Business Intelligence
Master Data-Science
ETL et Datawarehouse

Ludovic DENOYER - ludovic.denoyer@lip6.fr

UPMC

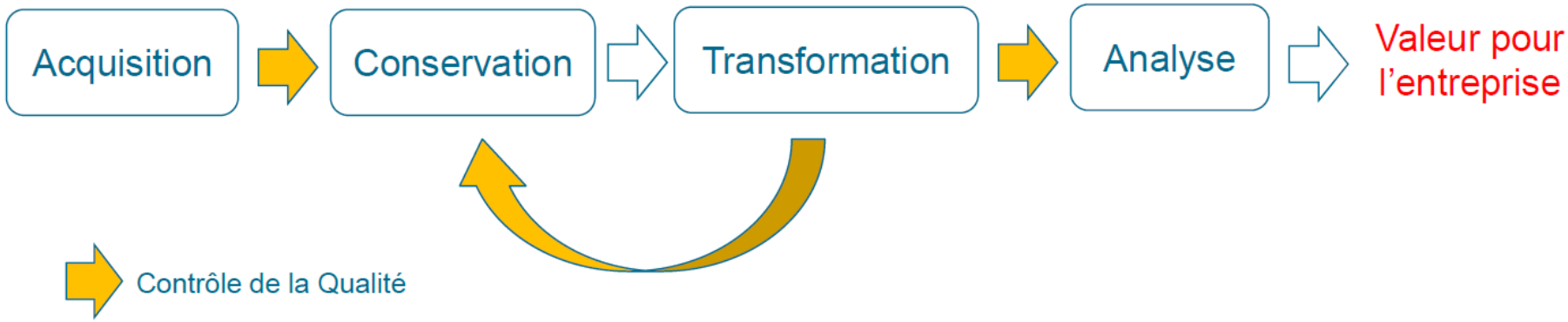
17 janvier 2016

Rappel

L'Informatique Décisionnelle (ID), en anglais Business Intelligence (BI), est l'informatique à l'usage des décideurs et des dirigeants des entreprises. Les systèmes de ID/BI sont utilisés par les décideurs pour obtenir une connaissance approfondie de l'entreprise et de définir et de soutenir leurs stratégies d'affaires, par exemple :

- d'acquérir un avantage concurrentiel,
- d'améliorer la performance de l'entreprise,
- de répondre plus rapidement aux changements,
- d'augmenter la rentabilité, et
- d'une façon générale la création de valeur ajoutée de l'entreprise.
- **...et à créer de nouveaux services...**

Les fonctions



Source : Groupe de travail CIGREF, 2014

Différents "métiers" :

- Data Integrator
- Data Analyst
- Data Scientist

+ **Data Steward (Responsable des données)**

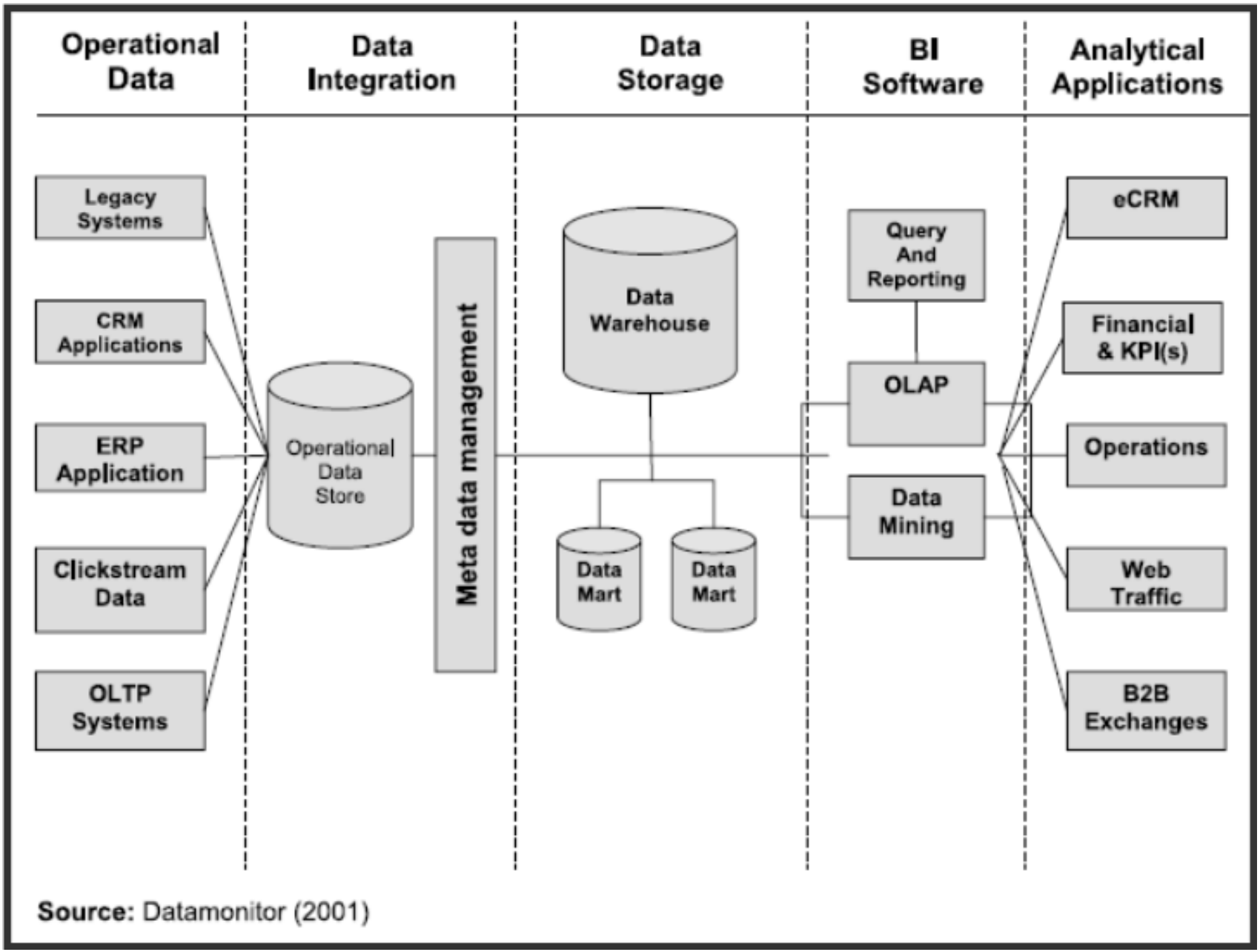
Les fonctions de la BI

Plan du cours

- Fonction de collecte de données
- Fonction d'intégration
- Fonction de diffusion (ou distribution)
- Fonction présentation

Aujourd'hui : Collecte de données + intégration

L'architecture classique de la BI



Données de l'entreprise

Les données de l'entreprise sont stockées dans des systèmes transactionnels qui enregistrent les données quotidiennes.

Différentes sources de données :

- Fichiers Excel....
- ERPs
- Systèmes de CRMs
- Capteurs

Et aujourd'hui :

- Données du Web
- Données sociales
  - Twitter
  - ...
- Données des ojets connectés

Difficultés

- Sources diverses et disparates ;
- Sources sur différentes plateformes et OS ;
- Applications utilisant des BDs et autres technologies obsolètes ;
- Historique de changement non-préservé dans les sources ;
- Qualité de données douteuse et changeante dans le temps ;
- Structure des systèmes sources changeante dans le temps ;
- Incohérence entre les différentes sources ;
- Données dans un format difficilement interprétable ou ambigu.

Intégration de données

Définition

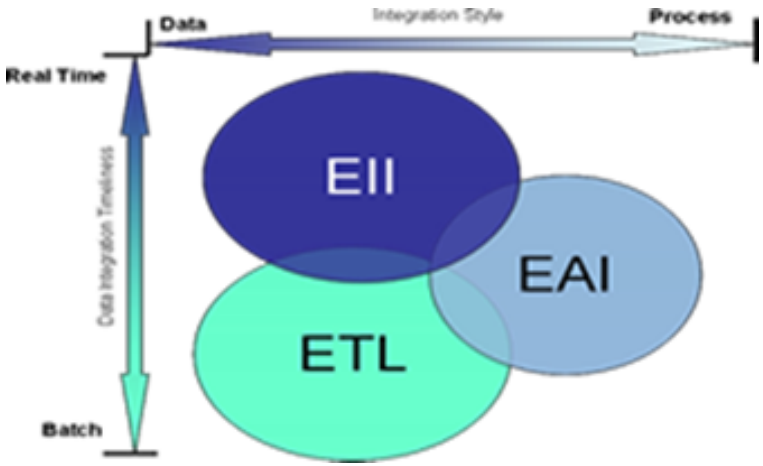
L'intégration de données appelé ETL (Extraction Transfer Loading) regroupe les processus par lesquels les données provenant de différentes parties du système d'information sont déplacées, combinées et consolidées. Ces processus consistent habituellement à extraire des données de différentes sources (bases de données, fichiers, applications, Services Web, emails, etc.), à leur appliquer des transformations (jointures, lookups, déduplication, calculs, etc.), et à envoyer les données résultantes vers les systèmes cibles.
Source : wikiversity.org

Il existe plusieurs système d'intégration de données :

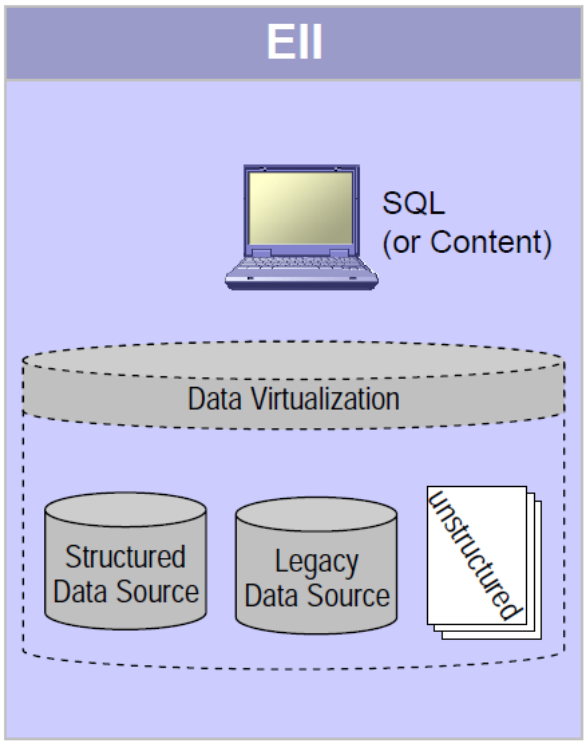
- La médiation au service de l'intégration de données d'entreprise (EII).
- L'intégration de données via les applications (EAI).
- L'intégration de données via les services Web (ESB, SOA).
- L'intégration de données en nuage (Data Cloud).
- L'ETL (Extract - Transform - Load)

Intégration de données

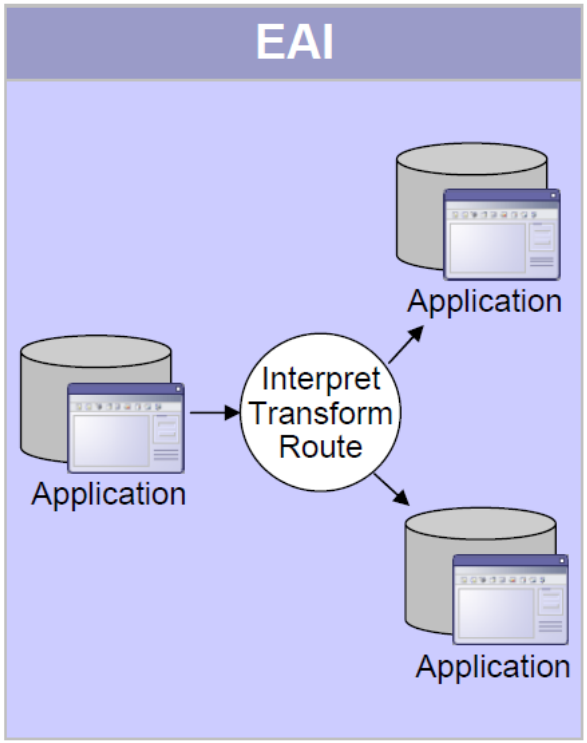
- La médiation au service de l'intégration de données d'entreprise (EII).
- L'intégration de données via les applications (EAI).
- L'ETL (Extract - Transform - Load)



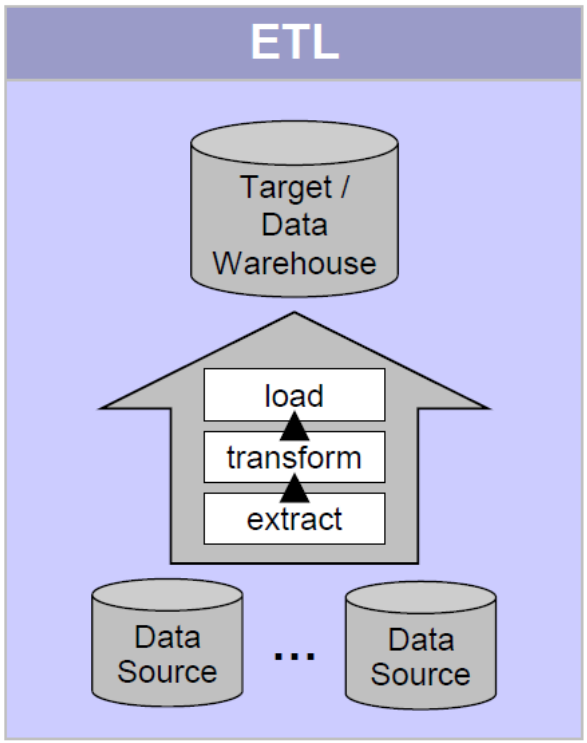
EII - EAI - ETL



- Real-time information access
- Federation of data from multiple sources
- Dynamic drill down
- Semi-structured & unstructured data



- Process based integration of application data
- Message-based, transaction-oriented processing
- Workflow and data orchestration, content-based routing



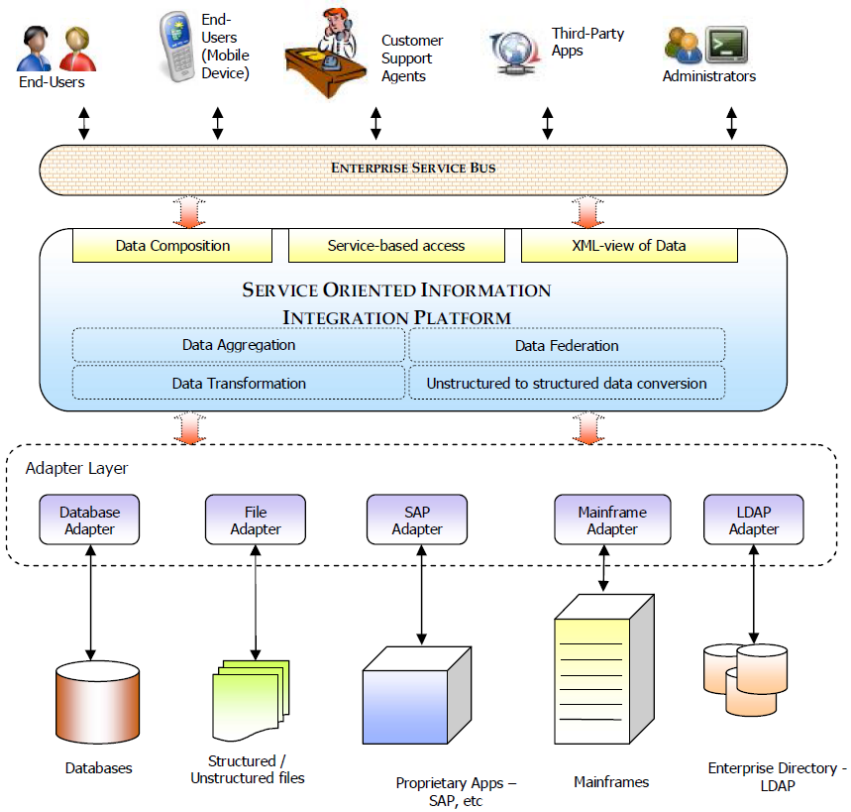
- Bulk data integration
- Set-based & hierarchical transformations
- High scale, batch-oriented data delivery

Source : IBM Software group

EII - Entreprise Information Intégration

Définition

Enterprise Information Integration (EII) est une approche d'architecture (voire d'urbanisme) permettant d'obtenir une vue unifiée des données informatiques de l'entreprise.
Source : Wikipedia



EII - Caractéristiques

En fonction des choix retenus, l'utilisateur aura la possibilité de :

- modifier les données (et non pas seulement un accès en lecture seule) ;
- agir en temps réel sur les données (et non pas en différé) ;
- accéder à des données structurées ;
- accéder à des données cohérentes ;
- accéder à des services ;
- remonter des informations jusque dans le modèle métier (objet) ;

EII

- EII Major Strengths
  - Relational access to non-relational sources
  - Ability to explore data before a formal data model and metadata are created
  - Quicker deployment
  - Can be reused by ETL and/or EAI further developments
  - Access in place data, meaning it avoids unnecessary movement of data.
  - Optimized for global access to remote sources
  - Event publishing technology provides a non-intrusive means to “listen” for particular changes (insert, update or deletes) that defined as being of interest.

Source : IBM Software Group

EII

- EII Main Challenges
  - Need Matching keys across sources
  - Data types mismatch
  - Data reconciliation
  - Possibly high resource utilization on the source system
  - Limited to hundreds of thousands of rows for remote result sets
  - Performance degradation when query pushdown is not used
  - Limited transformation – bounded by SQL capability and system capacity
  - May consume network bandwidth during peak hours
  - Multi-site updates require transactional control (2PC)

Source : IBM Software Group

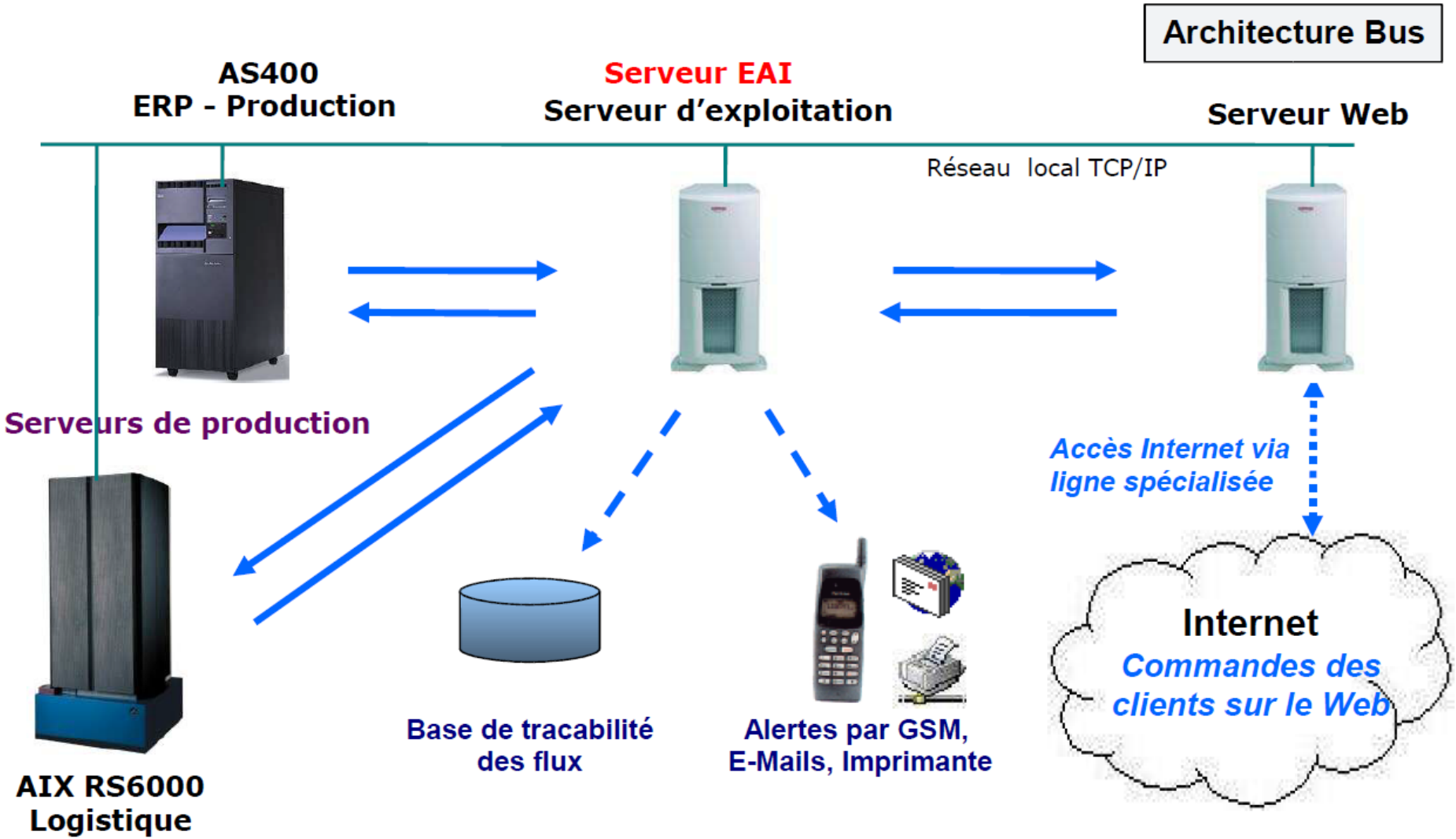
EAI - Entreprise Application Integration

Définition

L'intégration d'applications d'entreprise est une architecture intergicielle permettant à des applications hétérogènes de gérer leurs échanges. On la place dans la catégorie des technologies informatiques d'intégration métier (Business Integration) et d'urbanisation. Sa particularité est d'échanger les données en pseudo temps réel.
Source : wikipedia

- Logique de "Bus" ou de "Hub"
- Messages

Architecture EAI - Exemple



Source : Seralia

EAI

- EAI Major Strengths
  - Optimized for API-based applications
  - Real-time (or near)
  - Move/send individual events or transactions
  - Some capability for simple and basic transformation and rules
  - Workflow controlled
  - Broker capabilities (subscriptions)

Source : IBM Software Group

EAI

- EAI Main Challenges
  - Limited transformation capability
  - Limited support for data aggregation
  - Limited to tens of records per transaction
  - Complexity for development
  - Longer Time to market
  - Limited reuse for transformations
  - Limited support for metadata (use, import and export)
  - Semantic integrity
  - May consume network bandwidth during peak hours

Source : IBM Software Group

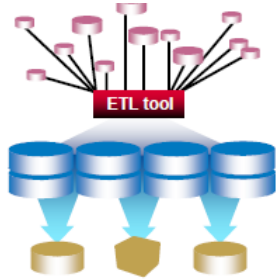
ETL - Etract, Transform, Load

Définition

Extract-Transform-Load est connu sous le terme ETL, ou extracto-chargeur, (ou parfois : datapumping). Il s'agit d'une technologie informatique intergicielle (comprendre middleware) permettant d'effectuer des synchronisations massives d'information d'une source de données (le plus souvent une base de données) vers une autre. Selon le contexte, on est amené à exploiter différentes fonctions, souvent combinées entre elles : « extraction », « transformation », « constitution » ou « conversion », « alimentation ».

Architecture ETL

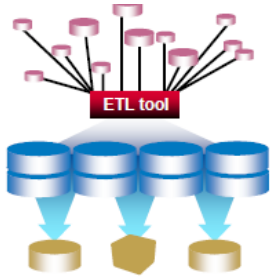
ETL



- ETL Major Strengths
  - Optimized for data structures
  - Periodic, batch-oriented (not intended for real-time)
  - Can move large volumes of data in one step
  - Enables complex data transformations requiring calculations, aggregations or multiple stages
  - Scheduling controlled by the administrator
  - Several GUI based tools available to increase productivity
  - High level of reuse of objects and transformations

Source : IBM Software Group

ETL



- ETL Main Challenges
  - Time to market
  - Change management
  - Data moved regardless of real need
  - Consumes storage systems
  - Data out-of-synch with the original source when it arrives in the DW
  - Large requirements for staging areas
  - Unidirectional
  - Lack of multi-site update support (2 phase commit)

Source : IBM Software Group

ETL/EII/EAI

	ETL	EII	EAI
Data Flow	Unidirectional – from source to target	Bidirectional	Bidirectional
Data Movement	Scheduled – batch Process managed	Query time Query (SQL) managed	Transaction triggered – asynchronous Transaction managed
Latency	Daily - Monthly	Real-time	Near real-time
Transformation, cleansing/enrichment Metadata process reuse	*Best* Generally high reusability of objects and processes	*Medium* Transformations embedded in views and database objects.	*Low* Transformations are done with ESQL. Metadata import limited with DB catalog information

Source : IBM Software Group

ETL/EII/EAI

	ETL	EII	EAI
Transport	FTP, direct database connection	Direct database connection	Messaging
Data Volume Processing	Very large (millions, billions of records)	Medium – access to 100's of thousands or few millions of remote records	Small (few records) – can handle several parallel pipes of few records
Complexity of Transformation	Any complexity	Transformations that can be expressed with SQL	Simple syntax transformations. Limited semantic transformations can be implemented via a broker.

Source : IBM Software Group

ETL/EII/EAI

	ETL	EII	EAI
Support for Event Monitoring	Very limited with high latency	Limited to data events and depended on trigger capability of data sources	Best – logic can be added to support true event propagation and not only data transaction movement.
Versioning	Full support	Limited support – custom build	Limited support – custom build
Workflow Control	Scheduling, dependencies and error or exception handling	None	Extensive – rules based

Source : IBM Software Group

Conception

Le rapatriement des données peut se faire de trois 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.

Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) 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.
Source : Wikipedia

Collecter : Récupérer l'information produite pr l'entreprise

Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) 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.
Source : Wikipedia

Ordonner : Structurer l'information dans le but de la prise de décision (structure différente des BDs opérationnelles)

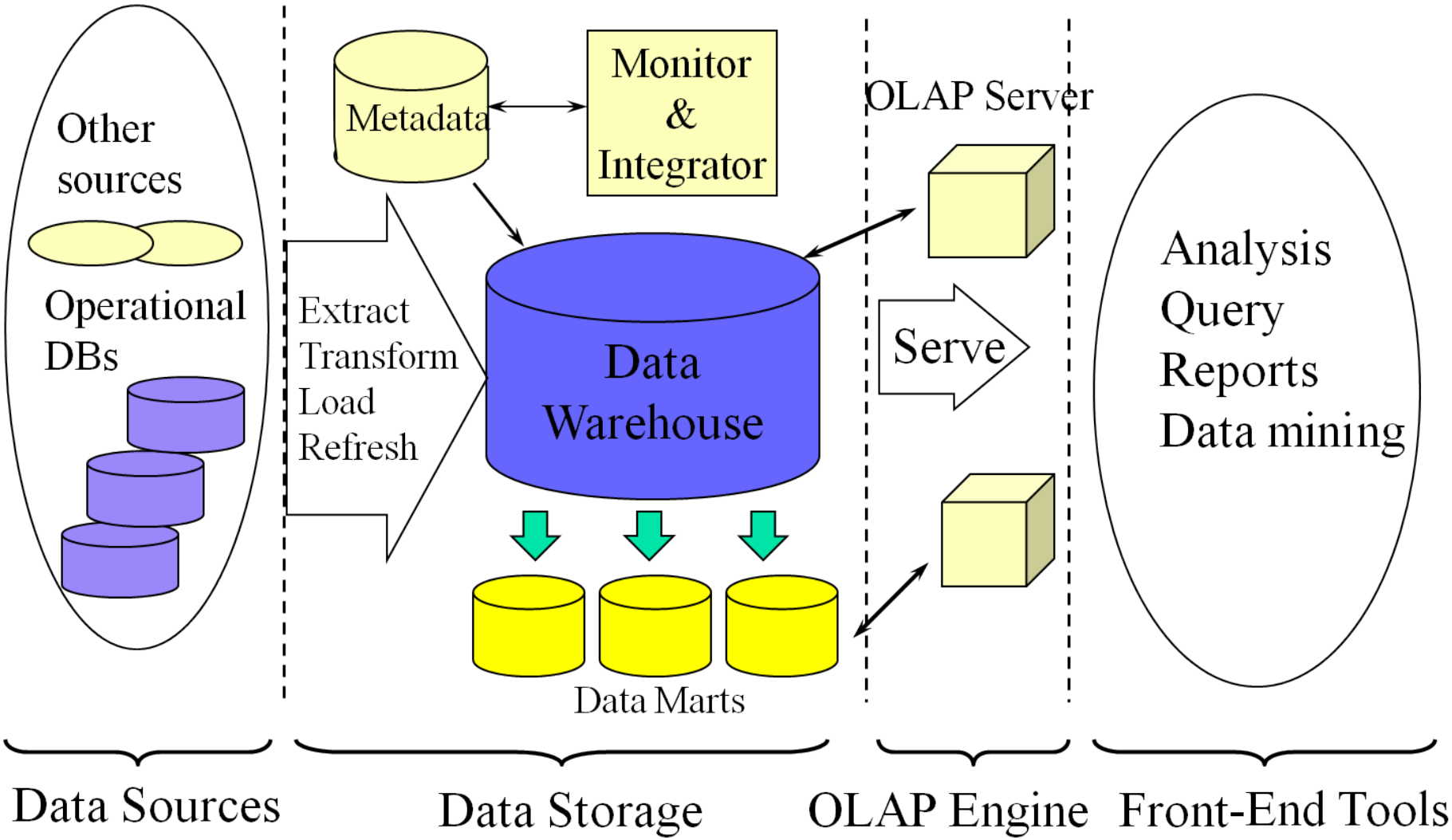
Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) 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.
Source : Wikipedia

Journaliser : Stocker l'historique des données

Datawarehouse



Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Rassembler des données provenant de sources multiples en une seule base de données afin qu'un moteur de requête unique puisse être utilisé pour présenter des données.
- Permettre l'éxécution de requête longues, bloquantes, sur des données opérationnelles
- Maintenir l'historique des données, même si les systèmes de transaction source ne le font pas
- Intégrer des données provenant de multiples systèmes sources, permettant une vue centrale dans l'entreprise. Cet avantage est particulièrement valable lorsque l'organisation est issue de fusions successives
- Améliorer la qualité des données

Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Présenter l'information de l'organisation
- Fournir un seul modèle de données commun pour toutes les données d'intérêt, indépendamment de la source de données
- Restructurer les données de sorte qu'elles prennent sens (décisionnel)
- Ajouter de la valeur aux applications métiers opérationnels, notamment la gestion de la relation client (CRM).
- Faire des requêtes d'aide à la décision plus faciles à écrire.

Datawarehouse vs BD opérationnelle

Table 11-1 Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

Datamart

Définition

Un DataMart (littéralement en anglais magasin de données) est un sous-ensemble d'un DataWarehouse destiné à fournir des données aux utilisateurs, et souvent spécialisé vers un groupe ou un type d'affaire. Techniquement, c'est une base de données relationnelle utilisée en informatique décisionnelle et exploitée en entreprise pour restituer des informations ciblées sur un métier spécifique, constituant pour ce dernier un ensemble d'indicateurs utilisés pour le pilotage de l'activité et l'aide à la décision.

Source : wikipedia

Le datawarehouse est **Général**, le datamart est **spécifique** à un métier.

Datamart vs datawarehouse

Table 11-2 Data Warehouse Versus Data Mart

Data Warehouse	*Data Mart*
Scope	*Scope*
• Application independent	• Specific DSS application
• Centralized, possibly enterprise-wide	• Decentralized by user area
• Planned	• Organic, possibly not planned
Data	*Data*
• Historical, detailed, and summarized	• Some history, detailed, and summarized
• Lightly denormalized	• Highly denormalized
Subjects	*Subjects*
• Multiple subjects	• One central subject of concern to users
Sources	*Sources*
• Many internal and external sources	• Few internal and external sources
Other Characteristics	*Other Characteristics*
• Flexible	• Restrictive
• Data-oriented	• Project-oriented
• Long life	• Short life
• Large	• Start small, becomes large
• Single complex structure	• Multi, semi-complex structures, together complex

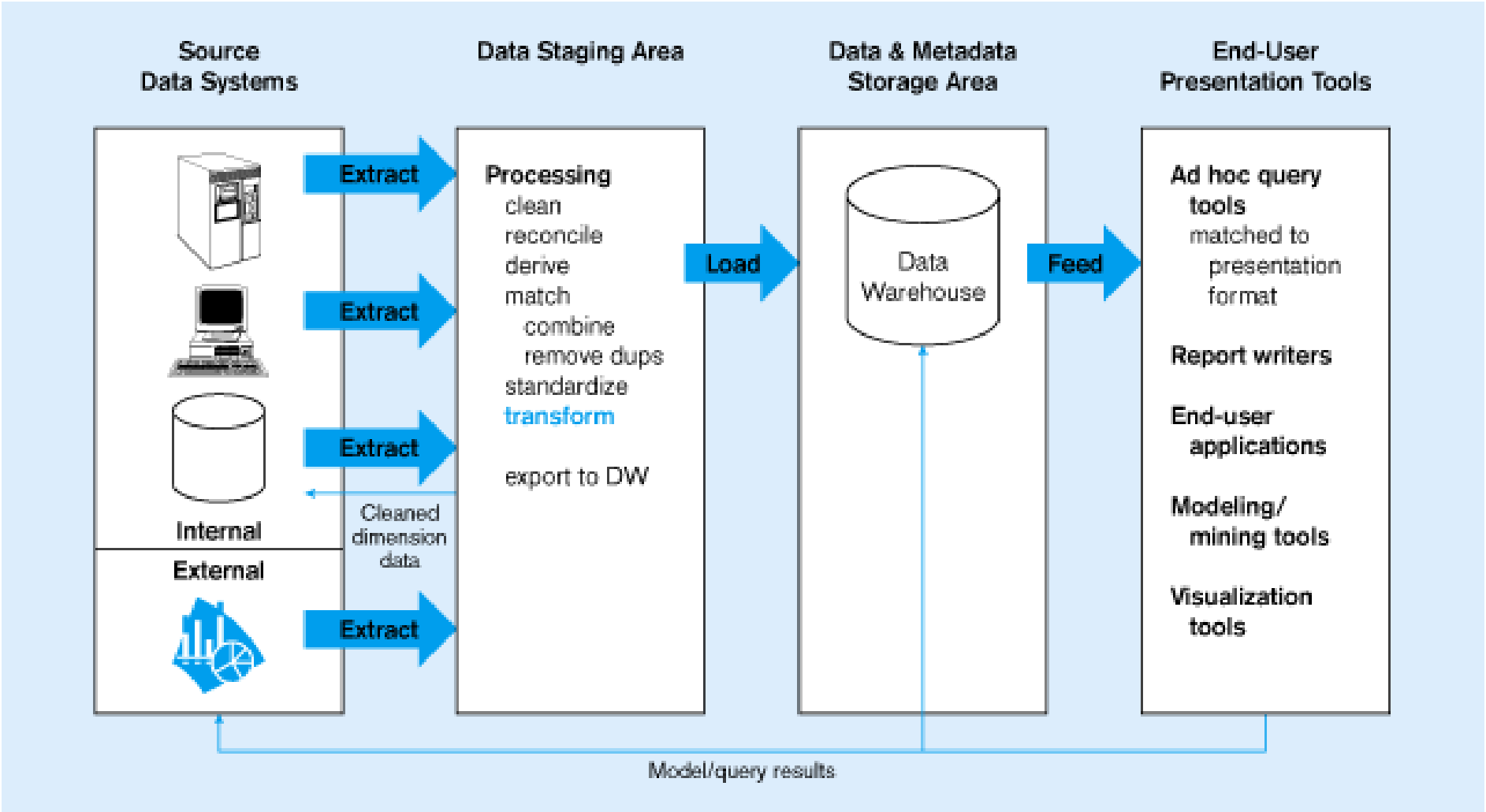
Adapted from Strange (1997)

Datamart vs datawarehouse

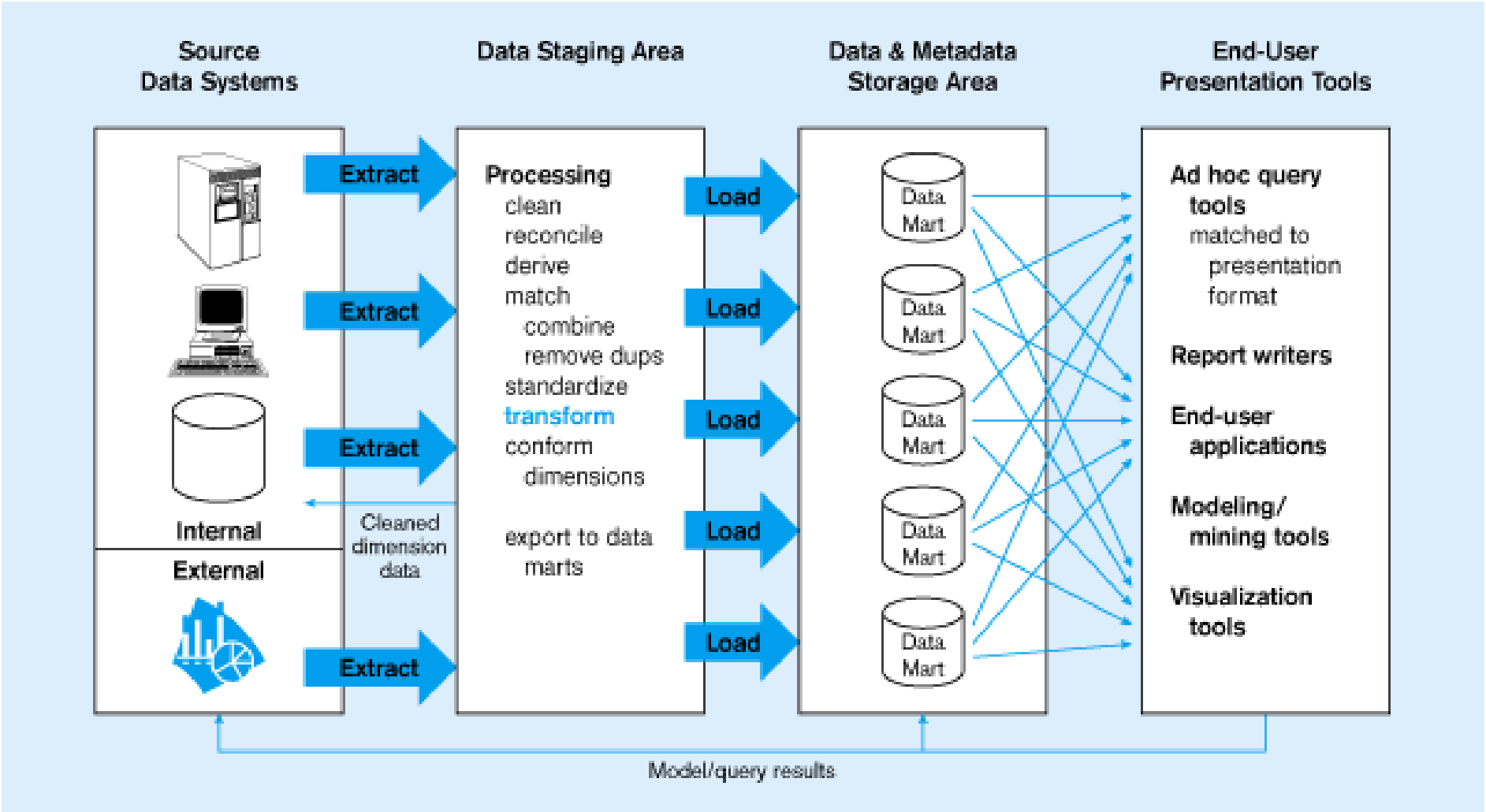
Deux conceptions existantes :

- Définition d'Inmon : Le DataMart est issu d'un flux de données provenant du DataWarehouse. Contrairement à ce dernier qui présente le détail des données pour toute l'entreprise, il a vocation à présenter la donnée de manière spécialisée, agrégée et regroupée fonctionnellement.
- Définition de Kimball : Le DataMart est un sous-ensemble du DataWarehouse, constitué de tables au niveau détail et à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier. L'ensemble des DataMarts de l'entreprise constitue le DataWarehouse.

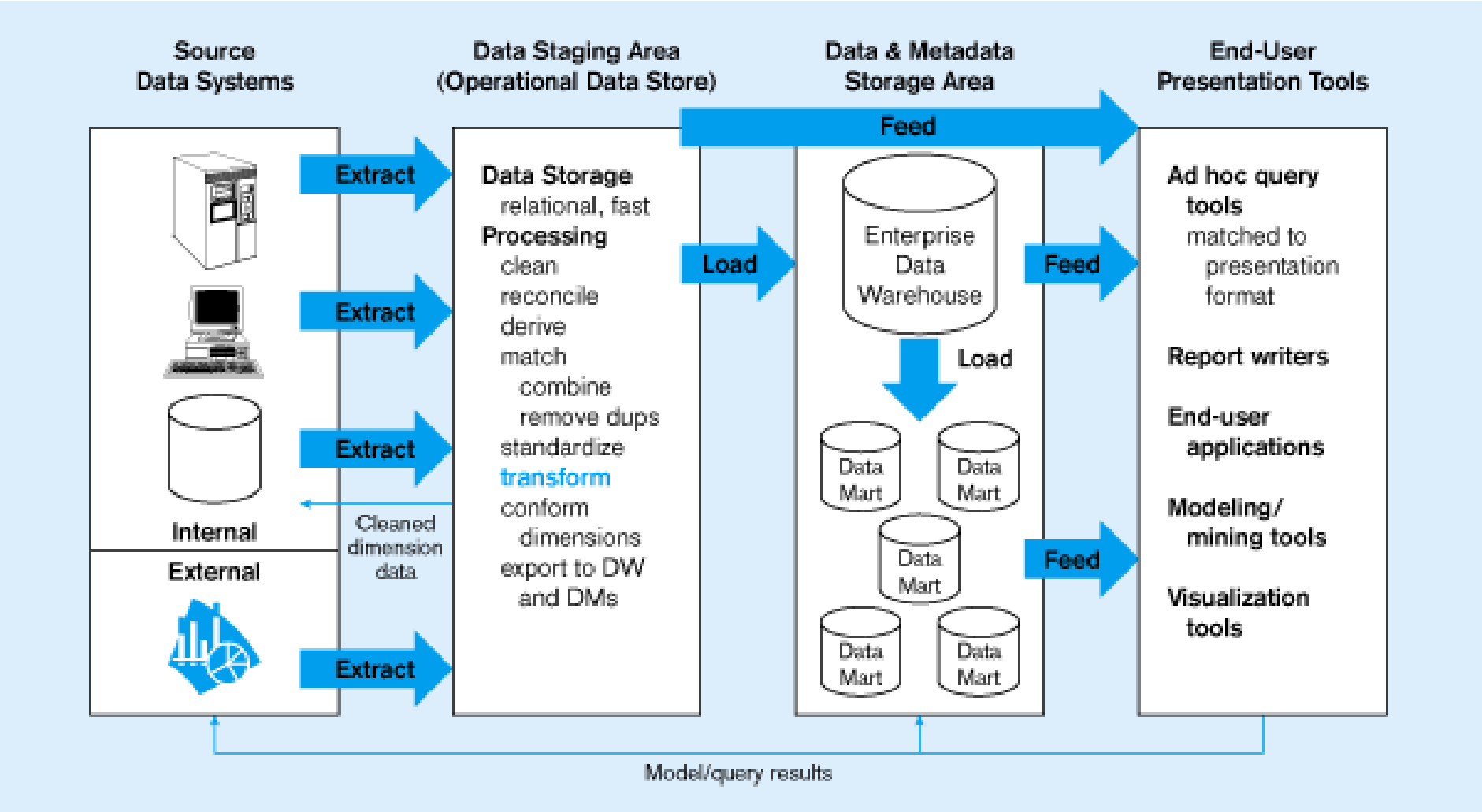
Différentes Architectures



Différentes Architectures

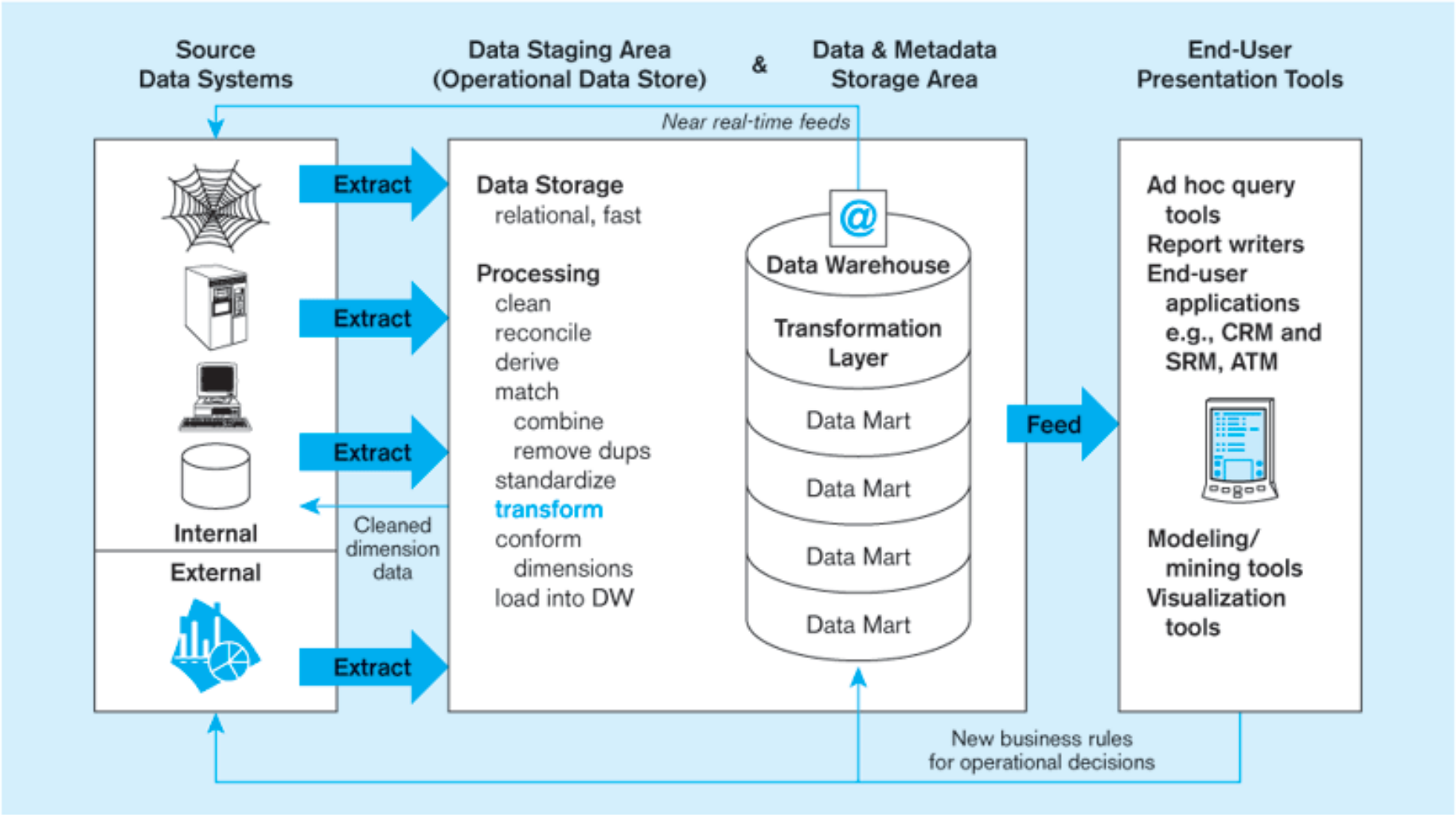


Différentes Architectures



Différentes Architectures

Figure 11-5 Logical data mart and @ctive warehouse architecture



Données orientées sujets

- En production : données organisées par processus fonctionnels
- Datawarehouse : données organisées autour de sujets majeurs
- Données structurées par thème, potentiellement transverses par rapport aux domains fonctionnels et organisationnelles

Examples (médecine) : Actes, Séjours vs Bases par services

Architecture

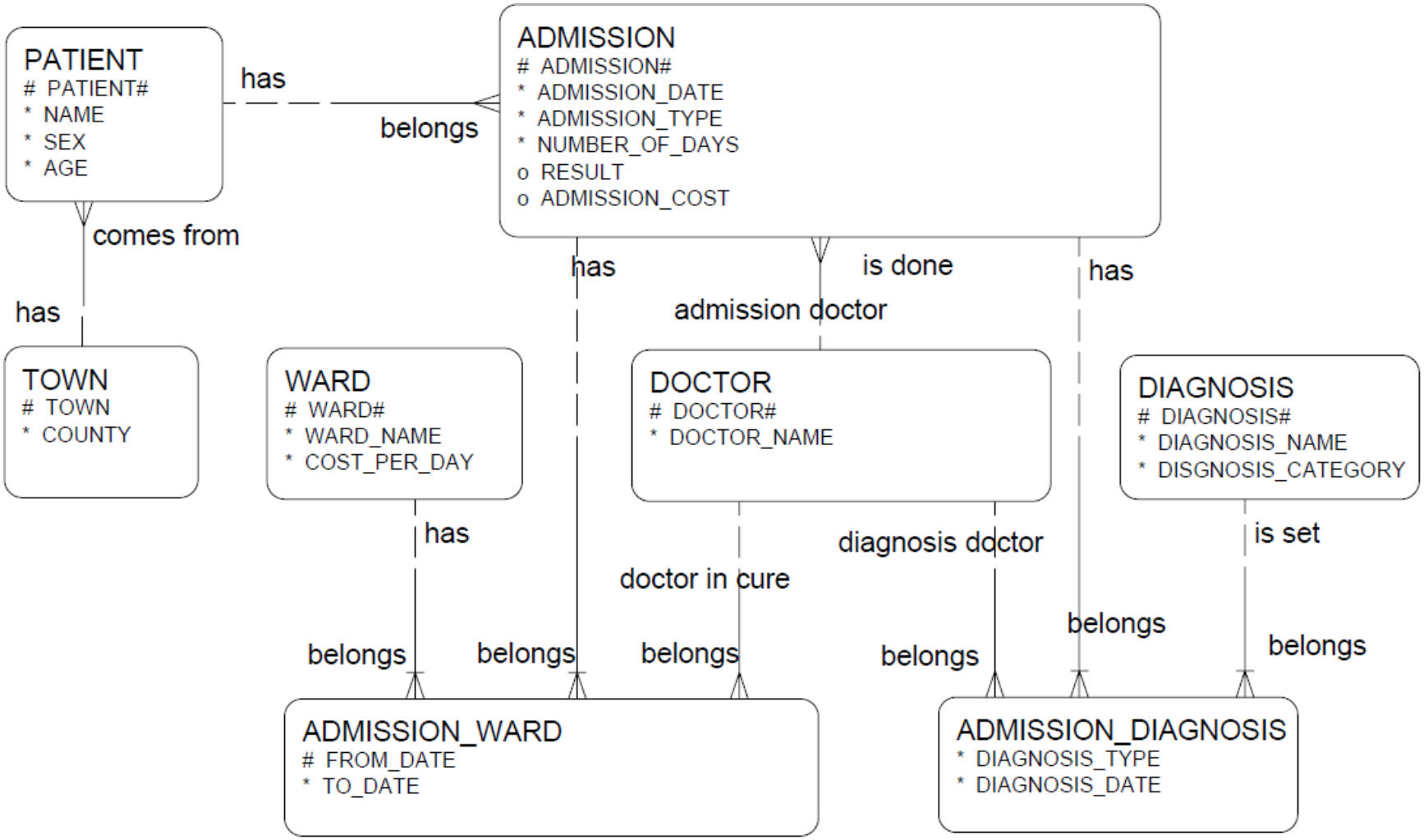
Good DW architecture

"It's not easy to describe a good design, but I'll know it when I see it"

Modèle relationnel

- Normalisation (3NF)
- Répond aux besoins transactionnels (OLTP)
- 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

Modèle relationnel



Le modèle relationnel n'est pas (très) approprié pour les DWs

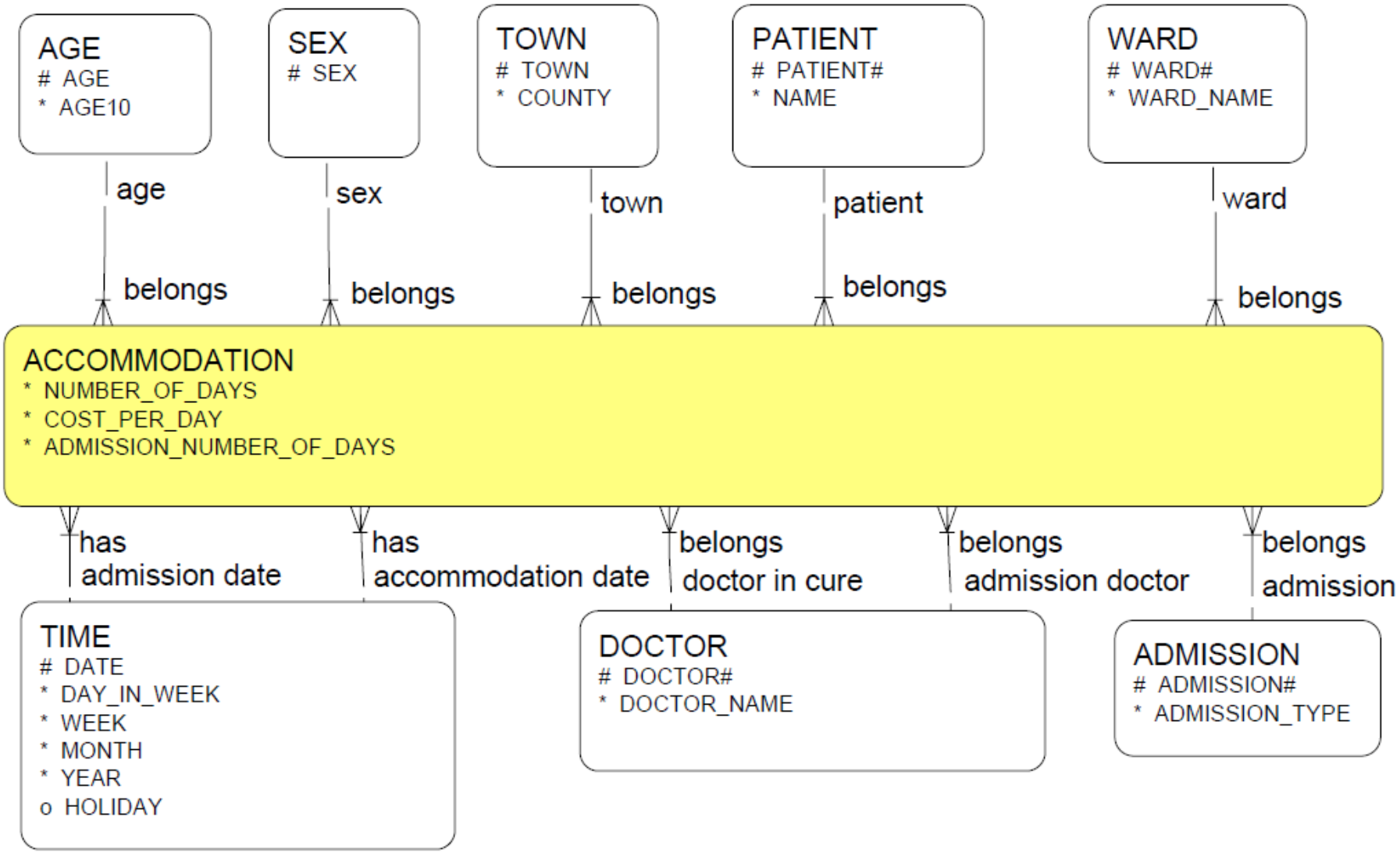
Modèle dimensionnel

Principes

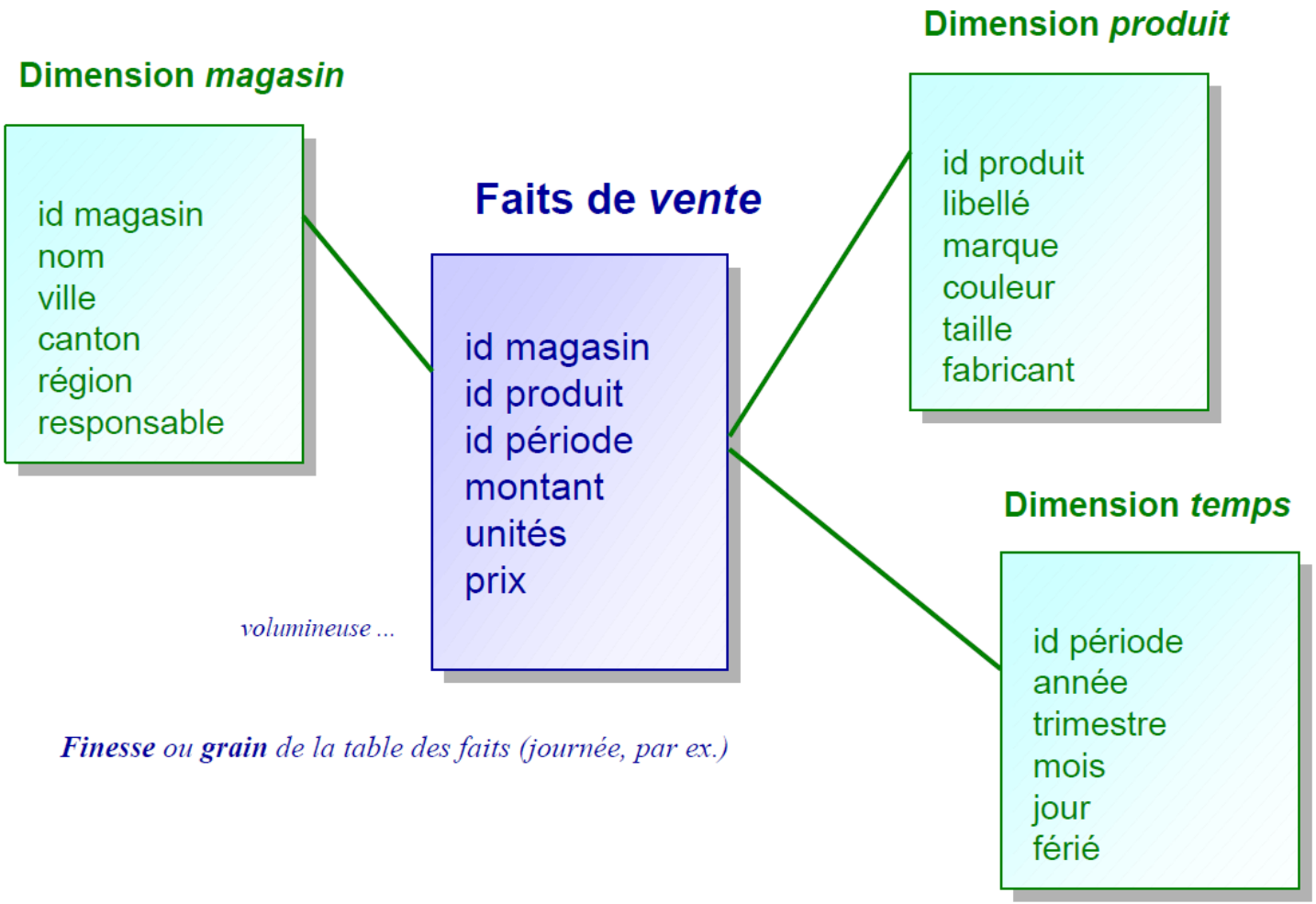
On va partir du besoin "client" (quel analyse ?). On va définir des **faits** et des **dimensions**.

- **Faits :** les faits représentent un sujet d'analyse. Les faits sont caractérisées par plusieurs informations
- **Dimensions :** les dimensions sont les critères selon lesquels on souhaite faire de l'analyse.

Modèle dimensionnel



Modèle dimensionnel



Aussi connu sous le nom de modèle en étoile