eFrei
PARIS PANTHÉON-ASSAS UNIVERSITÉ

Data Warehouse

Grp. LSI 1 – LSI 2

Khadija SLIMANI
Consultante data scientist
Enseignante chercheuse
✉ : khadija.slimani@qassil.com

Data Warehouses et bases de données décisionnelles

Plan

1. Décisionnel
2. Data warehousing
3. Disciplines participant au Big Data
4. Différence entre un DW et un système transactionnel
5. Implémentation du DW avec un SGBDR
6. Modélisation logique de données
7. Extraction, Transformation, Loading
8. L'analyse multidimentionnelle
9. Réalisation d'un Data Warehouse
10. Principales applications autour d'un ED

Objectifs

- ✓ Comprendre ce qu'est et à quoi sert un data warehouse.
- ✓ Comprendre les différences entre un data warehouse et une base de données transactionnelle.

I. Décisionnel

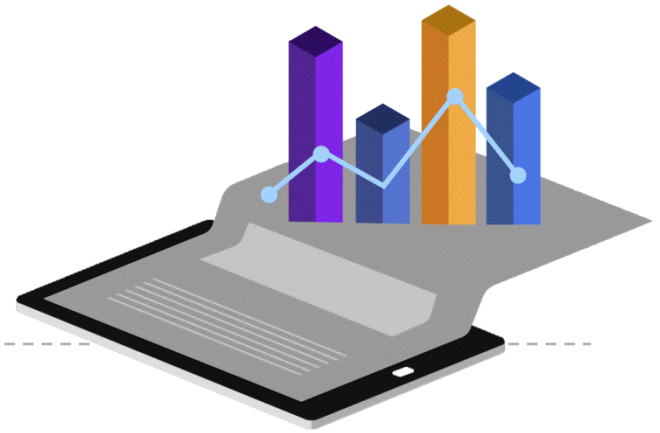
"Le système d'information décisionnel est un ensemble de données organisées de façon spécifiques, **facilement accessibles** et appropriées à la **prise de décision** [...].

La finalité d'un système décisionnel est le pilotage d'entreprise.

Les systèmes de gestion sont dédiés aux **métiers** de l'entreprise [...].

Les systèmes décisionnels sont dédiés au **management** de l'entreprise [...]. " (Goglin, 2001, pp21-22)

Synonymes : informatique décisionnelle, *business intelligence, BI*



2. Data warehousing

Un data warehouse (DW) est une base de données construite par copie et réorganisation de multiples sources, afin de servir de source de données à des applications décisionnelles :

- ✓ il agrège de nombreuses données de l'entreprise (intégration) ;
- ✓ il mémorise les données dans le temps (historisation) ;
- ✓ il les organise pour faciliter les requêtes de prise de décision (optimisation).

(Goglin, 2001, p27) [(Goglin, 2001)]

Synonymes : entrepôt de données, base de données décisionnelle

2. Data warehousing

L'objectif du data warehouse est de **permettre des requêtes** sur de **grands ensembles des données**,

la plupart du temps sous forme d'agrégats afin d'en obtenir une vision synthétique.

3. Disciplines participant au Big Data

- Big Data est un domaine pluridisciplinaire pour lequel on peut identifier 5 parties : On peut tout d'abord énumérer quatre parties clés :

  1. Une partie **Math-Info** comprend tout d'abord des mathématiques statistiques et probabilistes sur lesquelles sont fondés des algorithmes d'apprentissage numérique (ML), ainsi que des algorithmes de fouille de données et de graphes. Cette partie du Big Data est celle qui est souvent identifiée comme le **Data Science**.

3. Disciplines participant au Big Data

- Big Data est un domaine pluridisciplinaire pour lequel on peut identifier 5 parties : On peut tout d'abord énumérer quatre parties clés :

2. Une partie **d'informatique distribué** pour l'analyse de données large échelle. Il s'agit d'une forme d'algorithmique distribuée récente, visant à amener les traitements sur les machines où sont stockées les données. Cette approche permet des traitements de données à large échelle, voire à l'échelle du web. Une première mise en œuvre de cette approche utilisait le schéma **Map-Reduce** : un **schéma de calcul distribué** à première vue très contraint mais en fait assez générique.

3. Disciplines participant au Big Data

- Big Data est un domaine pluridisciplinaire pour lequel on peut identifier 5 parties : On peut tout d'abord énumérer quatre parties clés :

3. Une partie **d'informatique parallèle** à haute performance pour le data analytics et le machine learning visant à accélérer les calculs sur des machines parallèles. Par exemple, en utilisant un cluster de PC multi-cœurs ou un cluster de GPU, pour entraîner des réseaux de neurones profonds (DL).

3. Disciplines participant au Big Data

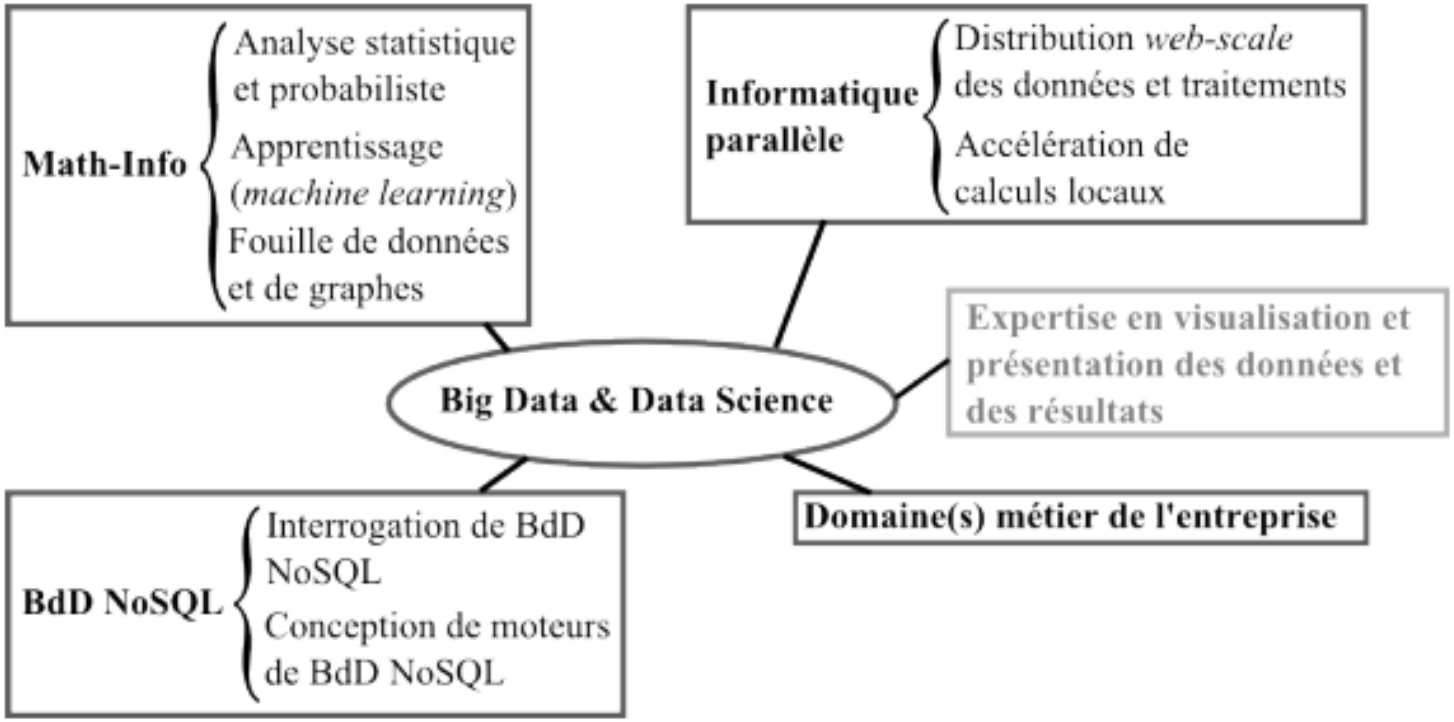
- Big Data est un domaine pluridisciplinaire pour lequel on peut identifier 5 parties : On peut tout d'abord énumérer quatre parties clés :

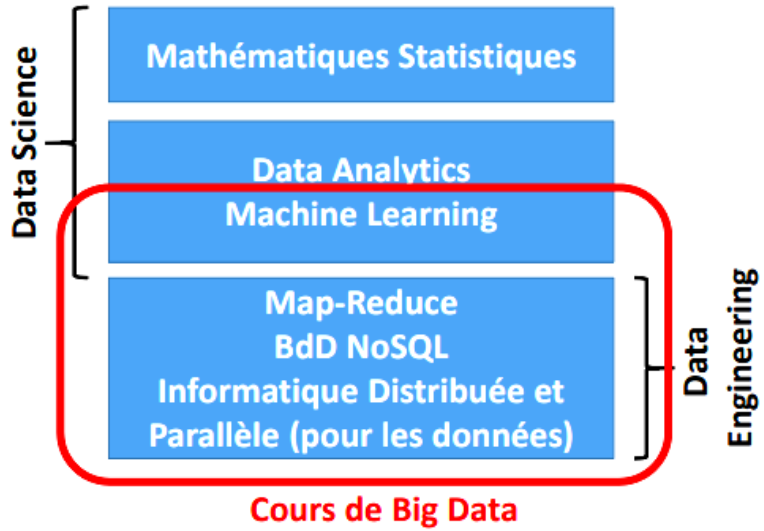
4. Une autre partie essentielle du Big Data réside dans la **conception et l'exploitation de bases de données** "not only SQL" (NoSQL). Elles permettent de stocker des données structurées complexes, ou au contraire de simples fichiers textes que l'on devra analyser en détail. Certaines BdD NoSQL ont été conçues pour un stockage distribué à très large échelle, d'autres pour favoriser la vitesse d'interrogation sur des données plus restreintes. Le domaine des BdD NoSQL est encore en pleine évolution.

3. Disciplines participant au Big Data

- Big Data est un domaine pluridisciplinaire pour lequel on peut identifier 5 parties : deux autres parties complètent l'aspect pluridisciplinaire du Big Data :
  - le **domaine applicatif considéré** : une connaissance du domaine d'activité de l'entreprise est nécessaire pour que le data scientist puisse donner du sens aux données, guider son analyse et interpréter les résultats de ses algorithmes.
  - Enfin, la **visualisation et présentation des données et résultats** : le data scientist doit aussi posséder une expertise en visualisation de gros volumes de données et en présentation synthétique/simplifiée des résultats. Cette facette de ces compétences et activités est essentielle pour aboutir à une prise de décision dans un contexte industriel.

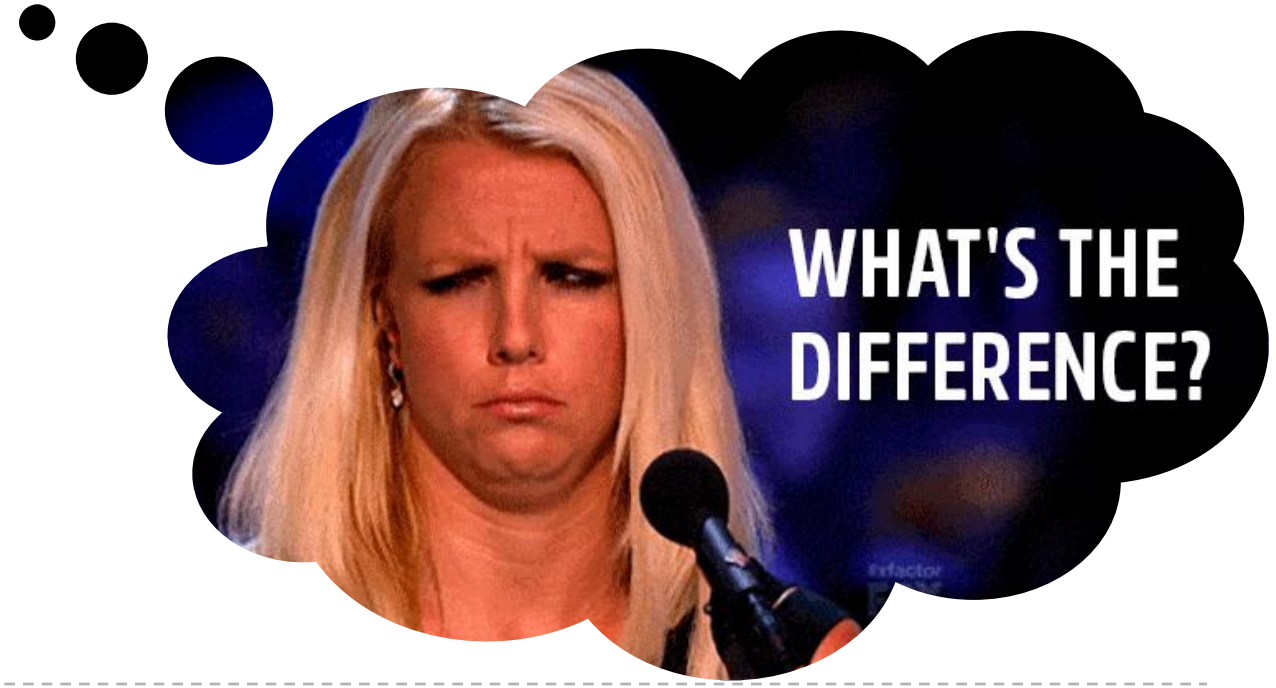
3. Disciplines participant au Big Data





4. Difference entre DW et un système transactionnel

Le **data warehouse** dédié au décisionnel est séparé du **système transactionnel qui est** dédié à la gestion quotidienne.



4. Difference entre DW et un système transactionnel

❑ BD transactionnelle

Une base données classique est destinée à assumer des transactions en temps réel :

- Ajout, mise à jour, suppression de données
- Questions sur des données identifiées ou questions statistiques

❑ Data Warehouse

Un DW est uniquement destiné à l'exécution de questions statistiques sur des données statiques (ou faiblement dynamiques).

5. Implementation du DW avec un SGBDR

Les deux problématiques fondamentales des DW sont **l'optimisation** et la **simplification**.

Comment rester **performant** et **lisible** avec de très **gros volumes de données** et des requêtes portant sur de nombreuses tables ???

5. Implementation du DW avec un SGBDR

On utilise massivement :

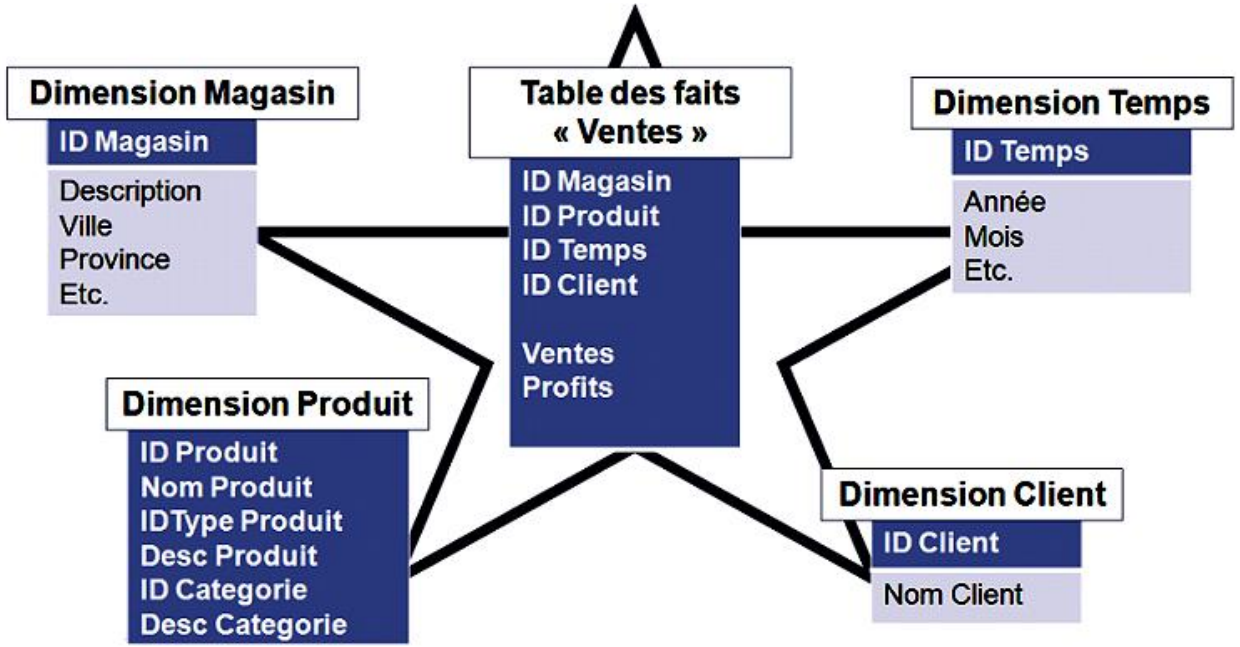
- ✓ **Les vues concrètes** : Un data warehouse procède par copie depuis le ou les systèmes transactionnels
- ✓ **La dénormalisation** : Un data warehouse est hautement redondant

Le caractère **statique** du data warehouse **efface** les inconvénients de ces techniques lorsqu'elles sont mobilisées dans des systèmes transactionnels.

6. Modélisation logique de données

Le modèle en étoile

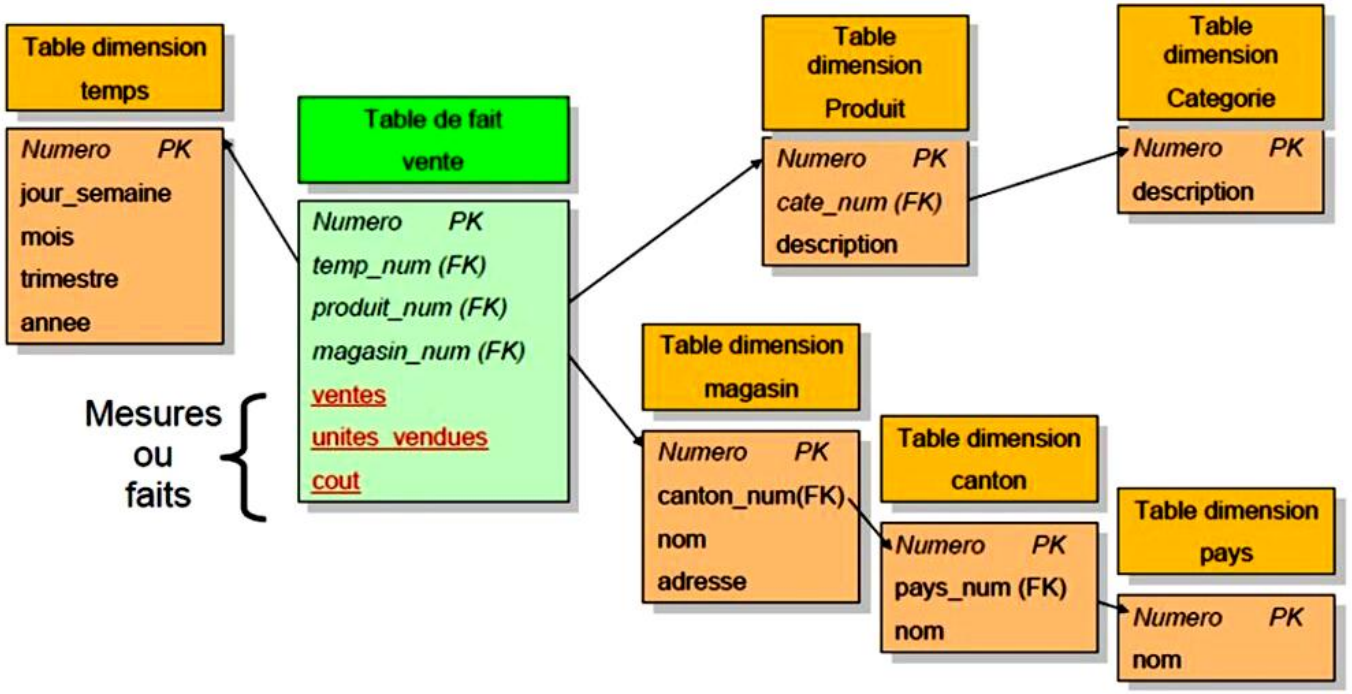
Le modèle en étoile est une représentation fortement **dénormalisée** qui assure un haut niveau de performance des requêtes même sur de gros volumes de données.



6. Modélisation logique de données

Le modèle en flocon

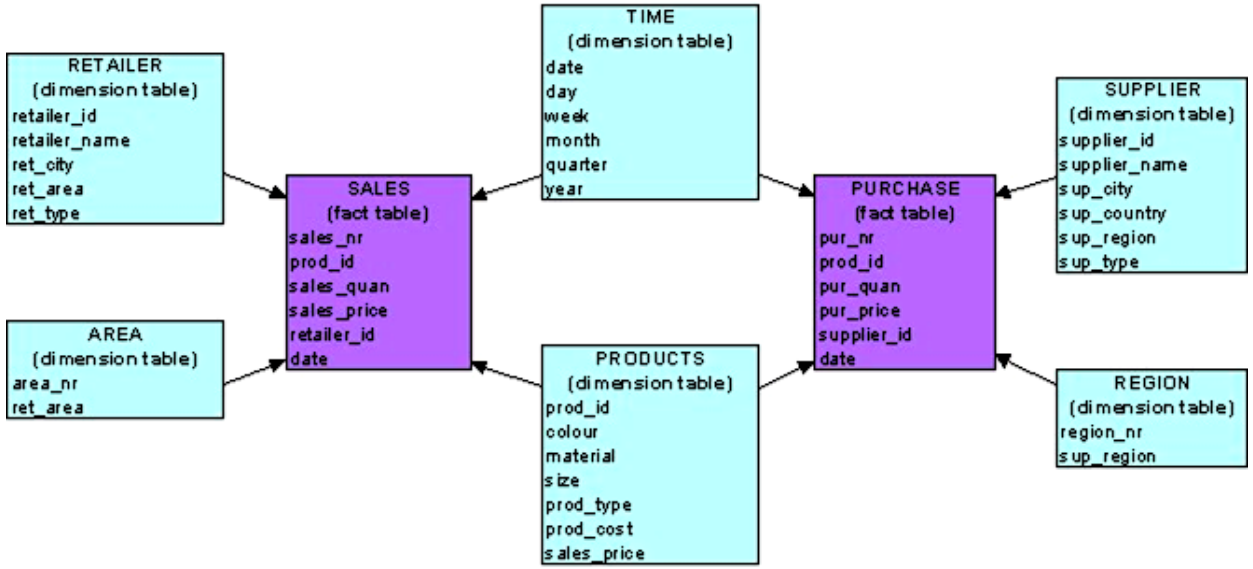
Le modèle en flocon est aussi un modèle **dénormalisée**, mais un peu moins que le modèle en étoile : il conserve un certain niveau de décomposition pour chaque dimension prise isolément.



6. Modélisation logique de données

Le modèle en constellation

Fact Constellation est un schéma de représentation de modèle multidimensionnel. Il s'agit d'une **collection de plusieurs tables de faits** ayant des tables de **dimensions communes**. Il peut être considéré comme une collection de plusieurs schémas en étoile et donc également connu sous le nom de **schéma Galaxy**.



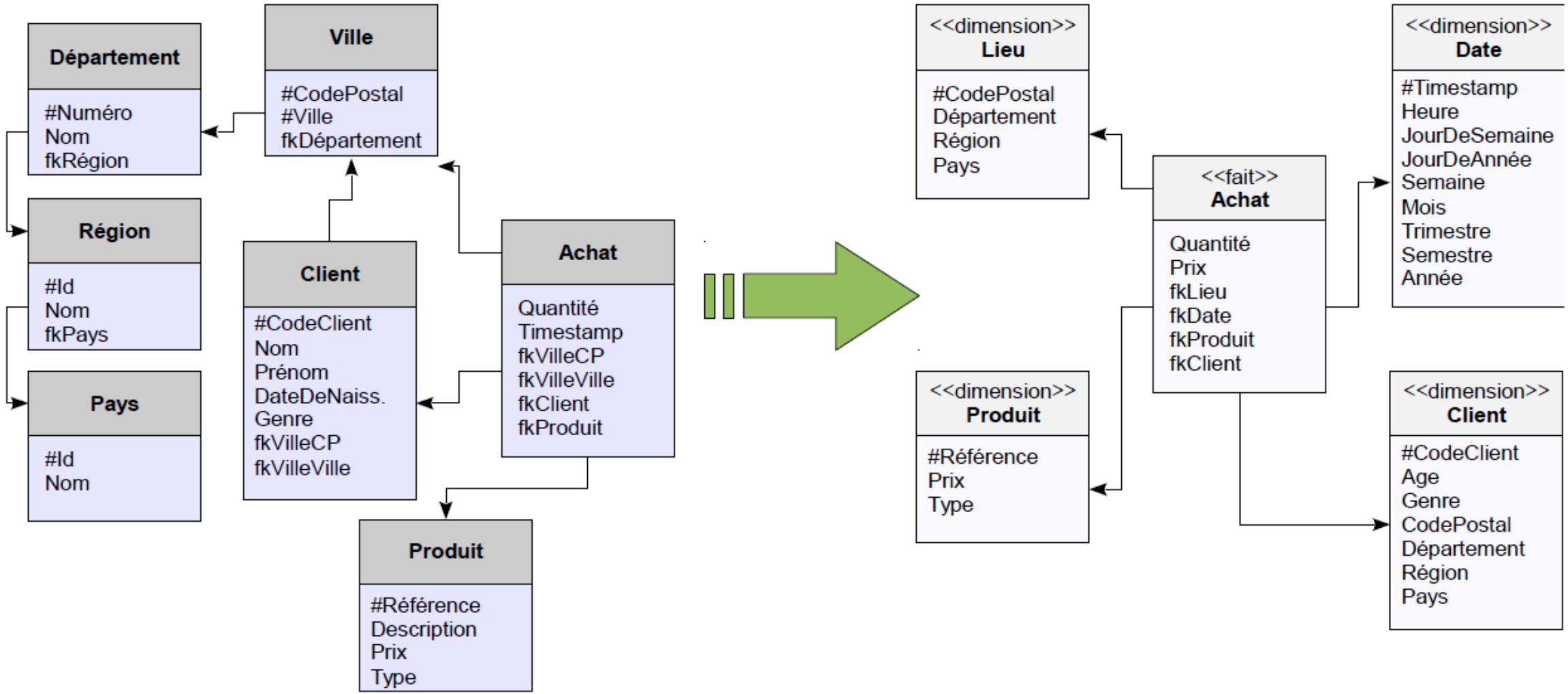
7. Extraction Transformation Loading

- **Définition ETL**

L'ETL (Extraction Transformation Loading) est le processus de copie des données depuis les tables des systèmes transactionnels vers les tables du modèle en étoile du data warehouse.

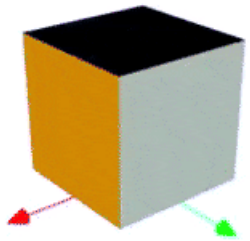
Les tables du modèle dimensionnel peuvent être vues comme des vues concrètes sur le systèmes transactionnel, à la nuance que des transformations (correction d'erreur, extrapolation...) peuvent avoir été apportées dans le processus ETL.

7. Extraction Transformation Loading



L'ANALYSE MULTIDIMENSIONNELLE

8. L'analyse multidimentionnelle



❑ Objectif

- ✓ obtenir des informations déjà agrégées selon les besoins de l'utilisateur: simplicité et rapidité d'accès.
- ✓ Modélisation multidimensionnelle des données facilitant l'analyse d'une quantité selon différentes dimensions :
  - Temps,
  - Localisation géographique,
  - …

❑ **HyperCube OLAP :** représentation de l'information dans un hypercube à N dimensions

8. L'analyse multidimentionnelle

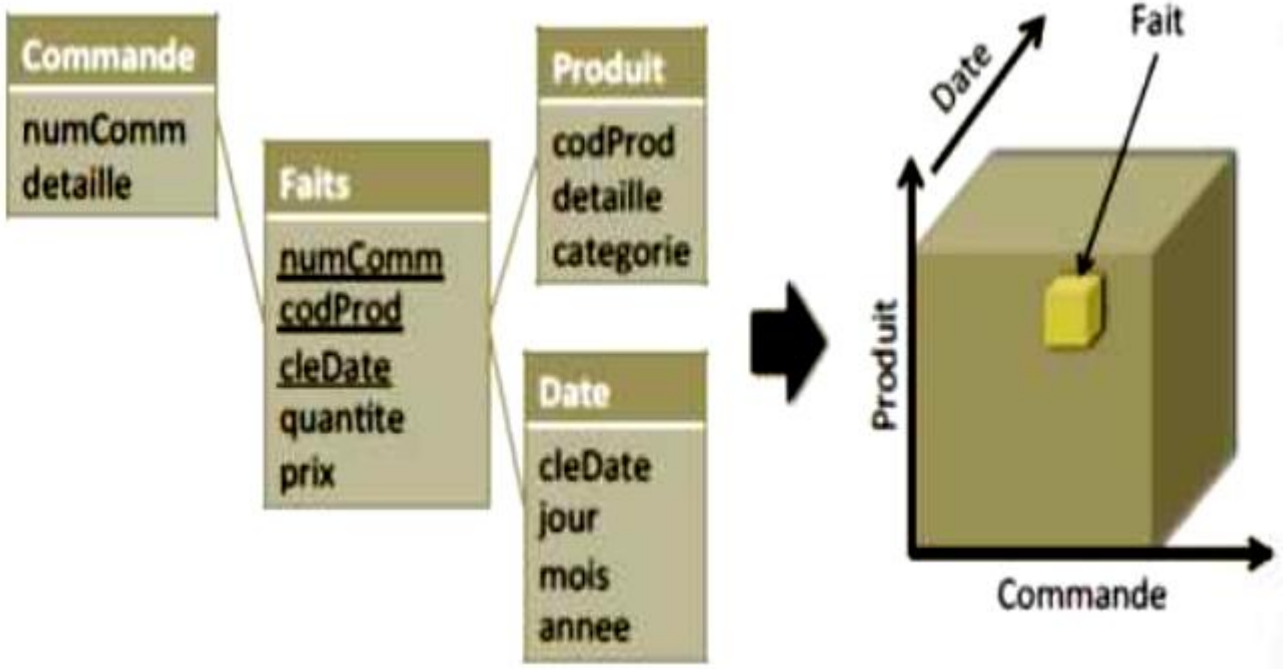
❑ **OLAP (OnLine Analytical Processing) :**

- ✓ Fonctionnalités qui servent à faciliter l'analyse multidimensionnelle : opérations réalisables sur l'hypercube…

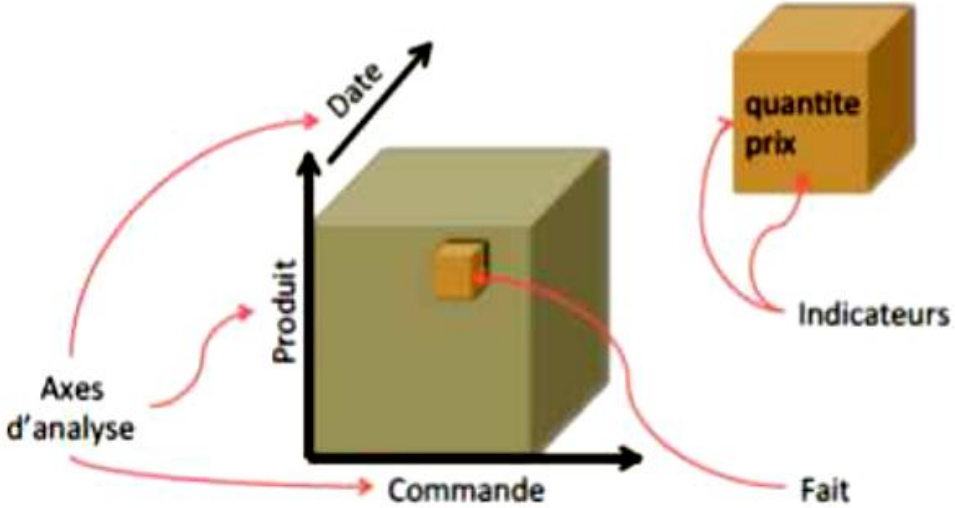
❑**Les calculs sont** réalisés lors du **chargement ou de la mise à jour du cube.**

8. L'analyse multidimentionnelle

❑ Cube de données



8. L'analyse multidimentionnelle

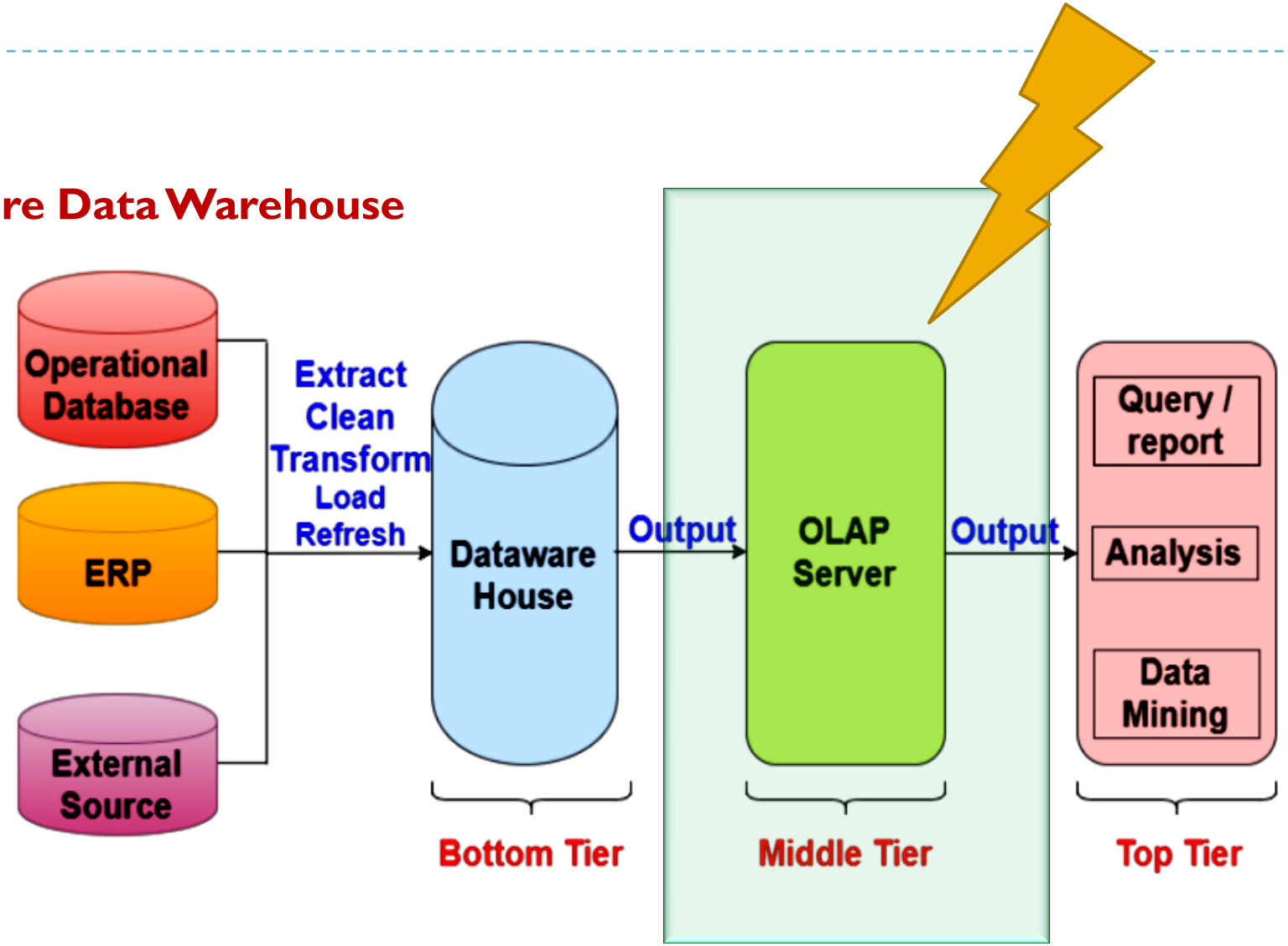


❑ Composantes d'un cube

- Chaque **cellule** du cube correspond a une occurrence du fait.
- Chaque cellule contient des **indicateurs** (variables, métriques ou mesures)
- Les axes d'analyse, également appelés **dimensions**, contiennent un ensemble de valeurs
- Des **hiérarchies** sont spécifiées sur les dimensions afin de Permettre une consolidation des indicateurs
- Chaque indicateur a une **fonction d'agrégat** afin d'être exploité sur la hiérarchie.

8. L'analyse multidimentionnelle

❑ Architecture Data Warehouse



8. L'analyse multidimentionnelle

❑ L'implémentation du OLAP

- Description de la base multidimensionnelle suivant la technologie utilisée :
  - ROLAP (Relational-OLAP)
  - MOLAP (Multidimensional-OLAP)
  - HOLAP (Hybrid-OLAP)

8. L'analyse multidimentionnelle

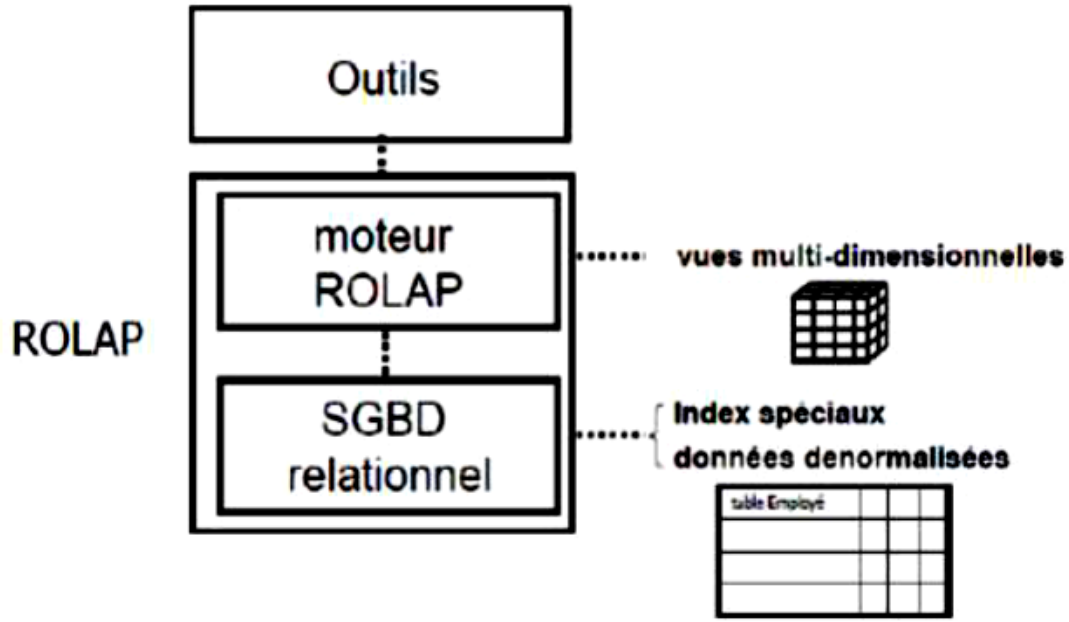
❑ L'implémentation du OLAP

- **ROLAP (Relational-OLAP)**
  - Les données sont stockées dans une BD relationnelle.
  - Le cube est stocké selon le modèle en étoile (flocon ou constellation).
  - Un moteur OLAP permet de simuler le comportement d'un SGBD multidimensionnel.

8. L'analyse multidimentionnelle

❑ L'implémentation du OLAP

- **ROLAP (Relational-OLAP)**
  - **Avantages :**
    - Facile à mettre en place
    - Peu couteux
    - Evolution facile
    - Stockage de gros volumes
  - **Inconvénients :**
    - Moins performant lors des phases de calculs



8. L'analyse multidimentionnelle

❑ L'implémentation du OLAP

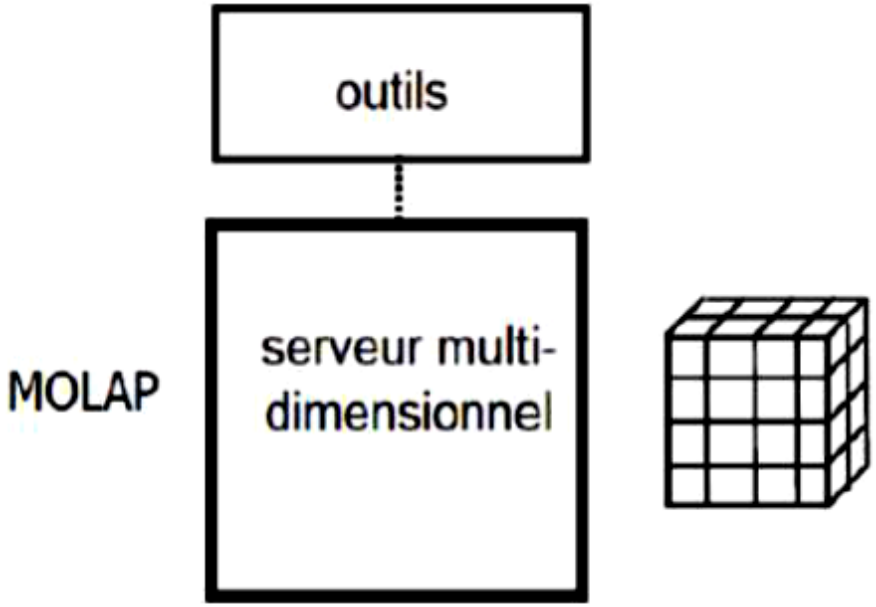
- **MOLAP (Multidimensional-OLAP)**
  - Les données sont stockées comme des matrices à plusieurs dimensions :

    Cube[1:m,1:n,1:p](mesure)
    - On trouve en colonne tous les axes, puis tous les indicateurs.
    - Chaque cellule du cube est stockée par une ligne dans la matrice.
    - Accès direct aux données dans le cube.

8. L'analyse multidimentionnelle

❑ L'implémentation du OLAP

- **MOLAP (Multidimensional-OLAP)**
  - **Avantages :**
    - Rapidité
  - **Inconvénients :**
    - Difficile à mettre en place
    - Formats souvent propriétaires
    - Ne supporte pas de très gros volumes de donnée



8. L'analyse multidimentionnelle

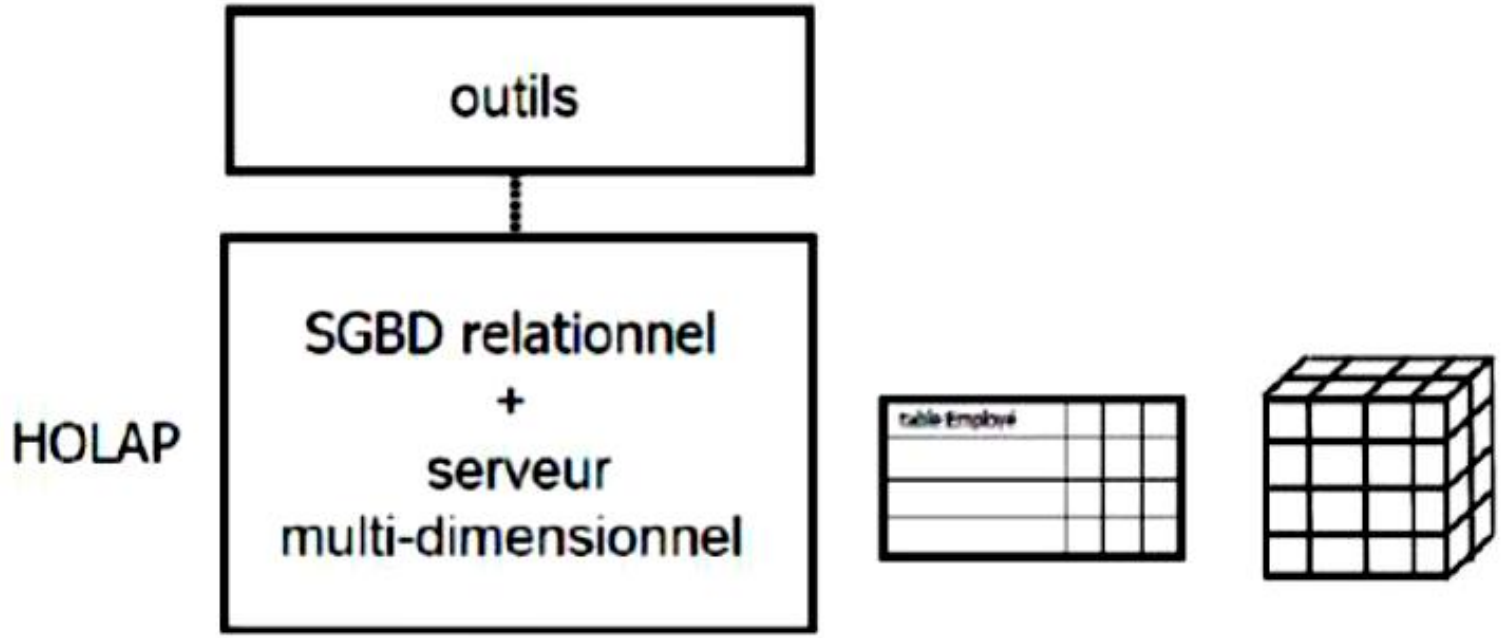
❑ L'implémentation du OLAP

- **HOLAP (Hybrid-OLAP)**
  - Solution hybride entre ROLAP et MOLAP.
  - Données de base stockées dans un SGBD relationnel (tables de faits et de dimensions) + données agrégées stockées dans un cube.
  - Avantages / inconvénients :
    - Bon compromis au niveau des coûts et des performances (les requêtes vont chercher les données dans les tables et le cube)

8. L'analyse multidimentionnelle

❑ L'implémentation du OLAP

- **HOLAP (Hybrid-OLAP)**



Réalisation d’un Data Warehouse

Réalisation d’un Data Warehouse

Les Techniques pour mettre en place un DW

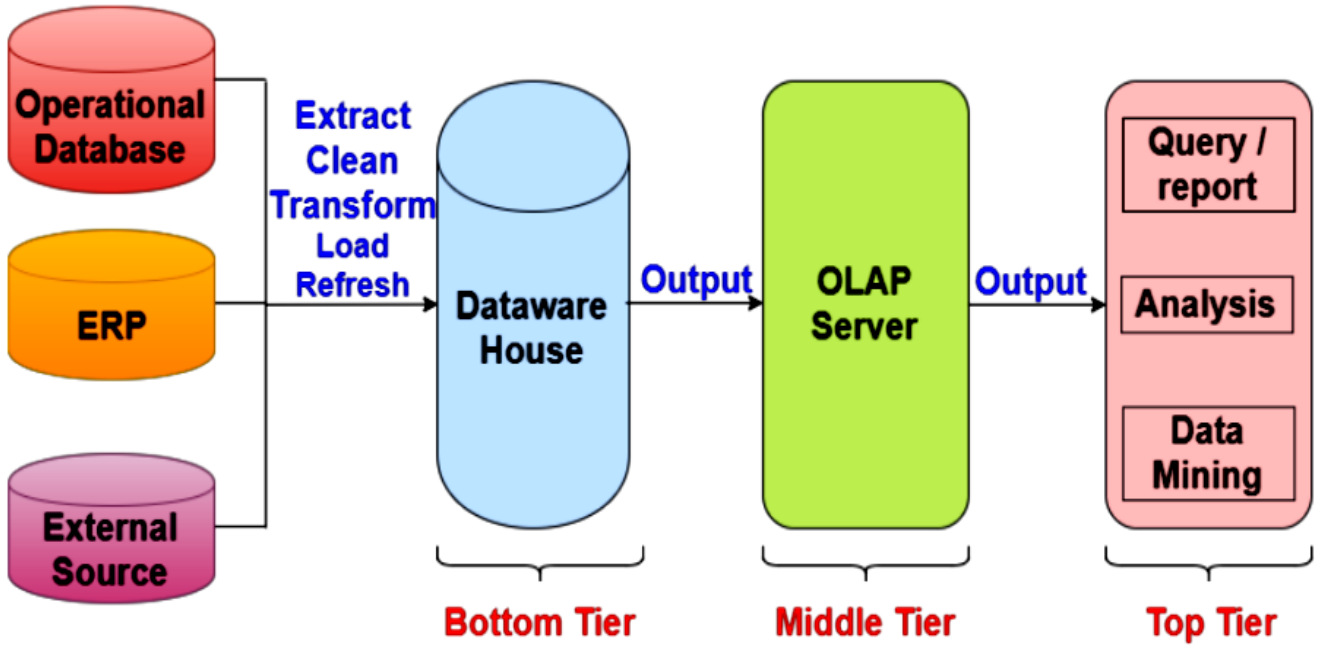
9. Réalisation d'un Data Warehouse

- Evolution des besoins et des sources

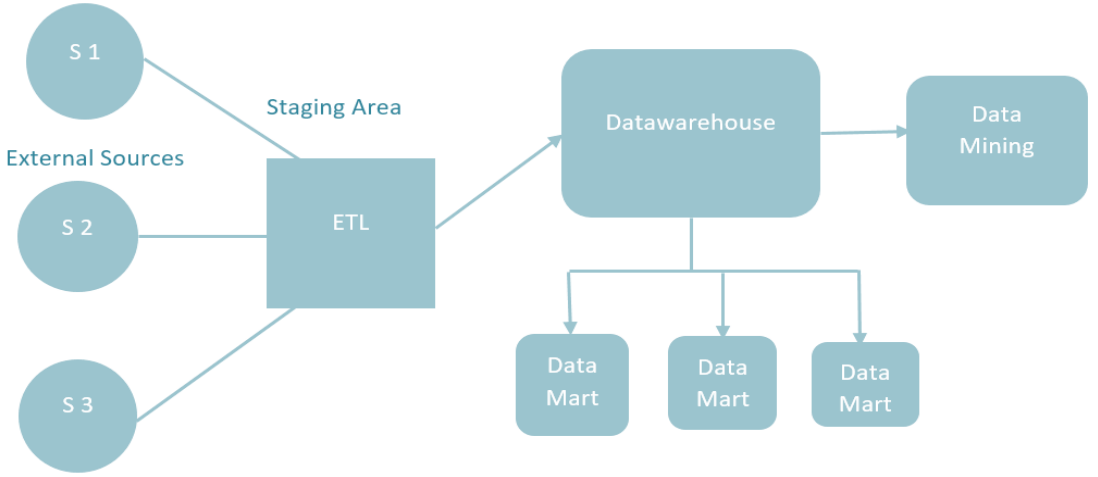
  → démarche itérative

- 3 techniques :

  1. Top-down
  2. Bottom-up
  3. Middle-out



9. Réalisation d'un Data Warehouse

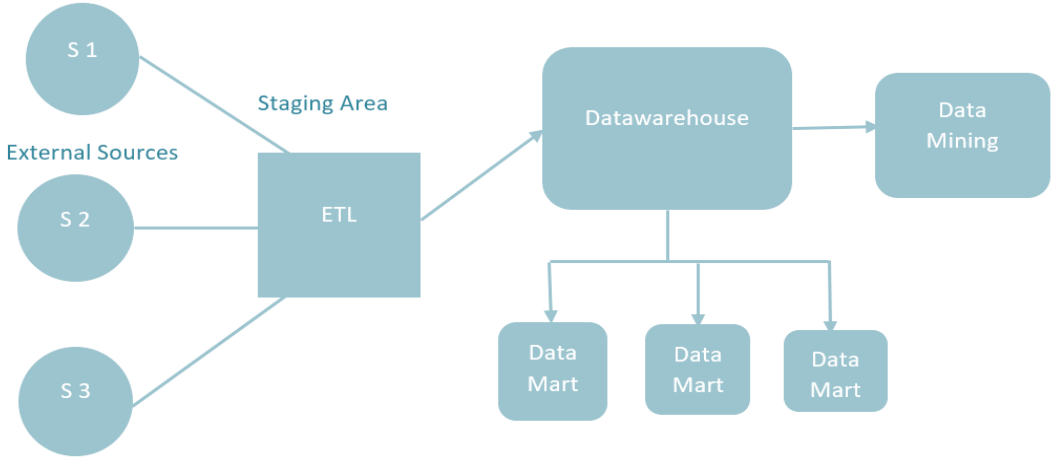


❑ **Trois techniques pour réaliser un DW**

1. Top-down

- Concevoir tout l'entrepôt intégralement : Il faut donc connaître à l'avance toutes les dimensions et tous les faits.
- **Objectif** : Livrer une solution technologiquement saine basée sur des méthodes et technologies éprouvées des bases de données.

9. Réalisation d'un Data Warehouse

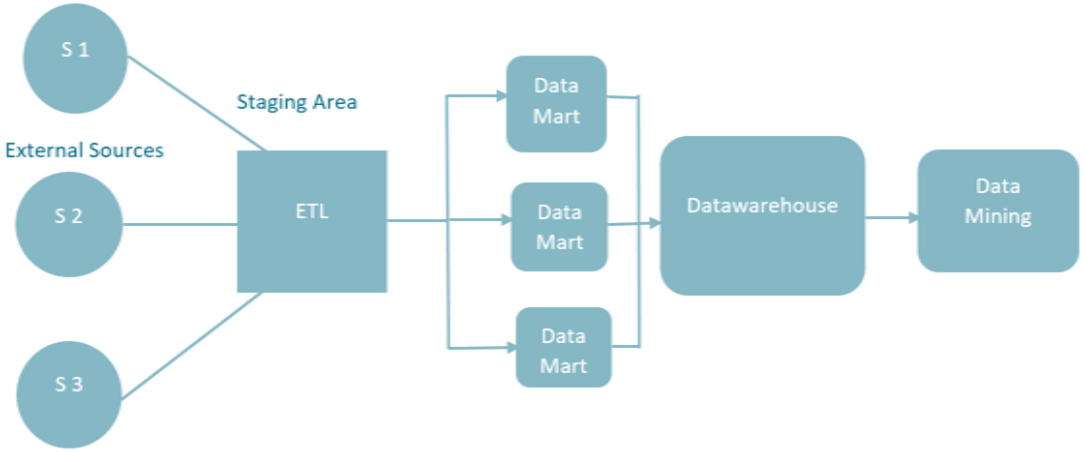


❑ Trois techniques pour réaliser un DW

1. Top-down

- **Avantages :**
  - Offrir une architecture intégrée : méthode complète
  - Réutilisation des données
  - Pas de redondances
  - Vision claire et conceptuelle des données de l'entreprise et du travail à réaliser
- **Inconvénients :**
  - Méthode lourde
  - Méthode contraignante
  - Nécessite du temps

9. Réalisation d'un Data Warehouse

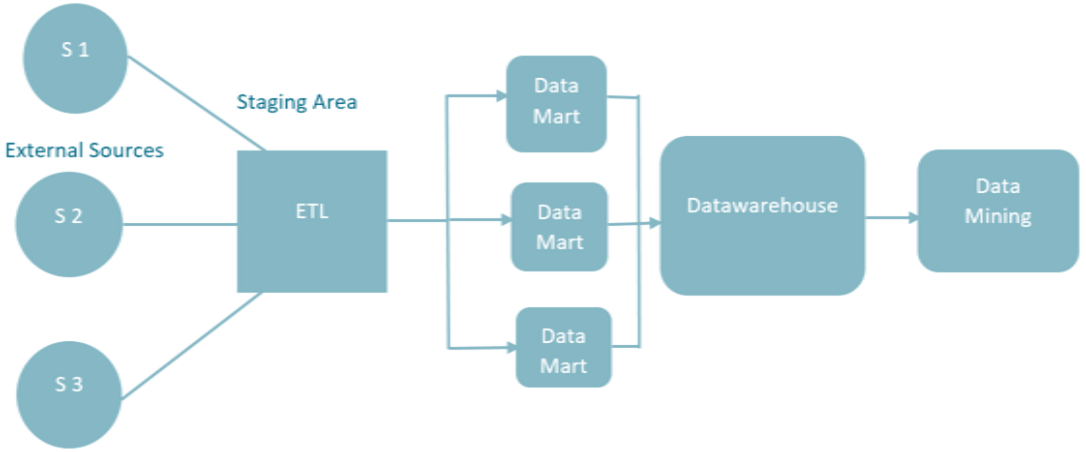


❑ **Trois techniques pour réaliser un DW**

2. Bottom-up (approche inverse)

- Créer les datamarts un par un puis les regrouper par des niveaux intermédiaires jusqu'à obtention d'un véritable entrepôt.
- **Objectif** : Livrer une solution permettant aux usager d'obtenir facilement et rapidement des réponses à leurs requêtes d'analyse

9. Réalisation d'un Data Warehouse



❑ **Trois techniques pour réaliser un DW**

2. Bottom-up (approche inverse)

- **Avantages :**
  - Simple à réaliser,
  - Résultats rapides
  - Efficace à court terme
- **Inconvénients :**
  - Pas efficace à long terme
  - Le volume de travail d'intégration pour obtenir un entrepôt de données
  - Risque de redondances (car réalisations indépendantes).

9. Réalisation d'un Data Warehouse

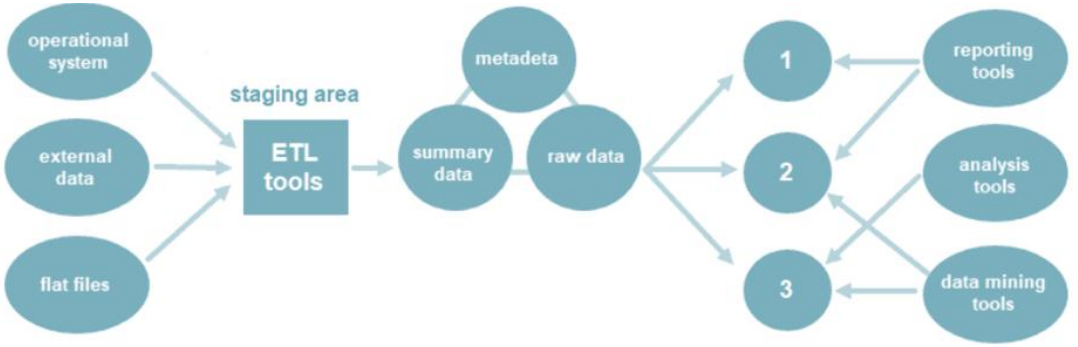
- **Trois techniques pour réaliser un DW**

3. Middle-Out (approche hybride)

- Concevoir intégralement l'entrepôt de données (toutes les dimensions, tous les faits, toutes les relations), puis créer des divisions plus petites et plus gérables..

9. Réalisation d'un Data Warehouse

❑ **Trois techniques pour réaliser un DW**



3. Middle-Out (approche hybride)

- **Avantages :**
  - Prendre le meilleur des 2 approches
  - Développement d'un modèle de données d'entreprise de manière itérative
  - Développement d'une infrastructure lourde qu'en cas de nécessité
- **Inconvénients :**
  - Implique, parfois, des compromis de découpage (dupliquer des dimensions identiques pour des besoins pratiques).

Réalisation d’un Data Warehouse

Les étapes pour mettre en place un DW

9. Réalisation d'un Data Warehouse

- **Cinq étapes importantes pour la réalisation d'un DW**

1. **Conception**
2. **Acquisition des données**
3. **Définition des aspects techniques de la réalisation**
4. **Définition des modes de restitution**
5. **Stratégies d'administration, évolution, maintenance**

9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

1. Conception

- **Définir la finalité du DW :**
  - Quelle activité de l'entreprise faut-il piloter?
  - Quel est le processus de l'entreprise à modéliser?
  - Qui sont les décideurs?
  - Quels sont les faits numériques? Qu'est ce qui va être mesurer?
  - Quelles sont les dimensions ?
    - *Comment les gestionnaires décrivent-ils des données qui résultent du processus concerné?*

9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

1. Conception

- **Définir le modèle de données :**
  - Modèle en étoile / flocon ?
  - et/ou Cube?
  - et/ou Vues matérialisées?

9. Réalisation d'un Data Warehouse

❑ **Cinq étapes importantes pour la réalisation d'un DW**

2. Acquisition des données

- **Pour l'alimentation ou la mise à jour de l'entrepôt**
  - **Mise à jour régulière**
  - ***Besoin d'un outil pour automatiser les chargements de l'entrepôt.***

ETL (Extract, Transform, Load)

9. Réalisation d'un Data Warehouse

❑ **Cinq étapes importantes pour la réalisation d'un DW**

ETL (*Extract, Transform, Load*)

2. Acquisition des données

- Modèle entité-relation (BD de production) : Modèle à base de dimensions et de faits
- ***Outil :***
  - Offrant un environnement de développement.
  - Offrant des outils de gestion des opérations et de maintenance.
  - Permettant de découvrir, analyser, et extraire les données à partir de sources hétérogènes.
  - Permettant de nettoyer et standardiser les données.
  - Permettant de charger les données dans un entrepôt.

9. Réalisation d'un Data Warehouse

❑ **Cinq étapes importantes pour la réalisation d'un DW**

ETL (Extract, Transform, Load)

2. Acquisition des données

- **Extraction :**
  - Depuis différentes sources (bd, fichiers, …)
  - Différentes techniques :
    - ***Push : règles (triggers)***
    - ***Pull : requêtes (queries)***
  - Périodique et répétée : Dater ou marquer les données envoyées
  - Difficulté : Ne pas perturber les applications OLTP

9. Réalisation d'un Data Warehouse

❑ **Cinq étapes importantes pour la réalisation d'un DW**

ETL (*Extract, Transform, Load*)

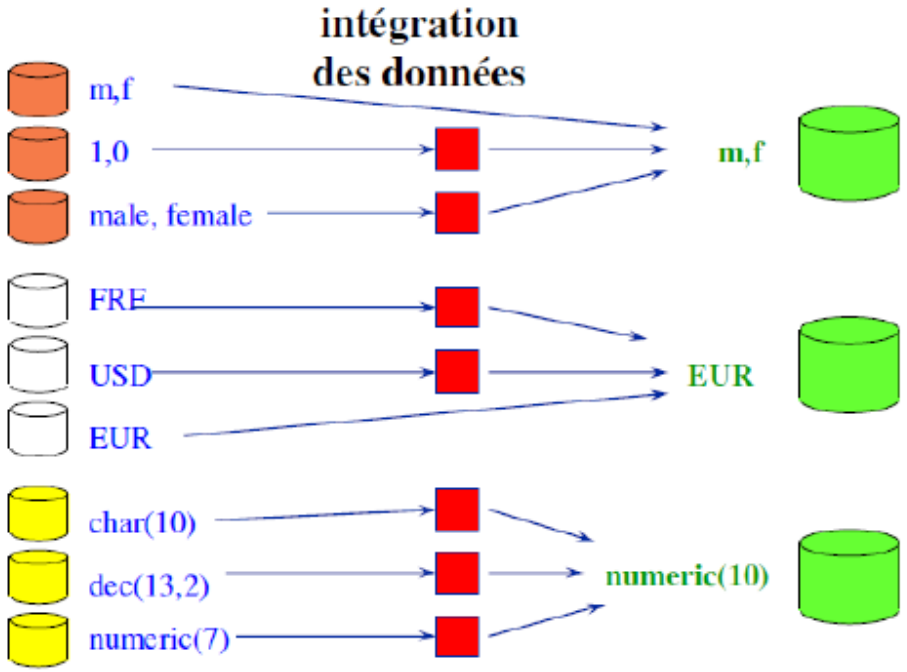
2. Acquisition des données

- **Transformation:** Etape très importante qui garantit la cohérence et la fiabilité des données
  - Rendre cohérentes les données issues de différentes sources
    - Unifier les données.
    - Trier, Nettoyer.
      - *Eliminer les doubles*
      - *Gestion des valeurs manquantes*
      - *Gestion des valeurs erronées ou inconsistantes*
    - Inspection manuelle de certaines données possible…

9. Réalisation d'un Data Warehouse

- **Cinq étapes importantes pour la réalisation d'un DW**

2. Acquisition des données



9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

2. Acquisition des données

- **ETL ≠ ELT**
  - L'approche ELT (Extraction, Loading, Transformation) génère du code SQL natif pour chaque moteur de BD impliqué dans le processus – sources et cibles

9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

2. Acquisition des données

- **ETL ≠ ELT**
  - Cette approche profite des fonctionnalités de chaque BD mais les requêtes de transformation doivent respecter la syntaxe spécifique au SGBD.

9. Réalisation d'un Data Warehouse

❑ **Cinq étapes importantes pour la réalisation d'un DW**

3. Aspects techniques

- **Contraintes**
  - logicielles
  - matérielles,
  - humaines,
  - …

9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

4. Restitution

- But du processus d'entreposage,
- Conditionne souvent le choix de l'architecture et de la construction du DW
- Toutes les analyses nécessaires doivent être réalisables !
- Types d'outils de restitution :
  - Requêteurs et outils d'analyse.
  - Outils de datamining

9. Réalisation d'un Data Warehouse

❑ Cinq étapes importantes pour la réalisation d'un DW

5. Administration, maintenance

- Toutes les stratégies à mettre en place pour l'administration, l'évolution et la maintenance

10. Principales applications autour d'un ED

- **Réalisation de rapports divers (Reporting)**
- **Réalisation de tableaux de bords (Dashboards)**
- **Analyse en ligne diverses (OLAP)**
- **Fouille de données (Data Mining)**
- **Visualisations autour d'un ED (Visualizations)**
- **Etc.**

10. Principales applications autour d'un ED

❑ **Réalisation de rapports divers (Reporting)**

- Ils sont crées pour les utilisateurs qui ont besoin d'un accès régulier à des informations d'une manière presque statique
  - ✓ Ex: les hôpitaux doivent envoyer des rapports mensuels à des agences nationales
- Un rapport est défini par une requête (plusieurs requêtes) et une mise en page.
- Les rapports peuvent être exécutés automatiquement ou manuellement

10. Principales applications autour d'un ED

- **Réalisation de rapports divers (Reporting)**

Report parameters

Auto-refresh: Off

Profit & Loss

Period: December, 2019
Department: AA:UK

REVENUE - Current Month	GROSS PROFIT - Current Month	PROFIT - Current Month
$11,874,844	$10,999,953	$8,505,653
▲18.77%	▲20.50%	▲20.70%

The Revenue this month was great at $11,874,844 versus Budget at $9,646,050, and the Operating Profit was $8,505,653, which is an increase of 20.7%.

		Current Month				YTD				Last Year YTD		
Account No	Description	Act Mth	Budget	Var	Var (%)	Act YTD	Budget	Var	Var (%)	Act LYTD	Var	Var (%)
REVENUE												
3030	Sales/Income	11,708,638	9,500,000	2,208,638	▲18.86%	68,828,141	65,000,000	3,828,141	▲5.56%	62,500,800	6,327,341	▲9.19%
3031	Sales/Income Type 1	29,917	15,300	14,617	▲48.86%	74,109,683	60,000,000	14,109,683	▲19.04%	69,423,738	4,685,945	▲6.32%
3032	Sales/Income Type 2	263	250	13	▲4.90%	1,052	1,000	52	▲4.90%	950	102	▲9.66%
3033	Sales/Income Type 3	-	-	-	-	12,166	12,300	(134)	▼1.10%	13,100	(934)	▼7.68%
3050	Discounts Allowed-Trade	500	500	-	▲0.00%	22,175	21,900	275	▲1.24%	18,620	3,555	▲16.03%
3070	Freight Income	125,526	120,000	5,526	▲4.40%	1,129,734	1,130,000	(266)	▼0.02%	1,135,226	(5,492)	▼0.49%
3080	Miscellaneous Income	-	-	-	-	45,411	22,800	22,611	▲49.79%	43,250	2,161	▲4.76%
3090	Customer Finance Charges	10,000	10,000	(0)	▼0.00%	90,000	90,000	(0)	▼0.00%	90,000	(0)	▼0.00%
3100	Miscellaneous Income 2	-	-	-	-	3,793,302	2,895,700	897,602	▲23.66%	2,598,650	1,194,652	▲31.49%
TOTAL REVENUE		$ 11,874,844	$ 9,646,050	$ 2,228,794	▲18.77%	$ 148,031,663	$ 129,173,700	$ 18,857,963	▲12.74%	$ 135,824,334	$ 12,207,329	▲8.25%
COST OF GOODS SOLD												
4030	Merchandise Purchases	874,855	890,000	(15,145)	▼1.73%	11,439,685	12,500,000	(1,060,315)	▼9.27%	12,630,214	(1,190,529)	▼10.41%
4040	Discounts Taken	36	-	36	-	13,498	12,000	1,498	▲11.09%	12,950	548	▲4.06%
4091	Federal Tax Expense	-	-	-	-	35,789	40,000	(4,211)	▼11.77%	35,988	(199)	▼0.56%

10. Principales applications autour d'un ED

❑ Tableaux de bords (Dashboards)

- Affichent une quantité limitée d'informations dans un format graphique facile à lire
- Fréquemment utilisé par les cadres supérieurs qui ont besoin d'un rapide aperçu des changements les plus importants
  - *Ex : un aperçu en temps réel d'évolutions*
- Pas vraiment utile pour une analyse complexe et détaillée

Conclusion

Good luck