- **Données, informations et connaissance**

- Données: Observations ou faits bruts concernant un phénomène, un domaine de connaissances ou les caractéristiques de certaines entités. Objet sans contexte.
- Exemple : Prix, poids,…
- Type :Alphabétique, numérique, texte, image, audio.

- Informations: Données qui ont été traitées et placées dans un contexte significatif et utile pour un utilisateur particulier.
- Le traitement de données ajoute de la valeur aux données brutes pour leur conférer une signification.

- La qualité de l'information

Temps

- Opportunité
- Fréquence
- Clarté
- Exactitude
- Pertinence
- Caractère exhaustif
- Concision

connaissance

- La connaissance est le résultat d'une analyse cognitif de plusieurs informations.
- La connaissance est humaine.

- Es ce que plus on s'informe et plus on a de connaissance?

Problème

a) Sous information

b) Sur information

Type de connaissance

a) connaissance explicite

b) Connaissance tacite

Plan

- Ce qu'est le data warehouse ?
- Un modèle multidimensionnel
- Architecture d'un data warehouse
- Implémentation d'un data warehouse
- Autres développements de la technologie data cube
- Data warehousing et data mining

Ce qu'est le data warehouse ?

- Différentes définitions
  - Une BD d'aide à la décision qui est maintenue séparément de la base opérationnelle de l'organisation.

- "Un data warehouse est une collection de données concernant un sujet particulier, varie dans le temps, non volatile et où les données sont intégrées."—W. H. Inmon
- Data warehousing:
  - Le processus qui permet de construire un data warehouse

Les caractéristiques des data warehouse

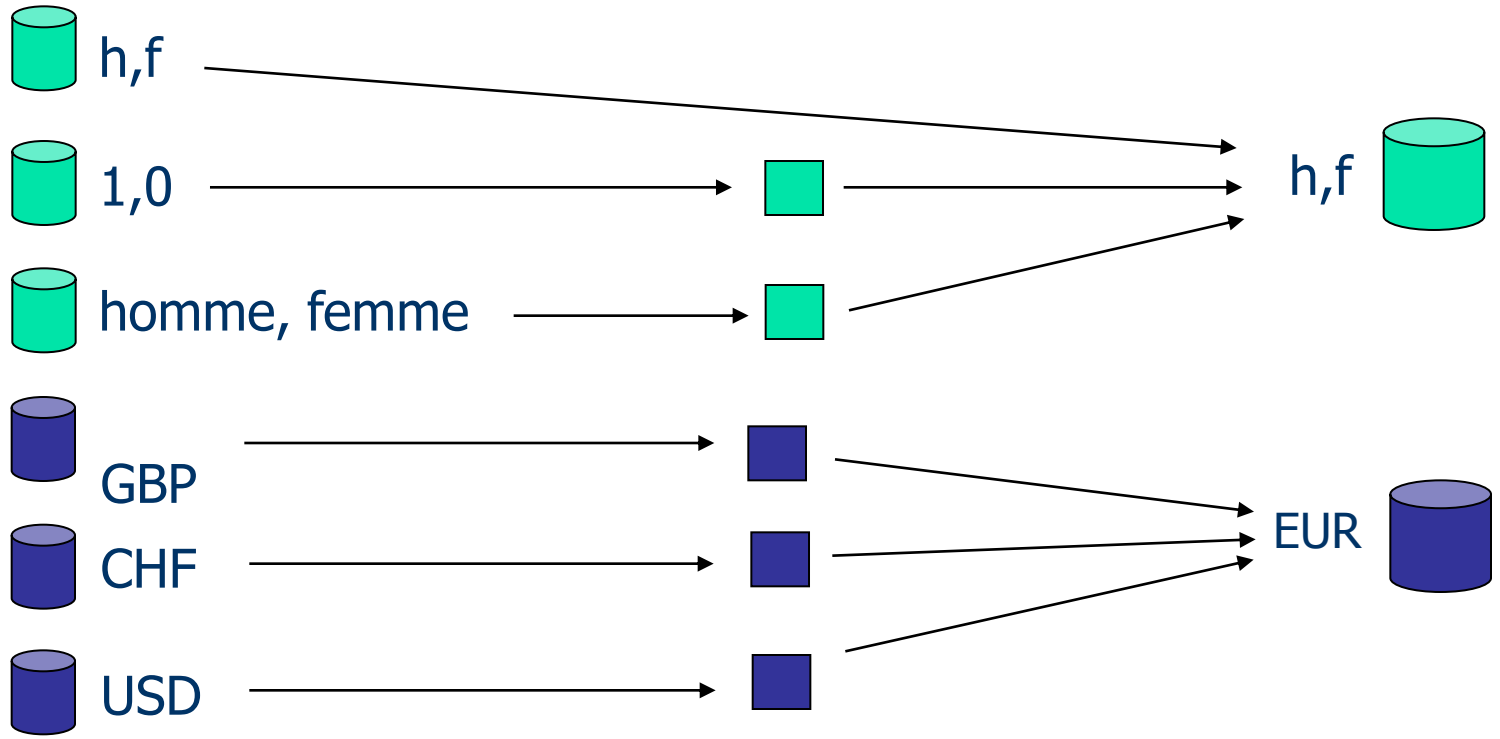
Orienté sujet

- Organisée autour d'un sujet bien précis, ex: client, produit, ventes.
- S'intéresse à la modélisation et l'analyse des données pour aider les décideurs, non pas pour des activités quotidiennes ou traitement transactionnel
- Fournit une vue simple et concise concernant un sujet particulier en excluant les données qui ne servent pas à la prise de décision

Les caractéristiques des data warehouse

Données intégrées:

- Normalisation des données
- Définition d'un référentiel unique



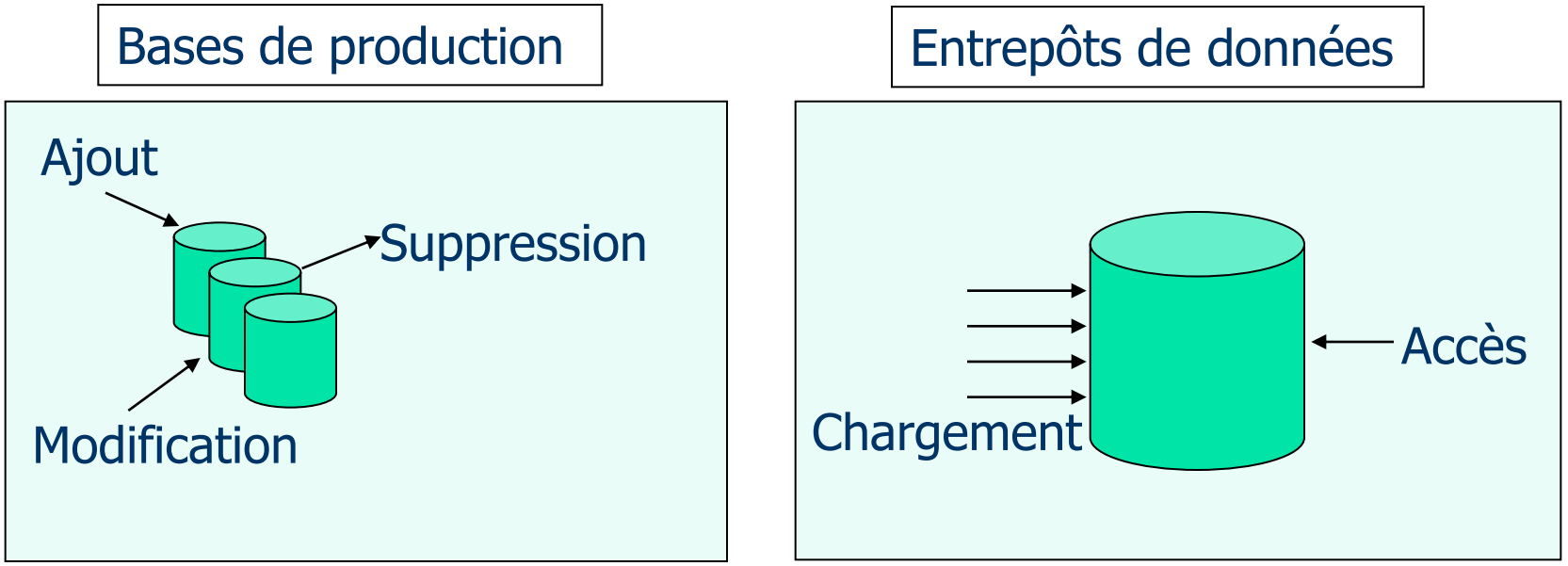
Données intégrées

- Construite en intégrant plusieurs sources de données possiblement hétérogènes
- Les techniques d'intégration et de nettoyage des données sont utilisées
  - Garantir la consistance des conventions de nommage (les attributs Nom et Nom_Famille dans BD1 et BD2 désignent la même chose)
  - structures de codage (l'attribut Nom est sur 15 char et 20 char sur BD1 et BD2; NSS est une chaîne dans BD1 et c'est un entier long dans BD2),
  - domaines des attributs (ex: cm vs pouce), etc.
  - C'est au moment où les données sont copiées dans le data warehouse qu'elles sont traduites

Les caractéristiques des data warehouse

Données non volatiles

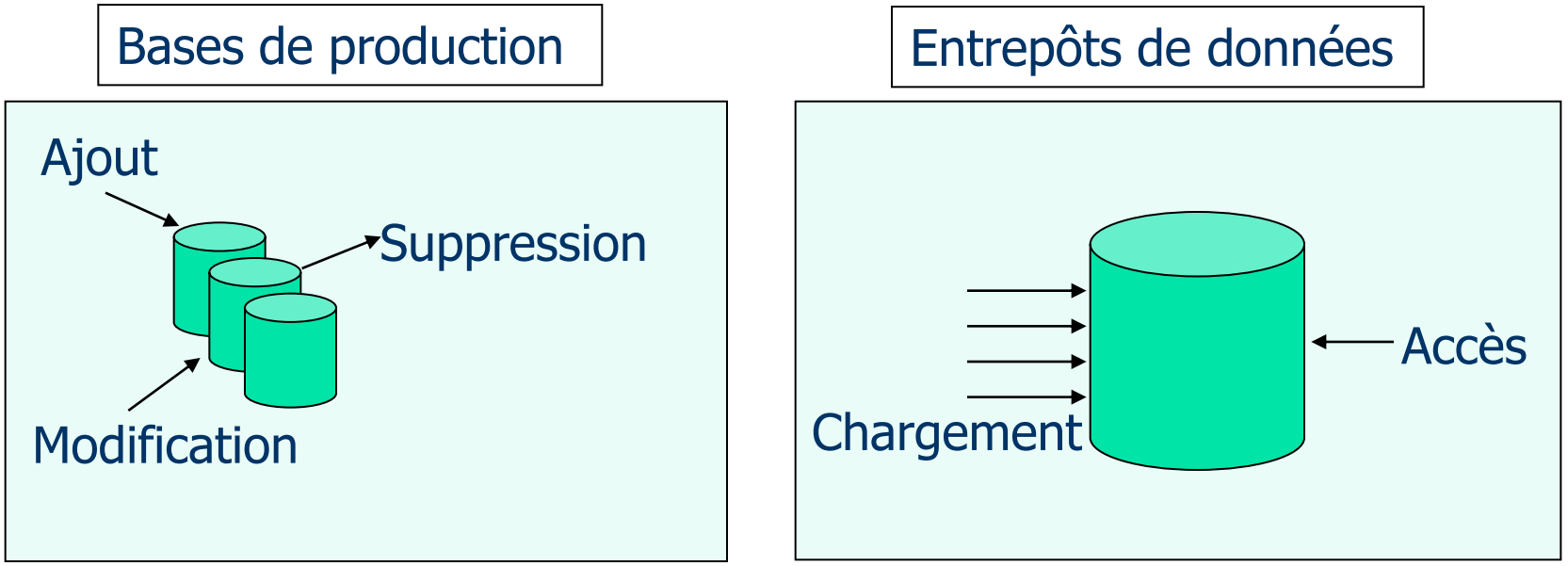
- Traçabilité des informations et des décisions prises
- Copie des données de production



- La portée temporelle des données dans un data warehouse est plus longue que celle des bases opérationnelles
  - Data warehouse: fournit des infos sous une perspective historique (ex: 5 à 10 dernières années)
- Dans un data warehouse, en général, chaque donnée fait référence au temps
  - Mais dans une base opérationnelle les données peuvent ne pas faire référence au temps

Les caractéristiques des data warehouse

3. Données non volatiles
   - Traçabilité des informations et des décisions prises
   - Copie des données de production



Data Warehouse est Non-Volatile

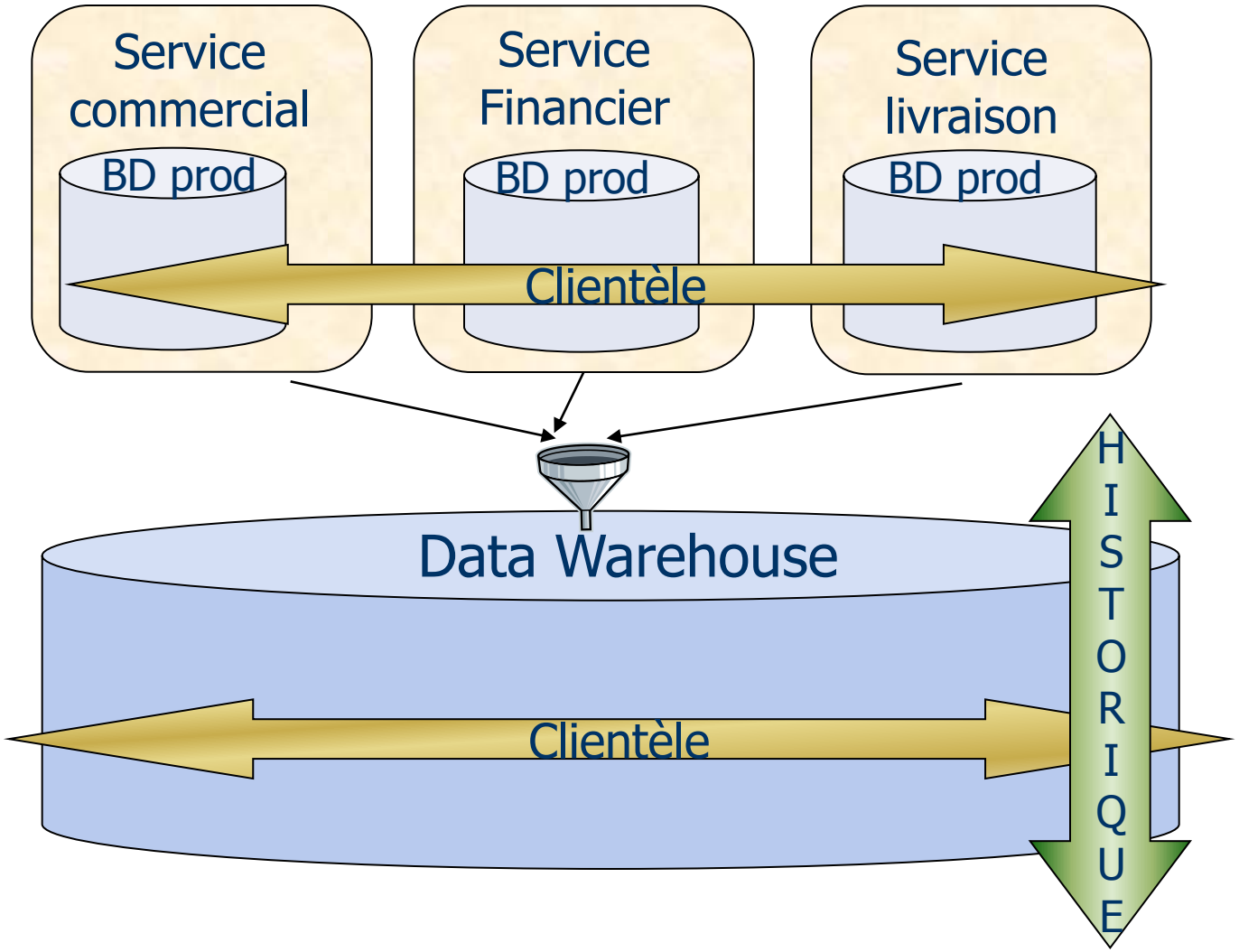
- Un support de stockage séparé
- Les mises à jour de la base opérationnelle n'ont pas lieu au niveau du data warehouse
  - N'a besoin que de deux opérations pour accéder aux données :
    - *Chargement initial des données et interrogation (lecture)*.

OLTP vs. OLAP

	OLTP	**OLAP**
utilisateurs	Tout le monde	décideurs
fonction	Opérations journalières	Aide à la décision
DB design	Orienté applications	Orienté sujet
data	courante, à jour, relationnel plat	historiques, résumés, multidimensionnelle intégrées
usage	répétitive	ad-hoc
accès	read/write index/hash sur clés	Beaucoup de scans
Unité de travail	Transactions courtes	Requêtes complexes
# enregistrement	dizaines	millions
# utilisateurs	Centaine(s)	Dizaine(s)
Taille BD	100MB-GB	100GB-TB
métrique	Exécution des transactions	Temps de réponse aux requêtes

Data Warehouse

OLTP: On-Line Transactional Processing



OLAP: On-Line Analitical Processing

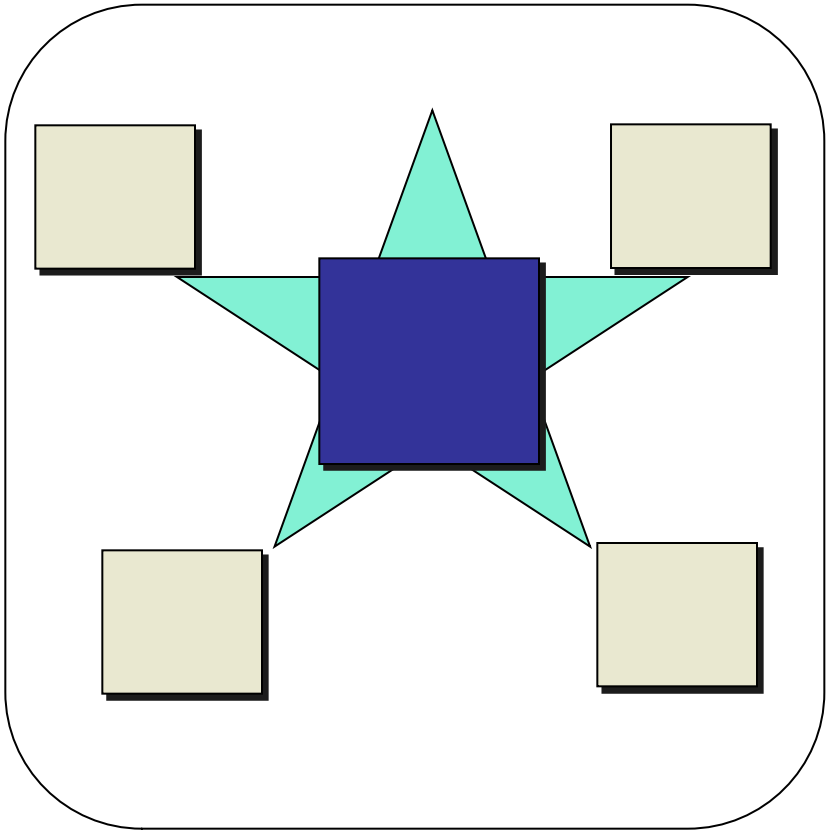
Des Tables aux Data cubes

- Un data warehouse est basé sur un modèle multidimensionnel où les données sont vues selon plusieurs dimensions
  - Les tables de dimension ex: Produit (nom_prod, marque, type), ou temps(jour, semaine, mois, trimestre, année)
  - La table de faits contient des mesures (ex: unités_vendues) et les clés externes faisant référence à chaque table de dimension

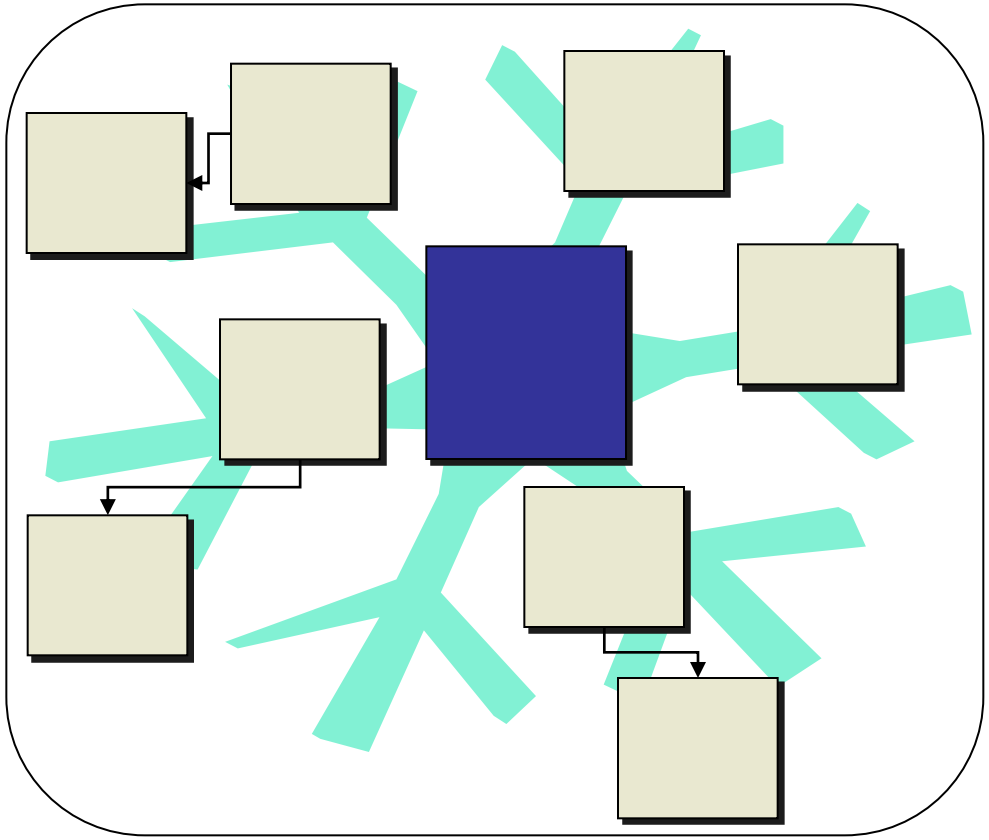
Modélisation Conceptuelle des Data Warehouses

- Dimensions & mesures
  - Schéma en étoile: Au milieu, une table de faits connectée à un ensemble de tables de dimensions
  - Schéma flocon de neige (snowflake): Un raffinement du précédent où certaines tables de dimensions sont normalisées (donc décomposées)
  - Constellation de faits: Plusieurs tables de faits partagent quelques tables de dimension (constellation d'étoiles)

Les types de modèles



Modèle en étoile



Modèle en flocon

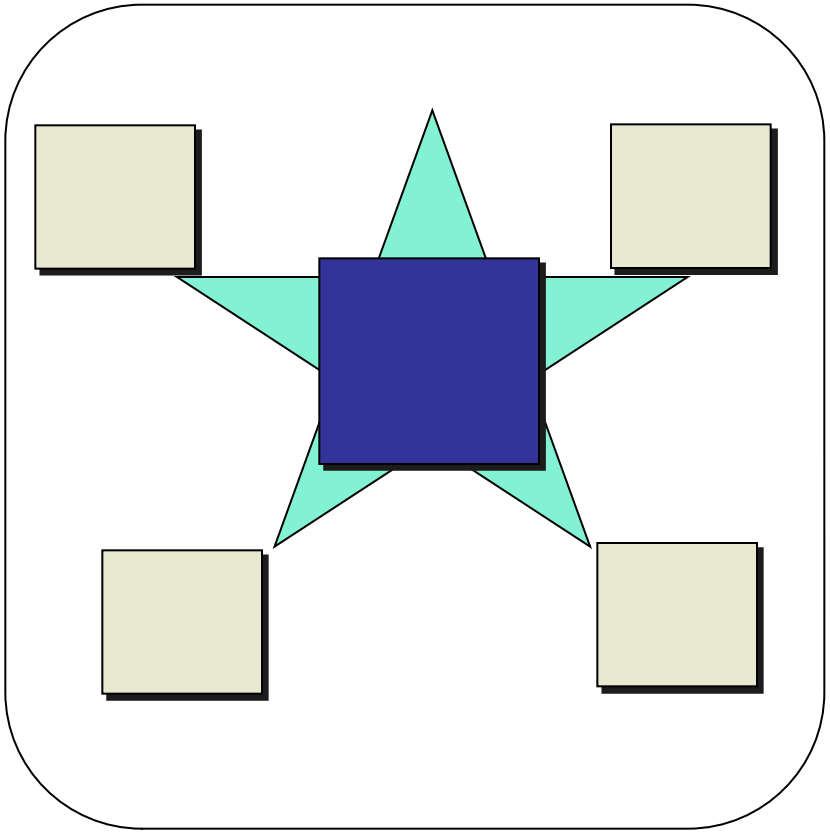
Modèle en étoile

- Une table de fait centrale et des dimensions
- Les dimensions n'ont pas de liaison entre elles
- Avantages:
  - Facilité de navigation
  - Nombre de jointures limité
- Inconvénients:
  - Redondance dans les dimensions
  - Toutes les dimensions ne concernent pas les mesures

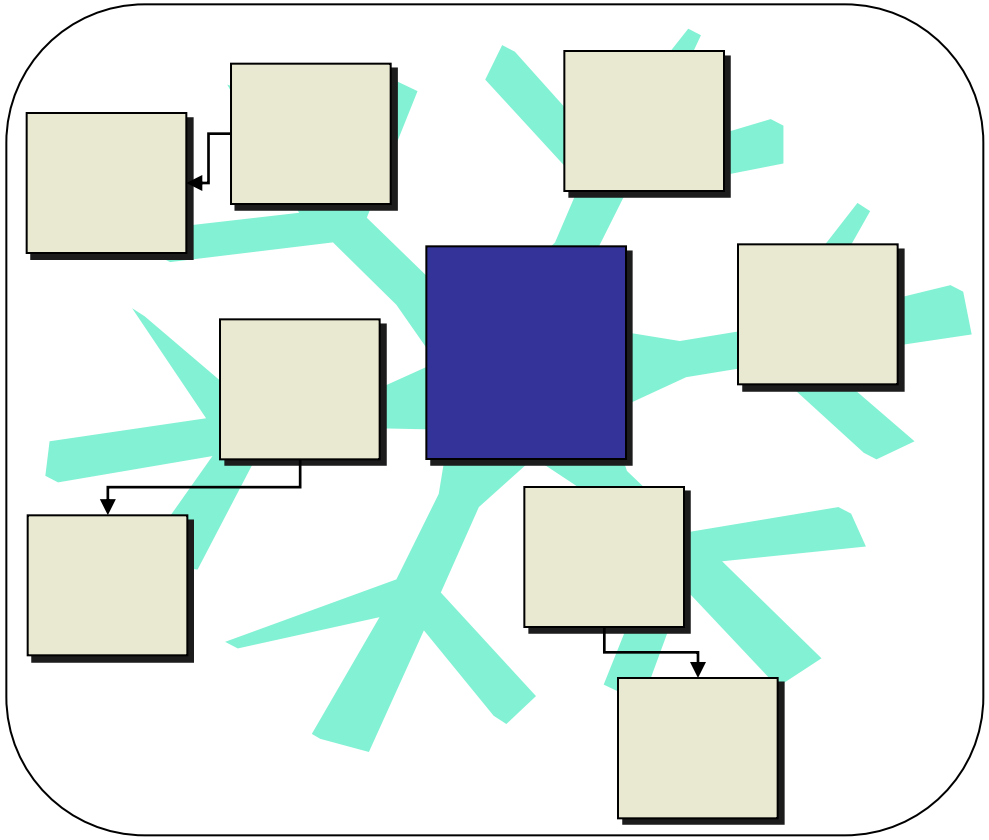
Modèle en flocon

- Une table de fait et des dimensions décomposées en sous hiérarchies
- On a un seul niveau hiérarchique dans une table de dimension
- La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine
- Avantages:
  - Normalisation des dimensions
  - Économie d'espace disque
- Inconvénients:
  - Modèle plus complexe (jointure)
  - Requêtes moins performantes

Les types de modèles



Modèle en étoile

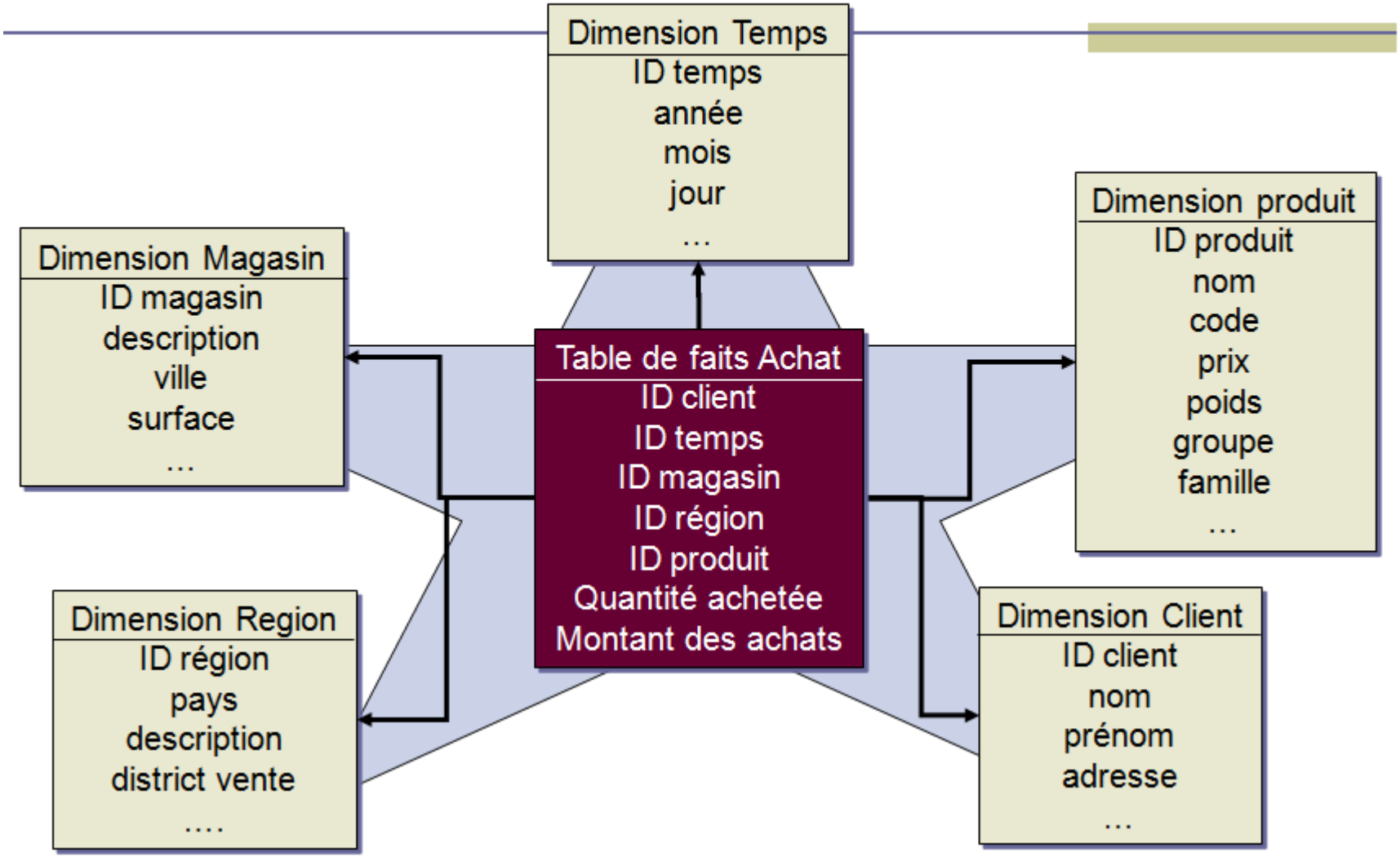


Modèle en flocon

Modèle en étoile

- Une table de fait centrale et des dimensions
- Les dimensions n'ont pas de liaison entre elles
- Avantages:
  - Facilité de navigation
  - Nombre de jointures limité
- Inconvénients:
  - Redondance dans les dimensions
  - Toutes les dimensions ne concernent pas les mesures

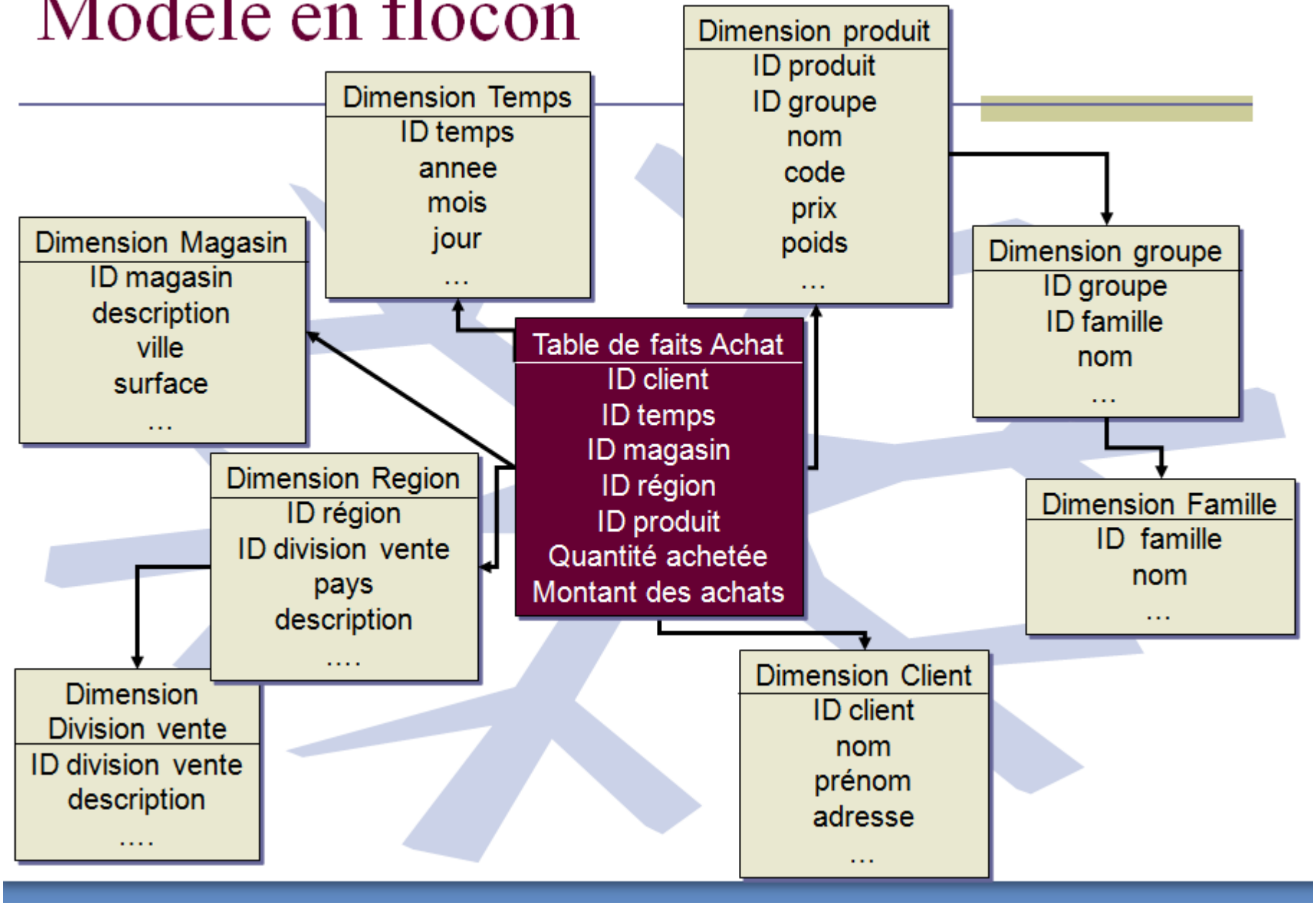
Modèle en étoile



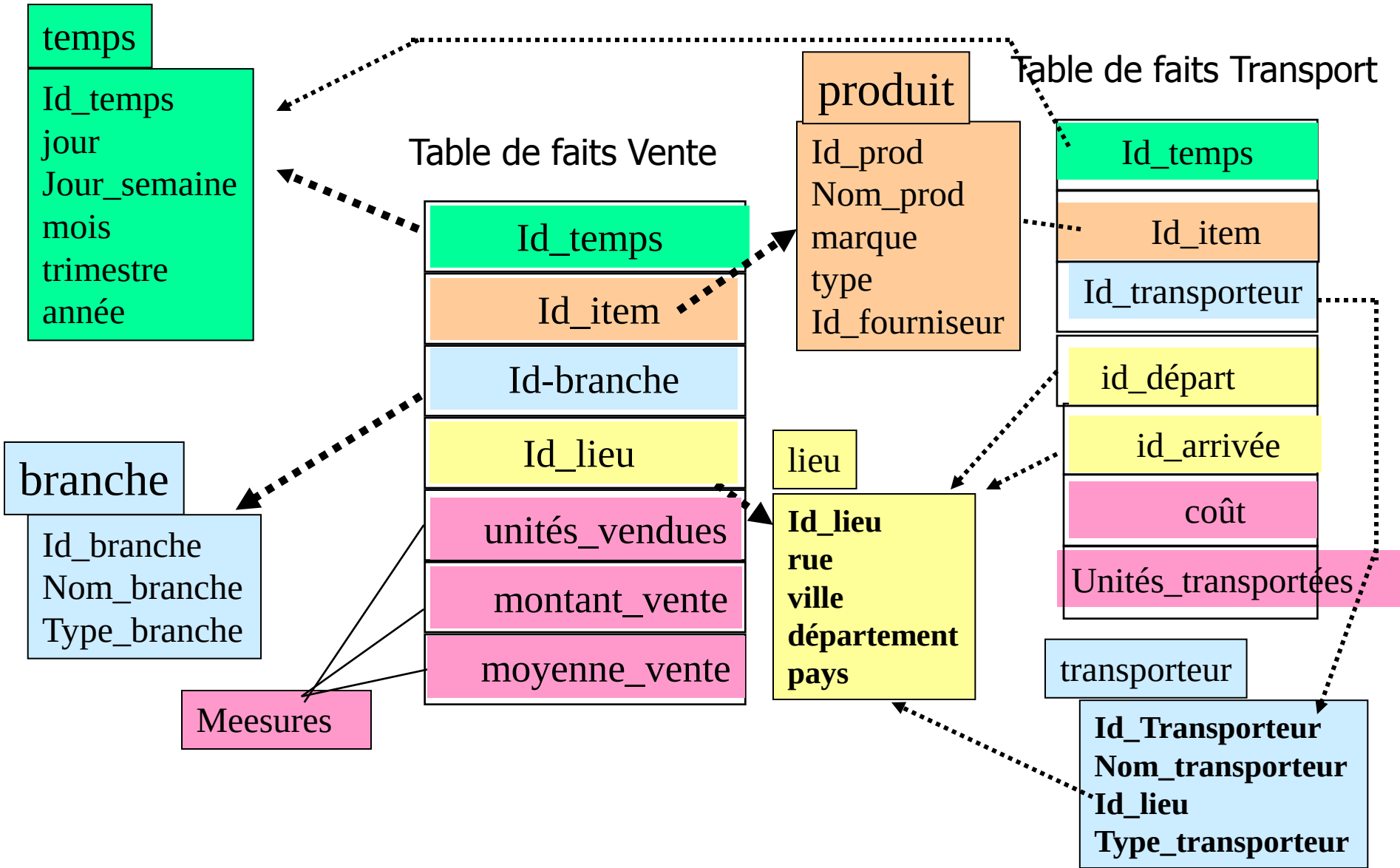
Modèle en flocon

- Une table de fait et des dimensions décomposées en sous hiérarchies
- On a un seul niveau hiérarchique dans une table de dimension
- La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine
- Avantages:
  - Normalisation des dimensions
  - Économie d'espace disque
- Inconvénients:
  - Modèle plus complexe (jointure)
  - Requêtes moins performantes

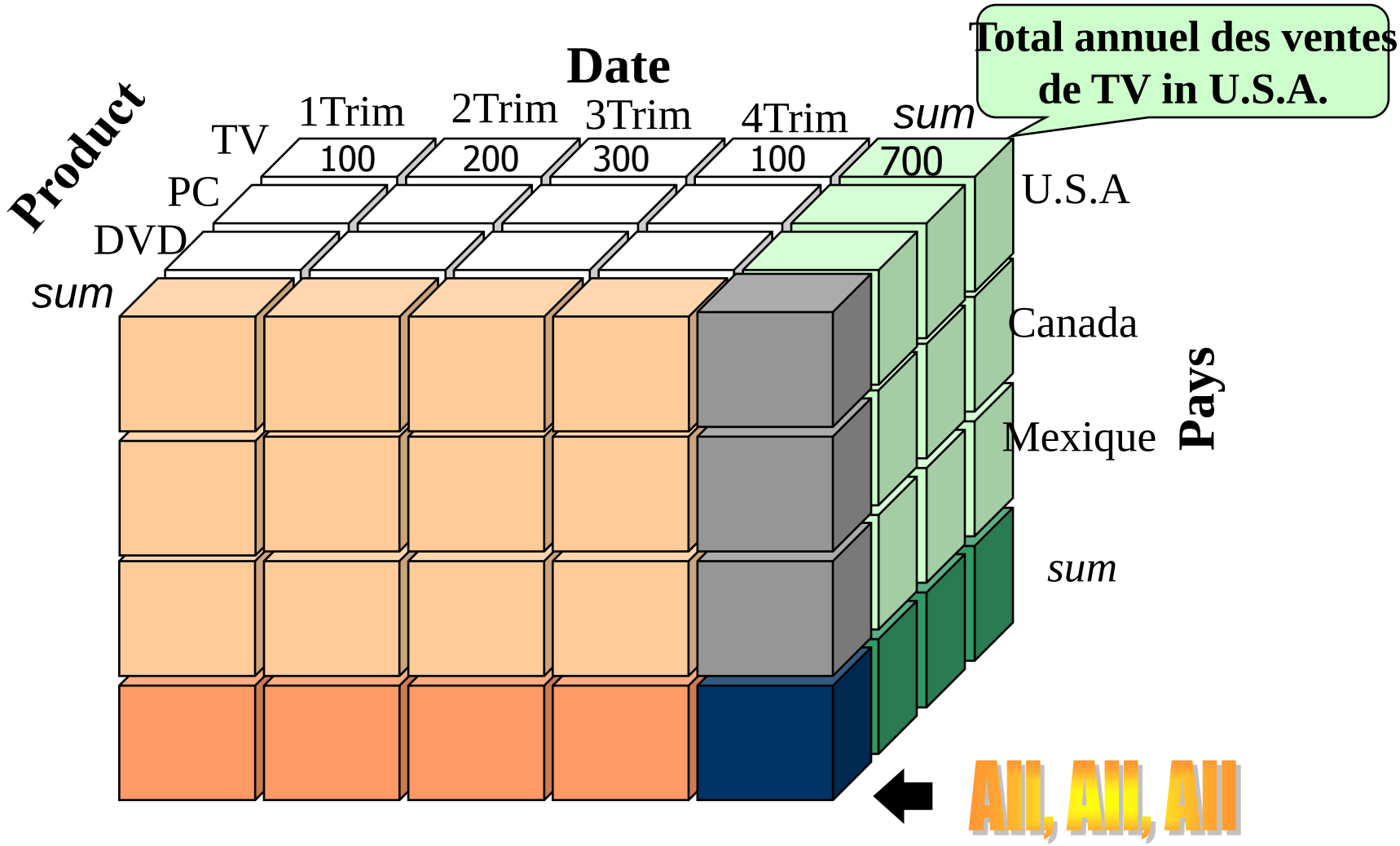
Modèle en flocon



Exemple de Constellation de faits



Un exemple de Data Cube



modèles de data warehouse

- Entreprise warehouse
  - Collecte de toutes les informations concernant les sujets traités au niveau de l'organisation
- Data Mart
  - Un sous ensemble d'un entreprise warehouse. Il est spécifique à un groupe d'utilisateurs (ex: data mart du marketing)