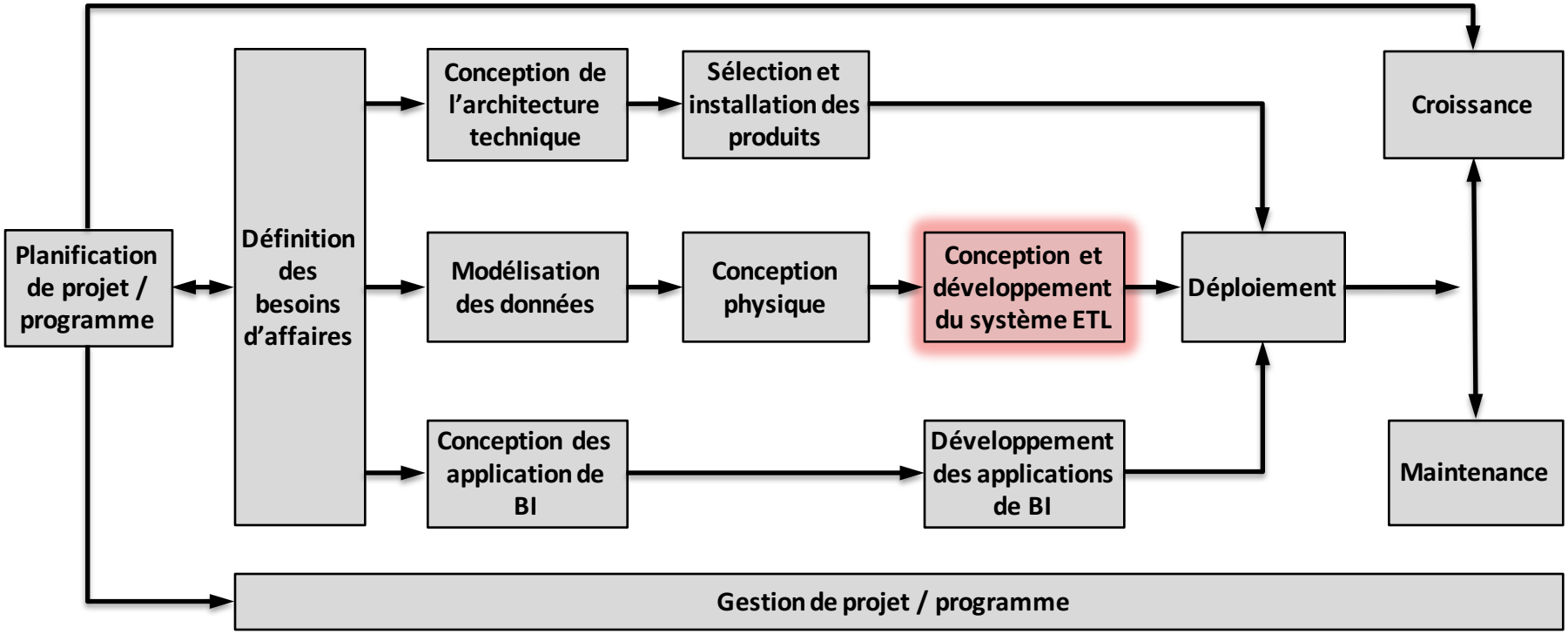
MTI820 – Entrepôts de données et intelligence d'affaires

Intégration des données et ETL

Le cycle de vie d'un projet en BI

- Diagramme de flux de travail:



Question

- Pourquoi est-il nécessaire de faire l'intégration des données?

Les problèmes des sources de données

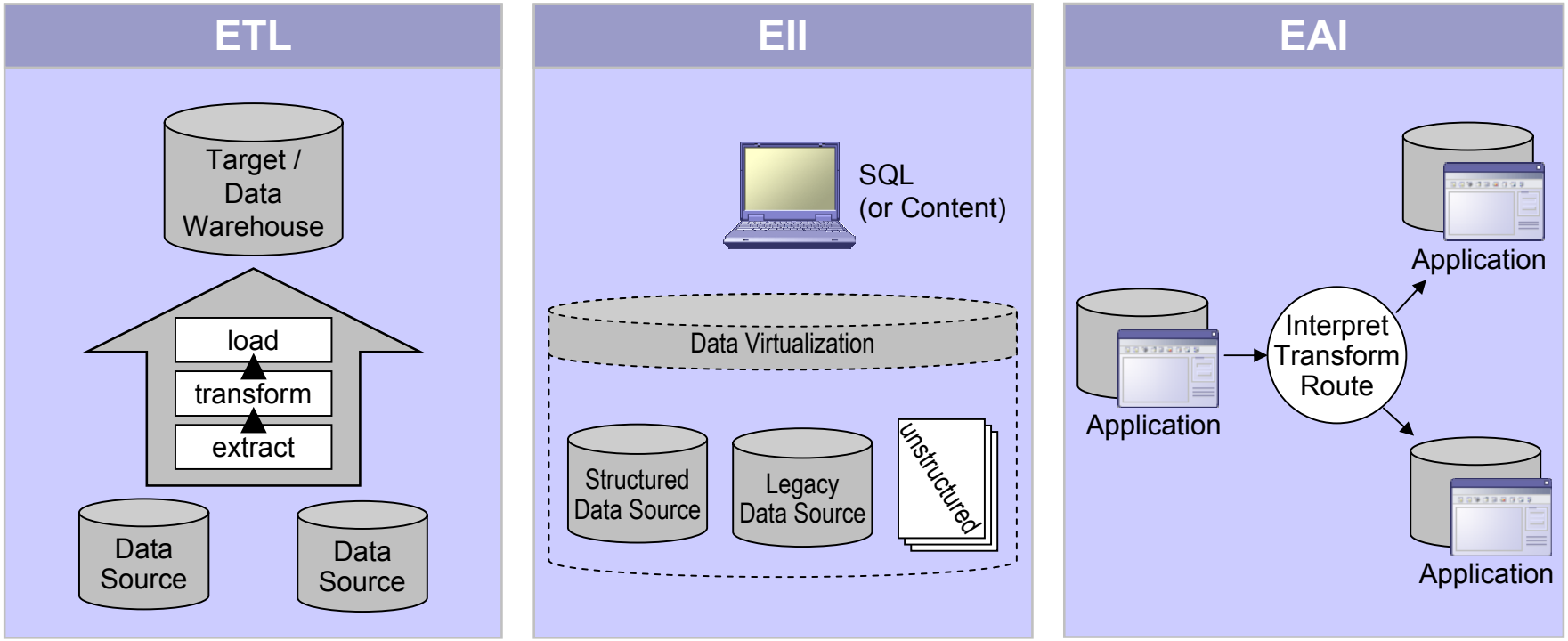
1. Sources diverses et disparates;
2. Sources sur différentes plateformes et OS;
3. Applications *legacy* utilisant des BD et autres technologies obsolètes;
4. Historique de changement non-préservé dans les sources;
5. Qualité de données douteuse et changeante dans le temps;
6. Structure des systèmes sources changeante dans le temps;
7. Incohérence entre les différentes sources;
8. Données dans un format difficilement interprétable ou ambigu.

Question

- Quelles sont les principales approches d'intégration et quels sont leurs principaux avantages/inconvénients?

Approches d'intégration

Source: "EII - ETL – EAI What, Why, and How!", Tom Yu, 2005



Extract, Transform and Load
- Intégration et livraison des données en lot
- Transformations appliquées sur les données

Enterprise Information Intergration
- Fédération de données provenant de plusieurs sources
- Accès temps-réel aux données
- Données structurées ou semi-structurées

Enterprise Application Intergration
- Processus d'intégration des données d'applications
- Basé sur l'échange de messages sur un bus commun

Extract, Transform and Load (ETL)

- Caractéristiques:
  - Permet la consolidation des données à l'aide des trois opérations suivantes:
    - Extraction: identifier et extraire les données de sources ayant subi une modification depuis la dernière exécution;
    - Transformation: appliquer diverses transformations aux données pour les nettoyer, les intégrer et les agréger;
    - Chargement: insérer les données transformées dans l'entrepôt et gérer les changements aux données existantes (ex: stratégies SCD).
  - Traite normalement de grande quantités de données en lots cédulés;
  - Est surtout utilisé avec les entrepôts de données et les comptoirs de données.

Extract, Transform and Load (ETL)

- Avantages:
  - Optimisé pour la structure de l'entrepôt de données;
  - Peut traiter de grandes quantités de données dans une même exécution (traitement en lot);
  - Permet des transformations complexes et agrégations sur les données;
  - La cédule d'exécution peut être contrôlée par l'administrateur;
  - La disponibilité d'outils GUI sur le marché permet d'améliorer la productivité;
  - Permet la réutilisation des processus et transformations (ex: *packages* dans SSIS).

Extract, Transform and Load (ETL)

- Inconvénients:
  - Processus de développement long et coûteux;
  - Gestion des changements nécessaire;
  - Exige de l'espace disque pour effectuer les transformations (*staging area*);
  - Exécuté indépendamment du besoin réel;
  - Latence des données entre la source et l'entrepôt;
  - Unidirectionnel (des sources vers l'entrepôt de données).

Entreprise Information Integration (EII)

- Caractéristiques:
  - Fournit une vue unifiée des données de l'entreprise, où les sources de données forment une fédération;
  - Les sources de données dispersées sont consolidées à l'aide d'une BD virtuelle, de manière transparente aux applications utilisant ces données;
  - Toute requête à la BD virtuelle est décomposée en sous-requêtes aux sources respectives, dont les réponses sont assemblées en un résultat unifié et consolidé;
  - Permet de consolider uniquement les données utilisées, au moment où elles sont utilisées (*source data pulling*).
  - Le traitement en-ligne des données peut cependant entraîner des délais importants.

Entreprise Information Integration (EII)

- Avantages:
  - Accès relationnel à des sources non-relationnelles;
  - Permet d'explorer les données avec la création du modèle de l'entrepôt de données;
  - Accélère le déploiement de la solution;
  - Peut être réutilisé par le système ETL dans une itération future;
  - Aucun déplacement de données.

Entreprise Information Integration (EII)

- Inconvénients:
  - Requiert la correspondance des clés d'une source à l'autre;
  - Consolidation des données plus complexe que dans l'ETL;
  - Surtaxe les système sources;
  - Plus limité que l'ETL dans la quantité de données pouvant être traitée;
  - Transformations limitées sur les données;
  - Peut consommer une grande bande passante du réseau.

Entreprise Application Integration (EAI)

- Caractéristiques:
  - Approche permettant de fournir à l'entrepôt des données provenant des sources (*source data pushing*);
  - Repose sur l'intégration et le partage des fonctionnalités des applications sources à l'aide d'une architecture SOA;
  - Généralement utilisé en temps réel ou en semi-temps réel (*Near Real Time*);
  - L'EAI ne remplace pas le processus ETL, mais permet de simplifier ce dernier.

Entreprise Application Integration (EAI)

- Avantages:
  - Facilite l'interopérabilité des applications;
  - Permet l'accès en (quasi) temps-réel;
  - Ne transfère que les données nécessaires;
  - Contrôle du flot de l'information.

Entreprise Application Integration (EAI)

- Inconvénients:
  - Support limité aux transformations et agrégations des données;
  - Taille des transactions limitée (en nombre de lignes);
  - Développement complexe;
  - Gestion complexe de l'intégrité sémantique des données (e.g., règles d'affaires);
  - Utilise la bande passante du réseau durant les heures de pointe.

Comparaison entre les approches d'intégration

	ETL	EII	EAI
Flot de données	Unidirectionnel (sources à l'entrepôt)	Bidirectionnel	Bidirectionnel
Mouvement de données	Lots cédulés	Au moment de la requête	Déclenché par la transaction
Latence	Journalier à mensuel	Temps-réel	Quasi temps-réel
Transformations/agrégations des données	Grande capacité	Moyenne capacité	Faible capacité
Volume des données	Grand (millions ou milliards de lignes)	Moyen (10,000 – 1,000,000 de lignes)	Petit (100-1000 lignes)

Quand utiliser les approches d'intégration

- Approche ETL:
  - Consolidation d'une grande quantité de données
  - Transformations complexes
- Approche EII:
  - Relier un entrepôt (EDW) existant avec des données de sources spécifiques
  - Données sources volatiles et accessibles à l'aide de requêtes simples (ex: SQL).
- Approche EAI:
  - Intégration de transactions
  - Requêtes analytiques simples
  - Sources non-accessibles directement

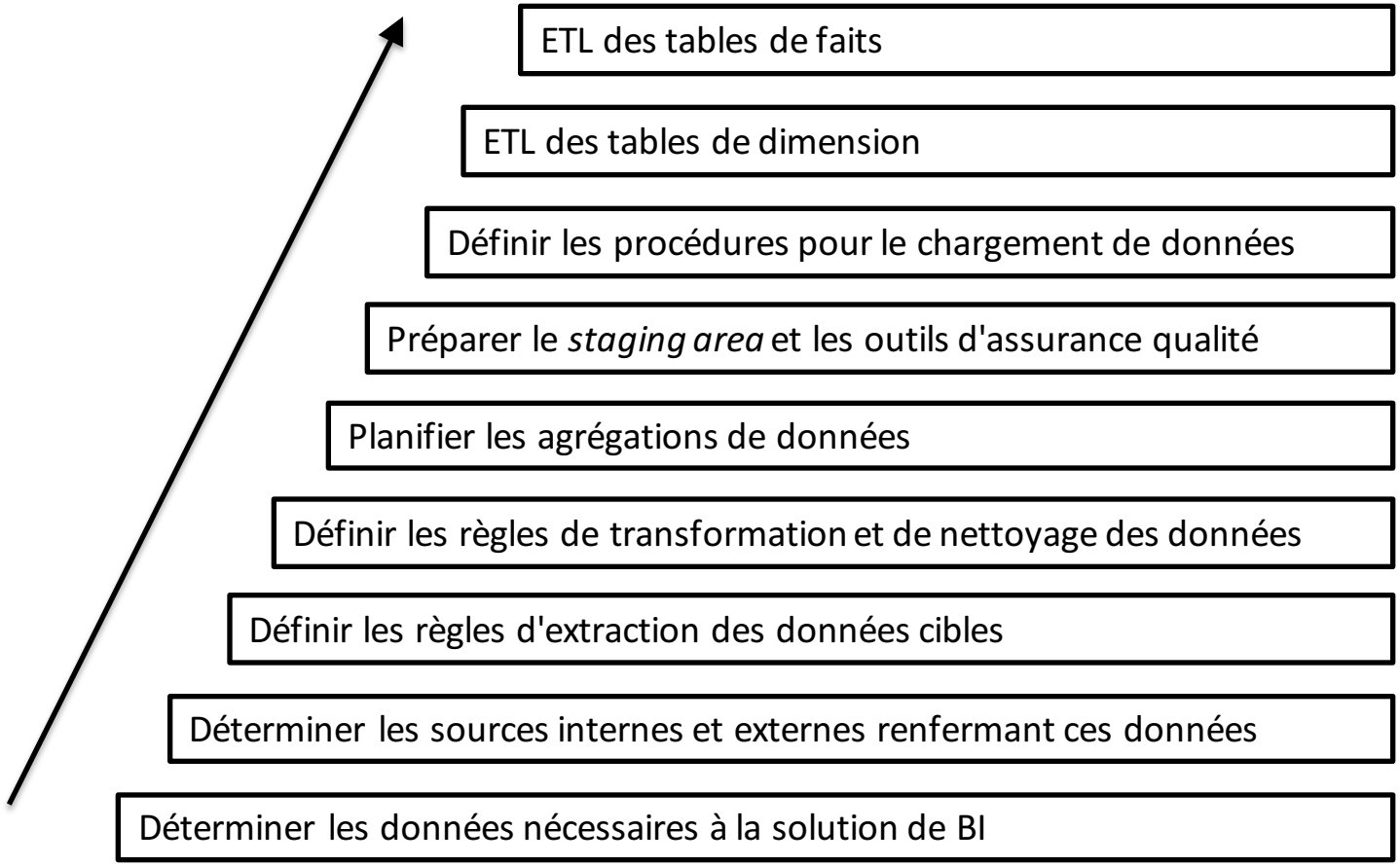
Exemples de produits commerciaux

- Outils ETL:
  - Oracle Warehouse Builder;
  - IBM Infosphere Information Server;
  - Microsoft SQL Server Integration Services (SSIS);
  - SAS Data Integration Studio.
- Outils EAI:
  - IBM WebSphere Message Broker;
  - Microsoft BizTalk Server;
  - Oracle SOA Suite.
- Outils EII:
  - SAP BusinessObjects Data Federator;
  - IBM WebSphere Federation Server.

Question

- Quelles sont les principales étapes dans le développement du système ETL?

Tâches et étapes de l'ETL



Extraction des données

- Identification des sources:
  1. Énumérer les items cibles (métriques et attributs de dimension) nécessaires à l'entrepôt de données;
  2. Pour chaque item cible, trouver la source et l'item correspondant de cette source;
  3. Si plusieurs sources sont trouvées, choisir la plus pertinente;
  4. Si l'item cible exige des données de plusieurs sources, former des règles de consolidation;
  5. Si l'item source referme plusieurs items cibles (ex: un seul champs pour le nom et l'adresse du client), définir des règles de découpage;
  6. Inspecter les sources pour des valeurs manquantes.

Extraction des données

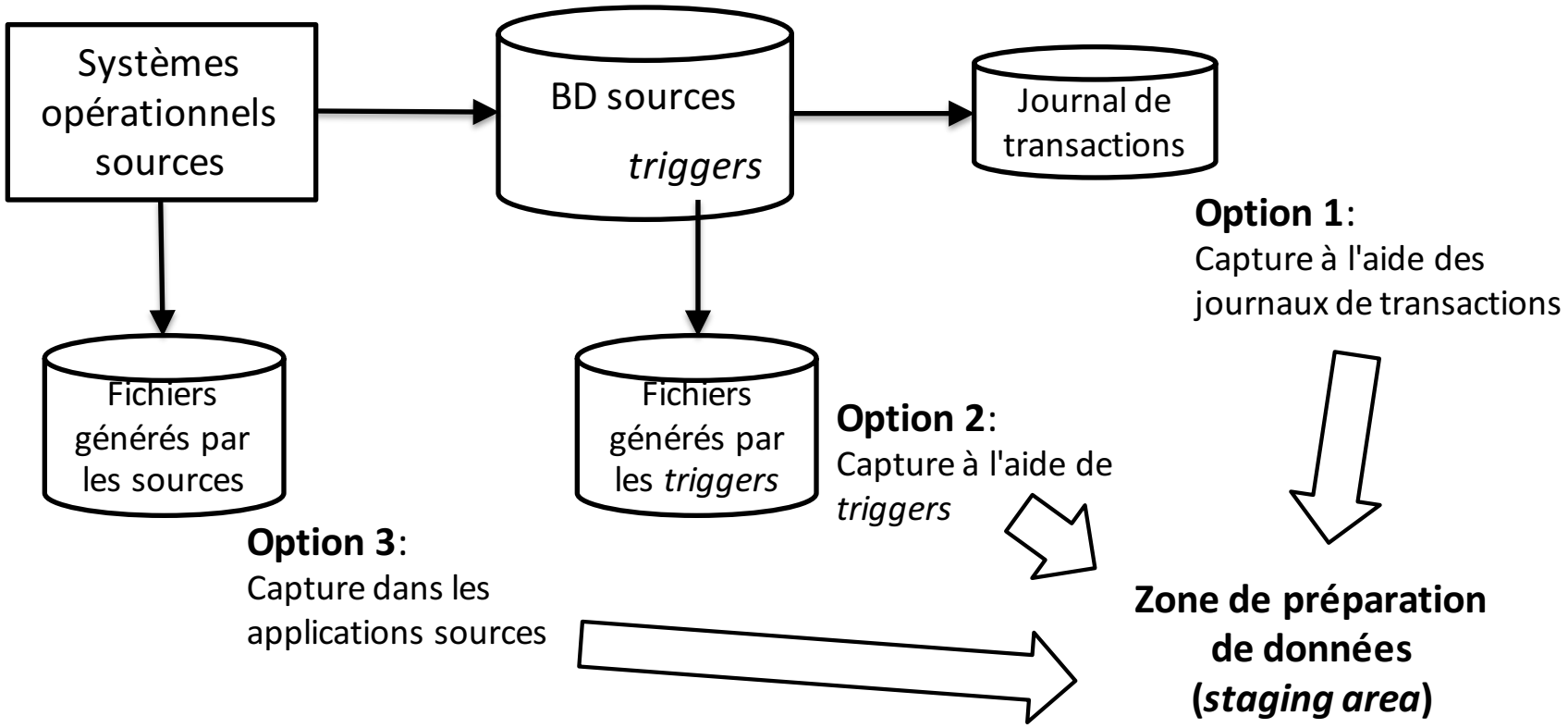
- Extraction complète:
  - Capture l'ensemble des données à un certain instant (*snapshot* de l'état opérationnel);
  - Normalement employée dans deux situations:
    1. Chargement initial des données;
    2. Rafraîchissement complet des données (ex: modification d'une source).
  - Peut être très coûteuse en temps (ex: plusieurs heures/jours).
- Extraction incrémentale:
  - Capture uniquement les données qui ont changées ou ont été ajoutées depuis la dernière extraction;
  - Peut être faite de deux façons:
    1. Extraction temps-réel;
    2. Extraction différée (en lot).

Question

- Comment peut-on extraire les données qui ont changées dans les sources:
  - En temps-réel?
  - En différé (lot)?

Extraction des données

- Extraction temps-réel:
  - S'effectue au moment où les transactions surviennent dans les systèmes sources.



Extraction des données

- Option 1: Capture à l'aide du journal des transactions
  - Utilise les logs de transactions de la BD servant à la récupération en cas de panne;
  - Aucune modification requise à la BD ou aux sources;
  - Doit être fait avant le rafraîchissement périodique du journal;
  - Pas possible avec les systèmes *legacy* ou les sources à base de fichiers (il faut une BD journalisée).

Extraction des données

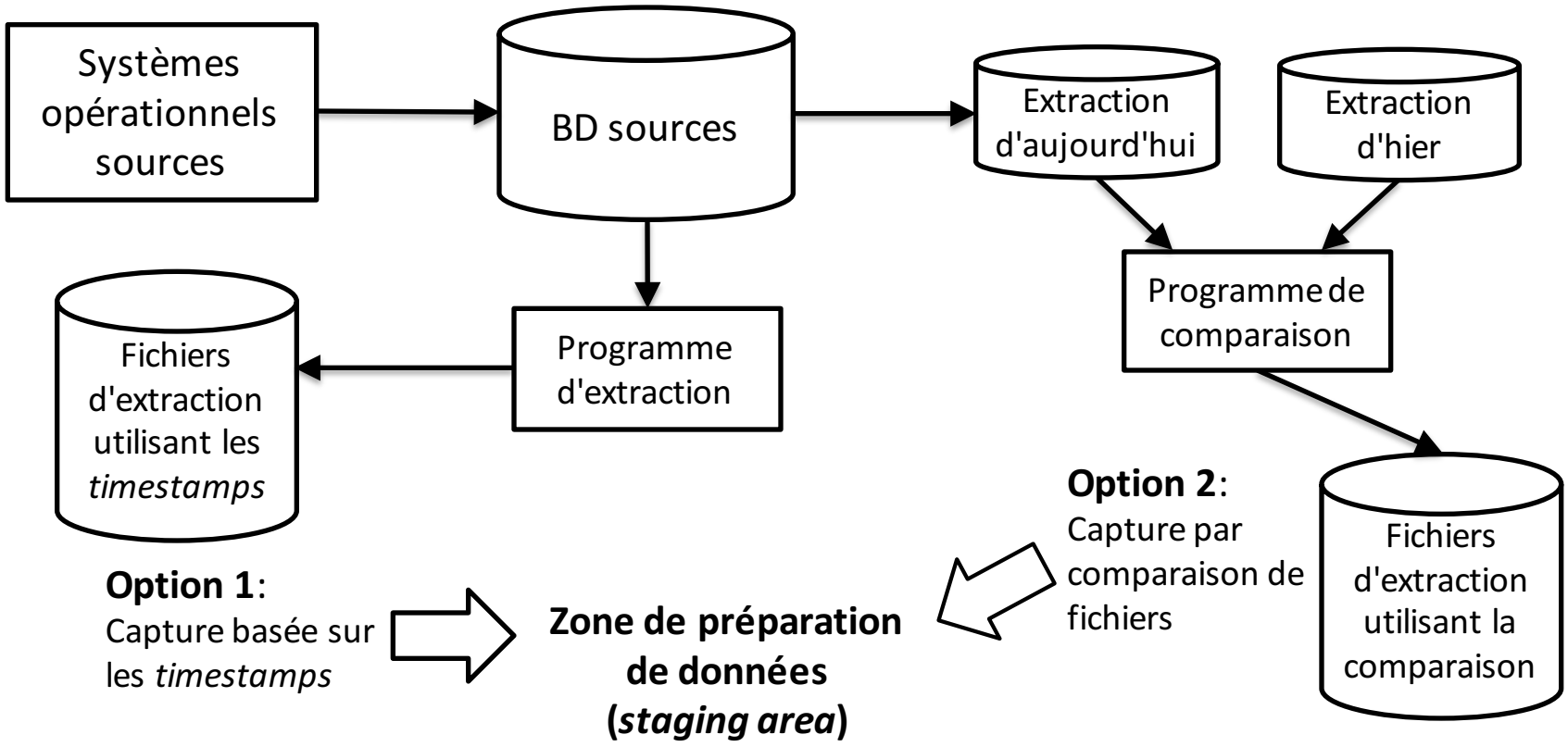
- Option 2: Capture à l'aide de *triggers*
  - Des procédures déclenchées (*triggers*) sont définies dans la BD pour recopier les données à extraire dans un fichier de sortie;
  - Meilleur contrôle de la capture d'évènements;
  - Exige de modifier les BD sources;
  - Pas possible avec les systèmes *legacy* ou les sources à base de fichiers.

Extraction des données

- Option 3: Capture à l'aide des applications sources
  - Les applications sources sont modifiées pour écrire chaque ajout et modification de données dans un fichier d'extraction;
  - Exige des modifications aux applications existantes;
  - Entraîne des coûts additionnels de développement et de maintenance;
  - Peut être employé sur des systèmes *legacy* et les systèmes à base de fichiers.

Extraction des données

- Extraction différée:
  - Extrait tous les changements survenus durant une période donnée (ex: heure, jour, semaine, mois).



Extraction des données

- Option 1: Capture basée sur les *timestamps*
  - Une estampille (*timestamp*) d'écriture est ajoutée à chaque ligne des systèmes sources;
  - L'extraction se fait uniquement sur les données dont le *timestamp* est plus récent que la dernière extraction;
  - Fonctionne avec les systèmes *legacy* et les fichiers plats, mais peut exiger des modifications aux systèmes sources;
  - Gestion compliquée des suppressions.

Extraction des données

- Option 2: Capture par comparaison de fichiers
  - Compare deux *snapshots* successifs des données sources;
  - Extrait seulement les différences (ajouts, modifications, suppressions) entre les deux *snapshots*;
  - Peut être employé sur des systèmes *legacy* et les systèmes à base de fichiers, sans aucune modification;
  - Exige de conserver une copie de l'état des données sources;
  - Approche relativement coûteuse.

Extraction des données

- Considérations pratiques:
  - Choisir, pour chaque source, la fenêtre temporelle durant laquelle sera faite l'extraction;
  - Déterminer la séquence des tâches d'extraction;
  - Déterminer comment gérer les exceptions.

Question

- Quelles sont les transformations à effectuer sur les données sources avant de les charger dans l'entrepôt?

Transformation des données

- Types de transformation:
  1. Révision de format:
     - Ex: Changer le type ou la longueur de champs individuels.
  2. Décodage de champs:
     - Consolider les données de sources multiples
       - Ex: ['homme', 'femme'] vs ['M', 'F'] vs [1,2].
     - Traduire les valeurs cryptiques
       - Ex: 'AC', 'IN', 'SU' pour les statuts *actif*, *inactif* et *suspendu*.
  3. Pré-calcul des valeurs dérivées:
     - Ex: *profit* calculé à partir de *ventes* et *coûts*.
  4. Découpage de champs complexes:
     - Ex: extraire les valeurs *prénom*, *secondPrénom* et *nomFamille* à partir d'une seule chaîne de caractères *nomComplet*.

Transformation des données

- Types de transformation (suite):
  5. Fusion de plusieurs champs:
     - Ex: information d'un produit
       - Source 1: code et description;
       - Source 2: types de forfaits;
       - Source 3: coût.
  6. Conversion de jeu de caractères:
     - Ex: EBCDIC (IBM) vers ASCII.
  7. Conversion des unités de mesure:
     - Ex: impérial à métrique.
  8. Conversion de dates:
     - Ex: '24 FEB 2011' vs '24/02/2011' vs '02/24/2011'.
  9. Pré-calcul des agrégations:
     - Ex: ventes par produit par semaine par région.
  10. Déduplication:
     - Ex: Plusieurs enregistrements pour un même client.

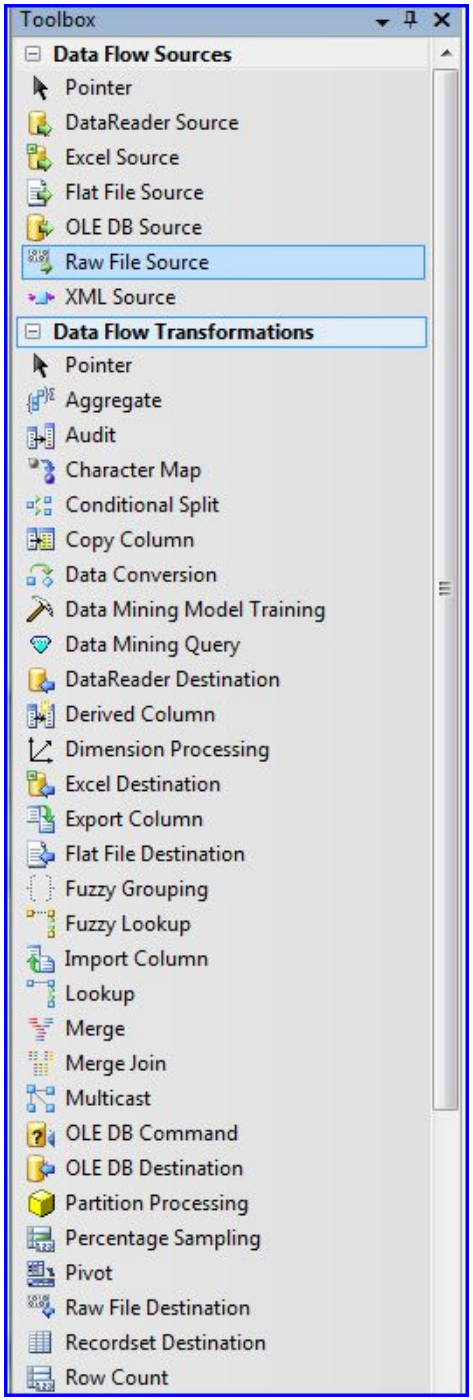
Transformation des données

- Problème de résolution d'entités:
  - Survient lorsqu'une même entité se retrouve sur différentes sources, sans qu'on ait la correspondance entre ces sources;
    - Ex: clients de longue date ayant un identifiant différent sur les différentes sources;
  - L'intégration des données requiert de retrouver la correspondance;
  - Approches basées sur des règles de résolution
    - Ex: les entités doivent avoir au moins N champs identiques (*fuzzy lookup / matching*).
- Problème des sources multiples:
  - Survient lorsqu'une entité possède une représentation différente sur plusieurs sources;
  - Approches de sélection:
    - Choisir la source la plus prioritaire;
    - Choisir la source ayant l'information la plus récente.

Transformation des données

- Gestion des changements dimensionnels:
  - Déterminer la stratégie de gestion des changements (SCD Type 1, 2 ou 3) de chaque attribut dimensionnel modifié;
  - Préparer l'image de chargement (*load image*) en conséquence:
    - **SCD Type 1**: ancienne valeur écrasée;
    - **SCD Type 2**: nouvelle ligne ajoutée;
    - **SCD Type 3**: déplacement de l'ancienne valeur dans la colonne d'historique et écriture de la nouvelle valeur dans la colonne courante.

Exemples de transformations: (SSIS)



Transformation des données

- Matrice de transformation:

Champs cible	Table cible	Champs source	Table source	Règle de transformation

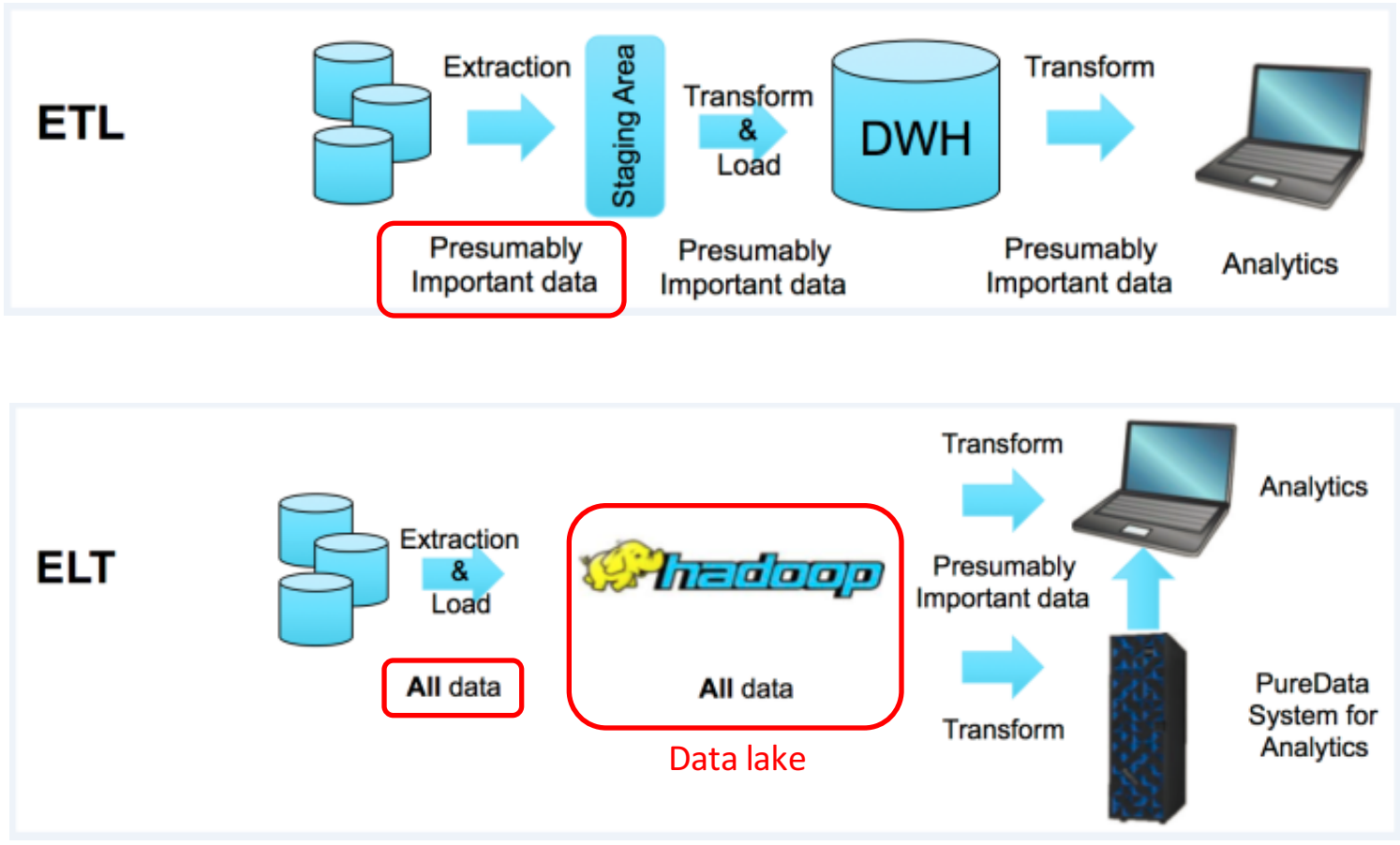
Chargement des données

- Types de chargement:
  - Chargement initial:
    - Fait une seule fois lors de l'activation de l'entrepôt de données;
    - Les indexes et contraintes d'intégrité référentielle (clé étrangères) sont normalement désactivés temporairement;
    - Peut prendre plusieurs heures.
  - Chargement incrémental:
    - Fait une fois le chargement initial complété;
    - Tient compte de la nature des changements (ex: SCD Type 1, 2 ou 3);
    - Peut être fait en temps-réel ou en lot.
  - Rafraîchissement complet:
    - Employé lorsque le nombre de changements rend le chargement incrémental trop complexe;
    - Ex: lorsque plus de 20% des enregistrements ont changé depuis le dernier chargement.

Chargement des données

- Considération additionnelles:
  - Faire les chargements en lot dans une période creuse (entrepôt de données non utilisé);
  - Considérer la bande passante requise pour le chargement;
  - Avoir un plan pour évaluer la qualité des données chargées dans l'entrepôt;
  - Commencer par charger les données des tables de dimension.

Extract, Load and Transform (ELT)



Source: Ralf Goetz, What is the fundamental difference between "ETL" and "ELT" in the world of big data?, 2017

Extract, Load and Transform (ELT)

- Problèmes avec ETL:
  - Traite les données importantes au moment de la **conception**;
  - Développement long et complexe.
- Solution ELT:
  - Utilise les technologies BigData (ex: Hadoop, Spark, etc.);
  - Chargements rapides, potentiellement temps-réels;
  - Lacs de données (data lakes) permettent les données non-structurées;
  - Transformations faites au moment de la requête;
  - Technologies moins matures que ETL et implémentation plus complexe (ex: code Java vs outil dédié).

Extract, Load and Transform (ELT)

	Entrepôts de données	Lacs de données
Données	Structurées, traitées	Semi-/non-structurées, brutes
Traitement	Schéma-à-l'écriture	Schéma-à-la-lecture
Stockage	Coûteux pour grandes quantités	Stockage à faible coût
Agilité	Moins agile, configuration fixe	Hautement agile, reconfiguration au besoin
Sécurité	Mature	Moins mature
Utilisateurs	Professionnels d'affaires	Experts en data science