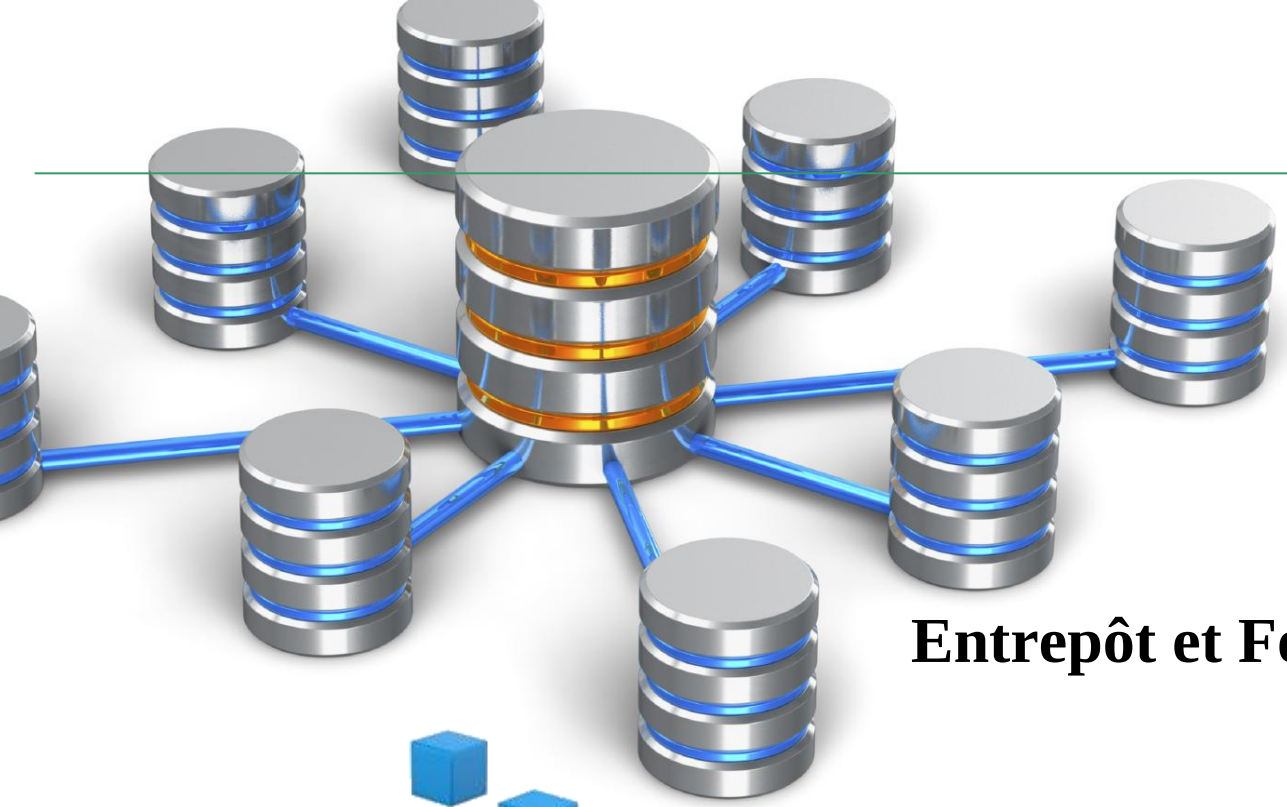
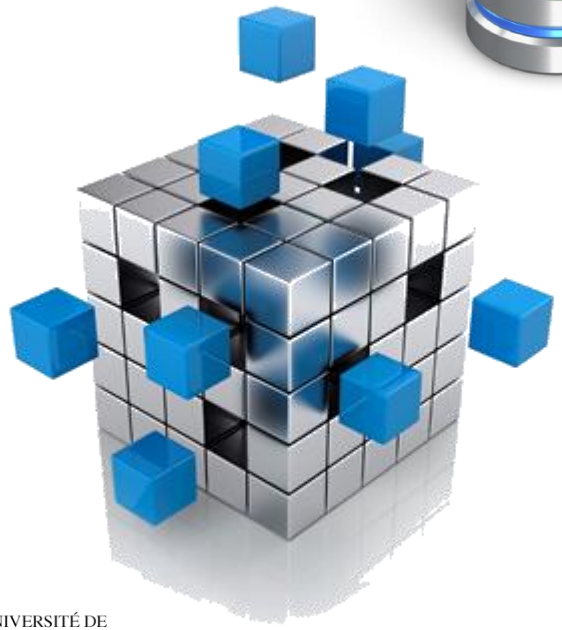


INF 735
Entrepôt et Forage de Données



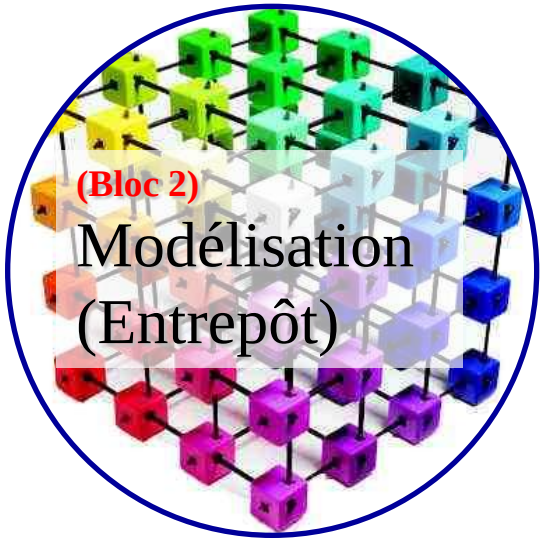
Bloc 5
Architecture et Méta données
par Robert J. Laurin

UNIVERSITÉ DE SHERBROOKE
Faculté des sciences
Centre de formation en technologies de l'information

Version: A18

Plan du cours – Les blocs

(Bloc1)
Introduction: Le besoin, concepts et définitions



(Bloc 5) Architecture et Méta données

(Bloc 6) Définition des besoins et gestion de projet

(Bloc 7) Techniques de réalisation et opération

Lectures relatives

- Suggéré:

  Data Warehousing Fundamentals, A Comprehensive Guide for IT Professionals,

  Paulraj Ponniah

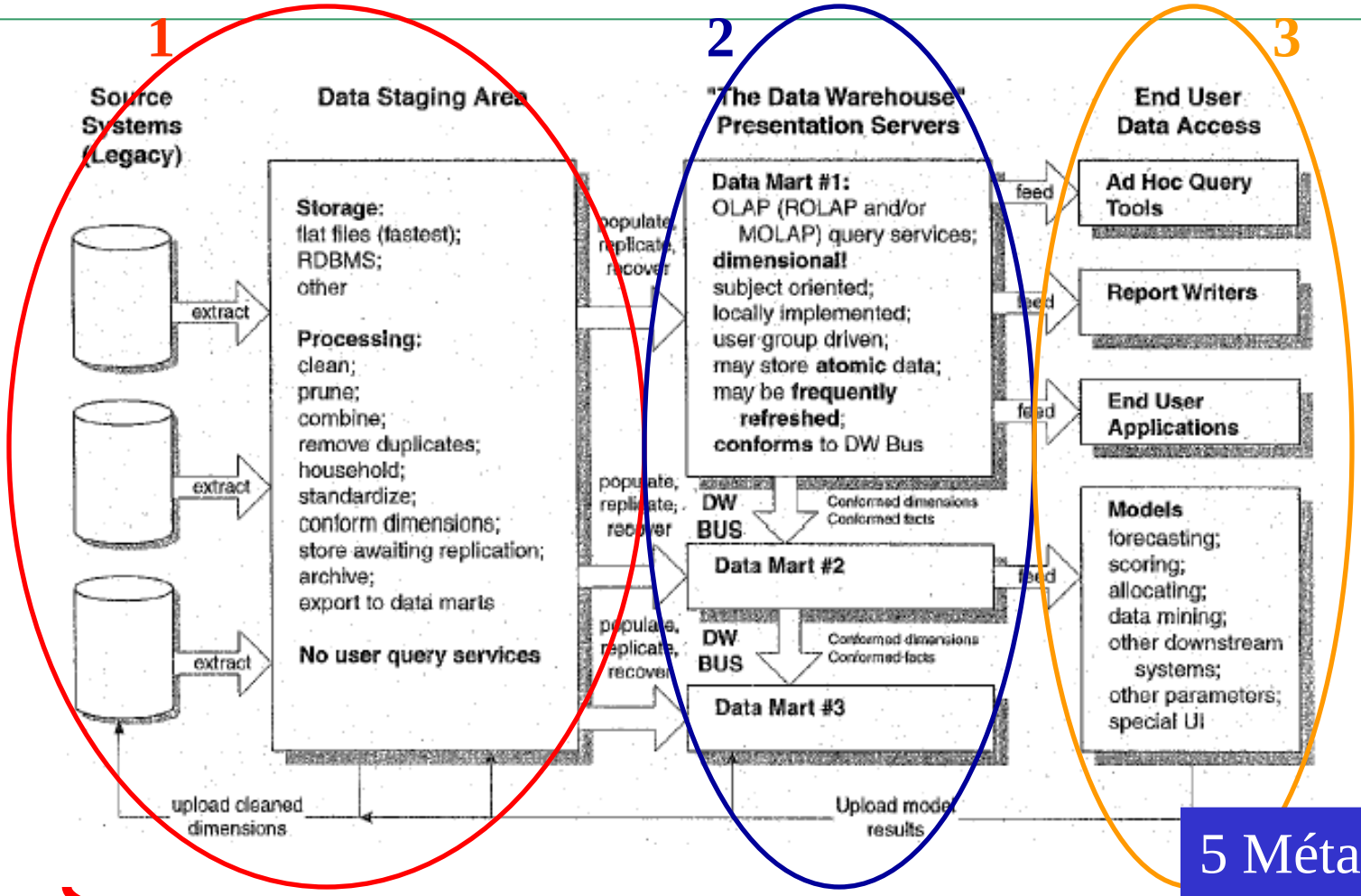
  – Chapitres 7 et 9

- Annexes:
  – Metadata Checklist
  – Source to target data map
  – Data source checklist
  – Data source definitions
  – Requirements Findings (template)

Concepts Clés du Bloc

- Architecture de l'entrepôt (en 4 sections + Méta information)
- La méta-donnée est au cœur de l'architecture.

Éléments d'architecture de l'E. D.



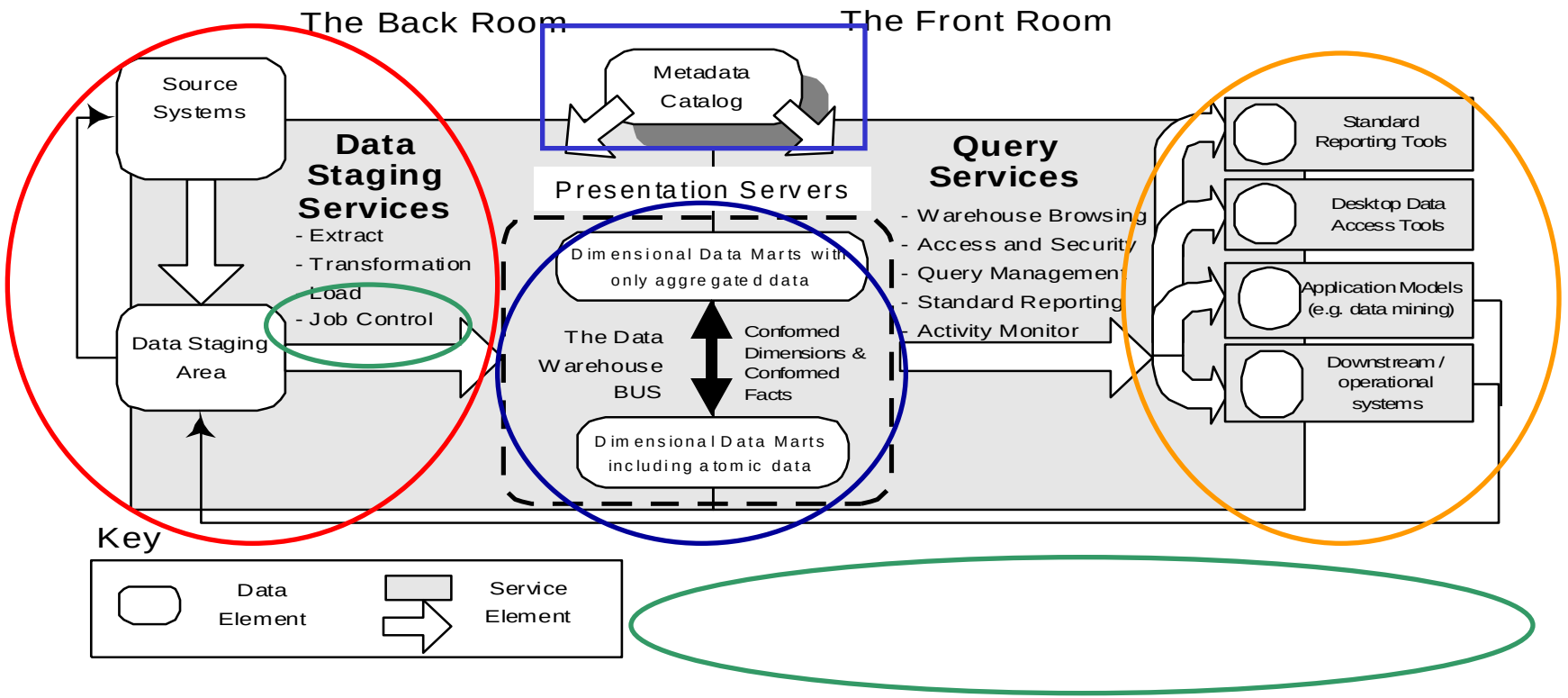
5 Méta Information

4 Gestion, contrôle et « monitoring »

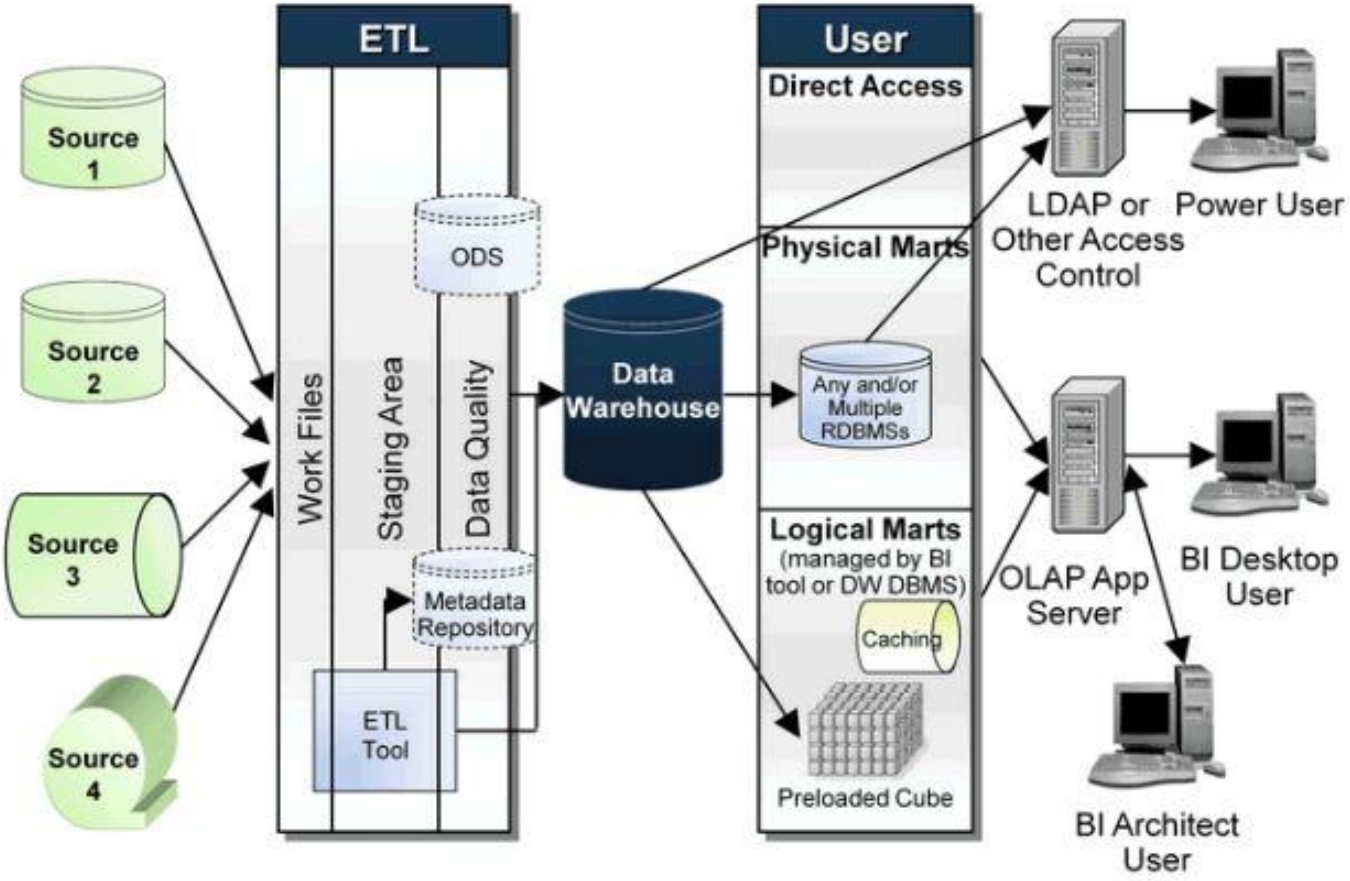
ARCHITECTURE

Architecture - Schéma

High Level Warehouse Technical Architecture



Architecture – vue différente



BI = business intelligence; DBMS = database management system; DW = data warehouse; ETL = extraction, transformation and loading; LDAP = Lightweight Directory Access Protocol; ODS = operational data store; OLAP = online analytical processing; RDBMS = relational database management system

Source: Gartner (August 2011)

Architecture - Efforts

- Trouver et décrire les sources de données
- Documenter la définition commune (acceptée et comprise par tous)
- Recouper l'information des sources et le besoin d'agrégation.
- Définir la structure dans la quelle la donnée doit se retrouver
- Prévoir la transformation de la donnée (macro)
- Définir le comportement de mise à jour et fréquence requise (+ Gestion)
- Déterminer quelle information sera diffusée à qui, sous quelle forme
- Voir détail au chap. 8

Architecture - raisons

- Communication
- Planification
- Flexibilité et maintenance
- Apprentissage
- Mesure de productivité et d'avancement

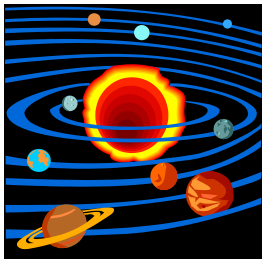
Architecture - Niveaux

Niveau de détail	Données (quoi)	Technique (comment)		Insfrastructure (où)
		Arrière boutique	Étalage	
Besoins d'affaire et vérification	Quelle information est nécéssaire pour prendre de meilleur décisions? Les données actuelles peuvent-elles être utilisables?	Comment trouver la source, la transformer et la rendre disponible? Comment faiosns-nous présentement?	Quelles sont les enjeux d'affaire? Comment les mesurer? Comment vouslons-nous analyser les données?	De quel hardware et software avons-nous besoin pour réussir? Qu'avons-nous déjà?
Modeles et documents	MODÈLE DIMENSIONEL: Quelles sons les entitées (faits et dimensions) qui composent cette information? Comment les liées et les structurées?	Quels seront les supermachés principaux? Où seront-ils? Quoi transformer pour quand?	Sous quelle forme de présentation pouvons nous rendre l'information utilisable? Quelles analyses et rapports produire - avec quelles priorités?	D,où viennent et où vont les données? Avons-nous les capacités de traitement et chargement? Qui en est responsable?
Modèles détaillés et spécifications	MODÈLE LOGIQUE ET PHYSIQUE: Quels sont les éléments de données cibles? Quelles sont les sources et les transformations pour atteindre la cible?	Quels standards et produits permettent le ETL et la sauvegarde des données? Quels sera le développement requis et les standards de développement?	Quels sont les requis des rapports - Titres, colonnes, rangées, filtres, etc... Qui les veux? Quand? Mode de distribution?	Comment interagissent ces logiciels et développements? Quels sont les API, appels, modules, etc?
Implantation	Créer la BD, Index, Backups Documentation.	Écrire les routines de ETL En automatiser le processus Documentation.	Implanter les outils de rapport et d'analyse de l'information - défénier les premiers rapports, former les utilisateurs. Documentation.	Installer et tester l'infrastructure . Connecter les sources et livrer l'information au poste de travail. Documentation.

L'ARCHITECTURE SELON L'APPROCHE

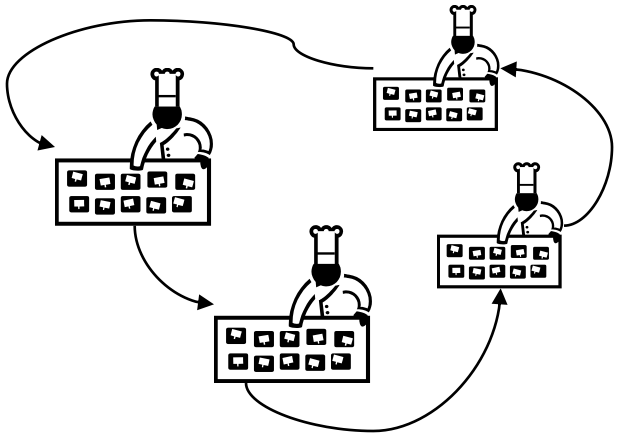
Approches d'implantation

- Big Bang (Top down)



Avantages	Désavantages
•Vue Corporative, effort entreprise •Harmonie globale dans le design •Entrepôt bien structuré / infrastructure unique •Contrôle et gestion central •Grand impact rapidement	•Long à réaliser/implanter •Risque élevé •Besoin de personnel avec vues globales multidisciplinaires •Se lancer complètement sans preuve de concept.

- Approche itérative par Supermarché(Bottom up)



Avantages	Désavantages
•Implantation rapide et plus facile de plus petits blocs •Retour sur investissement rapide •Preuve de concept •Moins risqué •Permet à l'équipe d'apprendre et évoluer	•Vision en silo •Risque de répétition de données par sujet avec des définitions différentes (pas communes) •Perpétue la tradition des données corporativement irréconciliables •Diversité d'interfaces et ETC.

Méthodes de réalisation

- **« Top-Down » :** Méthode lourde, contraignante et complète.
  - Conception de tout l'entrepôt (ie : toutes les étoiles) puis réalisation
  - Vision claire de l'entreprise et du projet
  - Implique: savoir à l'avance toutes les dimension et tous les faits de l'entreprise

- **« Bottom-Up » :** Simple, flexible mais souvent difficile d'intégration
  - Créer les étoiles une par une, par sujet
  - Les joindre jusqu'à obtention d'un véritable entrepôt avec une vision d'entreprise.
  - Travail d'intégration pour obtenir un entrepôt de données
  - Attention aux dimensions semblables avec définitions différentes

- **« Middle-Out" :** Approche hybride. → Recommandée
  - Conception totale de l'entrepôt (du moins conceptuellement- Dimension s et faits)
  - Découper le conceptuel par éléments en commun et réaliser sujet par sujet.
  - Implique: Compromis de découpage (dupliquer des dimensions identiques pour des besoins pratiques).

4 ÉLÉMENTS DE L'ARCHITECTURE

=> MÉTA INFORMATION (5IÈME ÉLÉMENT)

1 - ETC

Cette section comprend

- Sources
- Extraction
- Transformation
- Chargement

Outils ETC

3 approches

- Développement maison
  - Langage de programmation standards (Cobol, SQL, C, etc.)
  - Nouveaux standards interfaçage (XML)

- Les outils propriétaires
  - Outils complets par eux-mêmes
  - Utilise le produit sans programmation

- Les outils générateur de codes
  - Génère du code de programmation standard (Cobol, SQL, C, etc.)
  - Possibilité de programmation

➔ **Voir Bloc 4**

Outils ETC

- Extraire les données des systèmes opérationnels
- Les transférer sur la machine informationnelle (« Staging ») par FTP, passerelle ou autres
- Les nettoyer
- Les décoder
- Les dériver
- Les agréger
- Les sommariser
- Les historiser
- …

2 - Entrepôt

Architecture – Entrepôt (SGBD)

- Modélisation
  - Modèle étoile
  - Flocon (hiérarchisation)
  - Multidimensionnelle (cubes)

- Approche
  - Big bang ou comptoirs (« data marts »)
  - Centralisé ou décentralisé

➔ **Voir Bloc 2**

Modèles

- Dimensions
  - Noms
  - Définition/Raison d'être
  - Sources
  - Type de changement
  - Décisions
  - Responsable
  - Saisi par ?
- Spécial dimensions
  - Hiérarchies d'attributs
  - Hiérarchies de dimensions
  - N à N
- Faits
  - Comme dimensions +
  - Calculs
  - Additif, semi-additif, Non-additif
- Couches d'ajustements ?
  - Où
  - Champs touchés
  - Droits
  - Tracabilité

3 - Présentation

Architecture utilisateurs

- Objectifs:
  - Autonomie
  - Convivialité
  - Performance
  - Adapté

Outils d'accès

- Outils de base :
  - Méta-information
  - Navigateur

- Types d'outils :
  - Tableau de bord (indicateurs)
  - Multidimensionnel (OLAP, Tableaux croisés dynamiques)
  - Requêtes ad hoc
  - Rapports
  - Graphiques
  - Exploration ("Data Mining")
  - Navigation

➔ **Voir Bloc 3**

4 – Gestion, Contrôle et Monitoring

Architecture – Autres considérations (Gestion)

- Gestion des horaires de tâches
- Alarmes et reprises
- Réseaux fédératifs (« Backbones »)
- Télécommunication (sécuriser les transferts et interrogations)
- Sécurité
  - Backups à différents niveaux (une relève n'est pas un backup)
  - Antivirus
  - Zone de traitement sécurisé
  - Extraction et confidentialité
- Sources externes d'information
- Routines d'entretien
- Optimisation et indexation

Architecture – Autres considérations (Gestion)

- Retour sur l'opérationnel (modifications sur ETC ou post-analyse)
- Intégration avec l'opération
- Support et ententes de services
- Plan de relève / Continuité des affaires

MÉTA INFORMATION (5IÈME ÉLÉMENT)

5 - Méta données

Définition

« Méta Données: Ce sont toutes les données « physiques » et la connaissance **à propos** de l'entreprise et de ses procédés, des données, technologies, processus, règles et structure des données. »

David Marco

David Marco

Autorité en Méta données

David Marco est un expert internationalement reconnu dans le domaine des T.I. principalement dans l'architecture, les entrepôts de données, le «B.I. » et la sommité internationale en méta information.

- DMReview

Définition (prise 2)

- Données physiques
- Connaissances

<u>À propos de</u>

- Entreprise
- Procédés
- Données
- Technologies
- Processus
- Règles
- Structures des données

Méta données – Définition

Méta données importantes

- Définition de l'entrepôt
  - Structure
  - Schémas
  - Localisation
  - Etc.
- Affaires/Business
  - Propriété des données
  - Définitions des termes
  - Règles de calculs officiels
  - Politiques
  - Pratiques opérationnelles
- Techniques
  - Aspects techniques vus plus loin
- Cheminement des données

  ➔Cartographie/MAPPING

  (Fil conducteur)

Méta-Information - Rôles

1. Documente !

2. Informe!

3. Permet l'évolution en inventoriant les connaissances acquises sur les données comme l'entreprise

4. Augmenter la confiance dans les données

5. Précaution contre le taux de roulement important dans les projets d'entrepôt

Méta-Information – autres

- Public cible:

- Importance:

Méta-Information – autres

Exemple

NASA: projet sonde sur mars en 1999 – le Mars Climate Orbiter a été conçu avec une erreur d'interface pour la propulsion. Les ingénieurs ont calculés les poussées de propulsion des fusées pour les ajustements en vol en pieds par seconde. Le système informatique implanté calcule en Newton par seconde. Une différence de 4.4 pieds par seconde (plus d'un mètre). La propulsion d'ajustement avait lieu 12 à 14 fois par semaine durant le voyage de 9 mois.

Résultat: la sonde s'est écrasée sur Mars puisque la poussée pour freiner la descente a débuté beaucoup trop tard – 300 millions US écrasés.

Méta données appliquées à l'E.D.

En entrepôt - Connaissances sur:

➔ la cible du sujet

➔ les sources de données

➔ l'ETC

➔ La gestion, contrôle et monitoring

➔ La business → les règles, le comment, le pourquoi…

LA CONNAISSANCE COMMENCE ICI !

La gestion des métadonnées

“Documentation technique et d’affaires sur l’information corporative”

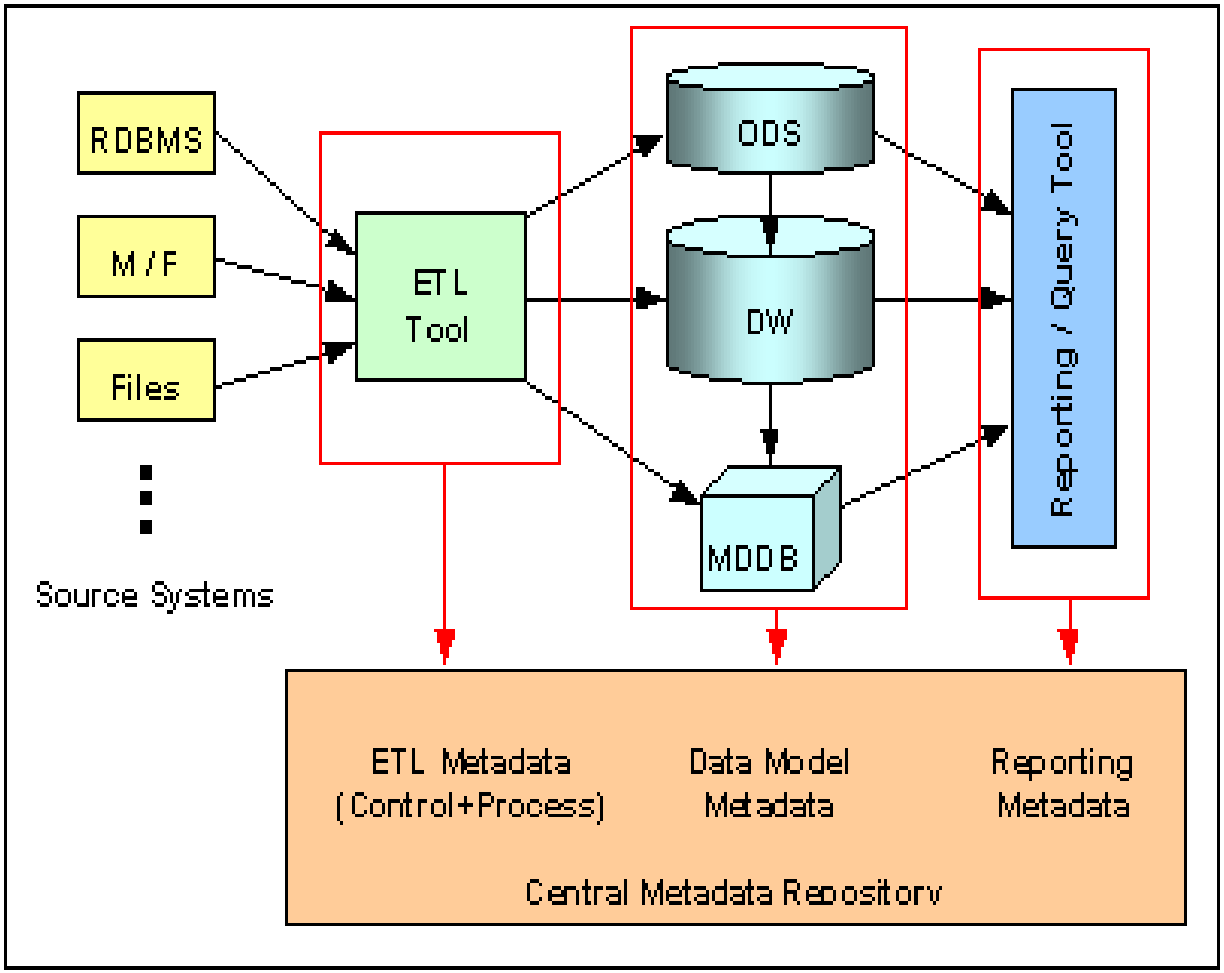
- Une des pièces d’une architecture informationnelle
- Il n’est pas requis de tout documenter
- Les objectifs: Uniformité, Optimisation, Réconciliation et Communication
- Doit être liée aux processus
- Le rôle du Gestionnaire des métadonnées
  - Les métadonnées sont complètes, à jour et de qualité

Source: Présentation de AgileDSS

Les étapes de mise en place

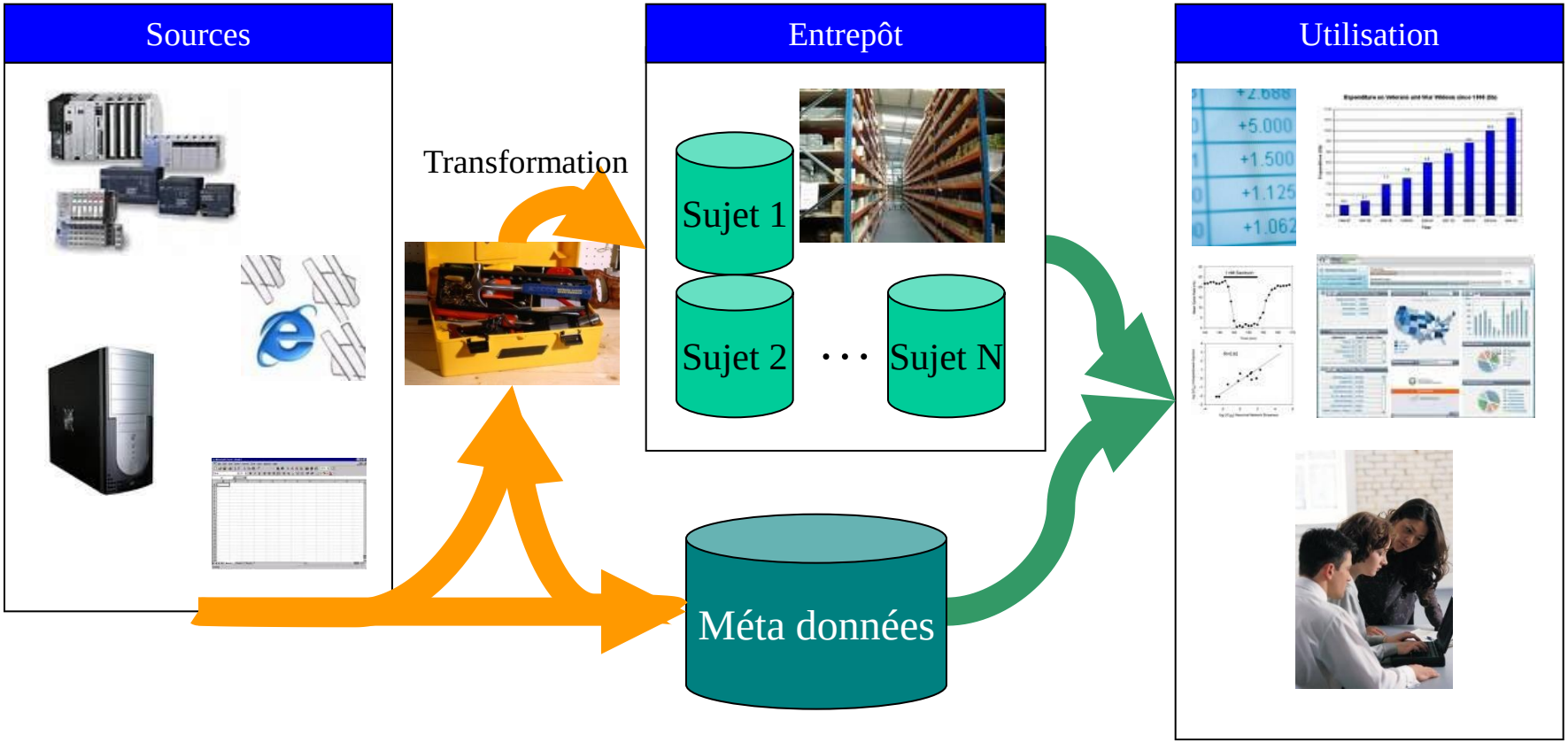
- Identifier les processus (ou sujets) prioritaires et le type de métadonnées requises
- Définir les standards de documentation
- Choisir l'architecture de solution (documentation et communication)
- Identifier un gestionnaire de métadonnées
- Identifier les propriétaires de données et les intendants de données
- Déployer par le biais d'un pilote

La gestion des métadonnées



Obtenue de présentation de AgileDSS, source inconnue

Méta données intégrées à l'E.D.



Méta-Information

- Types:
  - Organisationnelle
    - Décisions de standards
    - Décisions de sujets
    - Processus
  - Sources
    - Plate-forme
    - Modèle logique
    - Modèle physique
    - Définitions
  - Ravitaillement (ETC)
    - Méthode d'extraction (push, pull, push/pull, autre)
    - Règles de transformation
    - Règles de nettoyage
    - Sommaires
  - Traitement des rejets
    - Contrôle des logs
    - Corrections et reprises
  - SGBD (Cible)
    - Modèle de données (relationnel / Dimensionnel)
    - Lien de la source à la cible
  - Outils de livraison
    - Modèles par sujet
    - Méthode de navigation
    - Rapport pré-définis
    - Comment démarrer les outils et se connecter au sujet
  - Surveillance (« Job control »)
    - Alertes
    - Reprises

Méta-Information

- Types:
  - Organisationnelle
    - + Décisions de standards
    - + Décisions de sujets
    - + Processus
  - Sources
    - + Plate-forme
    - + Modèle logique
    - + Modèle physique
    - + Définitions
  - Ravitaillement (ETC)
    - + Méthode d'extraction (push, pull, push/pull, autre)
    - + Règles de transformation
    - + Règles de nettoyage
    - + Sommaires
  - Traitement des rejets
    - + Contrôle des logs
    - + Corrections et reprises
  - SGBD (Cible)
    - + Modèle de données (relationnel / Dimensionnel)
    - + Lien de la source à la cible
  - Outils de livraison
    - + Modèles par sujet
    - + Méthode de navigation
    - + Rapport pré-définis
    - + Comment démarrer les outils et se connecter au sujet
  - Surveillance (« Job control »)
    - + Alertes
    - + Reprises

La méta information comprendra tout ceci, et plus…

Méta-Information: Cible et source

- Cible:
  - Nom de la colonne / attribut
  - Dimension / fait
  - Date effective
  - Unité (devise, pouces, litres, etc..)
  - Règle d'affaire (ou de validité)
  - Formule si calculé
  - Précision
  - Valeur minimum/maximum
  - Statut (actif ou non)
  - Type (numérique, texte, etc)
  - Null possible?

- Pour la source, ajouter:
  - Système
  - Technologie (plate-forme) et lien d'accès
  - Modèle logique / structure
  - Fenêtre d'opportunité
  - Méthode d'approvisionnement (Push, Pull, combo, spécial)

Méta-Information: ETC

- ETC:
  - Stratégie de ravitaillement
    - Routine (nom, description)
    - Heure
    - Méthode de transfert
    - Conversions/comportement lors du transfert
    - Fenêtre d'opportunité
    - Stratégie de reprise
  - Transformations de la source à la cible
  - Toutes les opérations
  - Règles de validation
  - Traitement de rejets
  - Corrections et tolérances
  - Stratégie de chargement !!!
  - Validations pré-chargement
  - Règles d'affaires

PRÉSENTATION DE LA MÉTA INFORMATION

Méta-Information: Exemples

- Exemple sur la source:

EMPLOYÉS

Le fichier *Extraction_Employes_09-09-22.csv* contient la liste des employés avec leur date de naissance et d'embauche. Les informations sont délimitées par point-virgule. La première ligne du fichier contient le nom des colonnes. La première ligne devra être retirée lors de l'extraction.

Colonne	Information	Description	Type	Format
1	Number	Numéro d'identification de l'employé	Numérique	
2	Name	Prénom et nom de l'employé	Caractères	
3	Hired	Date d'embauche de l'employé	Date	mm-dd-yy
4	Birth	Date de naissance de l'employé	Date	mm-dd-yy
5	Sex	Genre (sexe) de l'employé	Caractère	M ou F

Source: TP INF735

Alain Bordeleau (88 028 272)
Laura Francheri (09 163 086)
Mondher Jarraya (03 440 967)

Méta-Information: Exemples

- Exemple sur la Transformation:
  - Date d'embauche de l'employé (f_Hired → Hired)
    - Vérifier que le champ contient une valeur
      - Si non, mettre l'enregistrement en rejet avec le message `'Colonne Hired vide dans le fichier source'` et passer à l'enregistrement suivant
    - Convertir le champ en type date (format mm-jj-aa)
      - Si la conversion a échouée, mettre l'enregistrement en rejet avec le message `'Colonne Hired n'a pas le format mm-jj-aa'` et passer à l'enregistrement suivant
    - Vérifier si la date est entre la date de démarrage de la société et la date du jour
      - Si non, mettre l'enregistrement en rejet avec le message `'Colonne Hired probablement invalide'` et passer à l'enregistrement suivant

Source: TP INF735

Alain Bordeleau (88 028 272)
Laura Francheri (09 163 086)
Mondher Jarraya (03 440 967)

Méta-Information: Exemples

- Exemple sur la Cible:

FAITS DE PRODUCTION

La table **FAIT_Production** est utilisée dans le dernier processus soit la génération des faits pour le fichier cible. Cette table contiendra le rapprochement et la sommarisation des informations de production obtenues, fusionnées et complétées depuis les différents fichiers sources.

Table : FAIT_Production

Nom du champ	Type	Description
Date	datetime	Numéro automatique d'identification de l'enregistrement
Quart	int	Quart de travail (donnée sommarisée)

Source: TP INF735

Alain Bordeleau (88 028 272)
Laura Francheri (09 163 086)
Mondher Jarraya (03 440 967)

Méta-Information: Exemples

- Exemple Source – transformation - cible:

Fichier source : Feuille_temps.csv **Table de destination** : Extraction_Feuille_temps_06-04-23 **Outil de transformation**: MS Access

1.Source				2. Extraction	3. Destination					
Source field	**Data Type**	**Size**	**Format**	**Details of the transformation**	**Destination field**	**Data Type**	**Size**	**Format**	**Details**	**Field description**
N/A	N/A	N/A	N/A	Création d'un numéro de feuille de temps	ID_Feuille_temps	LONG INTEGER	4	X to XXXX	NOT NULL Fixed size Auto-Increment	Ce champ contient uniquement des chiffres. Il est la clef primaire de la table. Il est automatiquement généré.
Date	DATE	8	AA/MM/JJ	Conversion du format de la date pour la standardiser	Date_Feuille_Temps	SHORT DATE	8	MM/JJ/AA	NOT NULL	Ce champ contient la date de la feuille de temps de l'employé.

Source: TP INF735

Maha Abdelhak (05 652 484)

Jihad Taher (05 722 382)

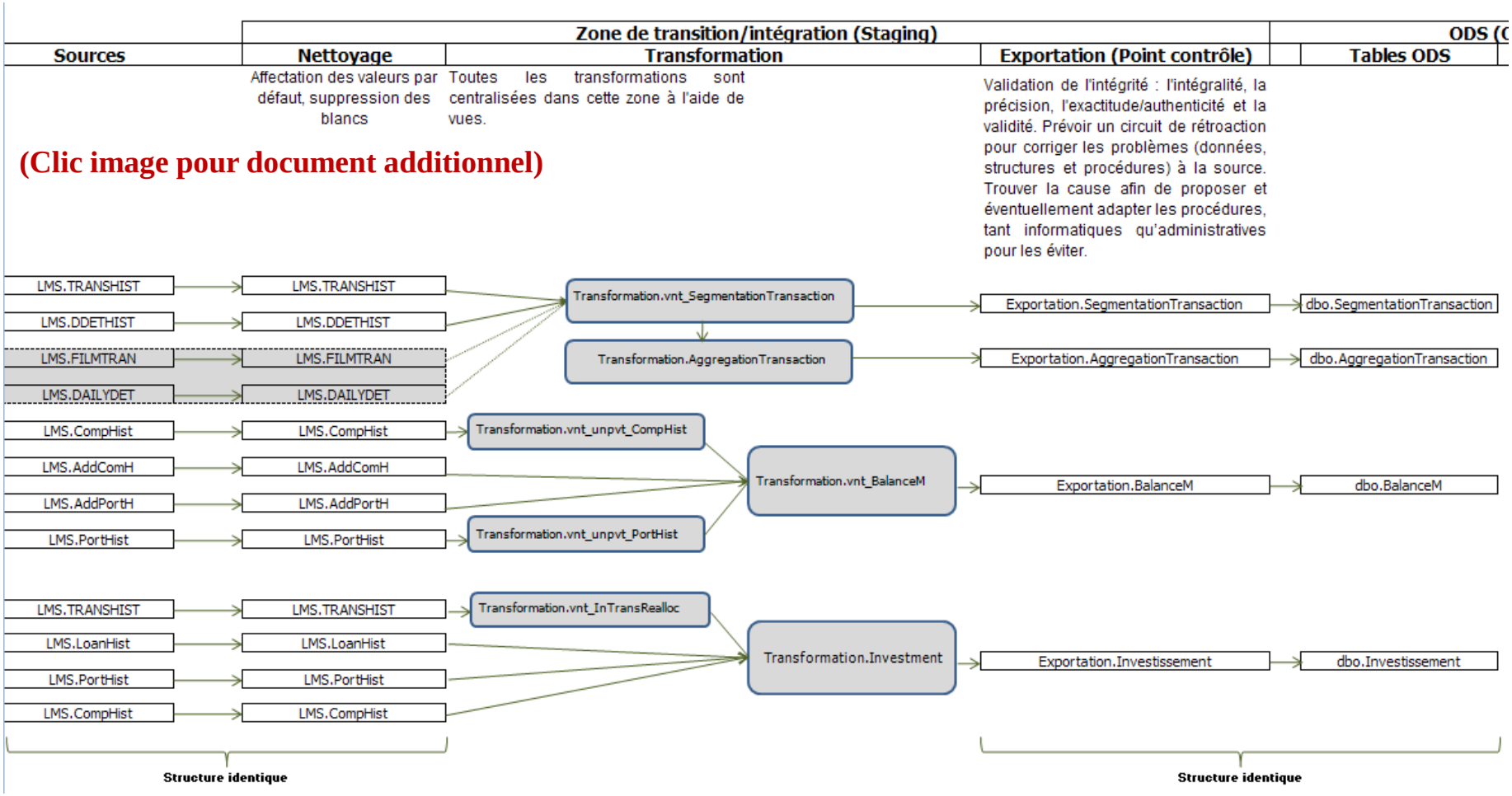
La grille inverse, soit

Cible (ou champs dimension/fait) ← Transformation ← Source

est toutefois plus classique…

Méta-Information: Exemples

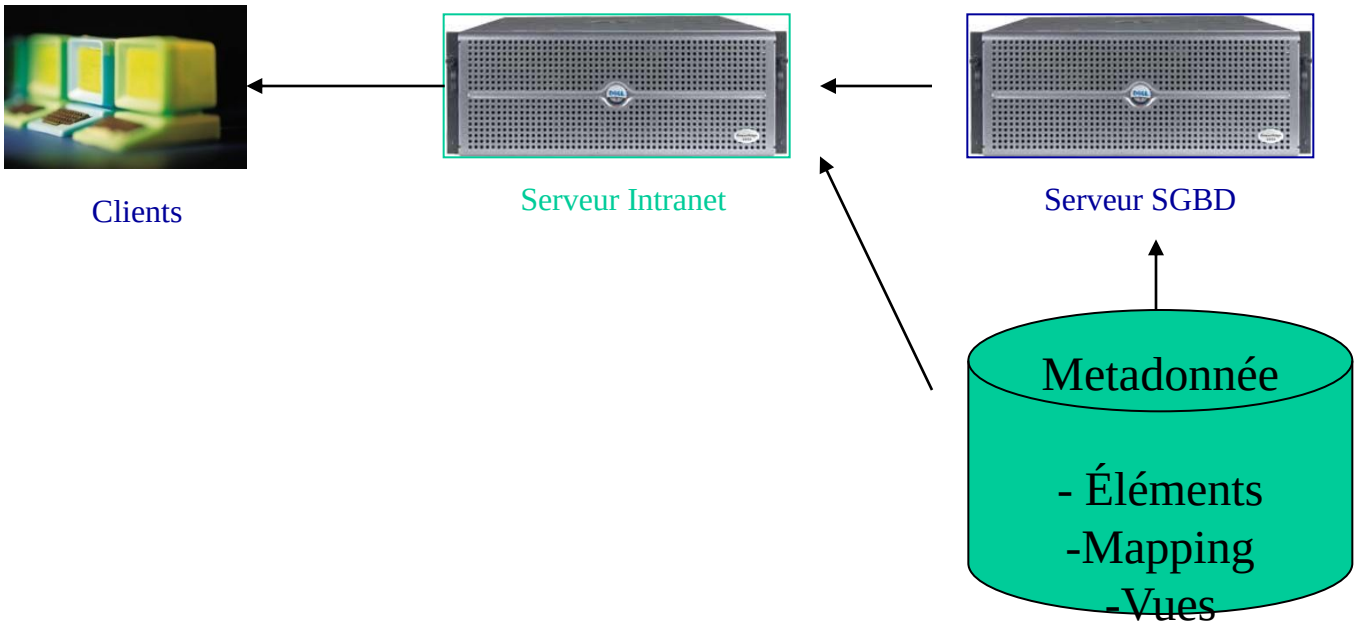
- Exemple Source – transformation - cible :



Source: Alimentation ODS, Lyne Fillion, Otéra Capital

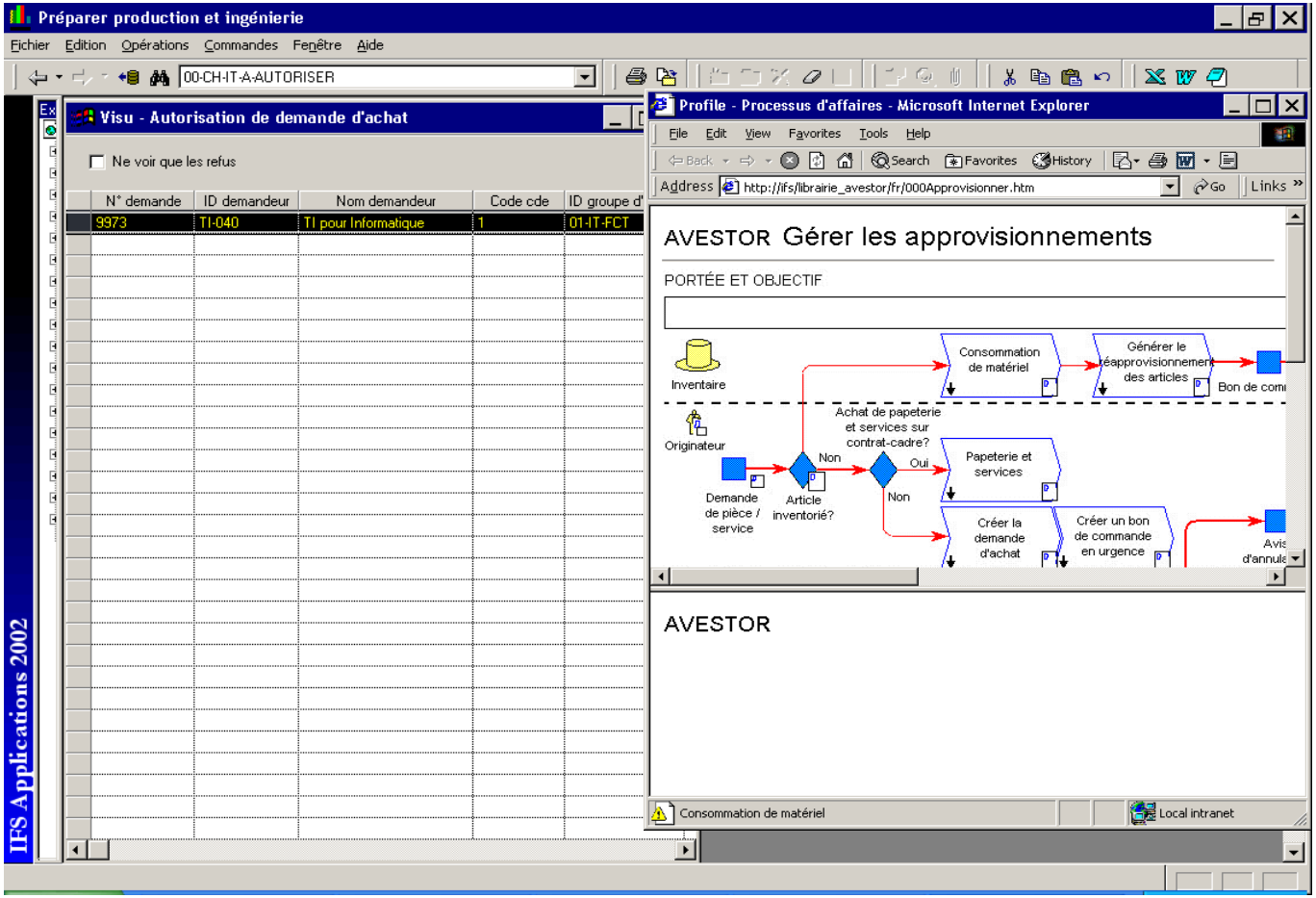
Méta-Information - Livraison

- Souvent Intranet



Méta-Information - Livraison

- Tendance applicative



Bell

Démo d'un site en construction

Méta-Information - Livraison

- Insertion dans les outils de création de modèles

  Ex: Business Object et l'Univers…

➔ S'assurer que la source est toujours l'entrepôt pour profiter de la méta information….

Méta-Information

- Réalité:
  - Environ 10% ont implanté un environnement de Méta-information avec succès.
  - Environ 15% y songent.

  (Data Warehousing Institute, 2008)

Votre livre (Paulraj Ponniah) dit 9% et 16%…