EPISEN

Ing2 2021-2022

Module BDM

BASES DE DONNÉES MULTIDIMENSIONNELLES ET NOMADES

LES SYSTÈMES D'INFORMATION DÉCISIONNELS

Contexte : Les S.I.D.

« L'informatique décisionnelle* désigne les méthodes, techniques et outils qui permettent de collecter, consolider, modéliser et restituer les données d'une entreprise en vue d'offrir une aide à la décision ...»

Bill Inmon 1994

Elle doit permettre aux décideurs et managers de :

- avoir une **vue d'ensemble de l'activité**, et
- anticiper sur des **sujets stratégiques** de l'entreprise : prise de part de marché, intégration des contraintes réglementaires, innovation, développement durable, optimisation des marges...

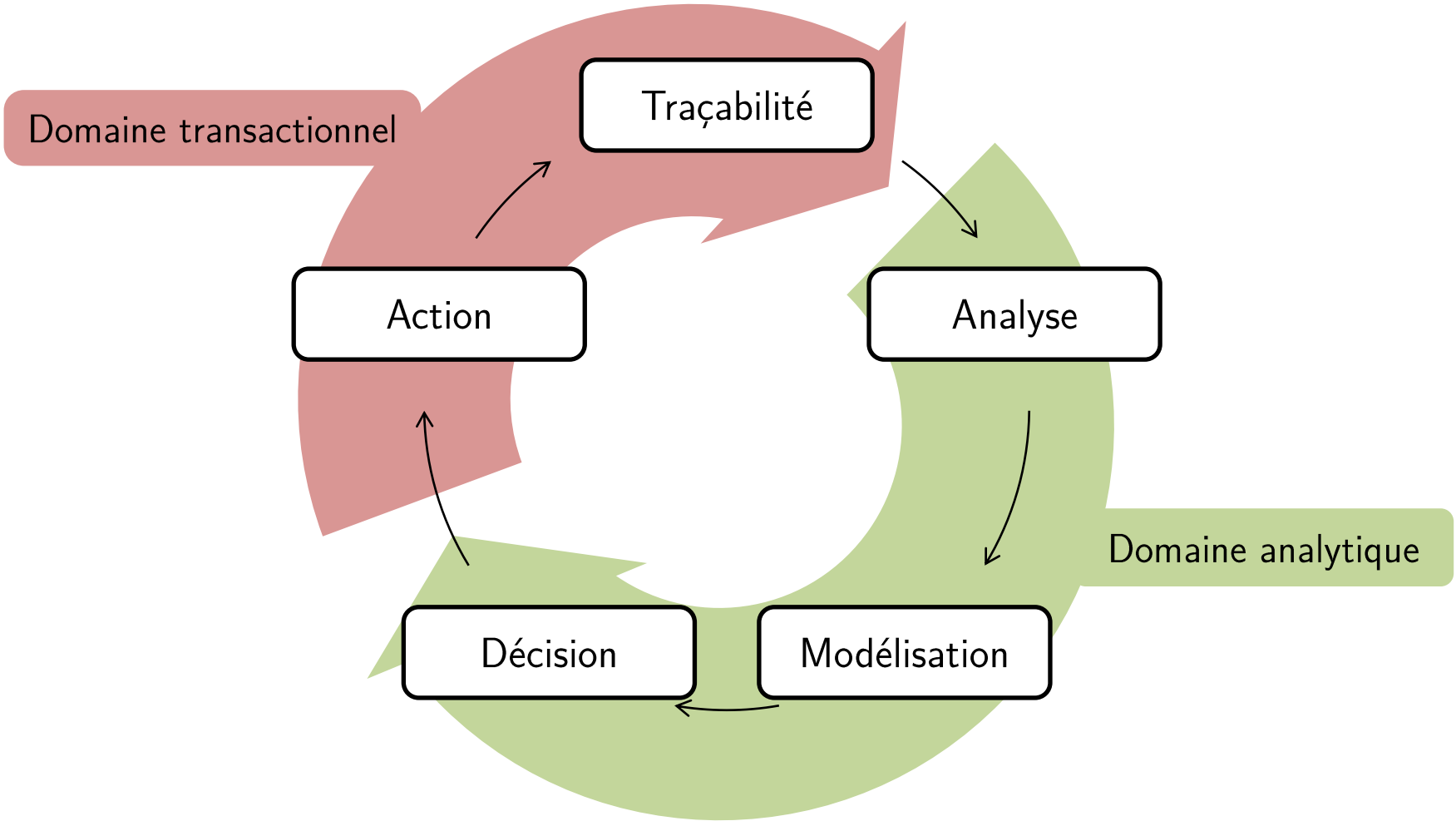
* en anglais : DSS pour Decision Support System ou encore BI pour Business Intelligence

Contexte : Les S.I.D.

- **Besoin** : prise de décisions stratégiques et tactiques
- **Pourquoi** : Soutenir le management et orienter la stratégie
- **Qui** : les décideurs et les gestionnaires
- **Comment** : répondre aux demandes d'analyse des données, dégager des informations qualitatives nouvelles
  - Pourquoi mon chiffre d'affaires a baissé?
  - Quel est le montant des ventes de la semaine X?
  - Les régions qui consomment le plus de viande?
  - Qui sont mes meilleurs clients?
  - ... etc.

Contexte : Les S.I.D.

Le cycle décisionnel dans l'entreprise

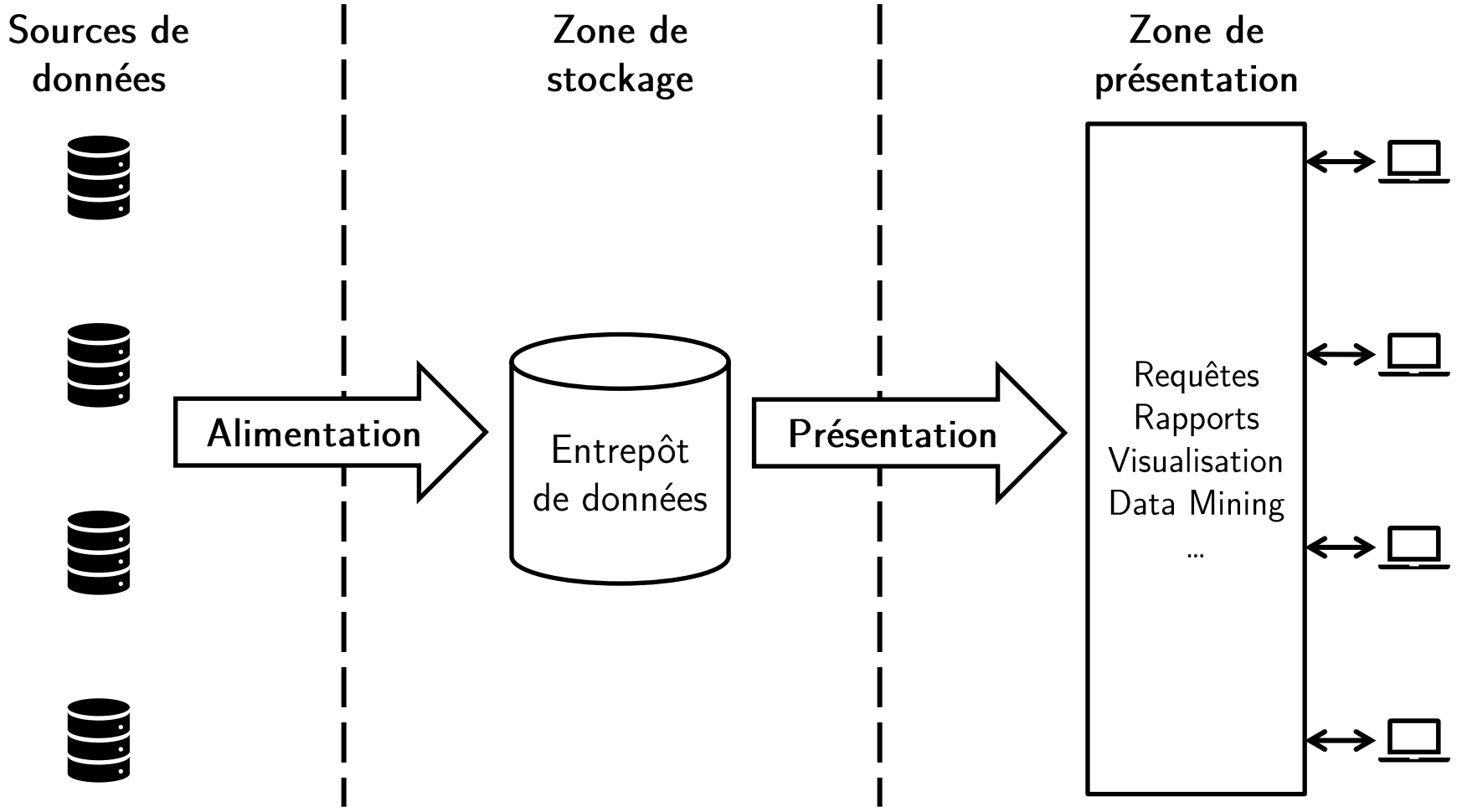


Contexte : Les S.I.D.

Comment répondre à ces demandes ?

- La mise en place d'un système d'information dédié **aux applications décisionnelles**
- Ce type d'application utilise en règle générale **un entrepôt de données** ('DataWarehouse' , DW)
- Un entrepôt de données stocke des données transverses provenant de plusieurs **systèmes opérationnels** et fait l'objet d'analyse avant la prise de décision.

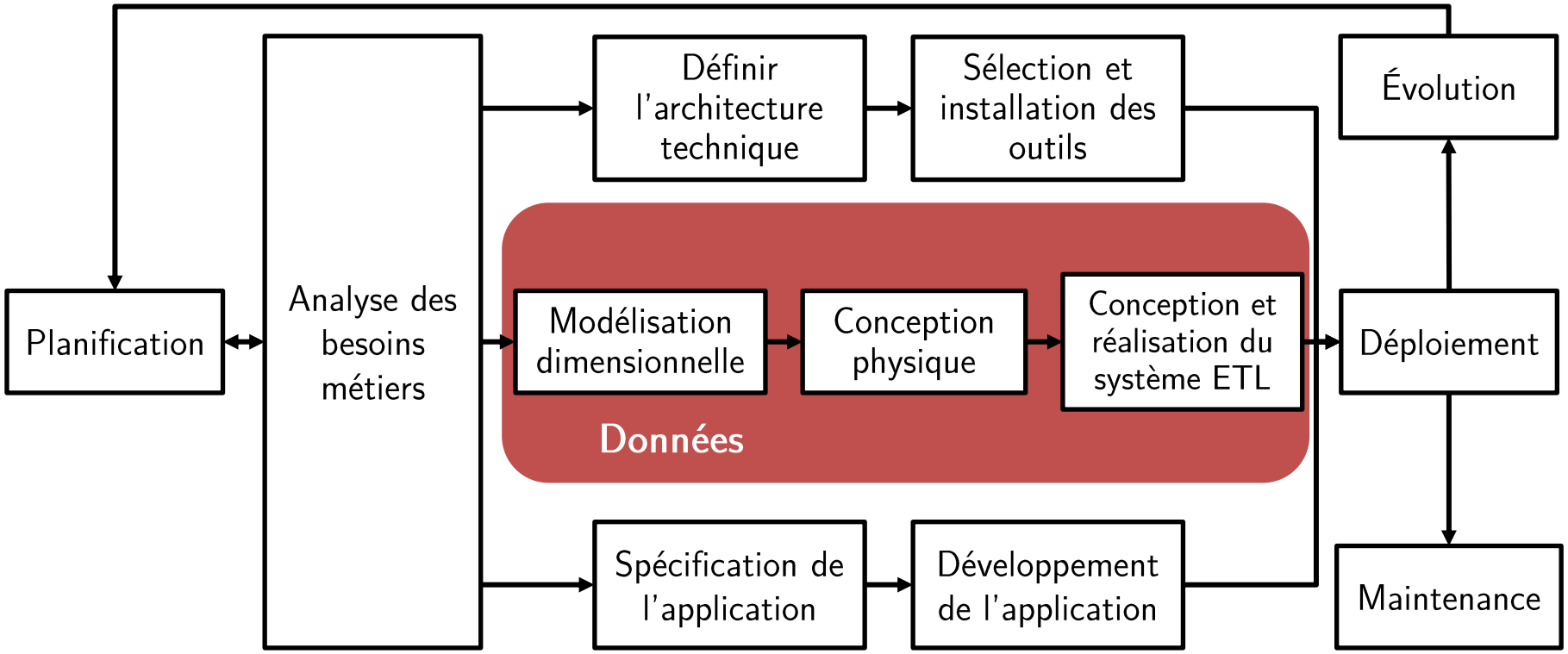
Architecture globale



L'objectif d'un E.D. est de rendre les données de l'organisation facilement accessibles et assurer que celles-ci sont cohérentes.

Cycle de vie d'un projet I.D.

Le cycle de vie définit les étapes de conception, développement et déploiement d'un Entrepôt de Données.



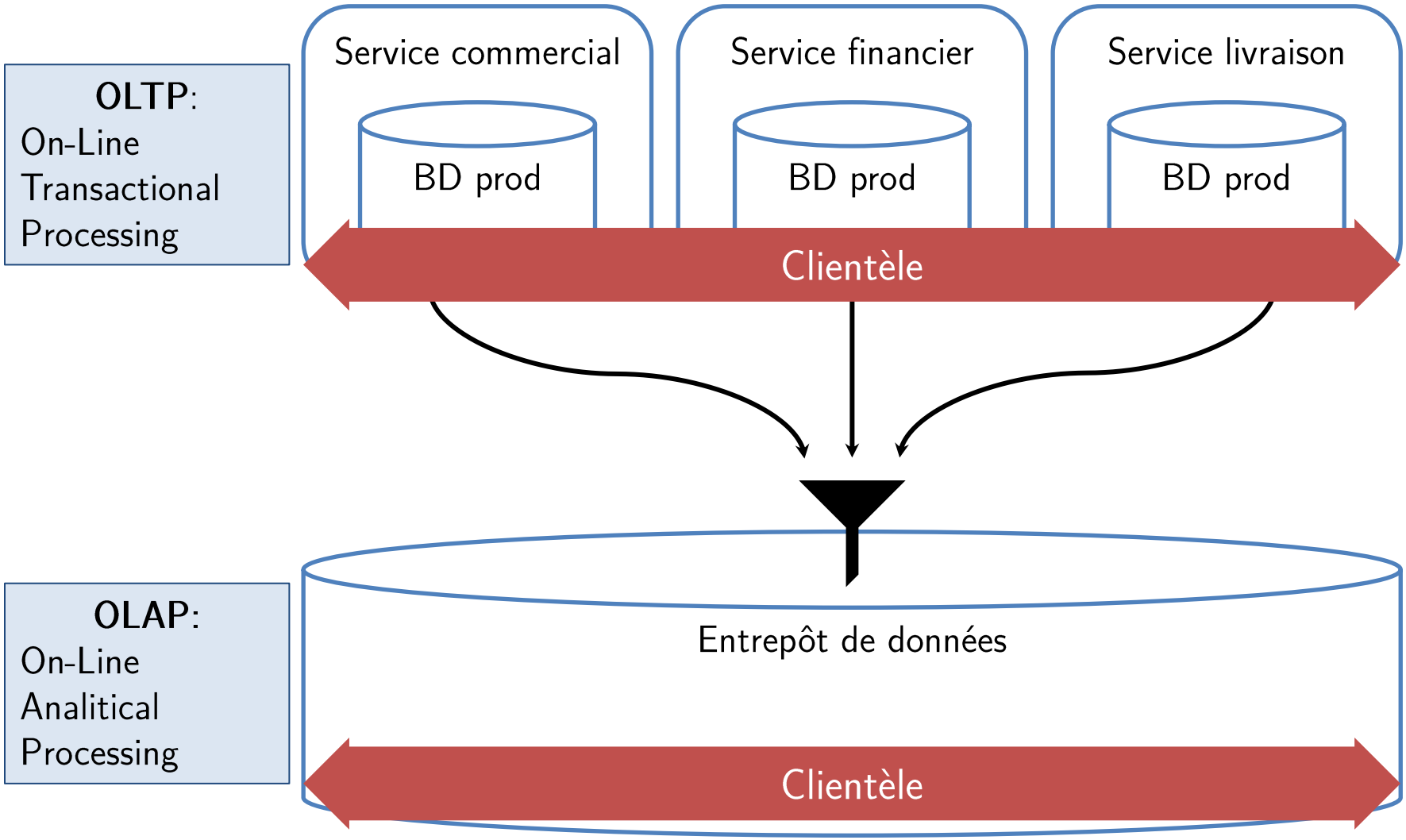
(Source Ralph Kimball)

3 composantes : humaine, technique et financière

Domaines d'utilisation des DW

- Banque
  - Risques d'un prêt, prime plus précise
- **Santé**
  - **Épidémiologie**
  - **Risque alimentaire**
- Commerce
  - Ciblage de clientèle
  - Déterminer des promotions
- Logistique
  - Adéquation demande/production
- Assurance
  - Risque lié à un contrat d'assurance (voiture)
- ...

B.D. production et B.D. décision



B.D. production et B.D. décision

OLTP	DW
Orienté transaction	Orienté analyse
Orienté application	Orienté sujet
Données courantes	Données historisées
Données détaillées	Données agrégées
Données évolutives	Données statiques
Utilisateurs nombreux, administrateurs/opérationnels	Utilisateurs peu nombreux, manager
Temps d'exécution: court	Temps d'exécution: long

OLTP : On-Line Transactional Processing
DW : Data Warehouse

ENTREPÔTS DE DONNÉES

Modélisation des entrepôts de données

Objectifs

- Quelle structure permet d'avoir les fonctionnalités requises pour un entrepôt de données?
- Quelles sont les techniques utilisées pour bien concevoir un entrepôt de données?
- Quels sont les indicateurs d'une bonne conception?

L'Entrepôt de Données (E.D.)

Définition

« une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » W. H. Inmon (1996)

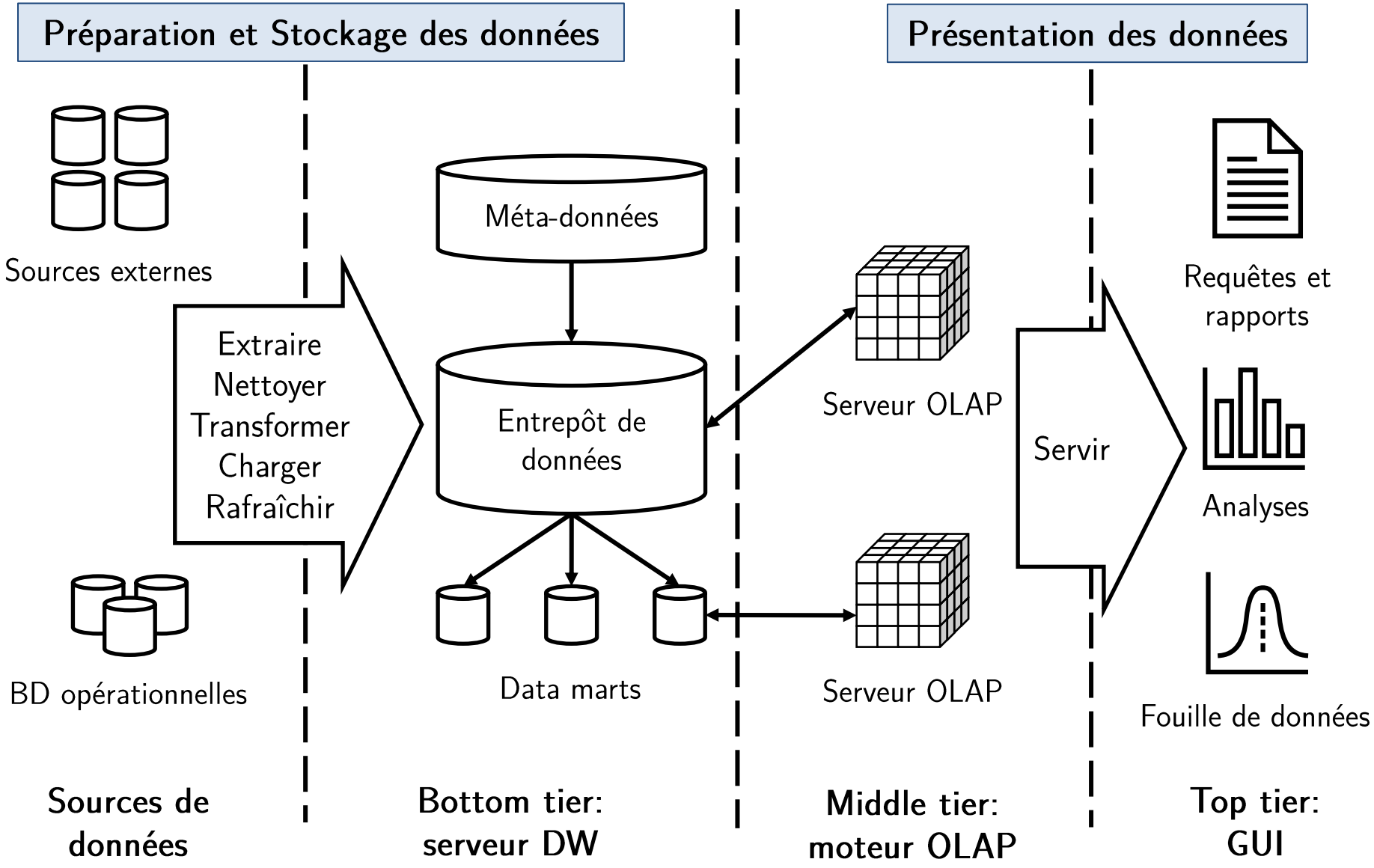
Trois fonctions essentielles :

1. **collecte** de données des bases existantes
2. **gestion** de données dans l'entrepôt
3. **analyse** de données pour la prise de décision

Les caractéristiques des E.D.

1. Données orientées sujet :
   - Regroupe les informations des différents métiers
   - Organisées par thème, par exemple les ventes
2. Données intégrées :
   - Données de plusieurs sources parfois hétérogènes
   - Mise en forme pour assurer la cohérence
3. Données non volatiles :
   - Conserver la traçabilité des informations et des décisions prises
   - Les données ne disparaissent pas et ne changent pas
4. Données datées :
   - Les données persistent dans le temps
   - Mise en place d'un référentiel temps

Architecture détaillée



Les caractéristiques des E.D.

En résumé

- Un E.D. est une structure informatique construite selon une approche **dimensionnelle** dans laquelle est stocké un volume important de données provenant des systèmes opérationnels.
- Le système informatisé gérant l'entrepôt de données (B.D. dimensionnelle) est un SGBD de type relationnel.

Modèle dimensionnel vs ER

- Modèle entité-relation (ER) – Merise :
  - Représente les données sous la forme d'entités (tables) et de relations (références ou tables);
  - Normalisation du schéma (ex: 1FN/2FN/3FN).
- Modèle dimensionnel:
  - Représente les données comme des faits et des dimensions;
  - Les dimensions ne sont pas normalisées.
- Avantages du modèle dimensionnel:
  - Compréhensibilité:
    - Les données sont regroupées selon des catégories d'affaires qui ont un sens pour les utilisateurs d'affaires;
  - Performance:
    - La dénormalisation évite les jointures coûteuses;
    - Autres optimisations (ex: index de jointure en étoile).

La modélisation dimensionnelle

Définition

Technique de conception logique permettant de structurer les données de manière à les rendre intuitives aux utilisateur d'affaires et offrir une bonne performance aux requêtes.

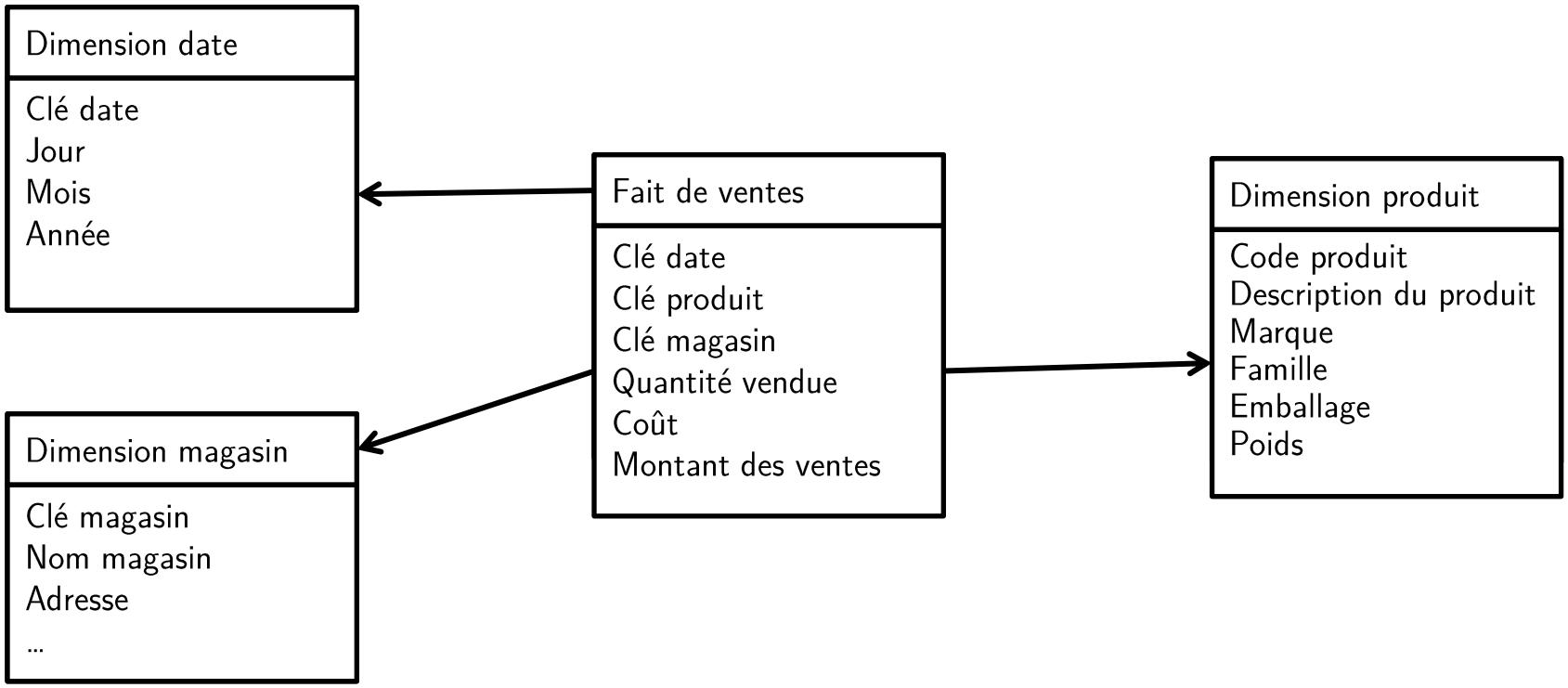
Caractéristiques :

- Divise les données en faits et dimensions;
- 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;
- Schéma en étoile: une table de faits entourée de plusieurs tables de dimension (normalement entre 8 et 15).
- Modélisation conceptuelle : peu dépendante de l'implémentation.

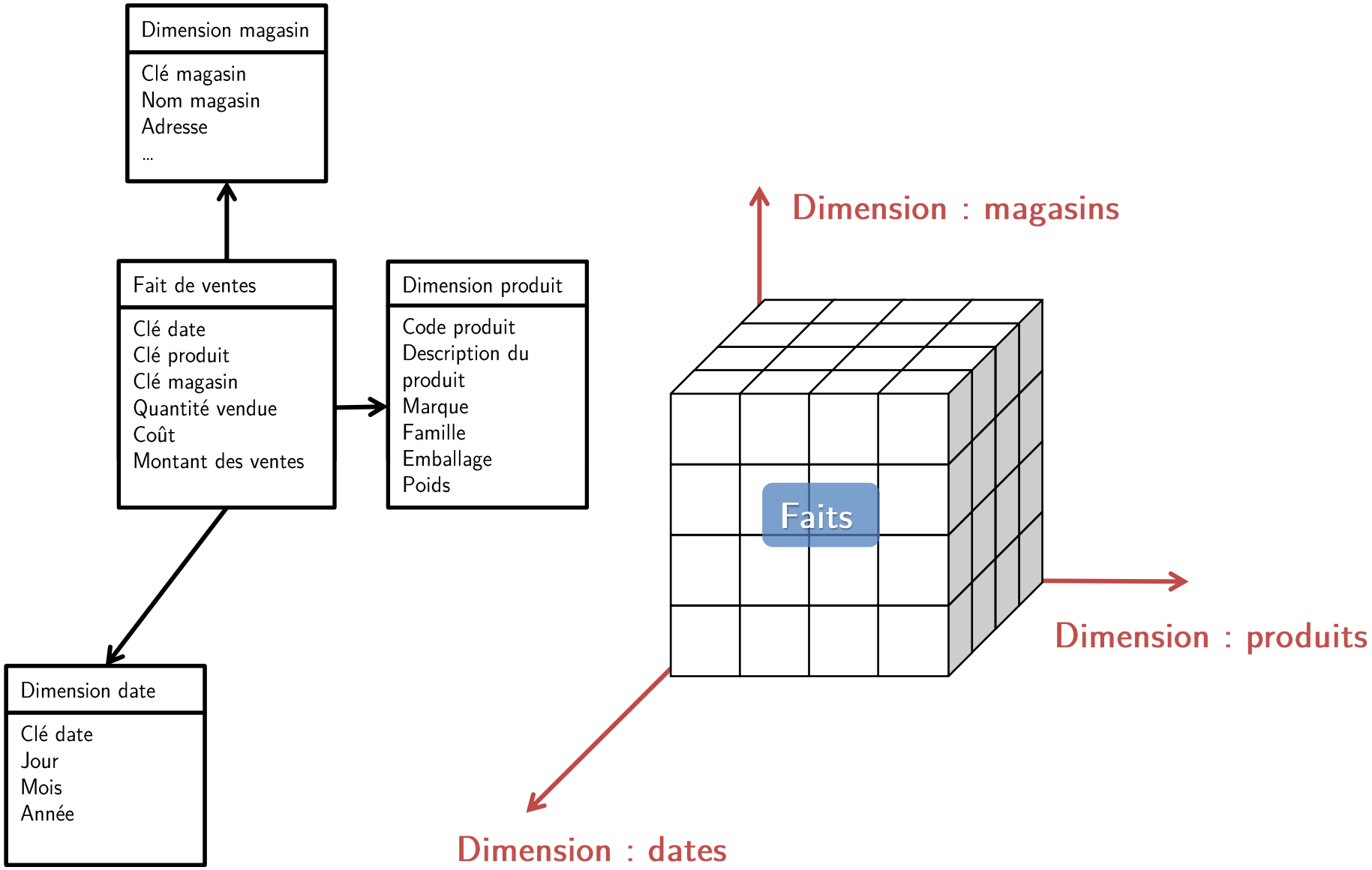
Le modèle dimensionnel

Définition

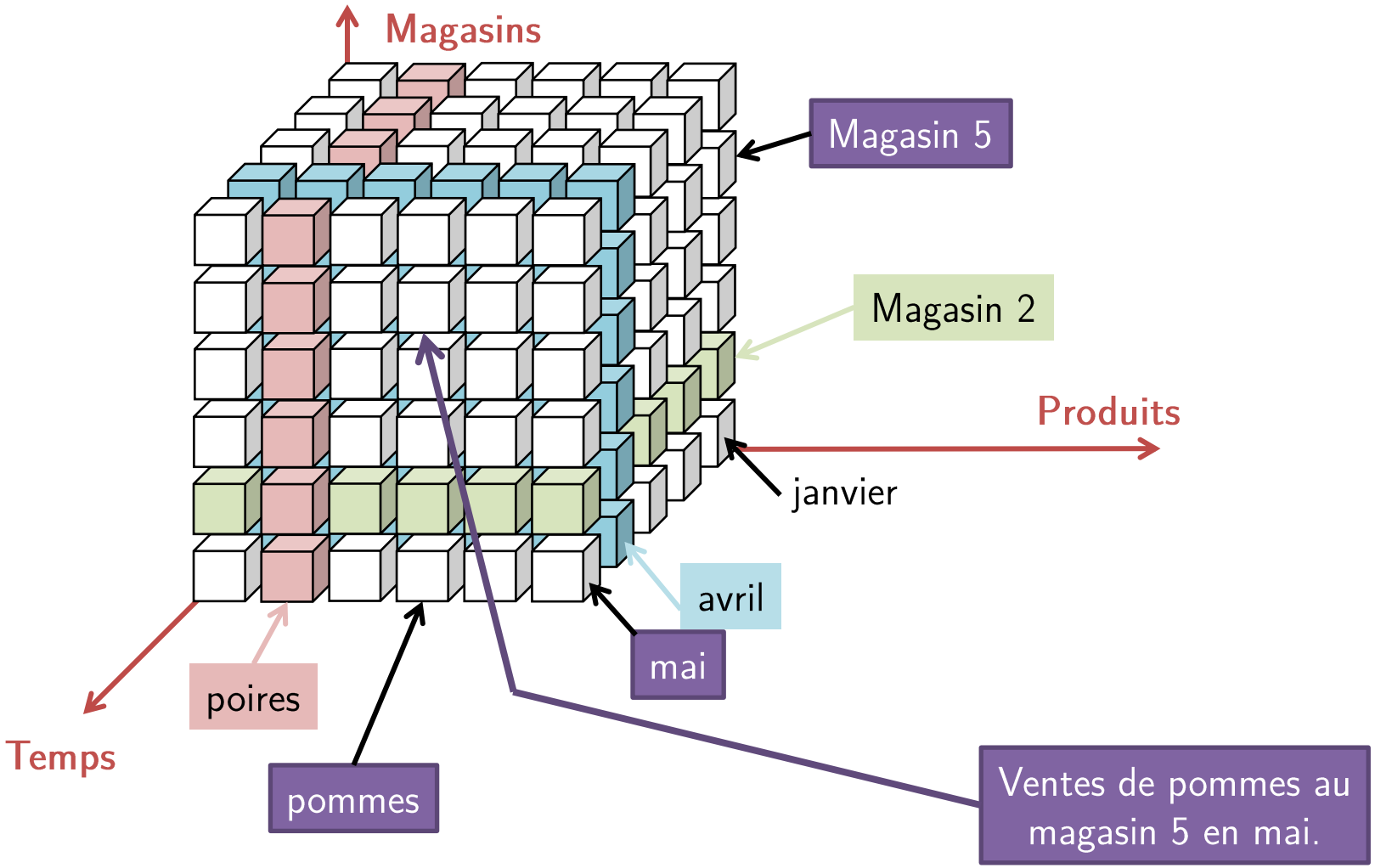
Un **modèle dimensionnel** est composé d'une **table de faits** contenant une clé multiple et d'un ensemble de **tables de dimensions** avec une clé primaire chacune



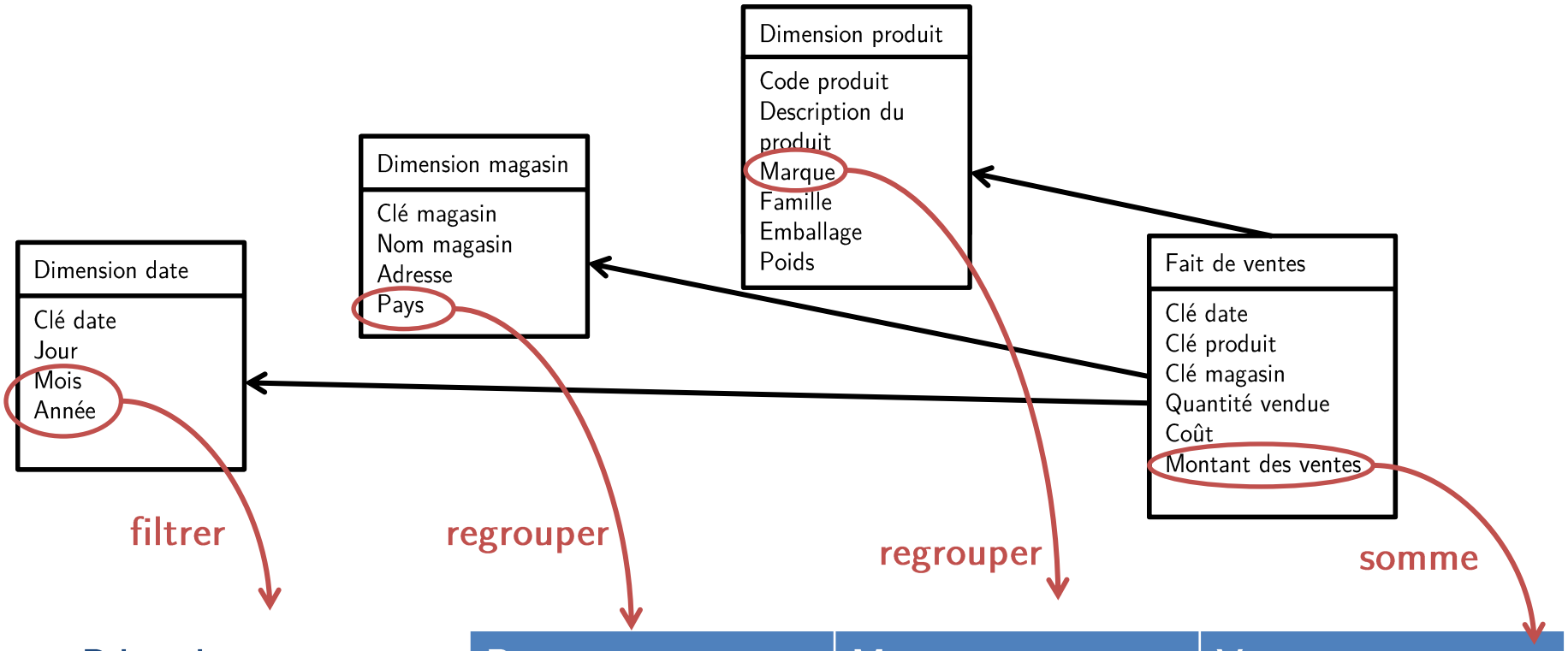
Modèle dimensionnel : intuition



Modèle dimensionnel : intuition



Modèle dimensionnel : rapports



Bilan des ventes pour janvier 2022

Pays	Marque	Ventes en euros
Allemagne	PowerClean	2 035
Allemagne	Zippy	707
France	Clean Fast	2 330
France	Zippy	527

Modèle dimensionnel : rapports

Attributs dimensionnels du rapport

```
SELECT
        store.country AS « Pays »,
        product.brand AS « Marque »,
        sum(sales_facts.sales_euros) AS « Ventes en euros »
FROM
        store, product, date, sales_facts
WHERE
        date.month_name = "January" AND
        date.year = 2022 AND
        store.store_key = sales_facts.store_key AND
        product.product_key = sales_facts.product_key AND
        date.date_key = sales_facts.date_key
GROUP BY
        store.country,
        product.brand
```

Métrique d'agrégation des faits

Tables impliquées dans la requête

Filtre du rapport

Jointures entre les tables de faits et de dimensions

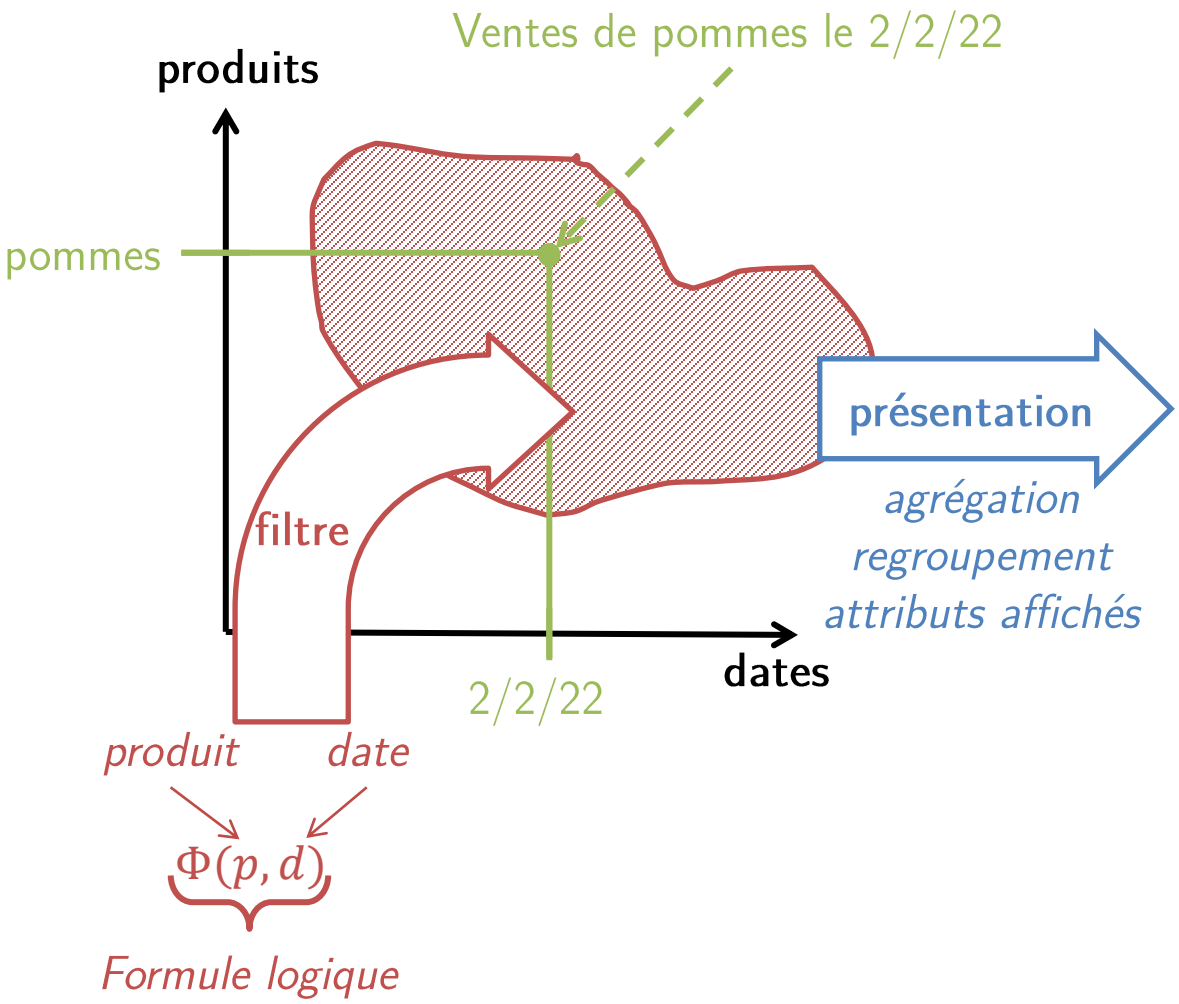
Agrégation au sein du rapport

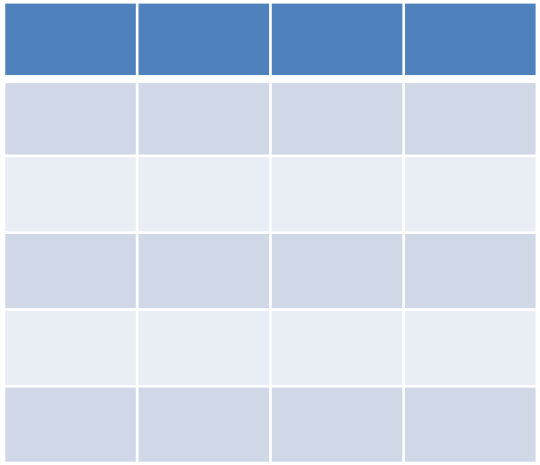
Bilan des ventes pour janvier 2022

Pays	Marque	Ventes en euros
Allemagne	PowerClean	2 035
Allemagne	Zippy	707
France	Clean Fast	2 330
France	Zippy	527

Modèle dimensionnel

Espace de faits





Le modèle dimensionnel

Définition

Un **modèle dimensionnel** est composé d'une **table de faits** contenant une clé multiple et d'un ensemble de **tables de dimensions** avec une clé primaire chacune

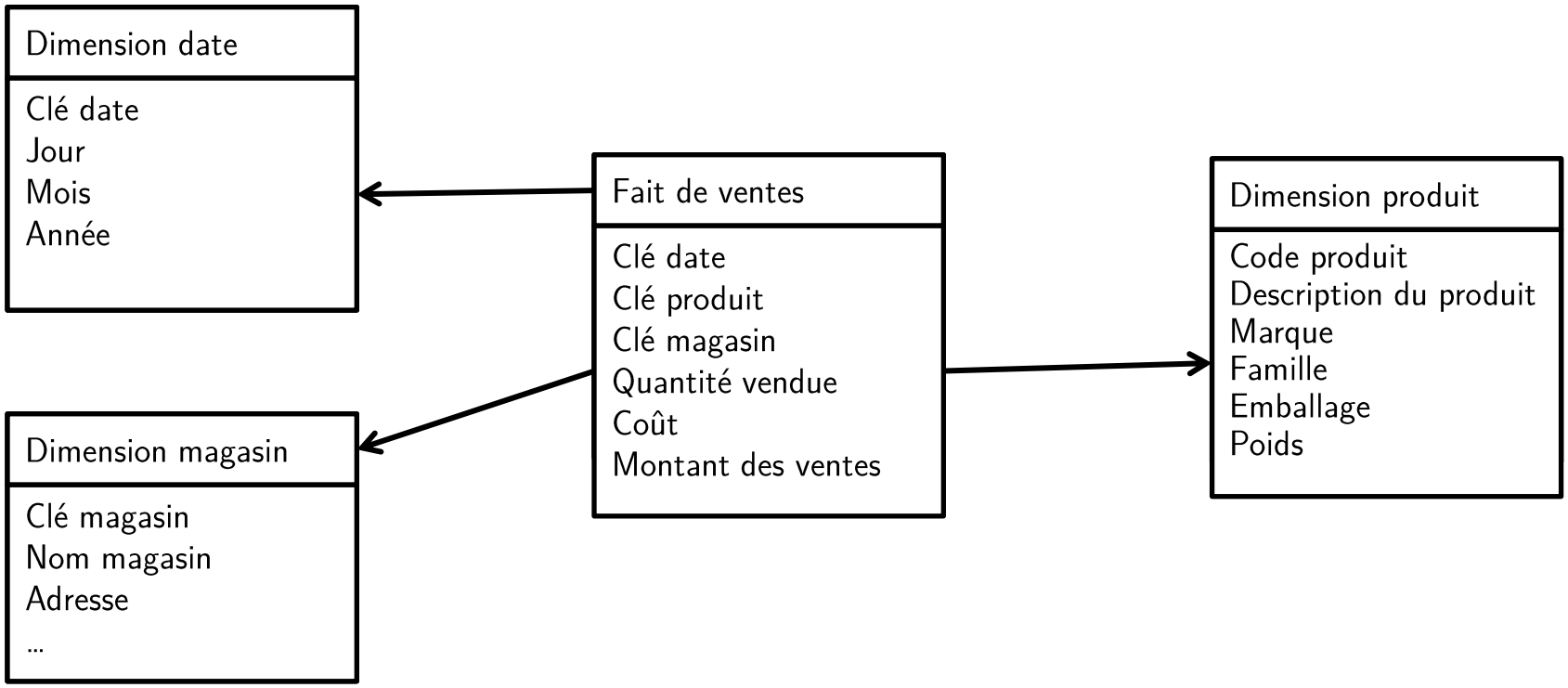
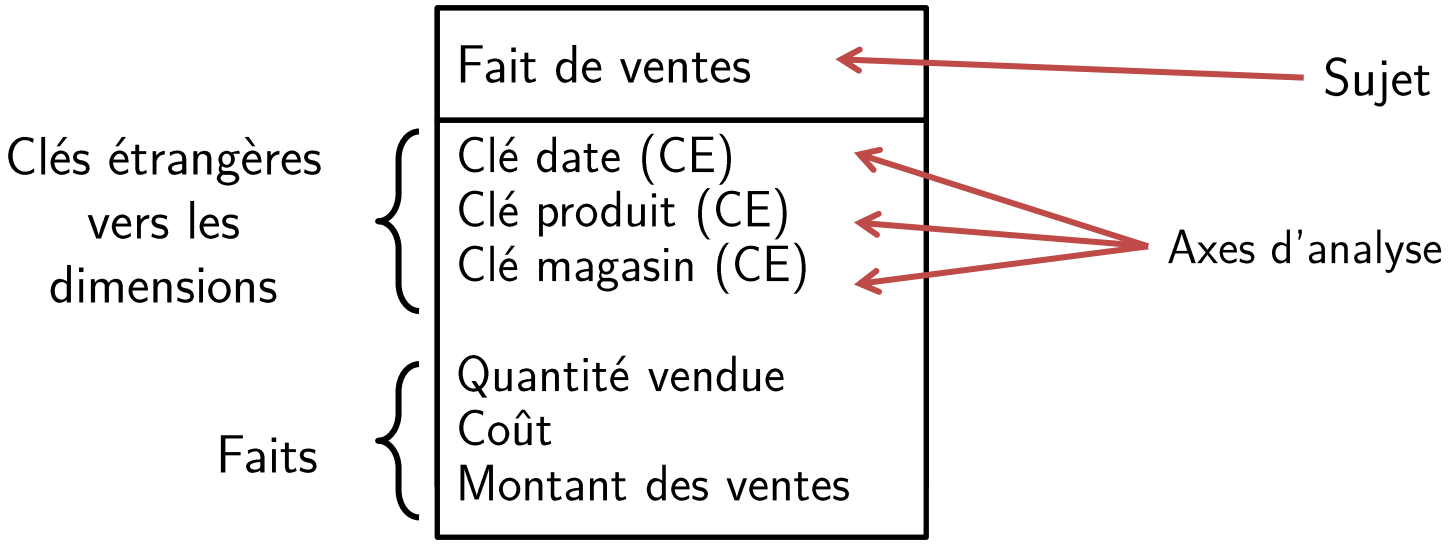


Table de faits

Contient les données observables (les faits) sur le sujet étudié selon divers axes d'analyse.



- Contient ce que l'on souhaite mesurer et les clés étrangères des axes d'analyse.
- Correspond à un seul processus d'affaires : 1 table de faits → 1 processus.
  Ex: ventes, commandes, livraisons effectuées
- Stockent des mesures générées par les événements du processus
  Ex: réception d'une commande, envoi d'une commande, etc.
- Les faits "prennent leur valeur" au moment où l'évènement d'affaires survient (aspect temporel important).

Types de mesure

- **Additive** : additionnable suivant toutes les dimensions

Exemple : quantité vendue

On peut parler de la quantité vendue en 2020 : $\sum_{f.date.year=2020} f.qt$

Ou bien de la quantité de lessive vendue : $\sum_{f.produit.catégorie=lessive} f.qt$

- **Semi additive** : additionnable suivant certaines dimensions

Exemple : solde d'un compte

On peut parler du solde combiné de tous les clients au début de l'année civile:

$$\sum_{f.date.detail="1/1/2022,00:00:00"} f.solde$$

Par contre, la somme des soldes d'un client n'a pas de sens : $\sum_{f.client.id=1231224} f.solde$

- **Non additive** : fait numérique non additionnable quelque soit la dimension

Exemple : prix unitaire, note au questionnaire satisfaction

L'addition sur n'importe quelle dimension donne un nombre dépourvu de sens.

Attention : des mesures non additives peuvent tout de même être agrégées, mais de manière plus fine qu'une somme (moyenne, min/max, valeur majoritaire,...).

Mesures et additions

- Le critère pour déterminer si une mesure est additive/semi-additive/non-additive est formulé en termes de sommes sur une dimension.
- Pourquoi ne pas considérer plutôt d'autres formes d'agrégation (moyenne, valeur maximale…) ?

- La somme est la meilleure des agrégations !
- Si on peut sommer, on peut aussi calculer des moyennes, ou n'importe quelle forme d'agrégation.
- La somme est compositionnelle :

$$(f_1 + f_2 + f_3 + f_4 + f_5) = (f_1 + f_2) + (f_3 + f_4 + f_5)$$

- La moyenne ne l'est pas :

$$moy(f_1, f_2, f_3, f_4, f_5) = \frac{f_1 + f_2 + f_3 + f_4 + f_5}{5}$$

$$\neq$$

$$moy\big(moy(f_1, f_2), moy\,(f_3, f_4, f_5)\big) = \frac{\frac{f_1 + f_2}{2} + \frac{f_3 + f_4 + f_5}{3}}{2}$$

Mesures versus attributs de dimension

- Mesures :
  - Dépendent d'un événement d'affaires;
  - Ont souvent des valeurs continues (ou un grand nombre de valeurs discrètes possibles);
  - Servent dans le calcul d'indicateurs de performance;
  - Ex: montant total et quantité d'une commande.
- Attributs (numériques) de dimension :
  - Indépendants des événements d'affaires;
  - Ont souvent des valeurs discrètes;
  - Servent à filtrer ou étiqueter les faits;
  - Ex: jour et heure d'une transaction, âge d'un client, etc.

Granularité de la table de faits

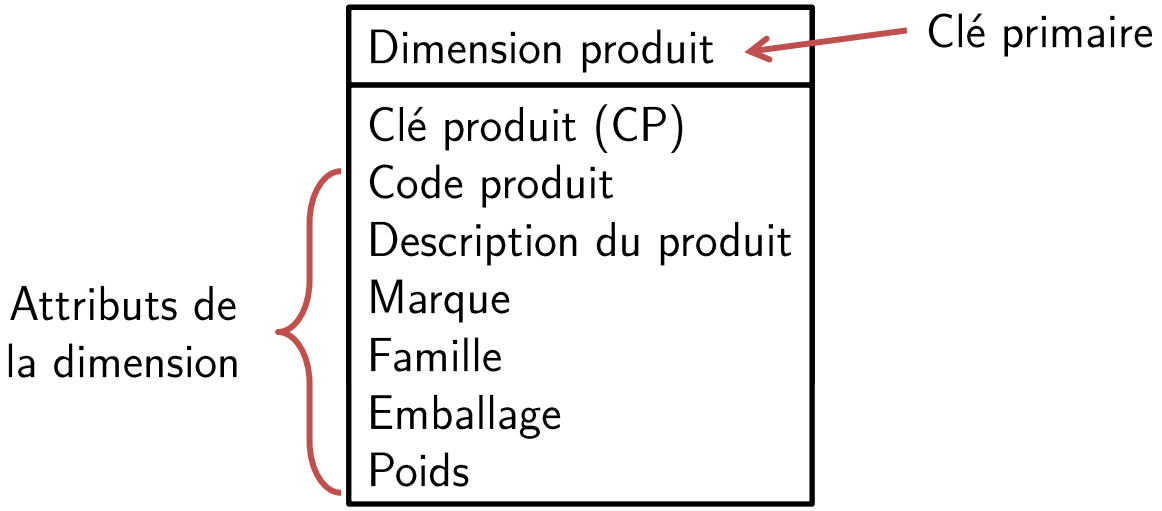
- **Répondre à la question** :

  Que représente un enregistrement de la table de faits?
- La granularité (grain) définit le niveau de détail d'une table :
  - Ex: une ligne de commande par client, par produit et par jour
  - Une transaction d'un point de vente par produit, par magasin, par jour et par heure
- Toutes les lignes de la table doivent avoir le même grain.
- Doit être le plus fin possible (atomique) pour le processus d'affaires:
  - Permet de faire des requêtes plus précises et imprévues;
  - Déterminé par les réalités physiques des sources de données.
- Détermine les dimensions du modèle.

Fait de ventes
Clé date (CE) Clé produit (CE) Clé magasin (CE) Quantité vendue Coût Montant des ventes

Table de dimensions

Contient des informations descriptives des axes selon lesquels vont être étudiées les données observables (faits).



- Ensemble hautement corrélé d'attributs regroupés selon les objets clés de l'entreprise :
  - Ex: produits, clients, employés, installations, etc.
- Rôles des attributs:
  - Filtrer / restreindre les requêtes (ex: ville, catégorie produit, etc.);
  - Étiqueter les résultats (ex: champs descripteurs).

Table de dimension

Dimension produit
Clé produit (CP)
Code produit
Description du produit
Marque
Famille
Emballage
Poids

- Dimension = axe d'analyse

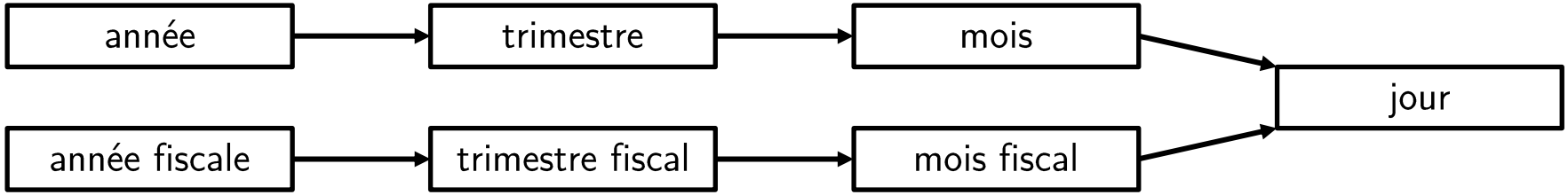
  Client, produit, période de temps...
- Contient souvent un grand nombre de colonnes

  l'ensemble des informations descriptives des faits
- Propriétés des attributs:
  - Descriptif (ex: chaînes de caractères);
  - De qualité (ex: aucune valeur manquante, obsolète, erronée, etc.);
  - Principalement des valeurs discrètes (ex: jour, âge d'un client);
- Contient beaucoup moins d'enregistrements (lignes) qu'une table de faits

- Les attributs sont la source de contraintes dans les requêtes SQL
- Les tables dimensionnelles sont le point d'entrée de L'Entrepôt de Données
- La puissance analytique de l'entrepôt est proportionnelle à la richesse et la qualité des attributs dimensionnels.

Hiérarchies dimensionnelles

- Un ensemble d'attributs ayant une relation hiérarchique (x est inclus dans y);
- Définissent les chemins d'accès dans les données (drill-down paths);
- Simples:
  - Temps: année → mois → semaine → jour → heure;
  - Produit: famille → catégorie → marque → produit;
  - Lieu: pays → province → région → ville → code postal.
- Multiples:



- Normalement dans **une seule** table de dimension.

Choix des dimensions

- Demande le jugement et l'intuition du modélisateur;
- Plus on a d'attributs non-corrélés dans une dimension plus la table correspondante aura de lignes (explosion combinatoire):
  - Ex: 10,000 produits x 100 magasins = 1,000,000 de lignes dans une table de dimension ProduitMagasin.
- Règles:
  - Les dimensions sont observables au niveau du grain de la table de faits (font partie de l'évènement d'affaires);
  - Les attributs non-corrélés vont dans des dimensions séparées.

La dimension Temps

- Commune à l'ensemble de l'entrepôt de données (DW)
- Reliée à toute table de faits

Dimension Temps
Clé temps (CP)
Jour
Mois
Trimestre
Semestre
Année
Num_jour_dans_année
Num_semaine_ds_année

La dimension Temps

La dimension suivante est-elle satisfaisante ?

Dimension Temps
Clé temps (CP)
Jour
Mois
Trimestre
Semestre
Année
Heure
Minute
Seconde

La dimension Temps

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 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/an contre 31 000 000 lignes/an.

- **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).

Processus de modélisation

1. Choisir le sujet d'étude
2. Choisir la granularité du processus
3. Identifier les dimensions
4. Choisir les attributs de la table de fait

Processus de modélisation : Étape 1

Choisir le(s) processus d'entreprise à modéliser

- Choisir l'activité ayant un intérêt pour les utilisateurs
  - Quelles sont les mesures de performance à analyser
    - Les achats? Les commandes? la gestion des stocks? la comptabilité générale?
- Un modèle dimensionnel par activité
- Se focaliser sur les activités et non sur les services

Processus de modélisation : Étape 2

Choisir la granularité du processus

- Spécifier exactement ce que représente une ligne de la table de faits
- Le grain représente le niveau de détail des mesures de la table de faits
- Quelques exemples :
  - Une ligne par ticket de caisse dans un centre commercial
  - Une ligne de prestation sur la note d'honoraires d'un médecin

Attention, c'est une étape critique

Processus de modélisation : Étape 3

Choisir les dimensions

- Les dimensions permettent de décrire les données du processus à analyser
- Qui s'appliquent à chaque ligne de la table de faits
- Lister les attributs par dimension
- En général, ce sont des champs textuels

C'est une étape où les utilisateurs jouent un rôle primordial

Processus de modélisation : Étape 4

Identifier les faits numériques

- Que mesure t'on?
- Tous les faits mesurés sont au grain défini dans l'étape 2.
- Les faits sont autant que possible des données numériques, additives.
- Quantité commandée, coût en euros.