Le data Warehouse

1. Généralités sur les data Warehouses
2. Eléments clés de la modélisation dimensionnelle
3. Processus de modélisation d'un Data Warehouse
4. Concepts avancés

1 Généralités sur les Data Warehouses

LE DATA WAREHOUSE

Introduction

- Dans le monde des affaires actuel, les données et l'analyse jouent un rôle indispensable dans le processus de prise de décision.
- La plupart des grandes entreprises créent donc des **entrepôts de données** (appelés aussi: Data Warehouse ou DW) à ***des fins de reporting et d'analyse***.
- Nous nous intéresserons donc dans ce module au **processus de mise en place d'un data Warehouse**
- A la fin de ce module vous serez capables de:
  - décrire les objectifs du data Warehouse
  - comparer une base de données transactionnelle à une base de donnée décisionnelle
  - décrire les différents éléments d'une architecture décisionnelle
  - modéliser un data Warehouse selon différents modèles (étoile, flocon de neige et constellation)

Le Data Warehouse

- Le *Data Warehouse*, ou entrepôt de données, est une base de données **relationnelle** dédiée au stockage de l'ensemble des données fonctionnelle d’une entreprise.
- Il est utilisé dans le cadre de la prise de décision et de l'analyse décisionnelle.
- Il est alimenté à partir des **bases de données de production**.
- Les utilisateurs, analystes et décideurs, accèdent aux données collectées et mises en forme pour étudier des cas précis de réflexion. Ils construisent des modèles d'étude et de prospective pour limiter la part d'incertitude lors du processus de prise de décision.

Définition du Data Warehouse

Définition de Bill Inmon (1996):**« Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision »**

- **Orienté sujet:**Au cœur du Data Warehouse, les données sont organisées par thème. Les données propres à un thème, les ventes par exemple, seront rapatriées des différentes bases OLTP de production et regroupées.
- **Intégré:**Les données proviennent de plusieurs sources différentes. Avant d'être intégrées au sein du data warehouse elles doivent être mise en forme et unifiées afin d'en assurer la cohérence. Cette phase est très complexe et représente une charge importante dans la mise en place d'un data warehouse.

Définition du Data Warehouse (suite)

- **Non volatile :** Un data Warehouse veut conserver la traçabilité des informations et des décisions prises. Les données ne sont ni modifiées ni supprimées. Une requête émise sur les mêmes données à plusieurs mois d'intervalles doit donner le même résultat.
- **Historisé :** Contrairement au système de production les données ne sont jamais mises à jour. Chaque nouvelle donnée est insérée. Un référentiel de temps doit être mis en place afin de pouvoir identifier chaque donnée dans le temps.

Un data Warehouse définit donc à la fois un **ensemble de données** et un **ensemble d'outils**.

- Il s'agit de données destinées aux décideurs, qui sont souvent une copie des données de production avec une valeur ajoutées (orientés objet, agrégés, historisées).
- C'est un ensemble d'outils permettant de regrouper les données des différentes sources, de les nettoyer et de les intégrer, ainsi que d'y accéder de différentes manières (requêtes, rapport, analyse, datamining).

Avantages liés aux Data Warehouses

Les entrepôts de données permettent de :

- Prendre de meilleures décisions
- Consolider des données provenant de sources différentes
- Posséder des données de qualité, cohérentes et précises
- Conserver un historique intelligent des données
- Séparer le traitement analytique des bases de données transactionnelles

Entrepôt de données VS Base de Données Transactionnelle

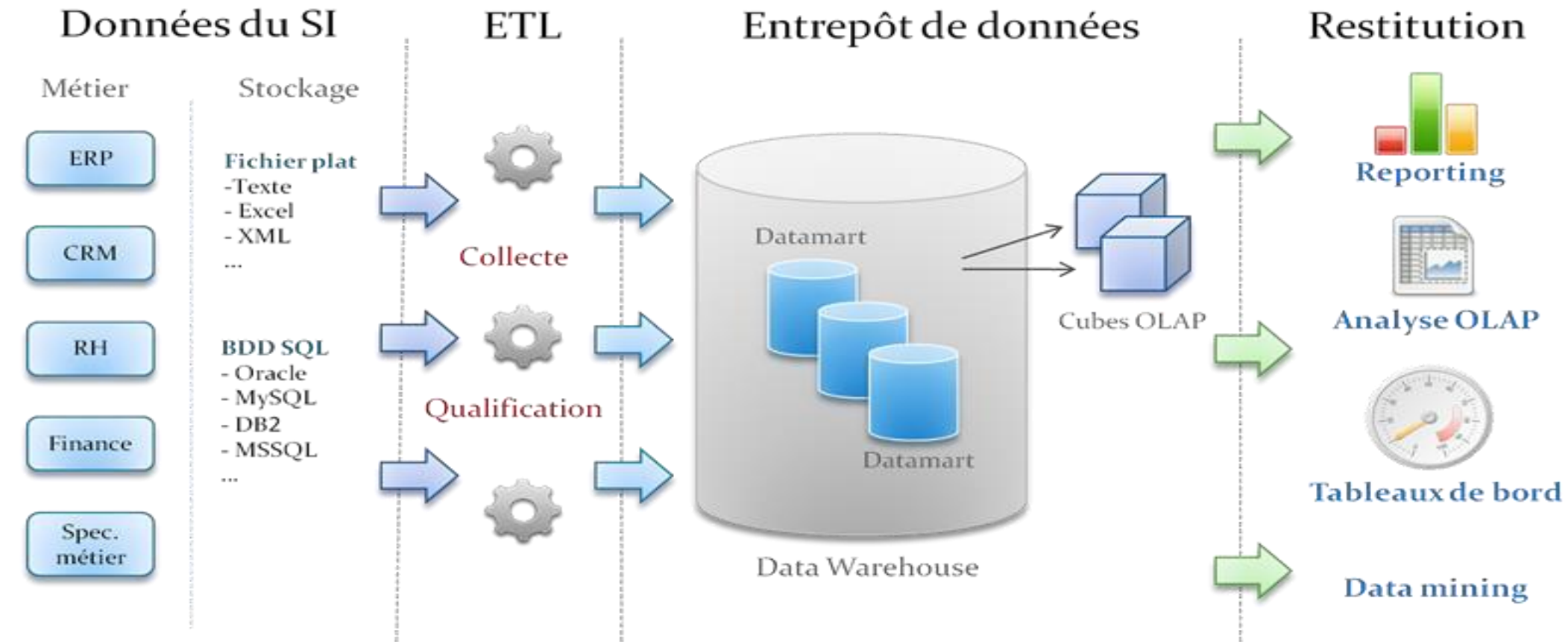
- Un **entrepôt de données** est conçu spécialement pour analyser des données ce qui Implique la lecture de grandes quantités de données dans le but de comprendre les relations et les tendances entre ces données.
- Une **base de données transactionnelle** sert à saisir et stocker les données Opérationnelles, en sauvegardant les détails liés à une transaction, par exemple.

Caractéristiques :

Caractéristiques	Entrepôt de données	Bases de données transactionnelles
Charges de travail adaptées	Analyses, rapport …	Traitement des transactions

Caractéristiques	Entrepôt de données	Bases de données transactionnelles
Source de données	Les données sont collectées et normalisées à partir de nombreuses sources	Les données saisies en l'état proviennent d'une source unique, comme un système transactionnel, par exemple
Saisie de données	Opérations d'écriture en masse sur un programme de lots prédéterminé	Idéale pour les opérations d'écriture en continu lorsque de nouvelles données sont disponibles afin d'optimiser le débit des transactions
Normalisation des données	Schémas dénormalisés, tels que le schéma en étoile ou en flocon	Schémas statiques normalisés
Stockage de données	Accès simple et rapide en termes de recherches grâce au stockage en colonnes	Idéal pour effectuer de nombreuses opérations d'écriture dans un bloc unique en colonne
Accès aux données	Idéal pour minimiser le nombre d'E/S et optimiser le débit des données	Très nombreuses petites opérations de lecture

Architecture générale d'un système d'information décisionnel (SID)



Etapes globales de mise en place d'un SID

1. **Etude des besoins :** Cette étape consiste à définir l'ensemble des axes d'analyse et des indicateurs ou mesures. Ces axes d'analyse et ces mesures serviront à analyser les données à travers des Rapports, Tableaux de bord, Cube OLAP …

2. **Collecte et Qualification :** Cette étape du projet est réalisée grâce à l'ETL (ExtractTransformLoad). Elle consiste à identifier les différentes sources de données puis à collecter ces données pour ensuite les transformer et les qualifier afin de les déposer dans un Data Warehouse dans un format adapté à l'analyse.

3. **Entrepôt de données :** L'entrepôt de donnée ou Data Warehouse est une base de données relationnelle avec des modèles en flocons ou en étoiles :

   L'entrepôt de données recevra les données collectées par l'ETL (ExtractTransformLoad). Des cubes OLAP seront créés à partir des données du Data Warehouse.

4. **Restitution :** De nombreuses solutions existent sur le marché et beaucoup sont performantes et permettent de répondre à la plupart des besoins.

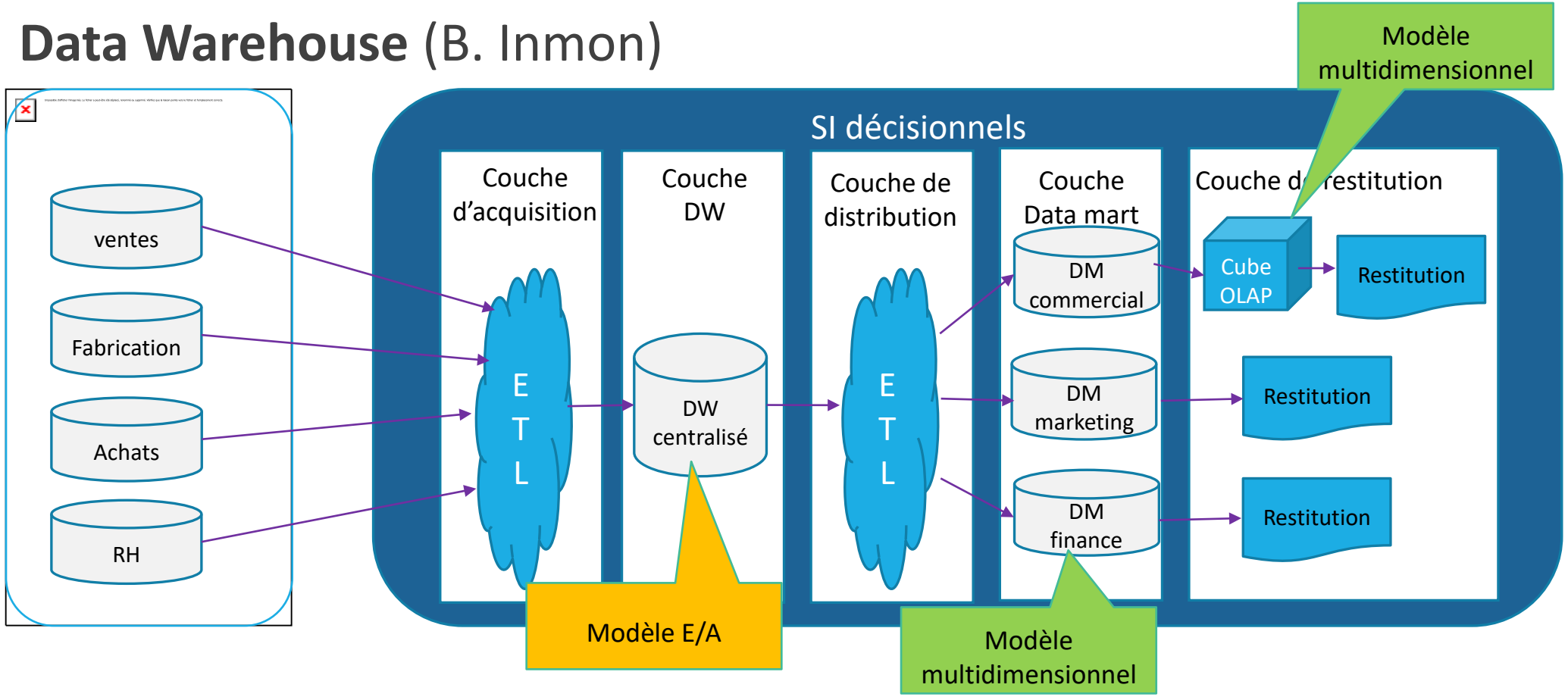
Classification des architectures des SID

- On peut trouver plusieurs variantes dans les architectures des SID.
- Néanmoins, dans la littérature, ces variantes appartiennent à l'une des deux familles suivantes :

1. **Architecture Corporate Information Factory** ou **Enterprise Data Warehouse** (B. Inmon)
2. **Architecture Dimensional Data Warehouse** ou **Bus Architecture** (R. Kimball)

Entreprise Data Warehouse

- **Architecture Corporate Information Factory (CIF)** ou **Entreprise Data Warehouse** (B. Inmon)



Entreprise Data Warehouse (suite)

Le SID est constitué de **5** couches:

- **La couche d'acquisition:**permet d'extraire, de transformer et de charger les données de chacun des SIO vers le DW centralisé
- **La couche data Warehouse:** elle contient un DW centralisé, unique pour l'entreprise.
  - Le DW est une BD relationnelle conçue à partir d'un modèle de données d'entreprise préalablement défini, de type E/A (en 3FN)
  - Le DW ne sera pas interrogé par les utilisateurs; il servira de source unique de données pour d'autres applications destinées aux utilisateurs

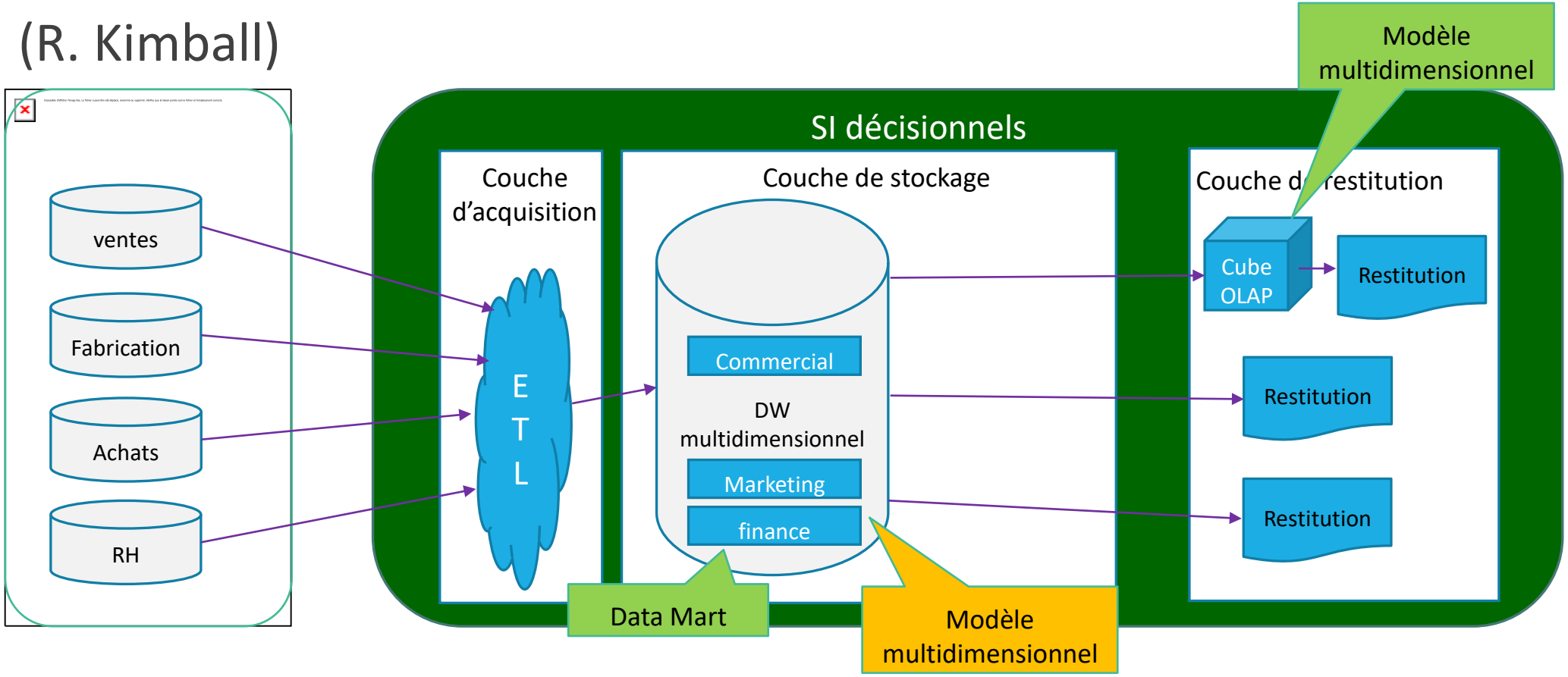
Entreprise Data Warehouse (suite)

- **La couche de distribution :** c'est une couche de traitement qui alimente les applications, utilisées par les utilisateurs, à partir du data Warehouse
- **La couche data marts :** c'est une couche de stockage contenant les données distribuées à partir de la couche de distribution.
  - Les DM servent à satisfaire les besoins de restitution des différents départements.
  - Les DM sont conçues en général avec la modélisation multidimensionnelle.
  - Ils contiennent en général des données agrégées par rapport à celles du DW
- **La couche de restitution** elle restitue aux utilisateurs du SID l'information contenue dans le SID sous forme de rapports ou de tableaux de bord.
  - Elle peut contenir un ou plusieurs cubes OLAP qui sont des espaces de stockages, optimisés, le plus souvent non relationnels

Bus Architecture

- **Architecture Dimensional Data Warehouse** ou **Bus Architecture**

(R. Kimball)


Modèle multidimensionnel
SI décisionnels
Couche d'acquisition
Couche de stockage
Couche de restitution
ventes
Fabrication
Achats
RH
E
T
L
Commercial
DW multidimensionnel
Marketing
finance
Cube OLAP
Restitution
Restitution
Restitution
Data Mart
Modèle multidimensionnel

Bus Architecture (suite)

Le SID est constitué de **3** couches:

- **La couche d'acquisition :** même principe que CIF
- **La couche stockage :** A la différence du CIF qui contient plusieurs couches de stockage, ici la couche de stockage est unique.
  - Elle contient un DW conçu avec la technique de modélisation multidimensionnelle.
  - Le DW est considéré comme un ensemble de data marts connectés par des dimensions et faits conformes.
- **La couche de restitution:** identique au CIF

Architecture Kimball vs Inmon

Architecture Kimball	Architecture Inmon
Le DW est modélisé avec le modèle multidimensionnel	Le DW est modélisé en 3^{ième} forme normale avec le modèle E/A
Le DW est accessible directement par les utilisateurs	Seuls les data marts sont accessibles aux outils de restitution.
Les data marts contiennent des données de détail	Les data marts contiennent le plus souvent des données agrégées
La modélisation multidimensionnelle est utilisée pour la conception des data marts	La modélisation multidimensionnelle est utilisée pour la conception des data marts

2 Eléments clés de la modélisation dimensionnelle

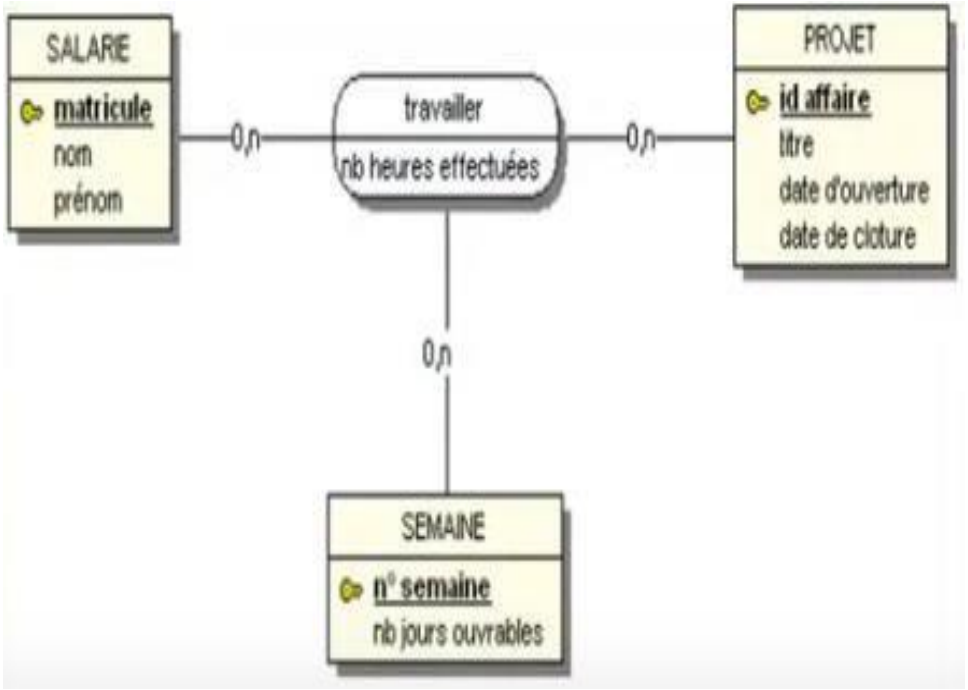
LE DATA WAREHOUSE

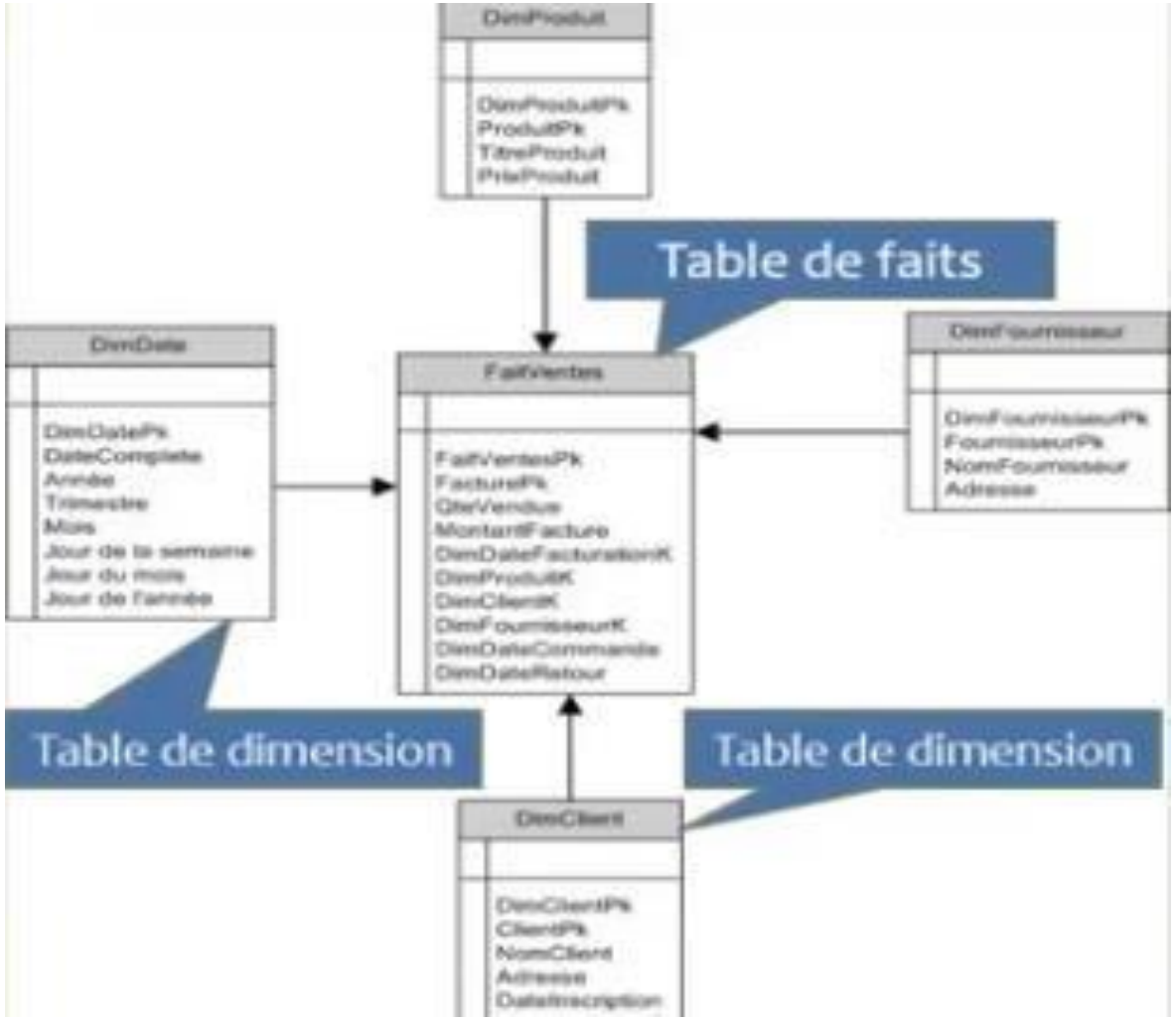
Concepts fondamentaux : Fait & Dimension

- La modélisation des bases de données relationnelles utilise

Les concepts d'entités et de relations afin de construire des tables

- En business Intelligence, la modélisation d'un data Warehouse utilise les notions de **table de faits** et de **table de dimension.**


SALARIE
matricule
nom
prénom
0,n
travailler
nb heures effectuées
0,n
PROJET
id affaire
titre
date d'ouverture
date de cloture
0,n
SEMAINE
n° semaine
nb jours ouvrables


DimProduit
Table de faits
DimDate
FaitVentes
DimFournisseur
Table de dimension
Table de dimension
DimClient

Qu’est-ce qu’une table de faits?

- La **table de faits** est *la table centrale du modèle dimensionnel*. Elle contient les informations observables (les mesures) sur ce qu’on veut analyser: table de faits des ventes par exemple.
- Une ligne d’une table de faits correspond à une mesure. Ces mesures sont généralement des valeurs numériques et additives;
- Une table de faits assure les liens plusieurs à plusieurs entre les dimensions. Elles comportent des clés étrangères, qui ne sont autres que les clés primaires des tables de dimension.

- Exemples de faits pour les certains processus métier
  - Pour les ventes, on peut avoir les faits suivants: chiffre d'affaire net, quantités et montants commandés, quantités facturées, quantités retournées, volumes des ventes, etc.
  - Pour la gestion de stock : nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock, taux de roulement d'une zone, etc.
  - Pour la gestion des ressources humaines : performances des employés, nombre de demandes de congés, nombre de démissions, taux de roulement des employés, etc.).

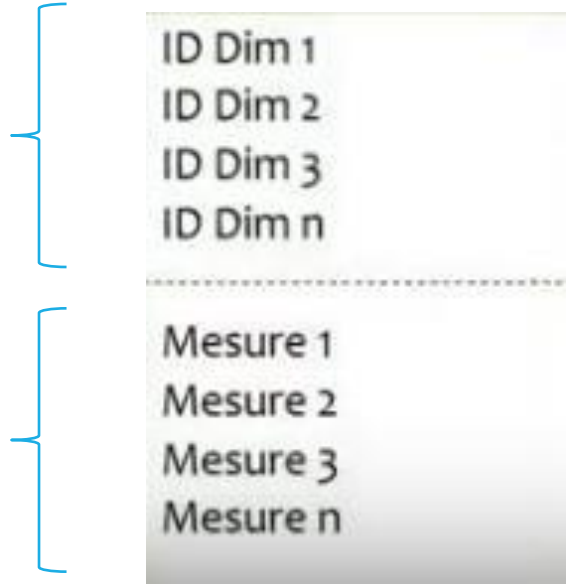
Qu’est-ce qu’une table de faits?(suite)

Exemple d’une table de faits:

Clés étrangères vers table de dimension

Faits ou mesures

Structure d’une

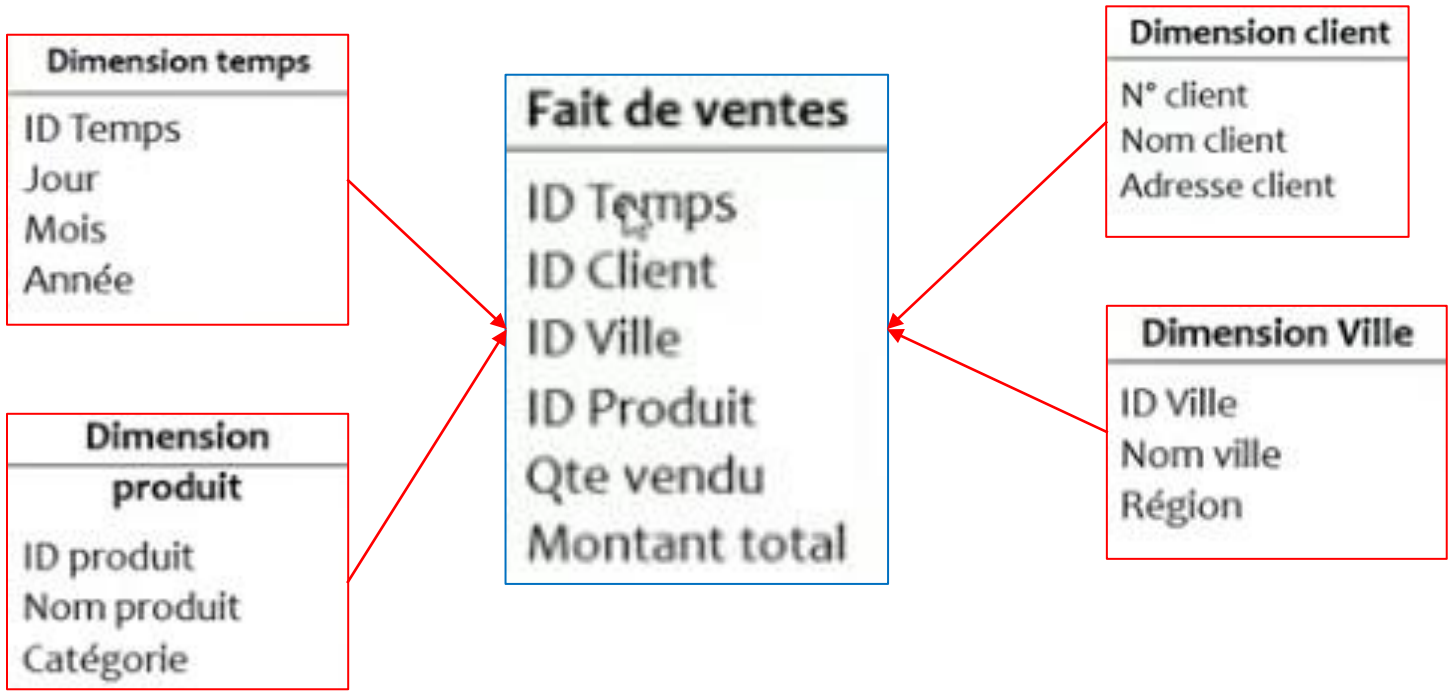


Qu’est-ce qu’une table de dimension ?

- Une **table de dimension** *représente un axe d’analyse* : dimension de temps, dimension géographique, dimension client, etc.
- Les tables de dimension sont les tables qui raccompagnent une table de faits, elles contiennent la description textuelle de l’activité. Une table de dimension est constituée de nombreuses colonnes qui décrivent une ligne. C’est grâce à cette table que l’entrepôt de données est compréhensible et utilisable, elles permettent des analyses en tranches et en dés.
- Une dimension est généralement constituée: d’une clé artificielle, une clé naturelle et des attributs.

Qu’est-ce qu’une table de dimension ?

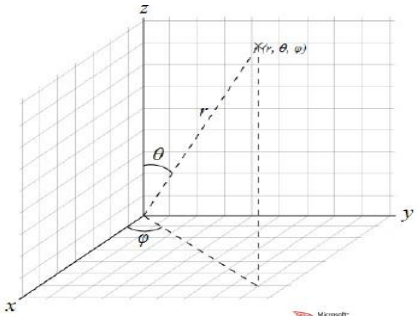
Exemple d’une table de dimension :



Concepts fondamentaux : Mesure, Fait & Dimension (Récap.)

- **La mesure:** Il s'est passé quelque chose, et on l'a *mesuré*!

- **Les dimensions:** On l'a mesuré selon notre *référentiel*
  (ça s'est passé quand, ça s'est passé où, etc)



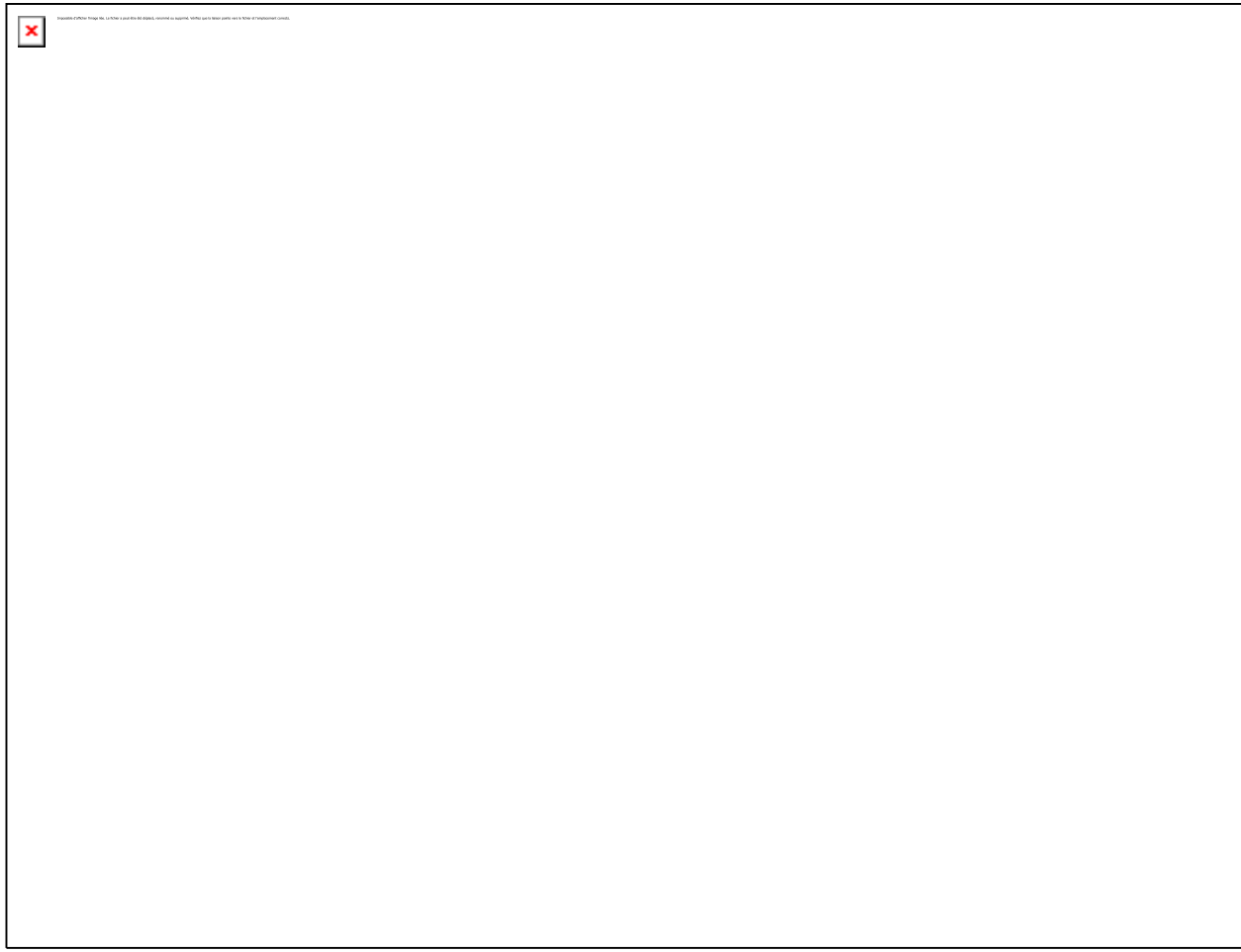
- **Les faits (actes, évènements):**Il s'est passé quelque chose, et on l'a mesuré selon notre référentiel, nos dimensions

Modélisation d’un Data Warehouse

- Trois modèles permettant la présentation d’un Data Warehouse :
  1. Modèle en étoile
  2. Modèle en flocon
  3. Modèle en constellation

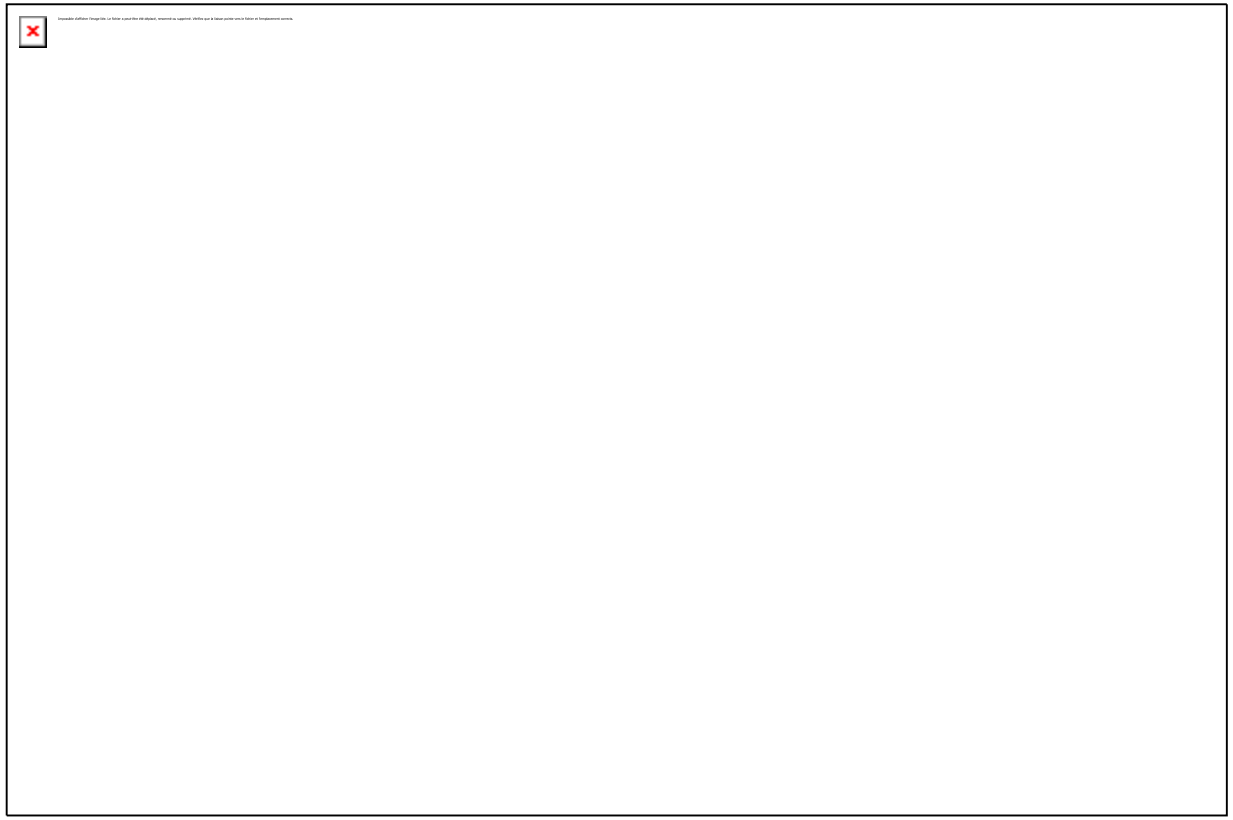
Modèle en étoile (R. Kimball)

- Ce modèle se présente comme une étoile dont le centre n’est autre que la table des faits et les branches sont les tables de dimension.
- La force de ce type de modélisation est sa lisibilité et sa performance



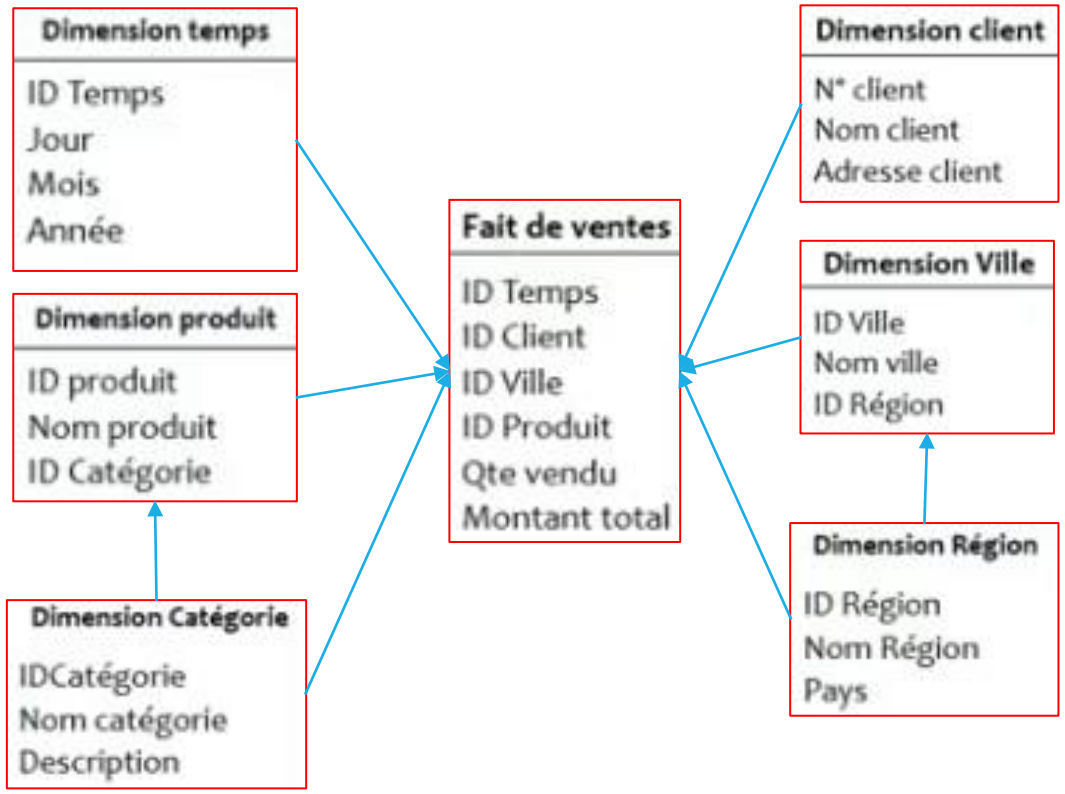
Modèle en flocon (Inmon)

- Identique au modèle en étoile, sauf que ses branches sont éclatées en hiérarchies.
- Cette modélisation est généralement justifiée par l'économie d'espace de stockage, cependant elle peut s'avérer moins compréhensible pour l'utilisateur final, très couteuse en terme de performance



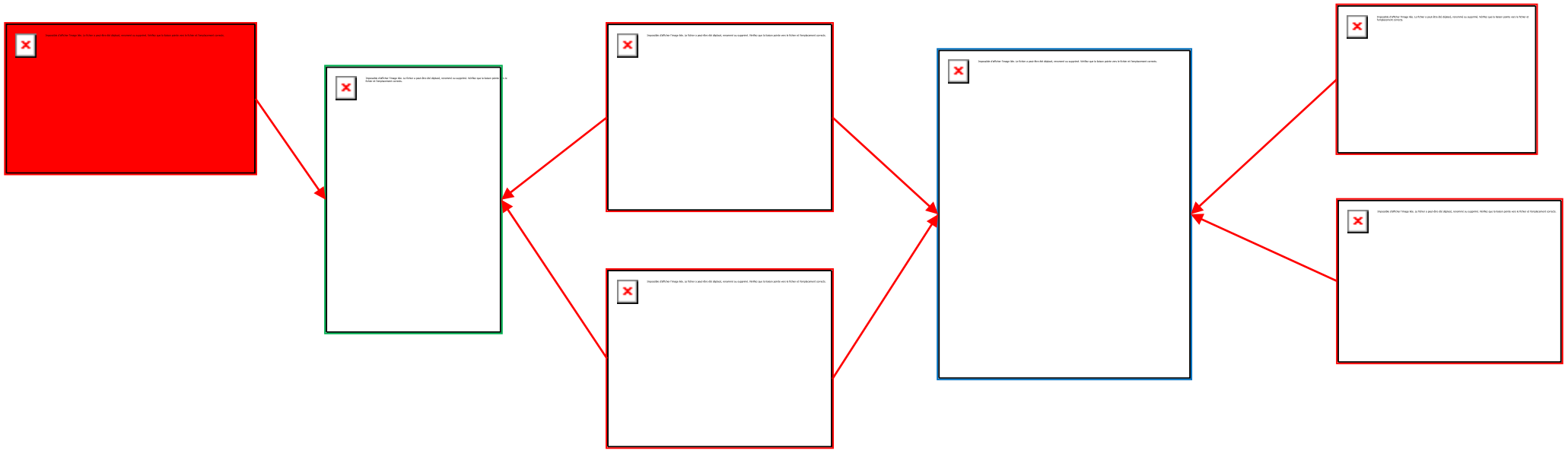
Modèle en flocon (Inmon) (suite)

Exemple :



Modèle en constellation

- Ce n’est rien d’autre que plusieurs modèles en étoile liés entre eux par des dimensions communes



3 Processus de modélisation Data Warehouse

LE DATA WAREHOUSE

L’avant modélisation

- Identifier **l’utilisateur :**
- C’est lui qui va utiliser le projet de data warehousing
- C’est lui qui connait le métier
- C’est lui qui va indiquer les règles de gestion
- Il faut savoir de quel reporting il a besoin pour pouvoir bien construire le modèle en étoile/ en flocon de neige

Processus de modélisation

1. **Sélection du processus métier à modéliser** (on modélise quoi? Quel processus métier « on va mesurer » ?)
   - *Quel acte est mesuré ?*
     - Une transaction : une vente, une commande, un voyage, etc.
     - Un stock : un *inventaire dans un magasin*, une position financière, etc.
   - Le choix du processus est fixé par les décideurs et responsables opérationnels

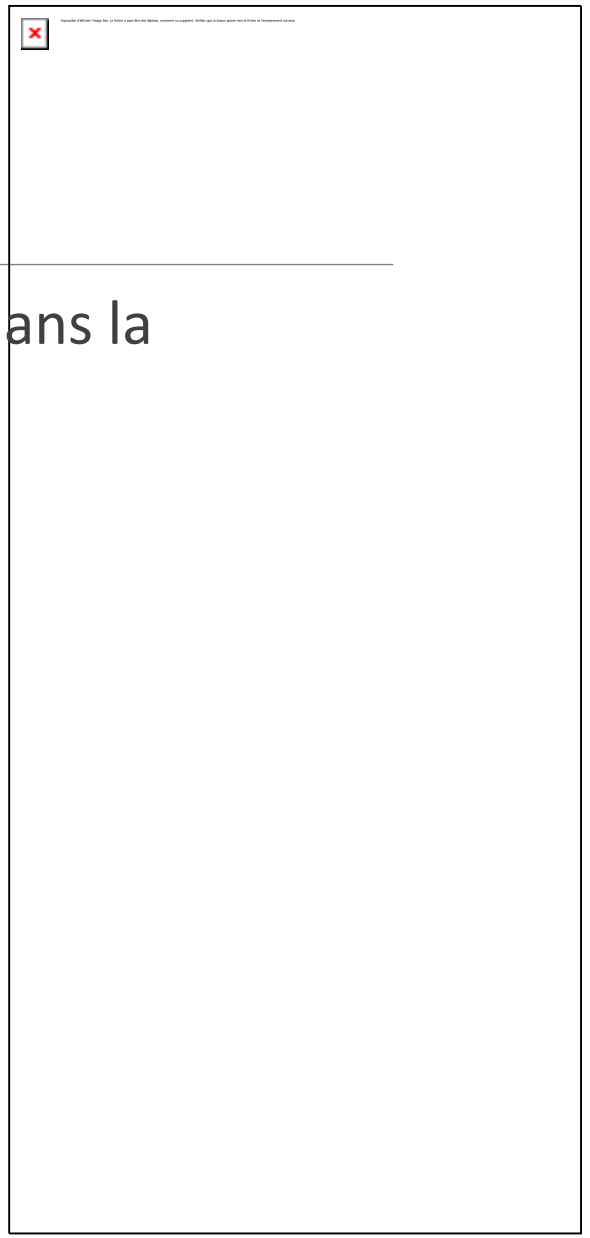
Processus de modélisation *(suite)*

Définition du grain du processus métier (que représente la ligne dans la table de fait?)

- Une ligne d'un ticket de caisse?
- Le total du ticket de caisse?
- Le stock en fin de semaine du magasin?

Exemple: Inventaire

- Comment considère t'on le « Où »:
  - Dans un rack?
  - Dans une rangée?
  - Dans l'entrepôt de quel magasin?
- Faire le choix du niveau de détail de l'information que l'on souhaite conserver



Processus de modélisation *(suite)*

Identification des dimensions (axes d'analyse) qui s'appliquent

- Choisir quels sont les axes d'analyse adéquats pour le processus en question
- Principe: Réutiliser les dimensions disponibles

	Date	Produit	Magasin	Entrepôt	Fournisseur	Promotion
Forecasting	x		x			x
Achats	x	x		x	x	
Commandes	x	x	x			x
Livraisons	x		x	x		

https://fleid.files.wordpress.com/2011/12/s221-modc3a9lisation-dimensionnelle-les-fondements-du-datawarehouse.pdf

4 Concepts avancés

LE DATA WAREHOUSE

Types de mesures

- **Mesures additives :** Peuvent être agrégées selon n'importe quelle dimension ;
  - Exemple : montant de vente, quantité commandée, etc.
- **Mesures semi-additives :** Peuvent être agrégées selon certaines dimensions seulement ;
  - Exemple : solde de compte agrégeable selon les clients, pas le temps.
- **Mesures non-additives :** Valeurs numériques ne pouvant être agrégées selon aucune dimension ;
  - Exemple: pourcentages ou ratios.

Types de mesures *(suite)*

- **Activité :** Mesure additive, semi-additive ou non-additive ?

---

1. Quantité en inventaire
2. Pourcentage de profit : 100*(vente – coût) / vente
3. Nombre d’items vendus 4. Produit en vente (valeur binaire)

Mesures vs Attributs

- Mesures:
  - Dépendent d'un événement métier;
  - Ont souvent des valeurs continues (ou un grand nombre de valeurs discrètes possibles);
  - Servent dans les calculs des requêtes; - Ex: montant total et quantité d'une commande.

- Attributs (numériques) *de dimension* :
  - Indépendants des événements métier ;
  - Ont souvent des valeurs discrètes; –Servent à filtrer ou étiqueter les faits; – Ex: âge d'un client, etc.

Dimensions conformes

- On parle de dimension conforme ou partagée lorsque la dimension est utilisée par les faits de plus qu'un data mart.
- L'exemple le plus courant est la dimension «Produit » qui est utilisée par différents data Mart «Finance », « Marketing »…

Dimension temporelle

- Centrale car la plupart des faits correspondent à des évènements d'affaires de l'entreprise
- Mettre ces valeurs même si la plupart peuvent être déduites des autres
- 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.

Dimension temporelle *(suite)*

- Solution 1: mettre le temps du jour (*time of day*) dans une dimension séparée:
  - **Dimension 1**: année → mois → semaine → jour;
  - **Dimension 2**: heure → minute → secondes; –

  86,400 + 365 lignes VS 31,000,000 lignes.

- Solution 2: mettre le temps du jour comme un fait et garder le jour, mois, année dans une dimension;
  - La solution 2 est normalement préférable à moins d'avoir des attributs supplémentaires (ex: descripteur texte).

Dimensions dégénérées

- La dimension dégénérée est une clé de dimension dans la table de fait qui est en général sans attribut.
- Exemple : No d'interruption de service. Dans ce cas, les utilisateurs veulent savoir par exemple « combien de fois un client a été interrompu dans une période de temps précise ».
- Vu qu'il s'agit d'une seule clé de dimension, nous évitons alors de créer une table de dimension, ce qui fait que cette table de dimension a dégénéré dans la table de fait, c'est pour cette raison que cette clé est appelée « dimension dégénérée »

MERCI DE VOTRE ATTENTION
Et bonne continuation,

Développez toujours.