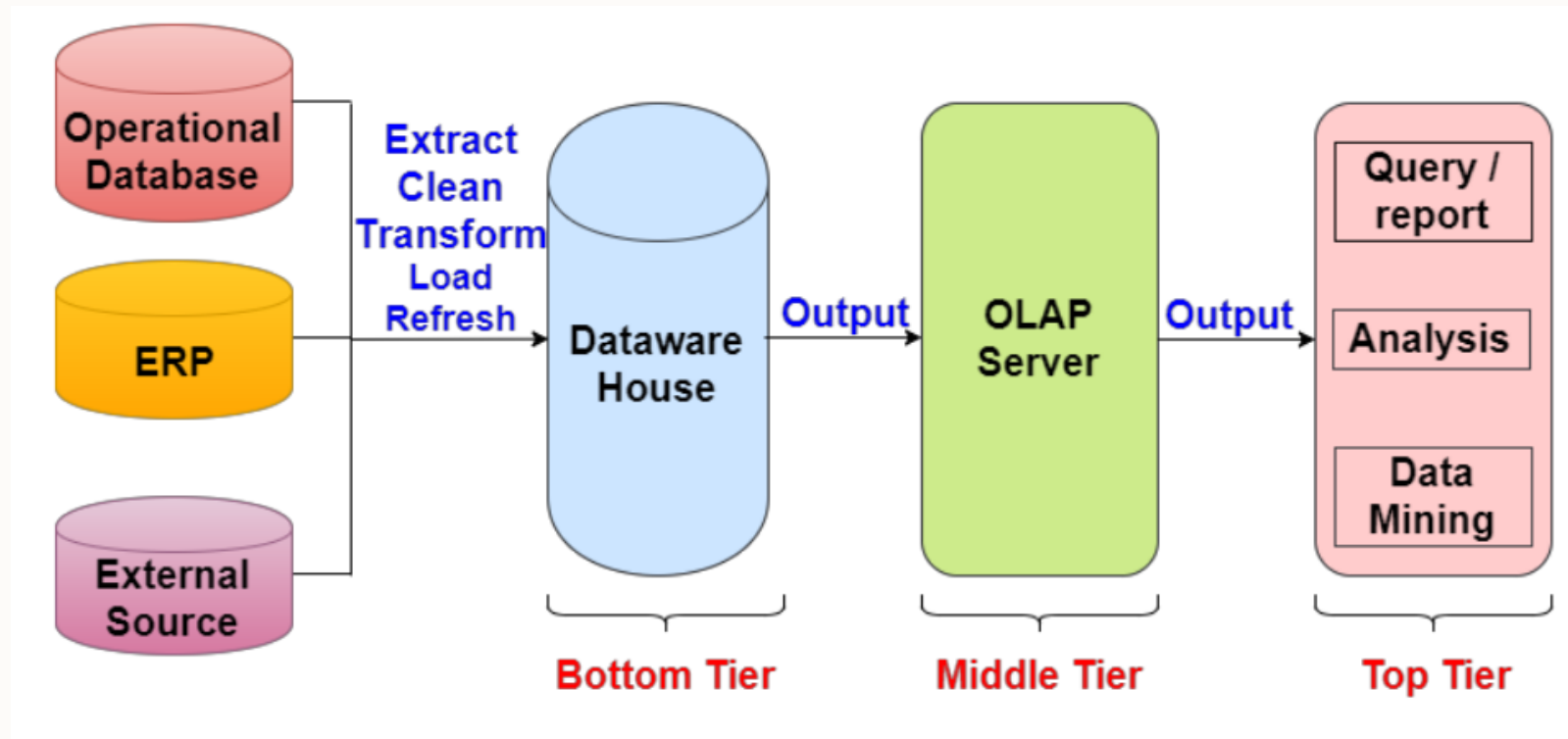


Chapitre 3 :

Le système de Data Warehouse et ses composants (Architecture d'un Data Warehouse)

Le système de Data Warehouse et ses composants

Architecture d'un Data Warehouse



Architecture générale Data Warehouse

Le système de Data Warehouse et ses composants

Architecture d'un Data Warehouse

Généralement, un entrepôt de données adopte une architecture à trois niveaux :

- **Niveau inférieur** ou **Bottom Tier** : composé généralement du système de base de données relationnel de l'entrepôt. Les programmes d'applications et les utilitaires ETL sont utilisés pour fournir les données au niveau inférieur.
- **Niveau intermédiaire** ou **Middle Tier** : le niveau où se trouve le serveur OLAP implémenté par deux modèles OLAP relationnel (ROLAP) et OLAP multidimensionnel (MOLAP).
- **Niveau supérieur** ou **Top Tier** : c'est la couche client. Elle contient les outils de requête et les outils de génération de rapports, les outils d'analyse et les outils d'exploration des données

la **différence** la plus importante entre les deux est **que ROLAP** fournit des données, directement à partir de l'entrepôt de données(data warehouse) principal, alors **que MOLAP** fournit des données à partir des bases de données propriétaires MDDB(Multi Dimensional Data Base)

Le système de Data Warehouse et ses composants

Processus Extract, Transform, Load (ETL)

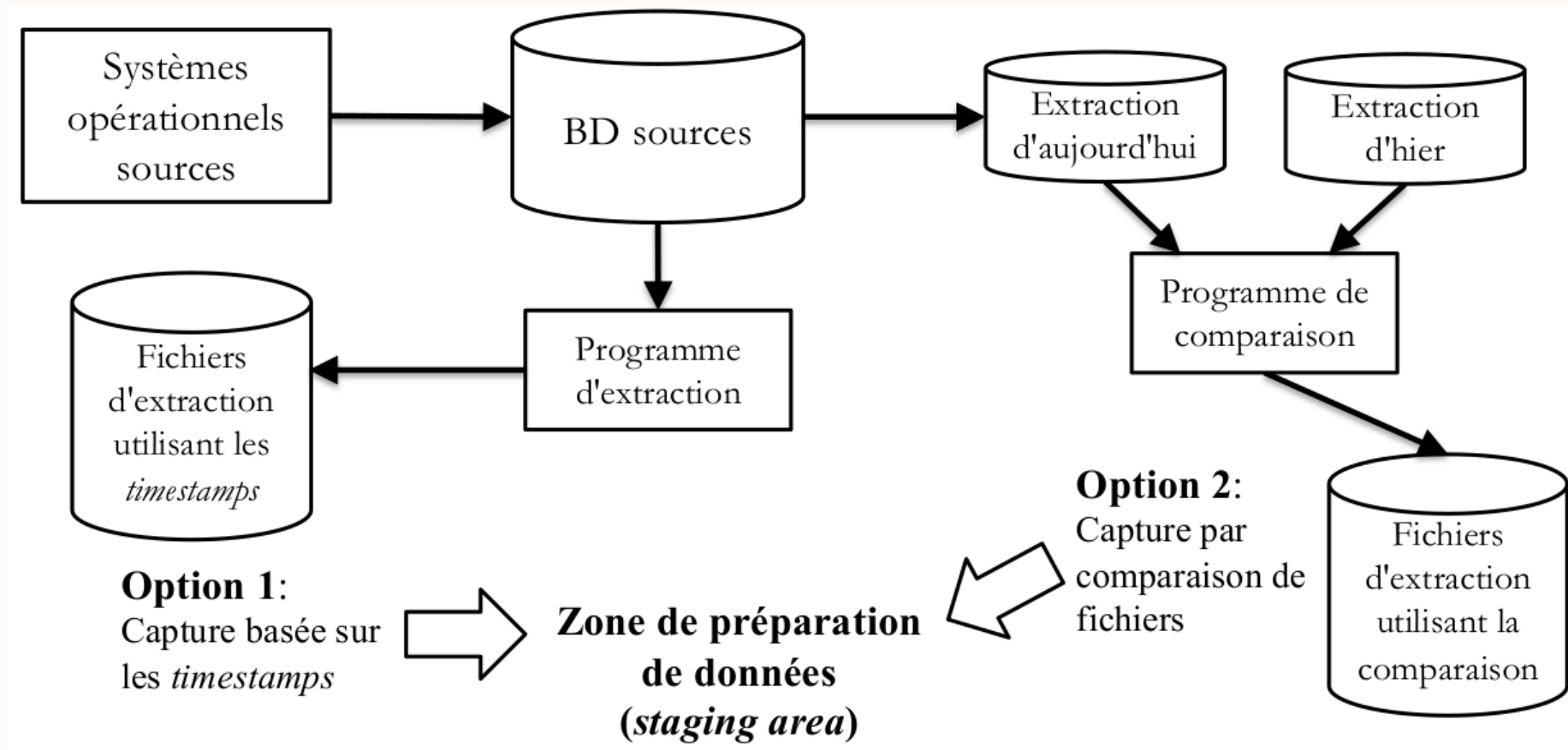
- **Extraire les données des sources hétérogènes (extract)**
 - Identifier les données sources utiles
 - Déterminer les données qui ont changé
- **Consolider les données (transform)**
 - Données redondantes, manquantes, incohérentes, etc.
 - Découpage, fusion, conversion, aggrégation, etc.
- **Charger les données intégrées dans l'entrepôt (load)**
 - Mode différé (batch) ou quasi temps-réel.
- Partie la plus longue du développement (jusqu'à 70% du temps total).

Le système de Data Warehouse et ses composants

Processus Extract, Transform, Load (ETL)

Extraction des données (différée)

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

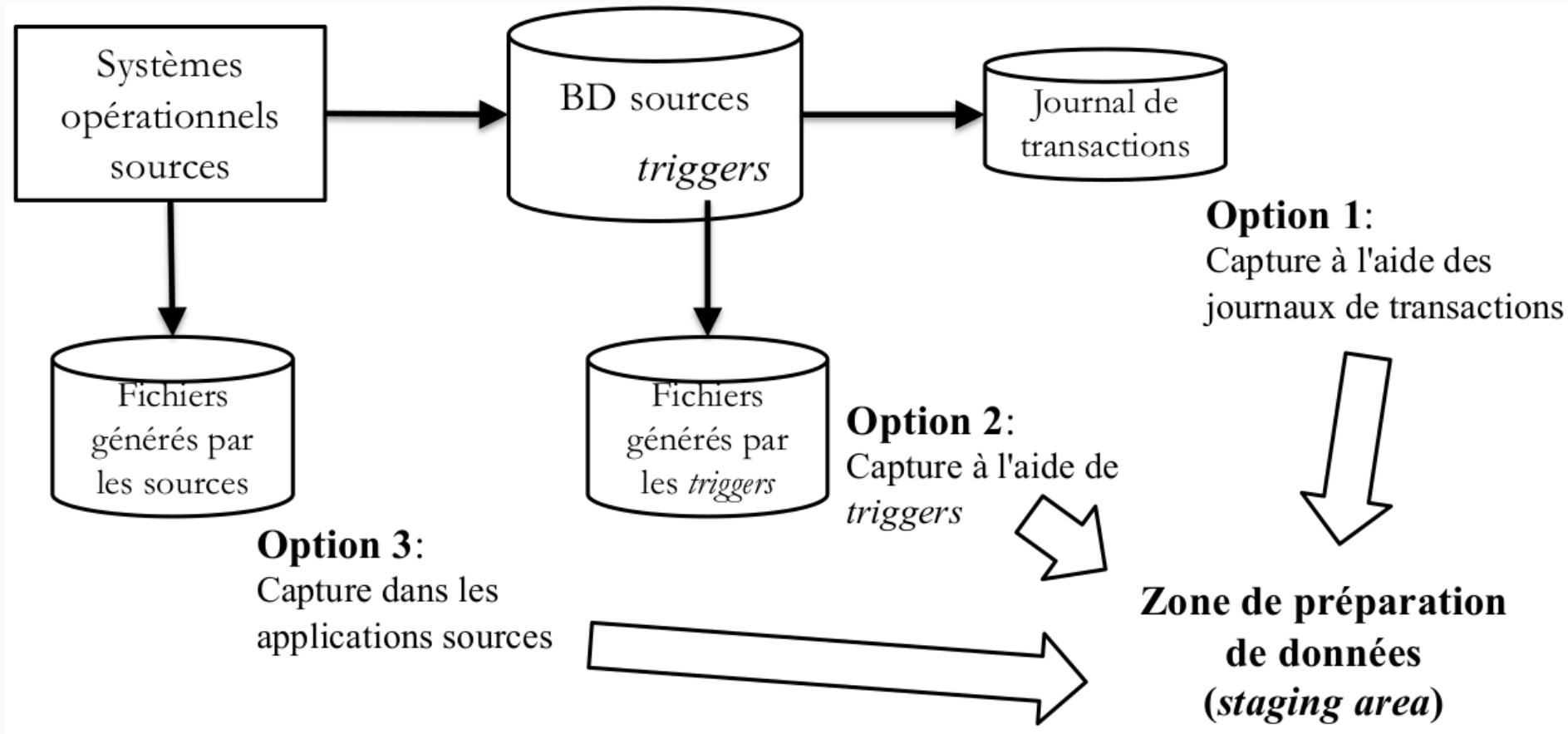


Le système de Data Warehouse et ses composants

Processus Extract, Transform, Load (ETL)

Extraction des données (temps-réel)

- S'effectue au moment où les transactions surviennent dans les systèmes sources.



Le système de Data Warehouse et ses composants

Processus Extract, Transform, Load (ETL)

Transformation des données

- Révision de format:
 - Ex: Changer le type ou la longueur de champs individuels.
- Décodage de champs:
 - Ex: ['homme', 'femme'] vs ['M', 'F'] vs [1,2].
- Pré-calcul des valeurs dérivées:
 - Ex: profit calculé à partir de ventes et coûts.
- Découpage de champs complexes:
 - Ex: extraire les valeurs prénom, secondPrénom et nomFamille à partir d'une seule chaîne de caractères nomComple.
- Pré-calcul des agrégations:
 - Ex: ventes par produit par semaine par région.
- Déduplication
 - Ex: Plusieurs enregistrements pour un même client

Le système de Data Warehouse et ses composants

Processus Extract, Transform, Load (ETL)

Chargement des données

- 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;
- Désactiver les indexes et clés étrangères lors du chargement.

Le système de Data Warehouse et ses composants

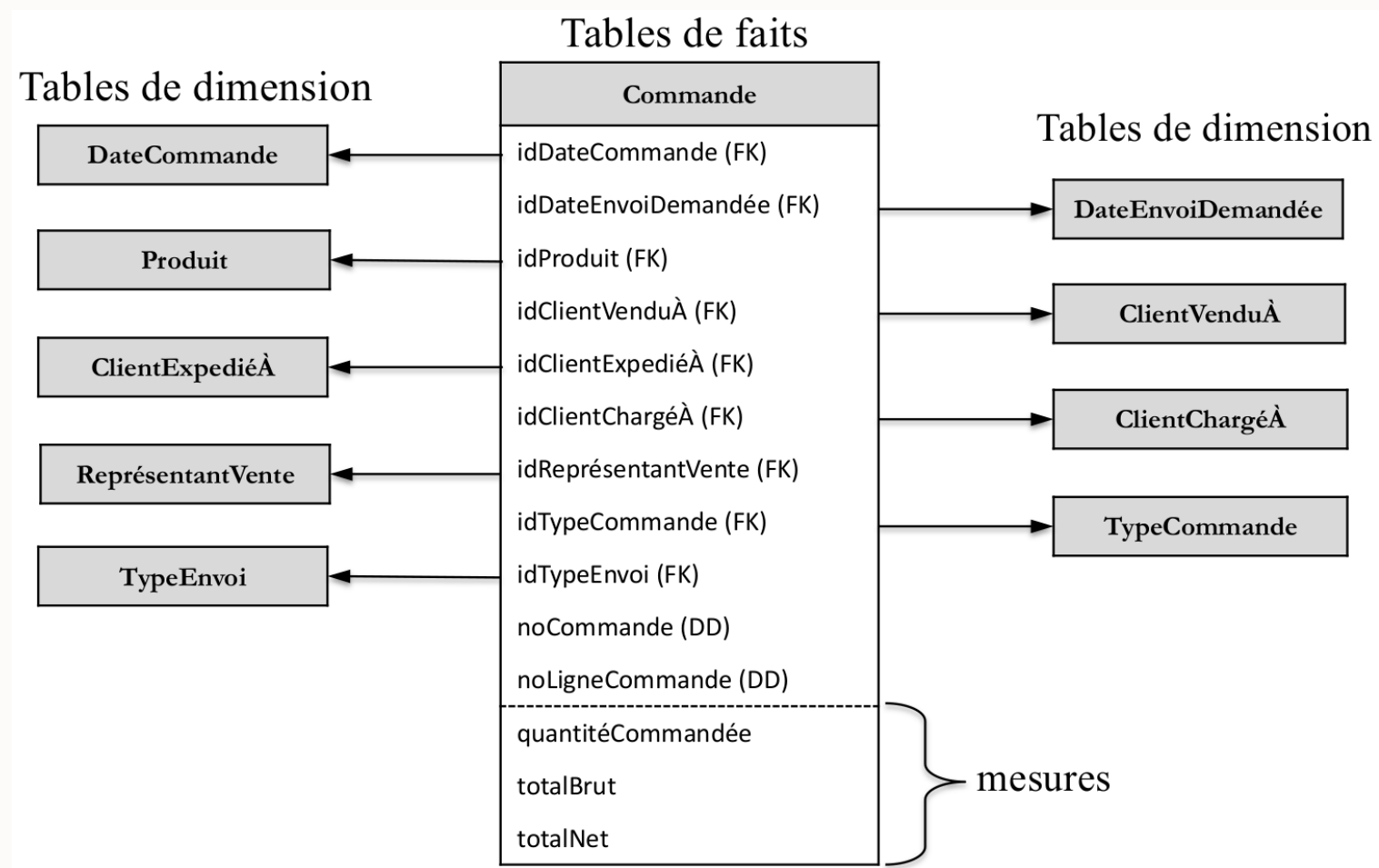
Processus Extract, Transform, Load (ETL)

Modélisation dimensionnelle

- Représente les données sous la forme d'un schéma en étoile:
 - Table de faits entourée de plusieurs tables de dimension (normalement entre 8 et 15)
- Les faits (mesures) sont généralement des valeurs numériques provenant des processus d'affaires;
- Les dimensions fournissent le contexte (qui, quoi, quand, où, pourquoi et comment) des faits;
- Les tables ne sont pas normalisées

Le système de Data Warehouse et ses composants

Exemple de schéma en étoile (Commande de produits)



Le système de Data Warehouse et ses composants

Exemple de schéma en étoile (Commande de produits)

Tables de dimension:

DateCommande	Produit	ClientExpédiéÀ
idDate (PK)	idProduit (PK)	idClient (PK)
date	description	nomFamille
jourDeSemaine	SKU	prénom
jourDuMois	marque	sexe
jourDeAnnée	sousCatégorie	dateNaissance
jourDansMoisFiscal	catégorie	dateAbonnement
jourDansAnnéeFiscale	département	forfaitAbonnement
congéFérié	poids	adresseRue
jourDeTravail	taille	adresseVille
semaineDuMois	couleur	adresseProvince
...

Tables de faits

- Correspondent à un événement d'affaires
 - Ex: achat d'un produit par un client, envoi du produit au client, commande de matériaux auprès d'un fournisseur, etc.
- Contiennent deux types de colonnes:
 - Des métriques associées à l'événement d'affaire:
 - Ex: total des ventes, nombre d'items commandés, etc.
 - Des clés étrangères vers les tables de dimension:
 - Ex: ID du client qui fait la commande, ID du produit commandé, etc.
- Contiennent typiquement un très grand nombre de lignes:
 - Jusqu'à plusieurs milliards de lignes;
 - Souvent plus de 90% des données du modèle.

Tables de dimension

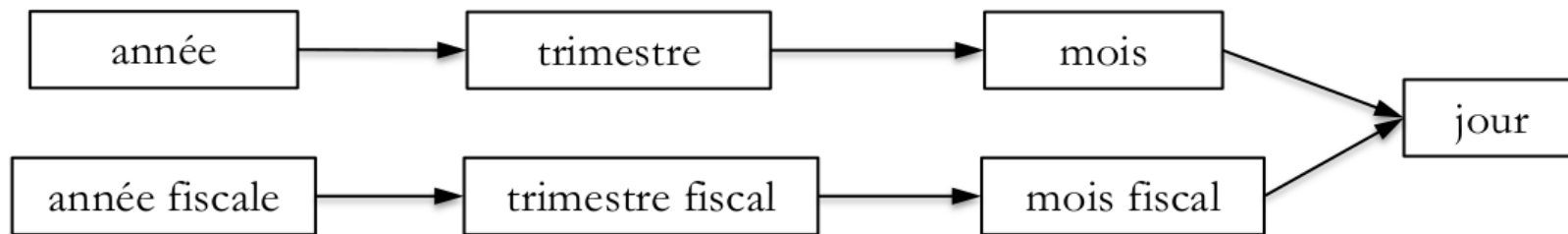
- Ensemble hautement corrélé d'attributs (jusqu'à plusieurs dizaines) regroupés selon les objets clés d'une entreprise:
 - Ex: produits, clients, employés, installations, etc.
- Propriétés des attributs:
 - Descriptif (ex: chaînes de caractères);
 - De qualité (ex: aucune valeur manquante, obsolète, erronée, etc.);
 - Valeurs discrètes (ex: jour, âge d'un client);
- Rôles des attributs:
 - Filtrer/agréger les données (ex: ville, catégorie produit, etc.);
 - Étiqueter les résultats (ex: champs descripteurs).

Le système de Data Warehouse et ses composants

Hiérarchies dimensionnelles

- Ensemble d'attributs d'une table de dimension ayant une relation hiérarchique (x est inclus dans y);
- Correspondent à des relations de type 1 à plusieurs;
- Définissent les chemins d'accès dans les données (drill-down paths);
- Peuvent être simples:
 - Produit : tous → catégorie → marque → produit;
 - Lieu : tous → pays → province → ville → code postal.

Ou multiples:



Le système de Data Warehouse et ses composants

Dimension temporelle

Dimension: Temps
idDate (PK)
date
jourDeSemaine
jourDuMois
jourDeAnnée
jourDansMoisFiscal
jourDansAnnéeFiscale
congéFérié
jourDeTravail
semaineDuMois
...

- Mettre toutes ces valeurs, même si la plupart peuvent être déduites d'une seule colonne;
- Pré-générer les lignes de la table (ex: 10 prochaines années) pour faciliter la référence et éviter les mises à jour

Le système de Data Warehouse et ses composants

Dimension temporelle

- Problème: avoir un grain trop fin dans la dimension temporelle (ex: temps du jour) peut causer l'explosion du nombre de rangées:
 - Ex: 31,000,000 secondes différentes dans une année.
- Solution: mettre le temps du jour dans une dimension séparée:
 - *Dimension Date*: année → mois → jour;
 - *Dimension TimeOfDay*: heure → minute → secondes;
 - 86,400 + 365 lignes au lieu de 31,000,000 lignes.
- Note: la dimension TimeOfDay est souvent modélisée comme un simple champs dans la table de faits.


Dimensions à évolution lente (SCD)

- Slowly Changing Dimensions (SCD);
- Même si elles sont plus statiques que les tables faits, les dimensions peuvent également changer:
 - Ex: adresse d'un client, catégorie d'un produit, etc.
- Stratégies d'historisation:
 - SCD Type 1: Écraser l'ancienne valeur avec la nouvelle
 - SCD Type 2: Ajouter une ligne dans la table de dimension pour la nouvelle valeur
 - SCD Type 3: Avoir deux colonnes dans la table de dimension correspondant à l'ancienne et la nouvelle valeur

Le système de Data Warehouse et ses composants

Dimensions à évolution lente (SCD)

Stratégie SCD Type 1



Product Key	Product Description	Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Education	ABC922-Z


Product Key	Product Description	Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Strategy	ABC922-Z

- Impossible de faire des analyses sur l'ancienne valeur;
- À utiliser seulement lorsque l'ancienne valeur n'est pas significative pour les besoins d'affaires;
- Exige de mettre à jour les données agrégées avec l'ancienne valeur.

Le système de Data Warehouse et ses composants

Dimensions à évolution lente (SCD)

Stratégie SCD Type 2

	Product Key	Product Description	Department	SKU Number (Natural Key)
	12345	IntelliKidz 1.0	Education	ABC922-Z
	25984	IntelliKidz 1.0	Strategy	ABC922-Z

- Permet de faire des analyses historiques;
- Demande l'ajout d'une nouvelle ligne par changement;
- À utiliser lorsque l'ancienne valeur a une signification analytique ou si le changement est une information en soi.

Le système de Data Warehouse et ses composants

Dimensions à évolution lente (SCD)

Stratégie SCD Type 3

Product Key	Product Description	Department	Prior Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Strategy	➔ Education	ABC922-Z

- Rarement employée;
- Profondeur de l'historique est de un seul changement;
- Utilisé lorsqu'on veut vouloir comparer les faits avec l'ancienne ou la nouvelle valeur;

Types d'entrepôts de données

1. Magasins de données
2. Entrepôts de données d'entreprise (EDW)
 - Bus de magasins de données (datamart bus)
 - Hub-and-spokes
 - Entrepôts de données fédérés

Le système de Data Warehouse et ses composants

Types d'entrepôts de données

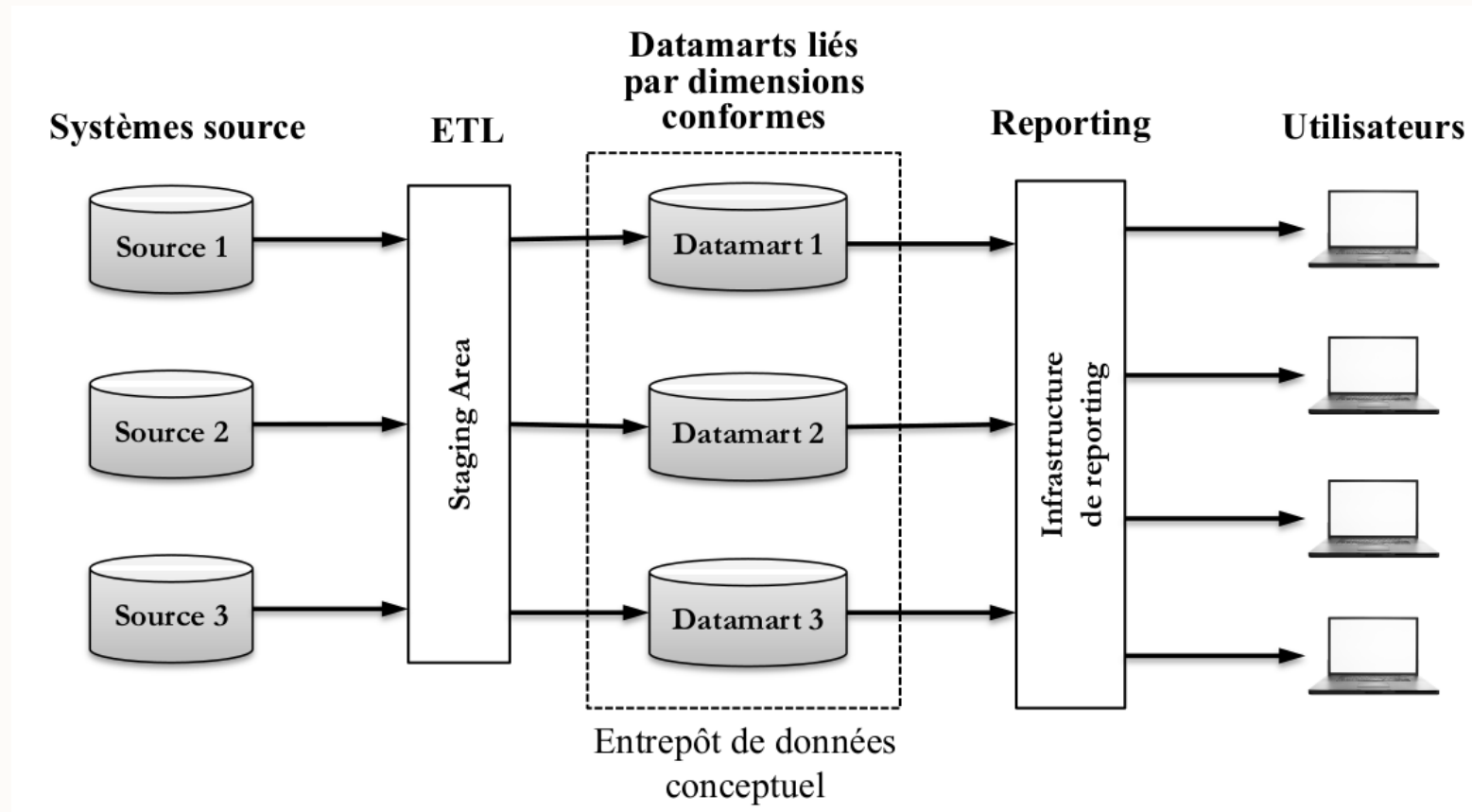
Magasins de données (datamart)

- Contiennent une portion du contenu de l'entrepôt de données;
- Se concentre sur un seul sujet d'analyse (ex: les ventes OU l'inventaire, mais pas les deux);
- Servent à faire des analyses simples et spécialisées (ex: fluctuations des ventes par catégorie de produits);
- Nombre de sources limitées, provenant la plupart du temps d'un même département;
- Modélisés sous la forme d'un schéma en étoile.

Le système de Data Warehouse et ses composants

Types d'entrepôts de données

Architecture Datamart bus



Le système de Data Warehouse et ses composants

Types d'entrepôts de données

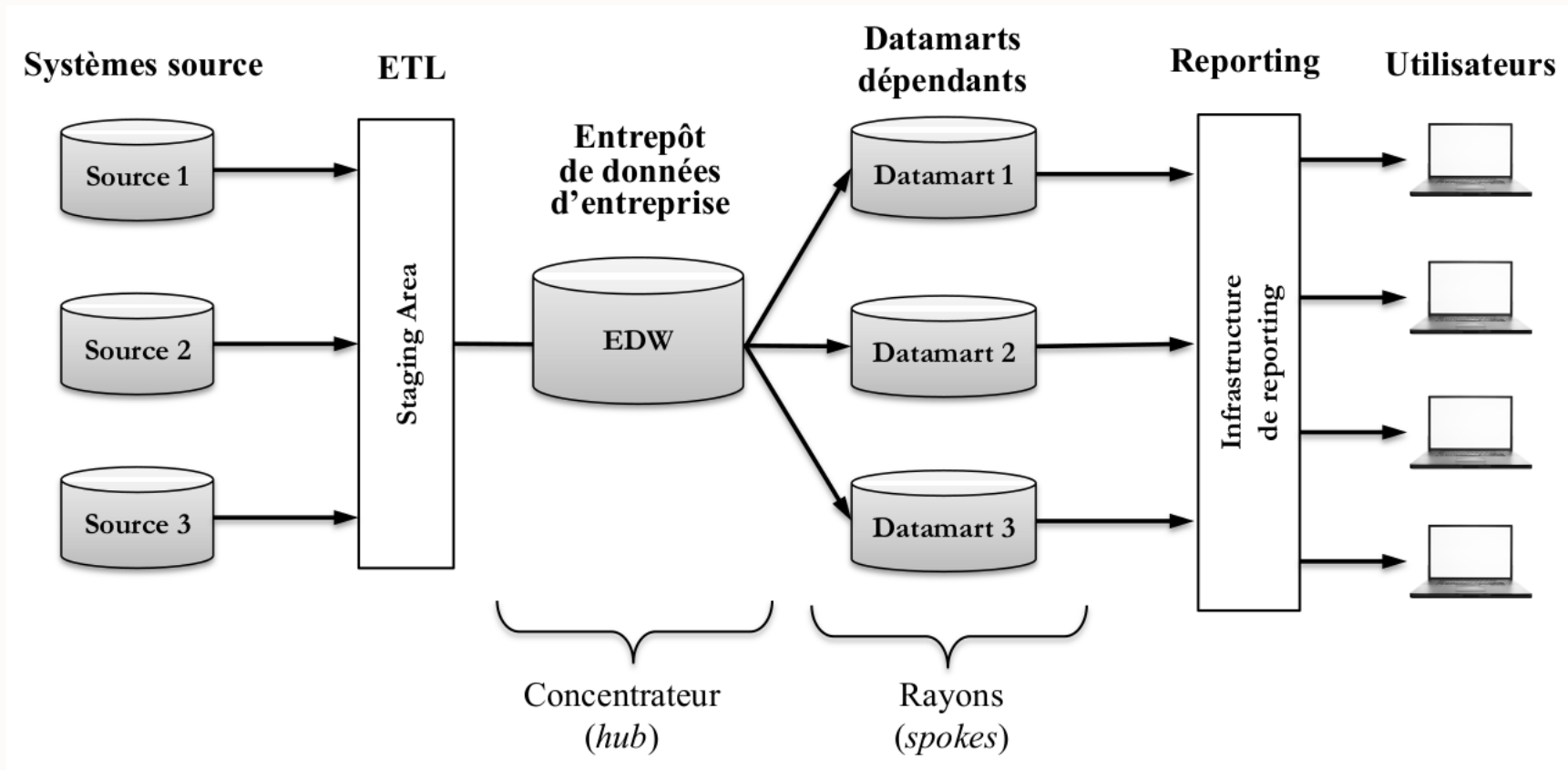
Architecture Datamart bus

- Approche bottom-up, où on construit l'entrepôt un datamart à la fois;
- Modélisation dimensionnelle (schéma en étoile) des datamarts, au lieu du diagramme entité-relation;
- Entrepôt de données conceptuel, formé de magasins de données inter-reliés à l'aide d'une couche d'intergiciels (middleware).
- Intégration des données assurée par les dimensions partagées entre les datamarts (i.e., dimensions conformes);
- Approche incrémentale qui donne des résultats rapidement (développement agile);

Le système de Data Warehouse et ses composants

Types d'entrepôts de données

Architecture Hub-and-spoke



Le système de Data Warehouse et ses composants

Types d'entrepôts de données

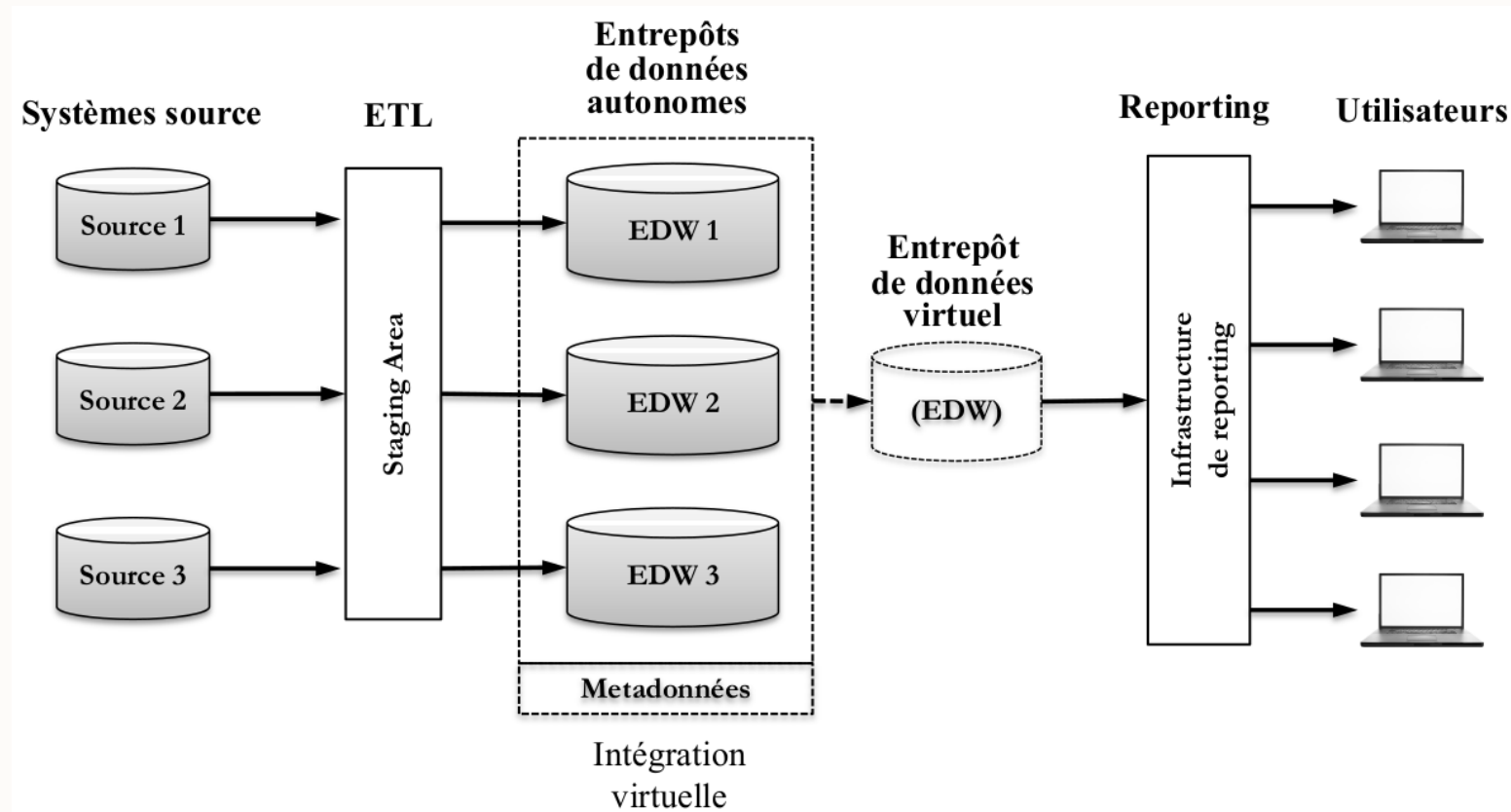
Architecture Hub-and-spoke

- Approche top-down, favorisant l'intégration et consolidation complète des données de l'entreprise
- Entrepôt (hub) contient les données atomiques (niveau de détail le plus fin) et normalisées (3FN);
- Les datamarts (spokes) contiennent principalement des données agrégées (pas atomique) et suivant le modèle dimensionnel;
- La plupart des requêtes analytiques sont faites sur les datamarts;
- Développement plus long, dû à la complexité du processus ETL et de la modélisation;
- Meilleure qualité de données que l'architecture par bus de datamarts.

Le système de Data Warehouse et ses composants

Types d'entrepôts de données

Architecture fédérée



Le système de Data Warehouse et ses composants

Types d'entrepôts de données

Architecture fédérée

- Entrepôt de données distribué sur plusieurs systèmes hétérogènes;
- Données intégrées logiquement ou physiquement à l'aide de méta-données (ex: XML);
- Opère de manière transparente (l'utilisateur ne voit pas que les données sont réparties);
- Utile lorsqu'il y a déjà un entrepôt en place (ex: acquisitions ou fusions de compagnies);
- Très complexe (synchronisation, parallélisme, concurrence, etc.) et faible performance.

FIN DE SEANCE

Résumé

1. Faites résumé détaillé sur l'architecture de Data Warehouse
2. Que-ce que vous avez compris des tables de dimensions et de faits ?
3. Faites un résumé sur les types d'un Data Warehouse