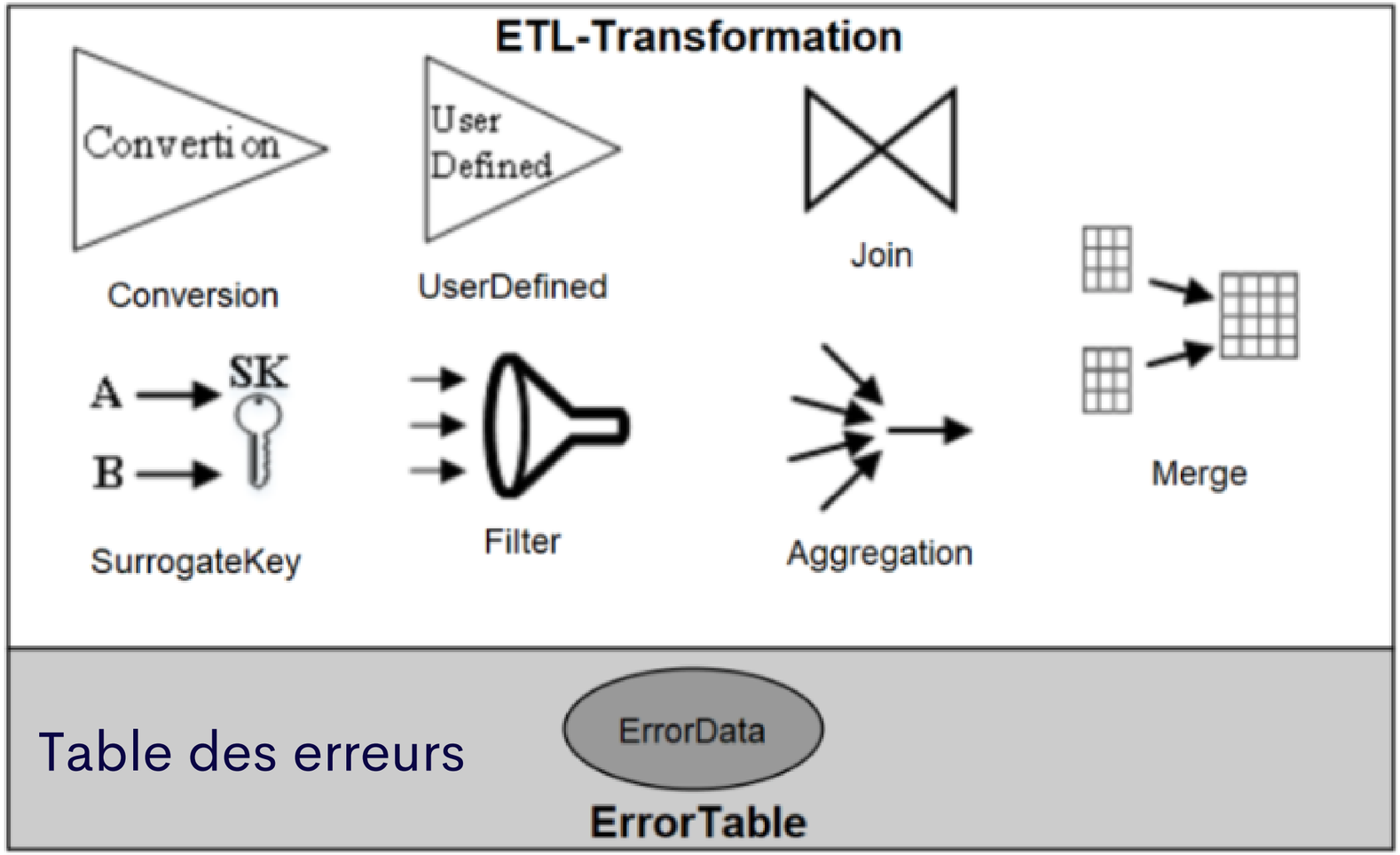
- ***Transform*** : les données extraites sont transformées dans un format spécifique, défini à l’avance. Cette étape facilite l’intégration et la consolidation des données pour l’étape finale

Problématique : définition et reconnaissance des formats à appliquer, prise en charge des nouvelles données, évolution des formats de données en fonction du temps, interopérabilité des formats

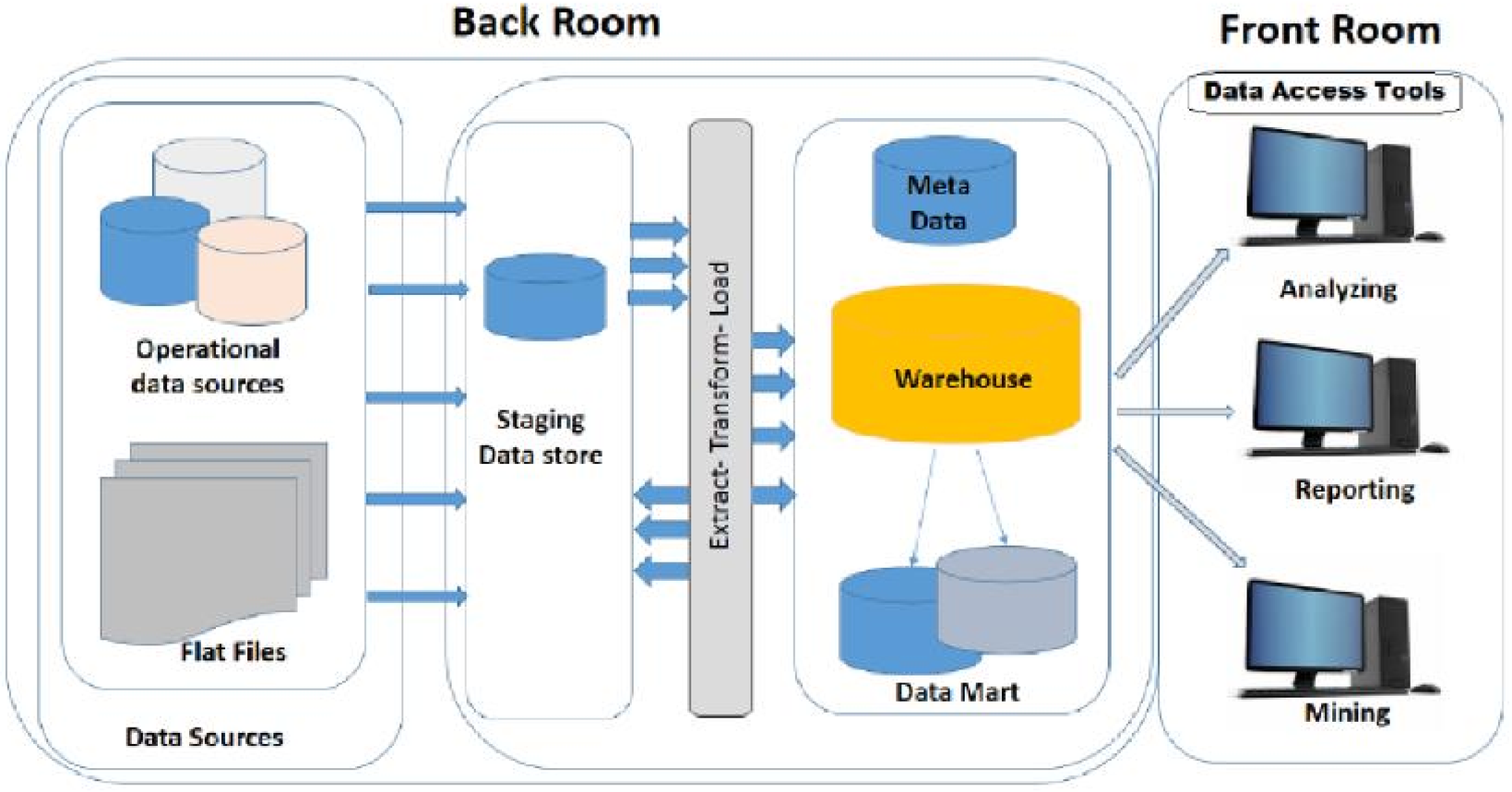
- ***Load*** : Les données sont transformées dans leurs formes/dimensions finales

Problématique : la gestion des « anciennes » données versus celles à mettre à jour

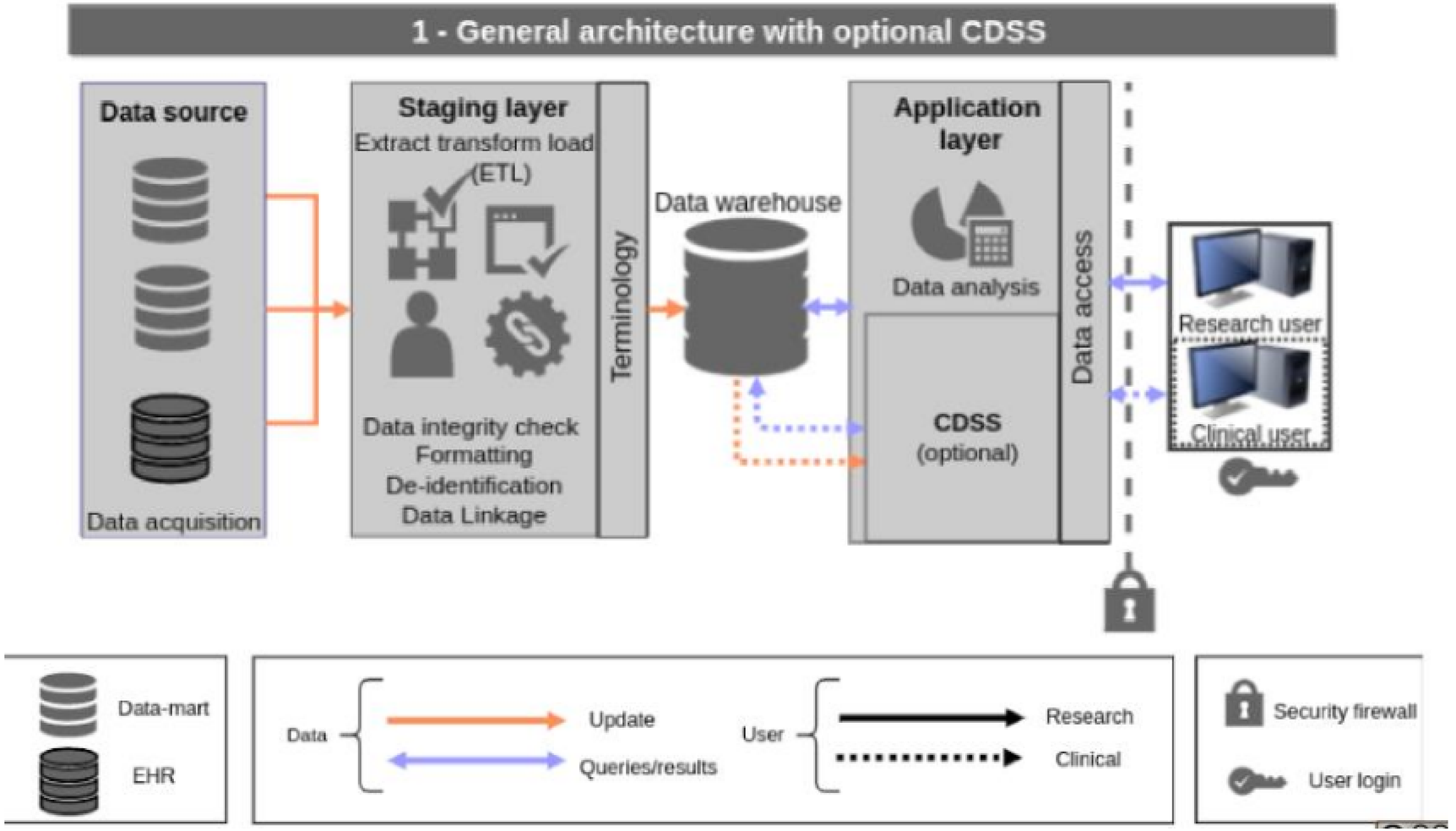

Source A
Source B
(Excel)
ETL-Job A
(ETL-Software)
Integration
target


ETL-Transformation
Convertion
Conversion
User
Defined
UserDefined
Join
A
B
SK
SurrogateKey
Filter
Aggregation
Merge
Table des erreurs
ErrorData
ErrorTable

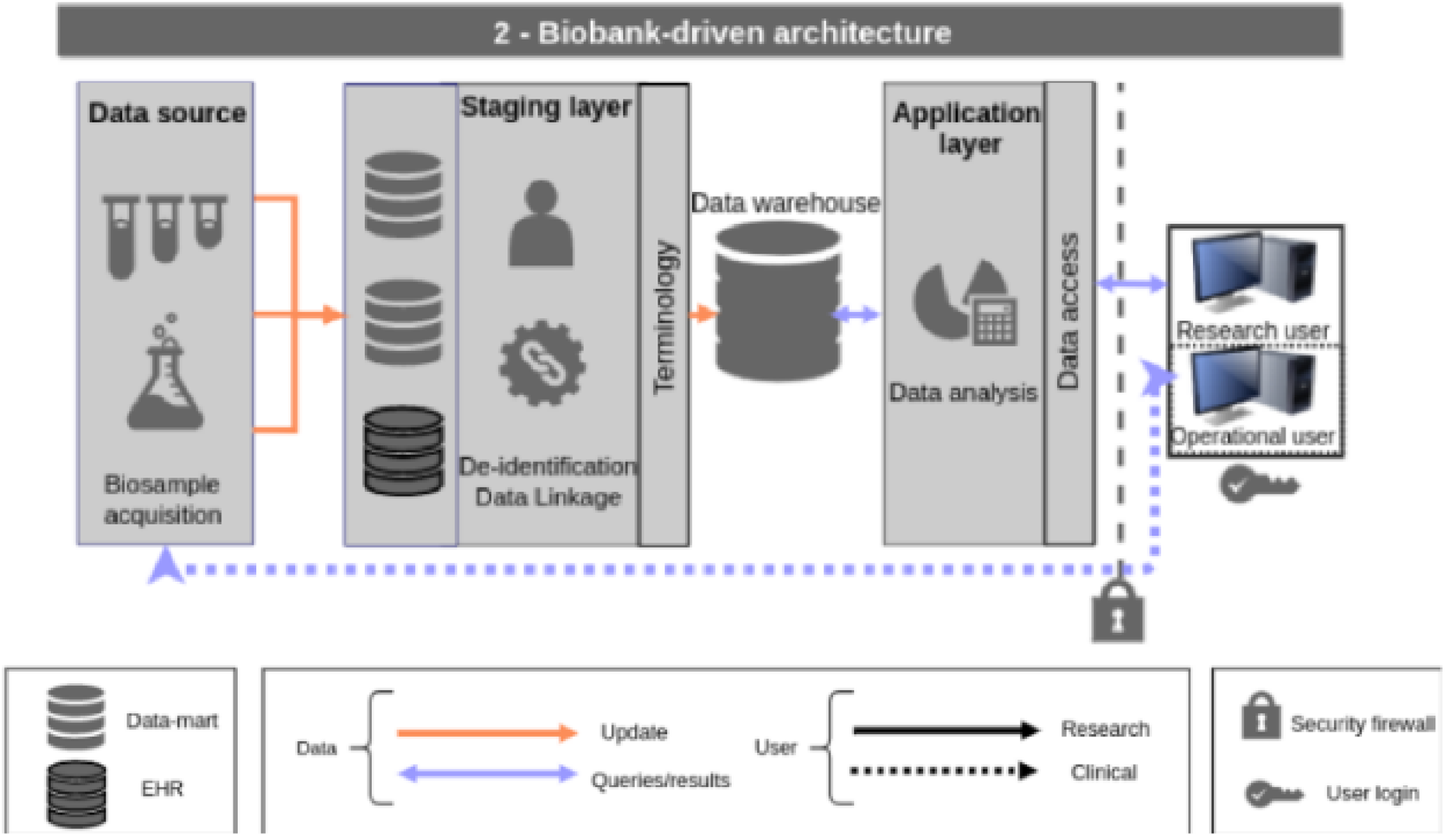
Architectures d'un entrepôt de données

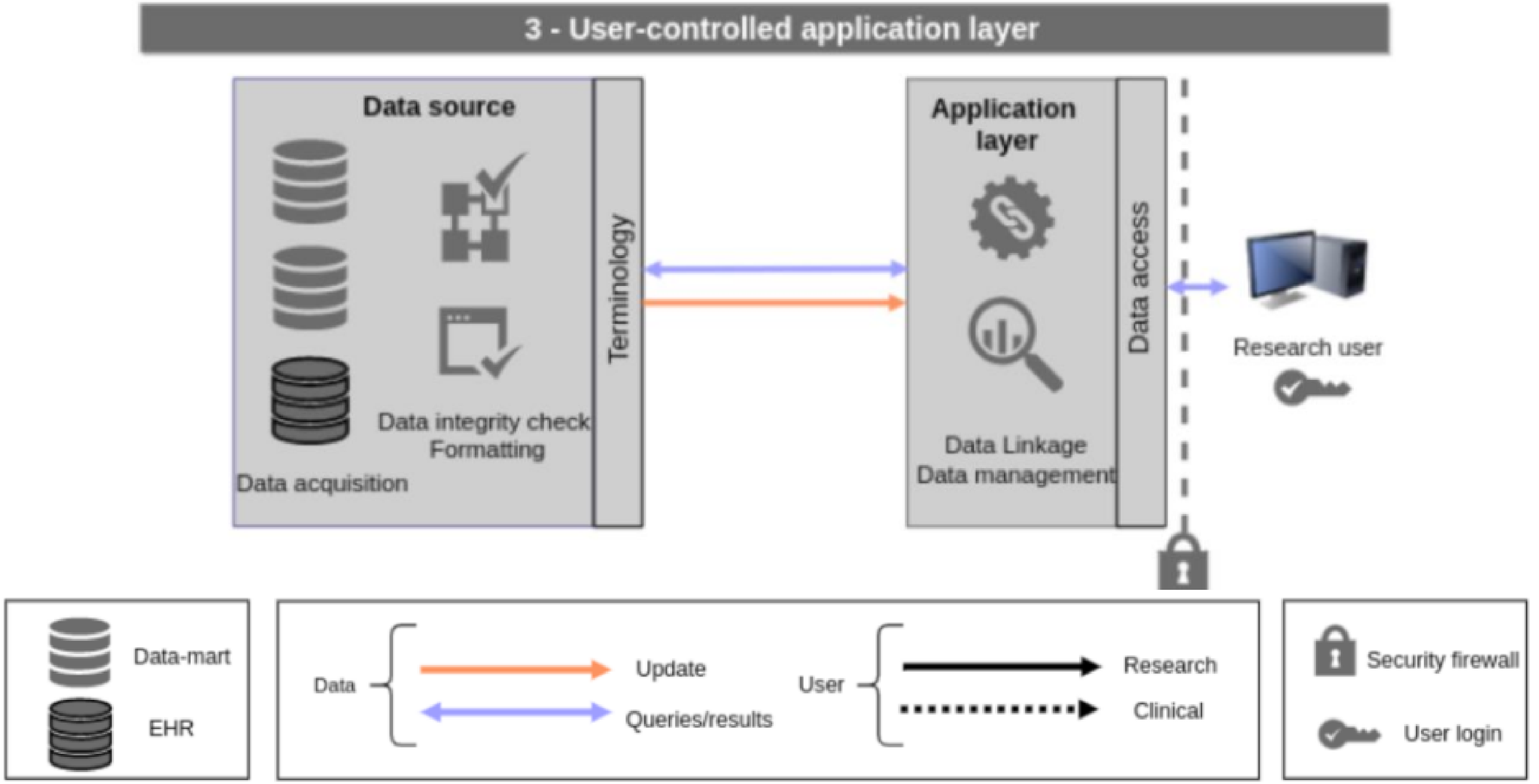


4 grands types d'architecture : General architecture

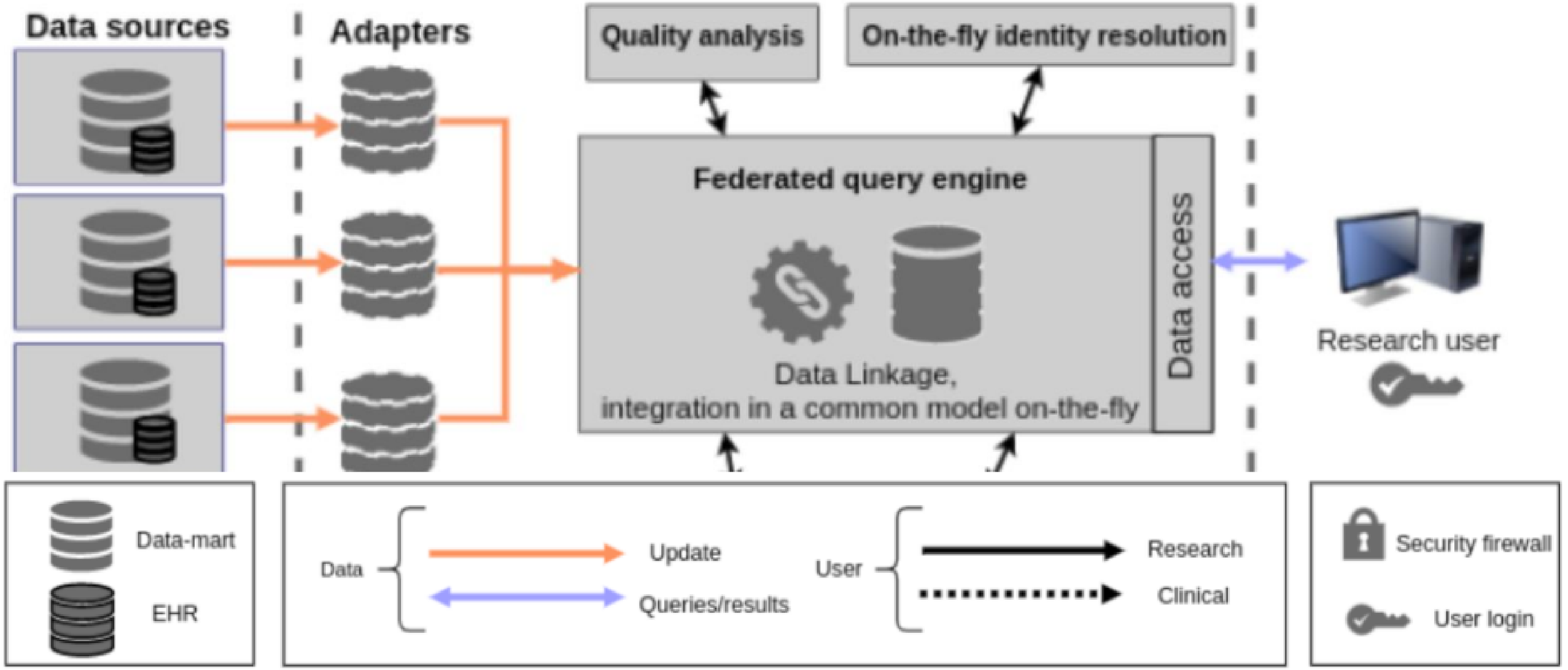


4 grands types d'architecture : Biobank driven




3 - User-controlled application layer
Data source
Data integrity check
Formatting
Data acquisition
Terminology
Application layer
Data Linkage
Data management
Data access
Research user
Data-mart
EHR
Data
Update
Queries/results
User
Research
Clinical
Security firewall
User login

4 - Federated architecture: inter-institutions data integration



❖ Chercheurs(-ses) : cherchent des traits cliniques qui permettent d'identifier des cohortes répondant à des questions précises → toutes les architectures leurs sont utiles

❖ **Médecins : aide à la prise de décision pour les traitements, interventions, risques pour un(e) patient(e) → La 1ère architecture avec CDSS est la plus appropriée.**

Toutes les architectures ne peuvent pas correspondre à tous les profils.

SOURCES ET DISPONIBILITÉS

chaque source de données cliniques est UNIQUE
Elles ne se ressemblent pas
Complétude + design des sources

FORMAT

DONNÉES

RÉCUPÉRATION

Le traitement des données suivant l'ETL est composé de plusieurs étapes :

1. Extraction (automatique ou manuelle) d
2. Anonymisation (optionnel) et attribution d'un identifiant unique.
3. Transformation et standardisation .
4. Mapping avec la terminologie standard utilisée.
5. Mapping des données entre les différentes sources.
6. Chargement dans la CDW (mise à jour ou réimport total).

STANDARDISATION ET INTÉGRATION

Common Data Model (CDM) : INTEGRATION BIOLOGY AND THE BEDSIDE (i2b2)

DÉFINITIONS :

- **Project management** : sécurité, identification des utilisateurs/trices, rôles.
- **Ontology management** : gère la terminologie.
- **Data repository** : gère les données structurées, permet l'interrogation et la visualisation des données.
- **File repository** : stocke les « gros » fichiers (images, puces)
- **Workflow Framework** : gère les intéractions entre les différentes « hives ».
- **Identity management** : anonymisation des patients.
- **Web client application** : permet aux utilisateurs/trices d'interroger le CDW.
- **Workbench** : application permettant d'analyse les données de façon plus précise.

SÉCURITÉ

Il est crucial de fixer les règles de sécurité

CONSEILS DU PROFESSEUR

Longétivité du projet

Architecture en adéquation des besoins

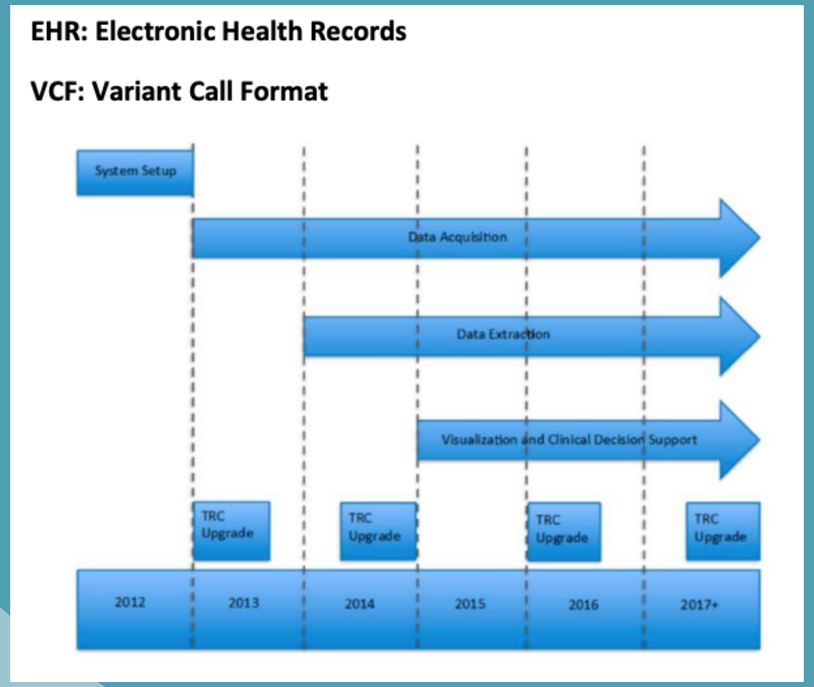
CDM déjà utilisé pour une meilleure approche

Terminologie

Mise à jour, détails du processus ETL, automatisation

information des utlisateurs

ENTRÊPOT DES DONNÉES



EXTRACTION DES DONNÉES

Variables d'intéret, nombre de valeurs, temporalité, effectuer

des évalutions de la qualité des données

Solèmostase

EXEMPLES :

George Pompidou University Hospital Clinical Data Warehouse

i2b2 + 3 niveaux d'accès aux données :

- 1° Seulement accès aux données agrégées répondant aux critères de sélection
- 2° : cohortes anonymes avec les données détaillées.
- 3° : Cohorte avec toutes les données, non anonyme.

CAL

Plateforme de données + Analyse des sources de données disponibles.
Mise en place d'une structuration automatique des données d'intelligence artificielle RUBY

RUBY

Language de programmation, facile de compréhension populaire !

Quelles sont les 3 caractéristiques des BIG DATA ?

VOLUME VITESSE ET VARIÉTÉ

Définition de ETL ?

Extract - transform - Load

Solèmostase

Quel est le type d’architecture utilisé par les médecins ?

CDSS (extrêment puissant, lien entre tous les hopitaux