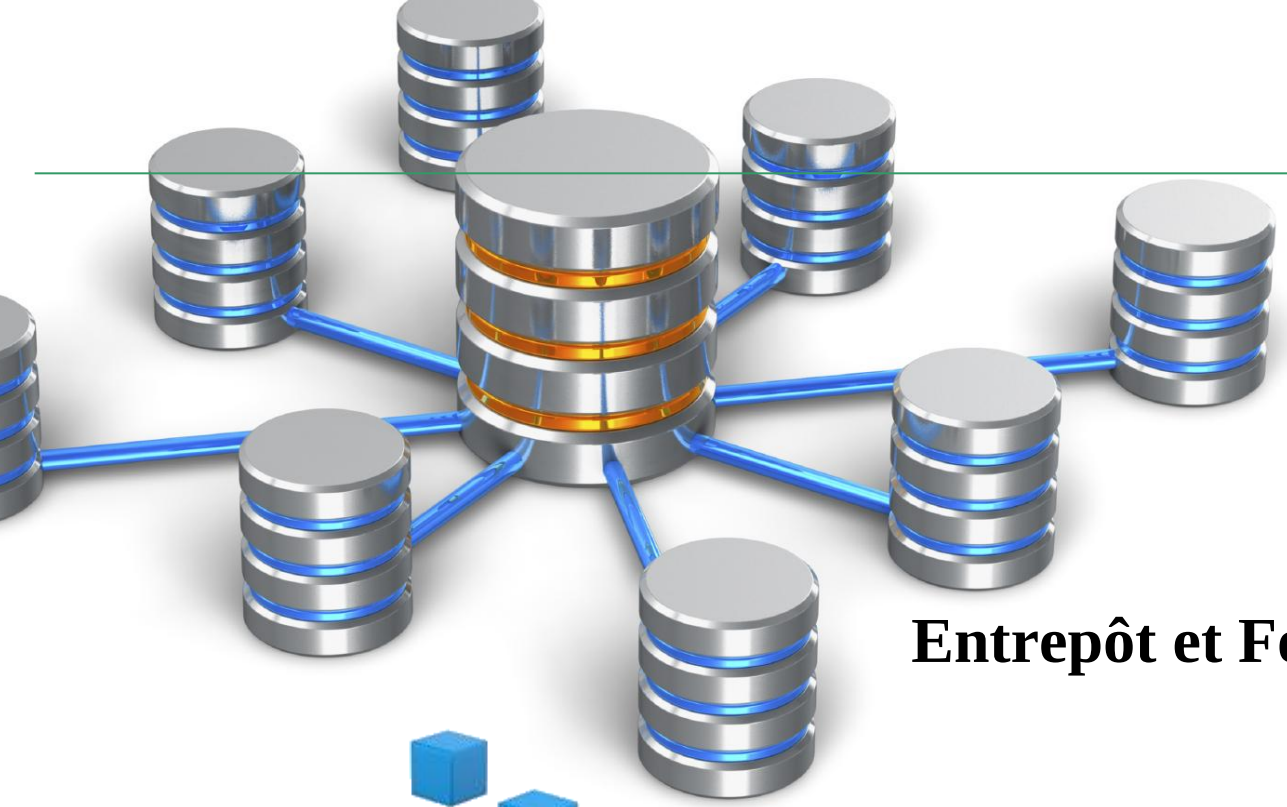
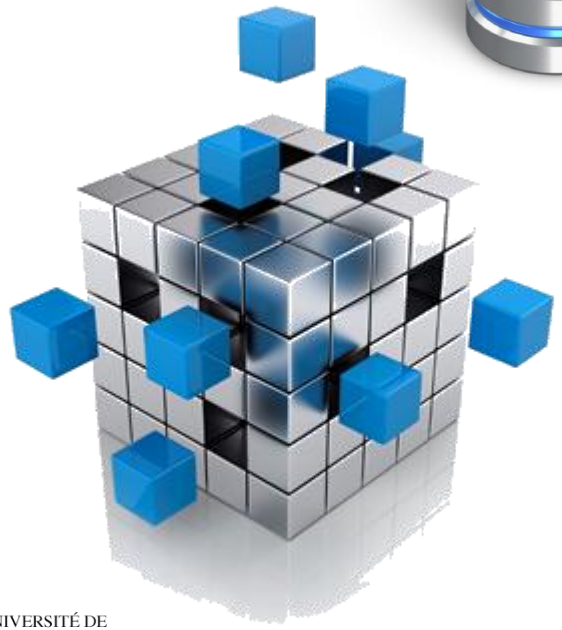


INF 735
Entrepôt et Forage de Données



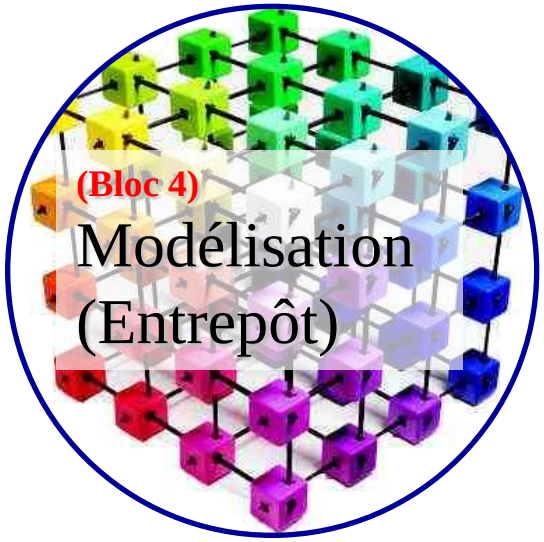
Bloc 4
ETC: Acquisition de données
par Robert J. Laurin

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 4) Architecture et Méta données

(Bloc 4) 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

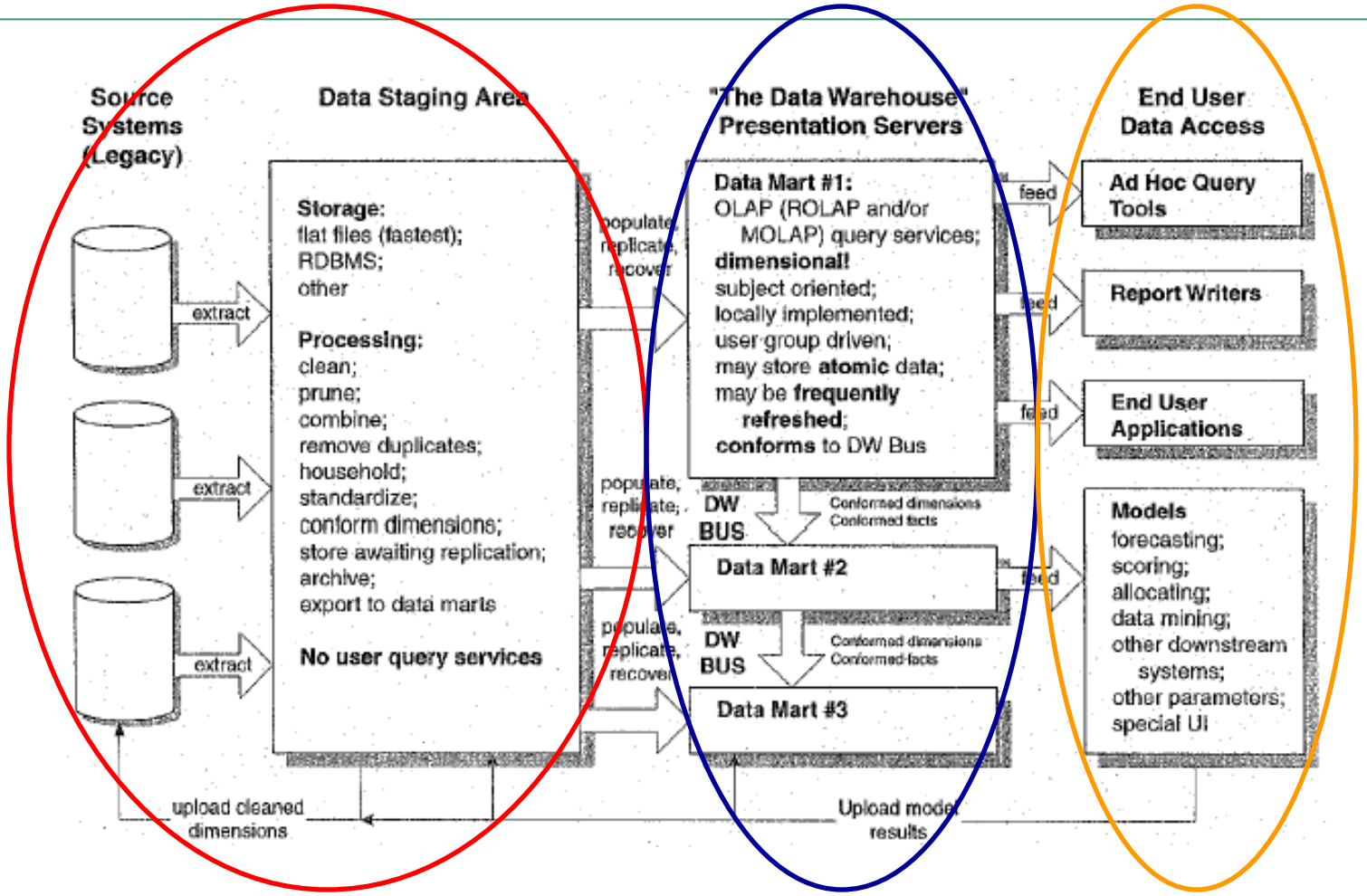
  – Chapitres 12 et 13

- Annexes:
  – Information systems data audit questionnaire
  – Data staging checklist
  – Data validation checklist
  – DBMS server code tree
  – La qualité des données, un facteur de succès en affaires (Direction Informatique)
  – Taking Data Quality to the Enterprise through Data Governance
  – Back room services

Concepts Clés du Bloc

- La donnée de qualité est durement obtenu.
- La phase d'ETC représente de loin le plus gros et plus important effort du projet global.
- Les systèmes opérationnels ne sont pas des historiens.

Phases



Bloc 4 Bloc 2 Bloc 3

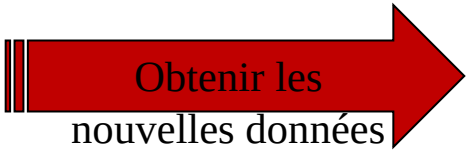
EXTRACTION, TRANSFORMATION, CHARGEMENT

REPRÉSENTE
jusqu'à de 70%
de l'effort total du projet

- Focus / Étapes:
  - Définir le besoin en information
    - + La cible pour combler le besoin
    - + quoi extraire (ne pas créer « une cour à scrape »)
  - Identifier les sources de donnée
  - Transformation requise
    - + Mapping des champs et formats
    - + Nettoyage
    - + Rassemblage, le montage et agrégation
    - + « Dé normaliser » au besoin pour le chargement
  - Définir le processus de mise à jour
    - + Comment identifier ce qui a changé – ce qui doit être chargé?

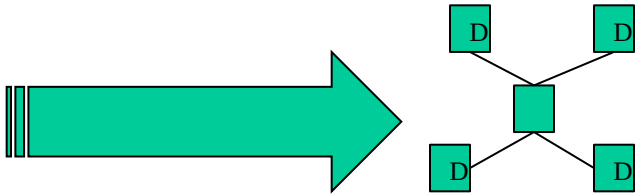
- E : Extraction

Source opérationnelle

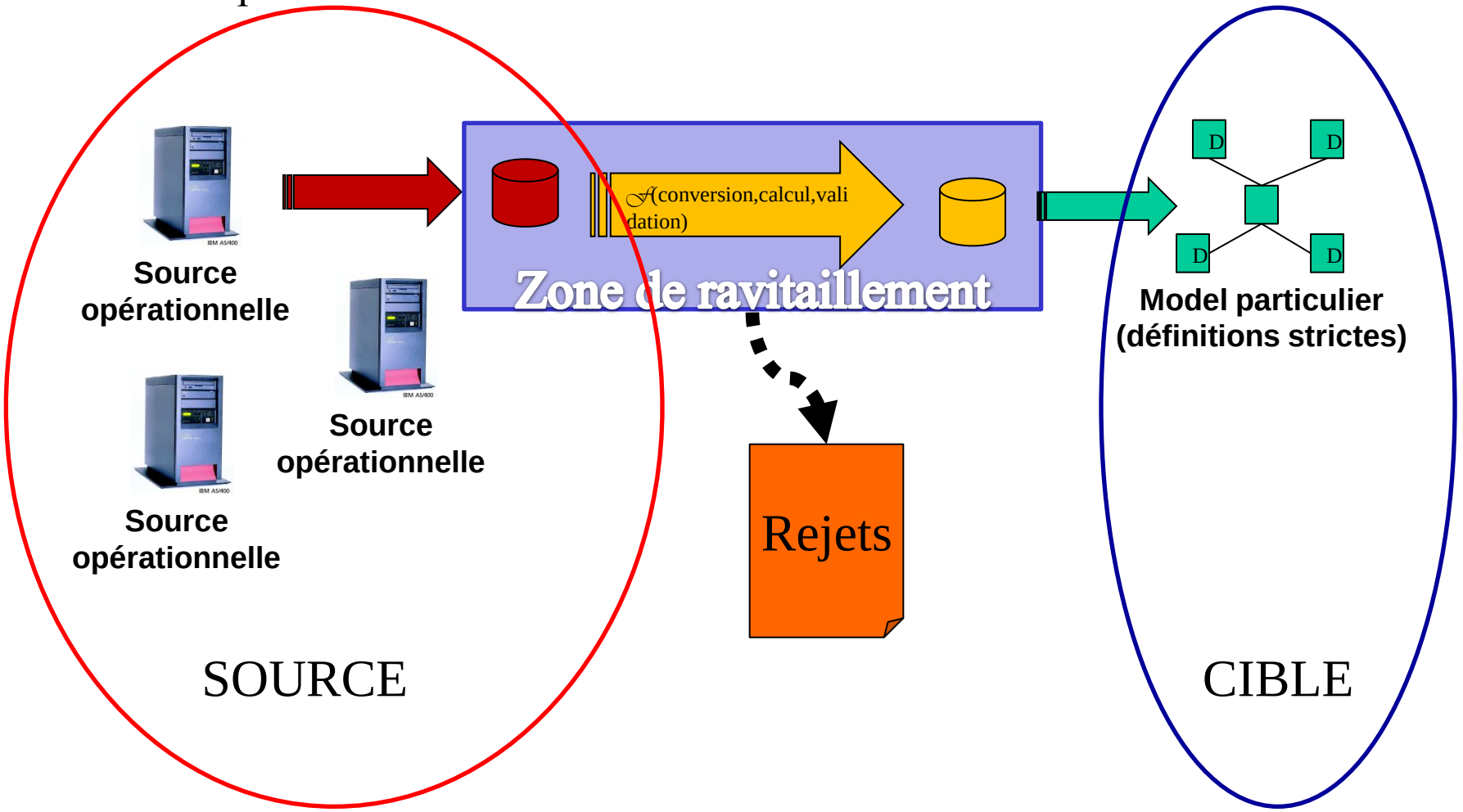


- T: Transformation

- C: Chargement

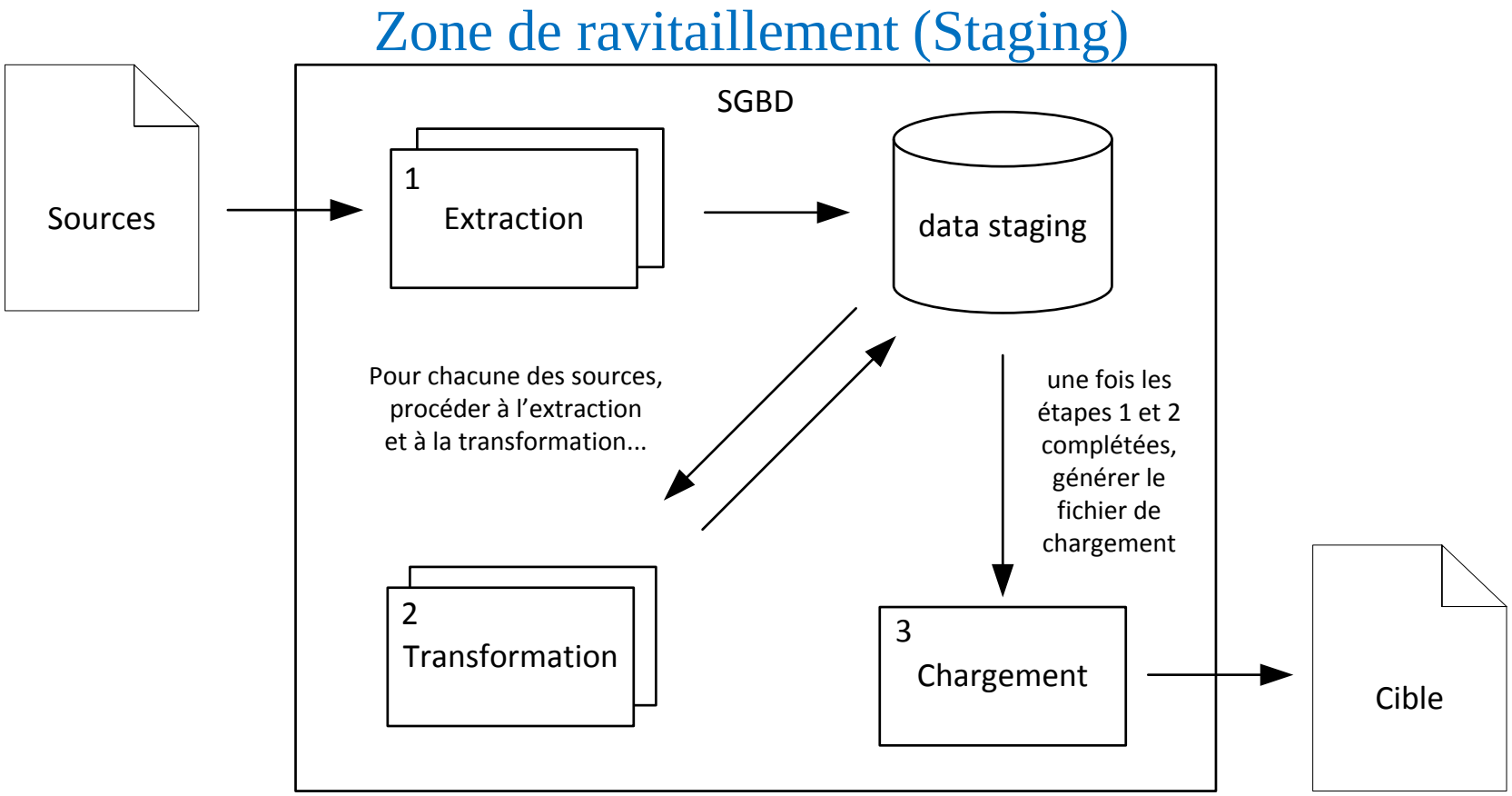


- ETC : processus d'amener la source à la cible



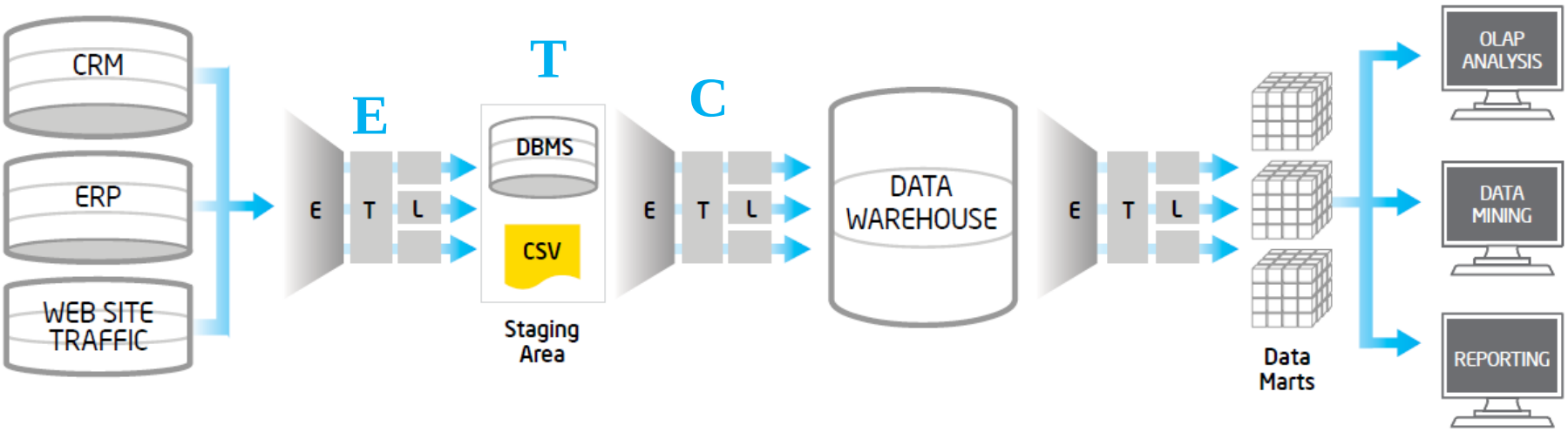
Étapes de l'ETC

Les étapes pour le chargement des données dans le « data mart » sont:



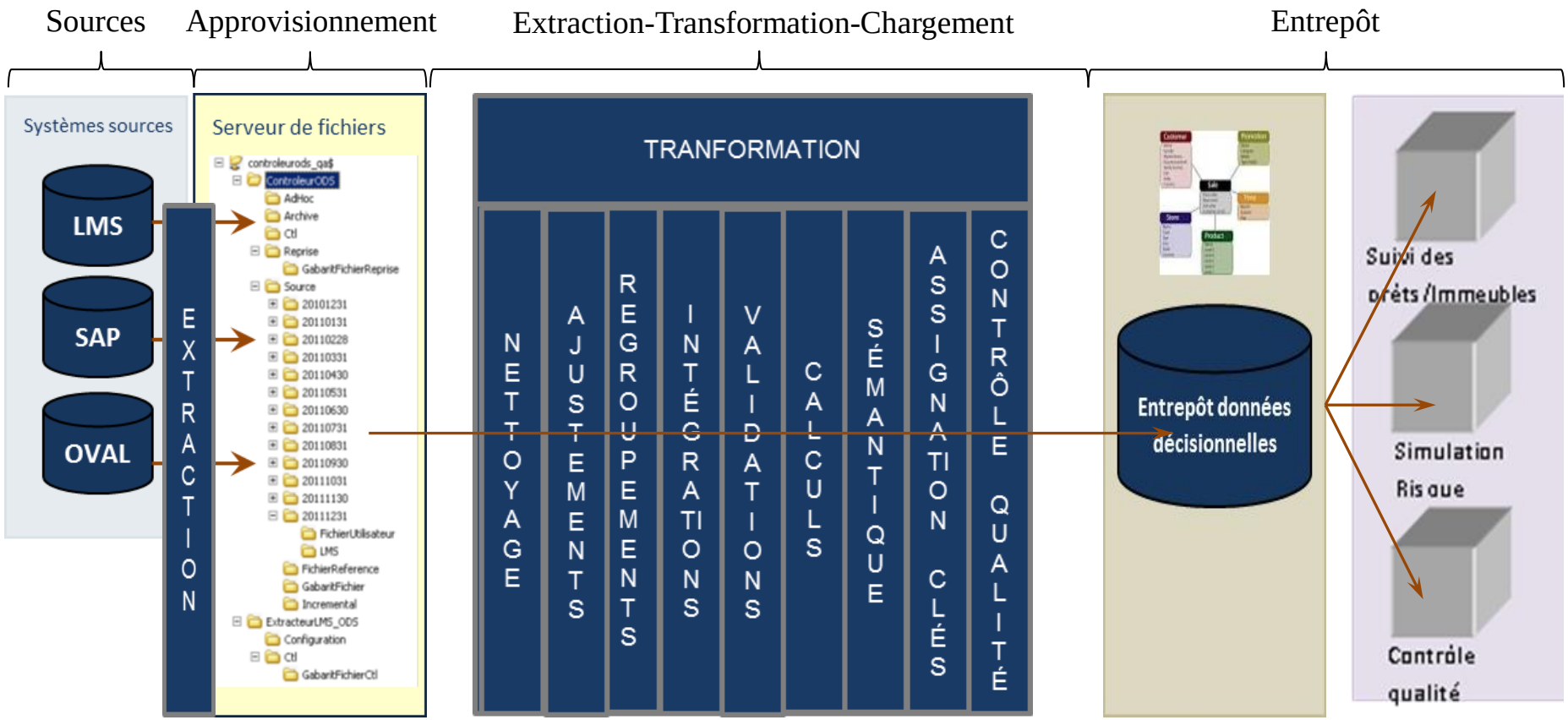
Source: TP1, Laura Francheri

ETC

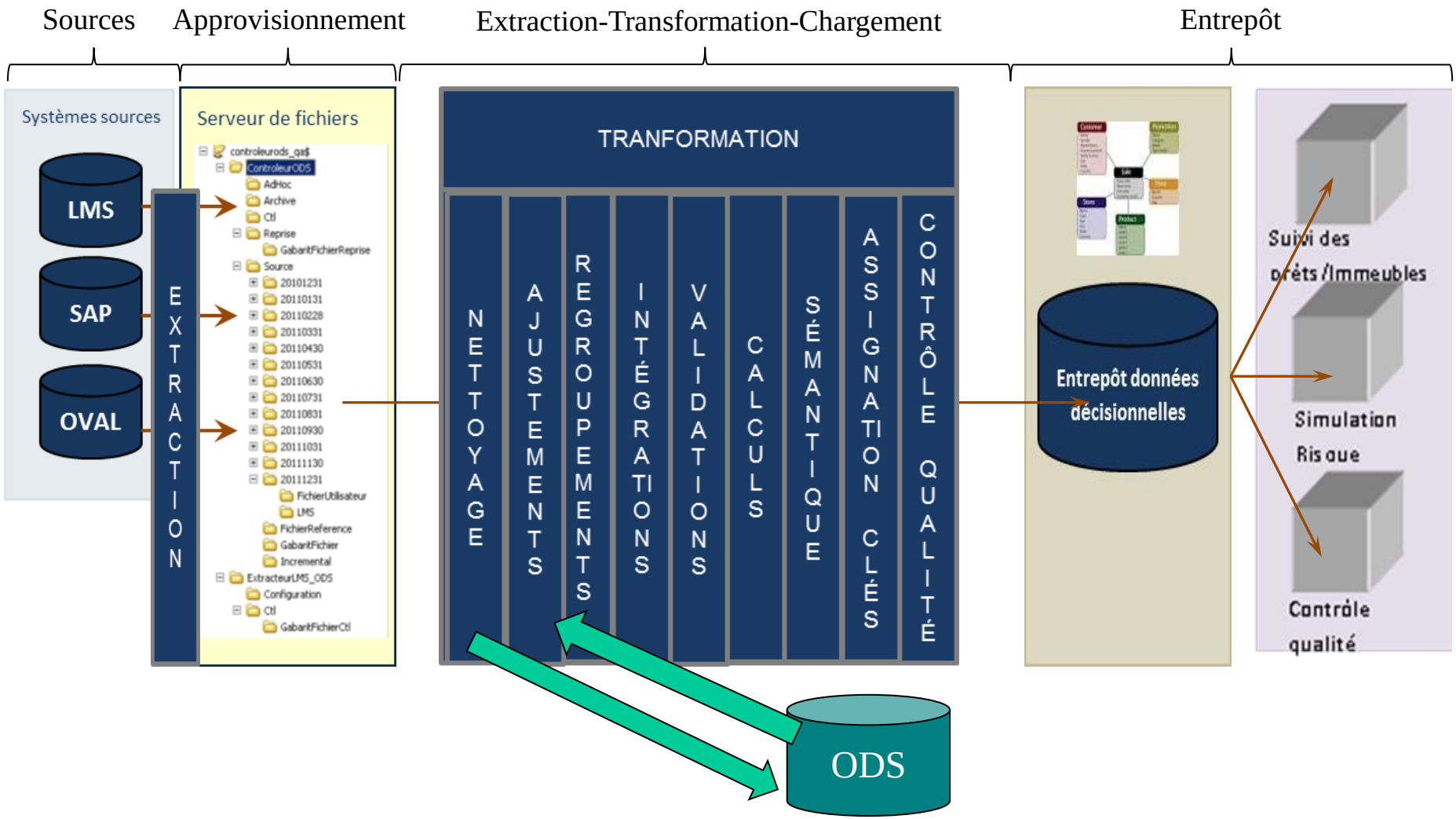


Source: Intel - White Paper, Big Data Analytics

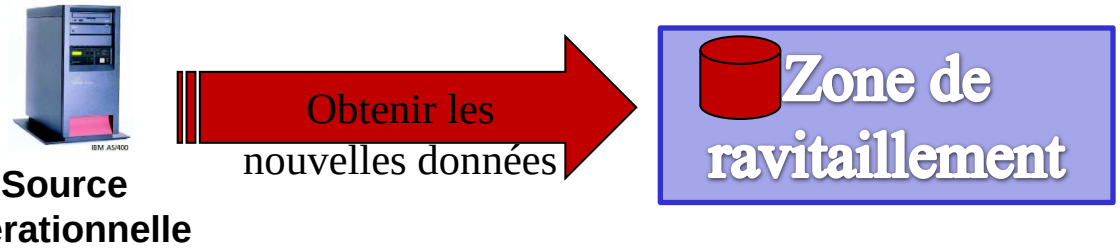
Flux de l'information



Flux de l'information



Extraction


Source opérationnelle
Obtenir les nouvelles données
Zone de ravitaillement

Extraction – Exemple d'architecture

- Ravitaillement des données (« Staging »)

Vancouver – Données de paye DB2

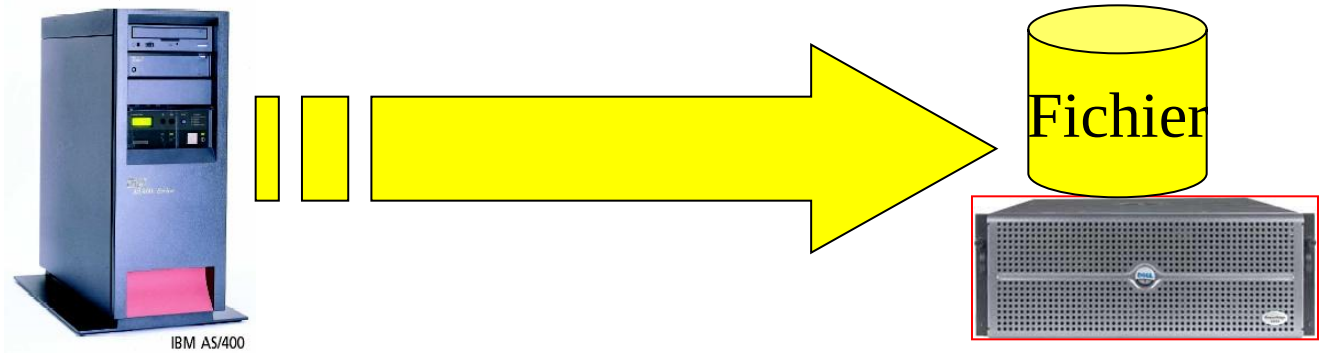
Sujet:
Qualification et développement des employés.

Extraire les heures de formation par sujet, employé, département, type, domaine
Pour les R.H.

Fichier de formation

Méthode techniques d'extraction – Acquisition de données

- **Pousse (« Push »)**



Vancouver – Données de paye DB2

Serveur Approvisionnement

•Avantages:

- •La source sait ce qui a changé…
- •La source le fait quand elle peut/veut
- •Sécurité de la source préservée

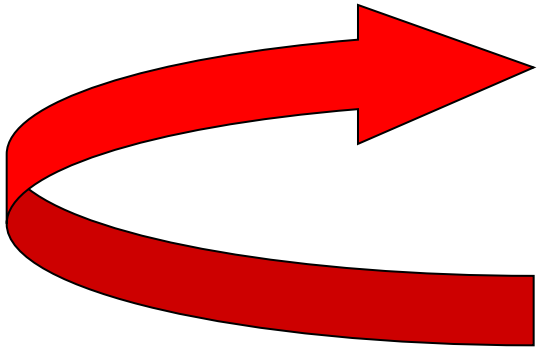
•Désavantages:

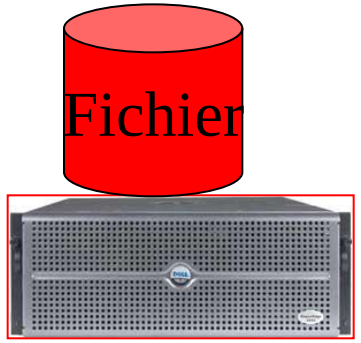
- •Réception pas nécessairement possible par zone de ravitaillement
- •Contrôle de la sécurité nécessaire de la zone de ravitaillement
- •Équipe « BI » Pas en contrôle si le fichier n'a pas été reçu (à temps)
- •Fardeau de la reprise pour la source

Méthode techniques d'extraction – Acquisition de données

- **Tire (« Pull »)**

Vancouver – Données de paye DB2





Serveur Approvisionnement

•Avantages:

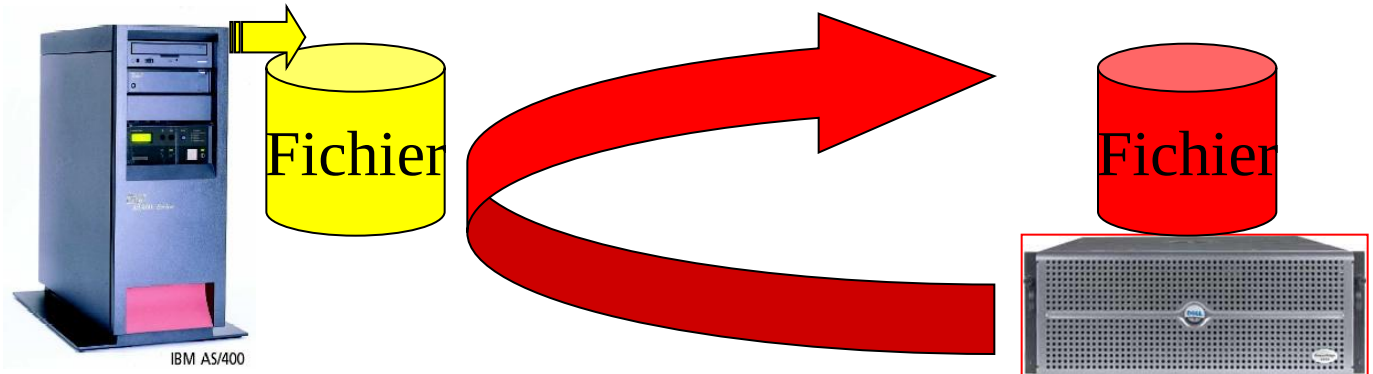
- •Agit selon l'horaire et contrôle pour ravitaillement
- •Meilleur possibilité de reprise en cas de problème

•Désavantages:

- •Source pas nécessairement disponible
- •ou plage horaire très serrée
- •Contrôle de la sécurité de la source
- •Possible impact sur la performance de la source
- •Comment savoir ce qui a changé ?

Méthode techniques d'extraction – Acquisition de données

- **Tire-Pousse (« Push/Pull »)**



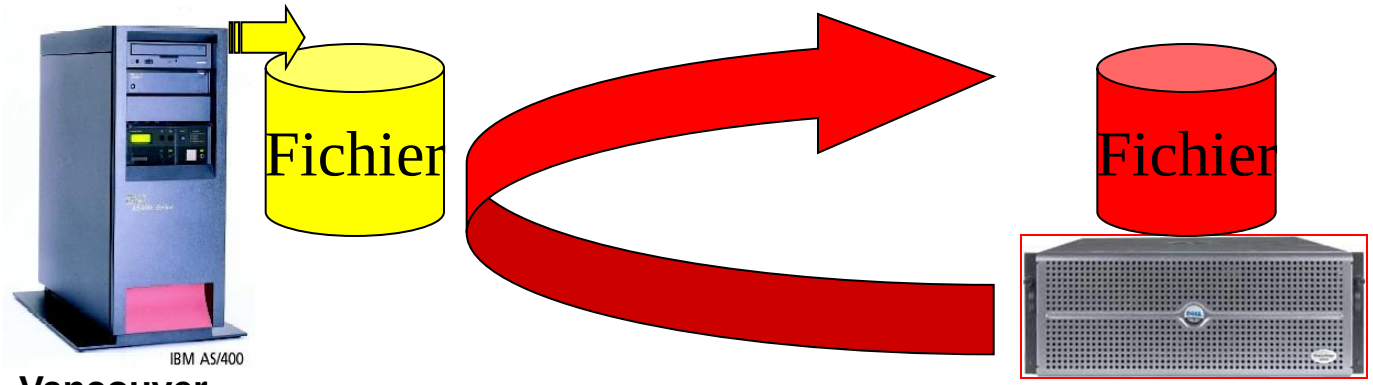
Vancouver – Données de paye DB2

Serveur Approvisionnement

- **Avantages:**
  - Tous les avantages

Infrastructure – Acquisition de données

- Préoccupations:



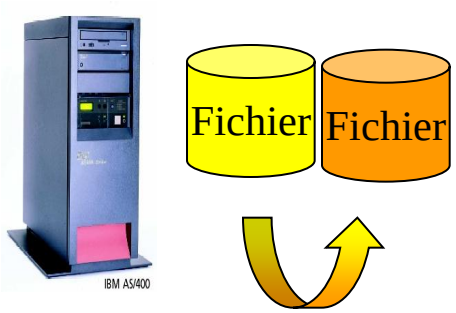
Vancouver – Données de paye DB2

Serveur Approvisionnement

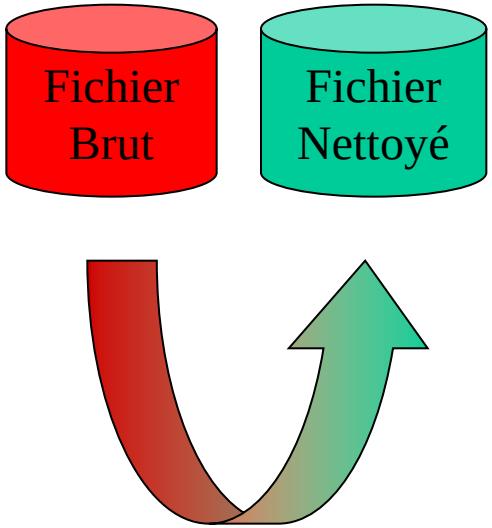
Extraction/Espace Horaire Sécurité Gestion des fichiers extraits	Télécommunications Réseau Protocoles Sécurité Bande passante	Espace Type de traitement Sécurité Gestion - Destruction - Alertes…

Infrastructure – Acquisition de données

- Nettoyage et traitement des données



Serveur Approvisionnement



Espace
Outils de traitement
SGBD ?
Sécurité
Gestion
- Destruction
- Alertes…

Types de sources

- Sources aussi diverses qu'il y a de procédés ou processus d'affaire
  - Automates (« PLC »)
  - Instruments
    + Balances
    + Systèmes météo
    + Systèmes téléphoniques
    + Contrôle d'Accès
    + Sondes
    + Binaires, linéaires, non-linéaires…
    + Etc…
  - Fichiers divers
  - BD/SGBD
    + Relationnels
    + Hiérarchiques
    + Dbase, btreive, autres PC et mac
    + BD spéciales ou sous contrôle applicatif (ex: SAP)
  - WEB
  - Aussi divers que techniquement possible

Extraction - Défis

- Sources disparates
- Plates-formes technologiques différentes
- Format de bd ou fichiers différents (sinon désuets).
- Le système opérationnel a changé et l'historique n'est pas conforme
- La qualité des donnée est correcte pour l'opération, mais pas pour l'historique
- Même donnée apparaissant sous formes différentes dans différents systèmes
- Sources en dehors des zones locales (« Politique » et « négociation »)
- Disponibilité (temps) pour extraction

Extraction – Identifier les sources de données

- Types Interne / Externe
- Systèmes
  - Architecture/Infrastructure
  - Nom des champs
  - Formats
  - Définition des valeurs (ex: 0 = oui, 1 = non)
  - Identifiant de changement
  - Plages critiques de disponibilité
- Sécurité de l'information
- Fréquence de mise à jour
- Stratégie de ravitaillement

Extraction – Identifier l'information

Raison:

- Identification du changement/ajout:
  - Extraction par l'application
  - Programmation (SGBD) Trigger
  - Log du SGBD
  - Log applicatif
  - Champs Date de modification au record
  - Comparaison des résultats
- Source: originale ou copie/relève

Chargement initial

.VS.

Mise à jour à différentes fréquences

Transformation

Transformation – Types

- Format et conversion
  - Dates
  - Texte .vs. Numérique
  - Points .vs. Virgules
  - Métrique .vs. Impérial
  - Nombre de décimales
  - EBCDIC .vs. ASCII
- Décodage
  - Male/Femelle
  - Oui/Non (0,1 – valeur ou nul)
  - Lettres représentant quelque chose…
- Calcul
- Séparation/Réunification de champs (ex: adresses, noms …)

- Amalgame de sources
- Regroupement et épuration de doublons
- Sommarisation (agrégation)
- Historisation (assurer la dimension temps)
- Ré-identification de la clé
- **Traitement des rejets**

AMENER SELON LA DÉFINITION STANDARDISÉE

Transformation – le cas des dates

- Les dates en valeur absolue pour ce qu'elles représentent (dans la dimension)

- Les dates en valeur relative calculée pour le fait.

Transformation – Ne pas normaliser, mais standardiser

1. L'entrepôt a une valeur historique
   (Variation d'un élément dans la dimension)

Vendeur	Quart	Vente
Bob	1-2003	100
Bob	2-2003	800
Bob	3-2003	500
Bob	4-2003	1200
Ray	1-2003	700
Ray	2-2003	650
Ray	3-2003	1400
Ray	4-2003	600
J.J.	1-2003	300
J.J.	2-2003	300
J.J.	3-2003	400
J.J.	4-2003	400

Vendeur	Région
Bob	Sud
Ray	Ouest
J.J.	Est

Extranction vente année

Région	Vente
Est	1400
Ouest	3350
Sud	2600

Vendeur	Région
Bob	Est
Ray	Ouest
J.J.	Sud

Modifié début Q3

Vendeur	Région
Bob	Sud
Ray	Ouest
J.J.	Est

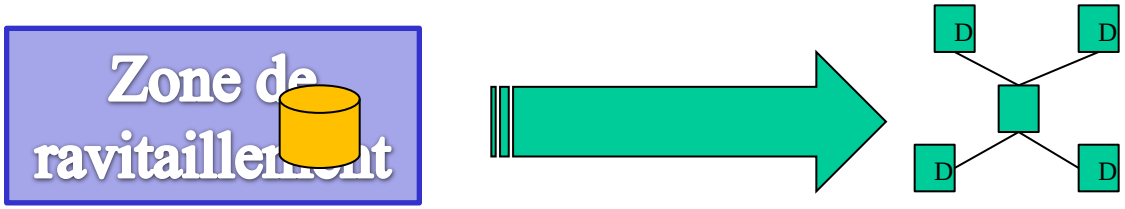
OU

Vendeur	Région	Quart	Vente
Bob	Est	1-2003	100
Bob	Est	2-2003	800
Bob	Sud	3-2003	500
Bob	Sud	4-2003	1200
Ray	Ouest	1-2003	700
Ray	Ouest	2-2003	650
Ray	Ouest	3-2003	1400
Ray	Ouest	4-2003	600
J.J.	Sud	1-2003	300
J.J.	Sud	2-2003	300
J.J.	Est	3-2003	400
J.J.	Est	4-2003	400

Extranction vente année

Région	Vente
Est	1700
Ouest	3350
Sud	2300

Chargement



Chargement

- Initial = Rafraîchissement complet

- Mise à jour (Faits):
  - Rafraîchissement (à chaque fois ou à intervalle avec ajouts)
  - Ajouts
  - Fusion
    + Destructive
    + Constructive (historisation du changement)

Historisation (reprise sur changement type 2)

- Client_X change de province du Québec à l'Ontario le 19 mai 1980…

Dim_Client

Clé_client	No_Client	Nom	Province	Actif	Date_changement
…	…	…	…	…	…
23456	Cl234	Client_X	Québec	N	1980-05-19
36680	Cl234	Client_X	Ontario	Y	1980-05-19
…	…	…	…	…	…

Fait_vente

Clé_client	Vente_$	*(date)*
…		
23456	100	*12/03/1979*
23456	355	*22/05/1979*
23456	233	*05/01/1980*
36680	545	*07/12/1985*
36680	666	*12/04/2001*

QUALITÉ DE LA DONNÉE

Qualité de la donnée

- Définition:
  - Exactitude
  - Dans les temps
  - Pour augmenter la confiance dans la prise de décision

Motivation de qualité

- Confiance dans les décisions
- Réduit le risque de mauvaises décisions
- Meilleur service à la clientèle
- Nouvelles opportunités de marketing
- Support à la ré-ingénierie des processus
- Augmentation de productivité
- Réduction des coûts
- Amélioration des opérations

« La crédibilité de l'entrepôt repose sur la qualité des données ! »

Quelques pistes de vérification de la donnée source

- Exacte
- Se réfère au domaine de valeurs acceptées
- Respect du type
- Constance
- Redondance
- Complète
- **Doublons**
- Données conforme aux règles d'affaire
- Les champs d'agrégation sont complets
- Anomalie de champ
- Claire
- Temporelle
- Utile
- Règles d'intégrité

Qualité de la donnée – Informationnel .vs. Opérationnel

- Valeurs nulles
- Données mal saisies
- Contournement de contraintes
- Dimension « free texte »
- La donnée voulue dépends de transactions peut-être saisies
- Donnée simplement pas disponible
- La validation propose un défaut ok pour l'opérateur
- Information contradictoire dans différentes sources
- Champs mal utilisés
- Abréviations non-connues
- Violation des règles d'affaire
- Clé en double !!!
- Absence de clé

« Si la donnée n'affecte pas l'opération, les chances sont qu'elle n'y soit pas ou qu'elle soit de mauvaise qualité ! »

Qualité – 2 types d'outils

- Découverte d'erreurs à la source
  - Doublons
  - Référence au domaine (valeurs possibles)
  - Règles d'intégrité
  - Sources multiples simultanées comparées

- Correction d'erreur
  - Normalisation
  - Regroupe des sources un peu différentes
  - Agrégation des doublons
  - Valeurs par défaut et validation

Il y aussi le SGBD qui fait office de filtre sur la « qualité » de la donnée.

Correction d'erreurs

- Corriger dans l'entrepôt ?
  - + Seulement sur erreur !
- Corriger dans l'ETC
  - – Programmation
    - + Programmation pour tous les cas
  - – Vérification utilisateur
    - + Routine de pré-chargement
    - + Chargement dans un environnement Contrôle Qualité (« QA »)
- Corriger la source?
  - + Impact sur l'opération
  - + Impact sur les interfaces
  - + Impact sur la sécurité
  - + Est-ce dans notre zone de contrôle?
  - + Reprise de l'extraction

Résistance à la qualité

- Investissement en outils
- Documentation système n'aide pas à la compréhension de la source
- Sans importance à la tâche (gratification)
- De toutes les priorités, ce n'est pas la plus haute pour un opérateur
- Qui veut vraiment soulever et brasser la M..?
- Doit reformer beaucoup de personnel
- L'alourdissement d'une tâche est rarement bien perçu
- Les coûts réels de l'amélioration de la qualité .vs. gains

« Il faut donner un niveau acceptable; 100% est rarement atteignable ! »

GOUVERNANCE

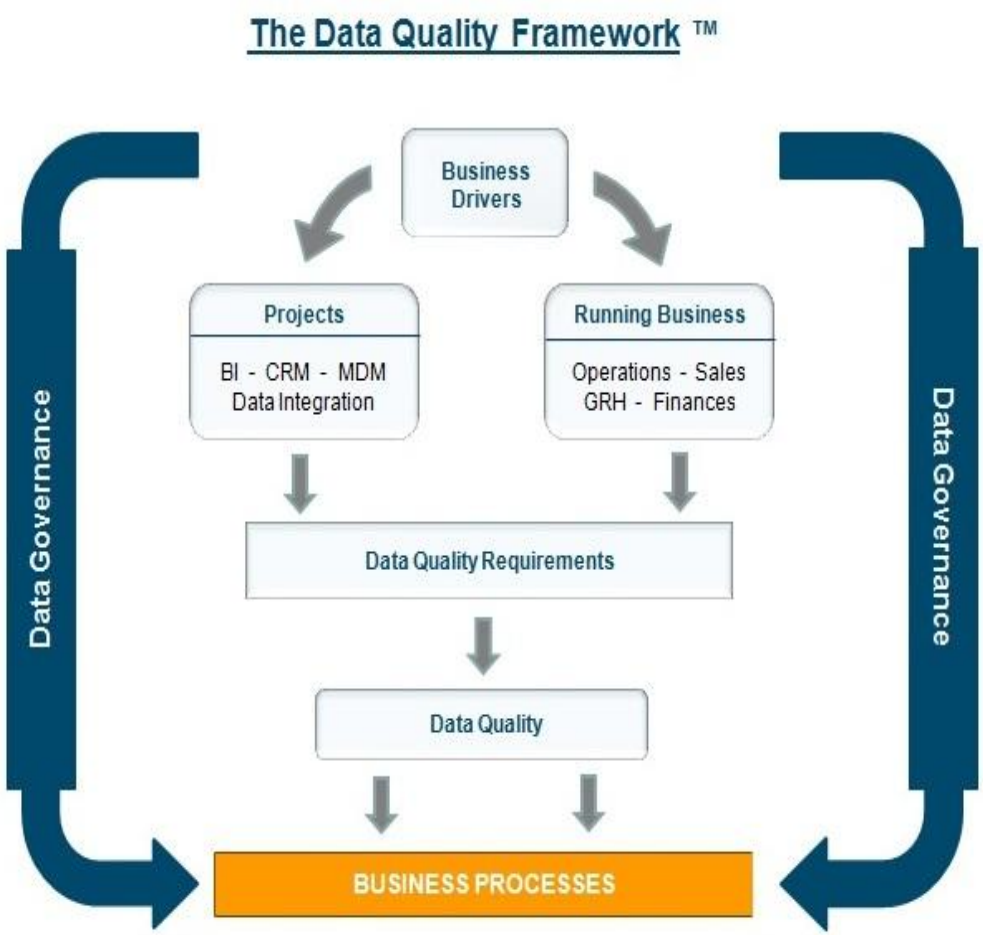
Rôles dans la Gouvernance de données

Rôles:

- **Qualité de données (DQ)**
- **Propriétaires de données (Data ownership)**
- **Intendance de données (Data stewardship)**
- **Gestion des métadonnées (Metadata management)**
- **Gestion des données maîtres (Master Data Management)**

Fondements d'une infrastructure de gestion de l'information (EIM)

Gouvernance et qualité de données



- La qualité de données (DQ) doit être mesurée et communiquée
- Les propriétaires de données (Data ownership) doivent être nommés et imputables
- L'intendance de données (Data stewardship) doit être facilitée par l'organisation
- Les métadonnées servent de support de communication à la gouvernance de données

COUCHES D'AJUSTEMENTS UTILISATEURS

Couches d'ajustements à l'ETC

1er cas: Couches d'Ajustements

Ajustement temporaire aux valeurs dans le fait

- Ajoute une couche prise en compte par la vue d'exploitation de l'étoile
- Ajuste les valeurs dans la génération des cubes
- Laisse un trace pour la vérification!

2ième cas Permettre à l'utilisateur de programmer les corrections d'erreurs.

- Ex: « Produit X » de l'opérationnel/ODS doit être « Produit Extraterrestre » dans l'entrepôt.

3ième cas: Permettre à l'utilisateur de contrôler les regroupements

- Associer un fait à une catégorie ou regroupement qui n'est pas à l'opérationnel (ex: Produits saisonniers, périssables, etc).

On peut ajouter 2 couches à la zone présentation, soit les calculs et les libellés personnalisés.

Couches utilisateurs – briser avec la tradition