Université de Sherbrooke

Bases de données dimensionnelles

Modèle dimensionnel

UdeS:BDD_00

Christina KHNAISSER *(christina.khnaisser@usherbrooke.ca)*

CoFELI/Scriptorum/BDD_00-Modele (v103), version 1.1.1.a, en date du 2025-04-05

— en vigueur —

Plan

Introduction

Le présent document a pour but de présenter les bases de données analytiques (entrepôts de données) et le modèle dimensionnel.

La présentation repose sur une connaissance des bases de fonctionnement d'une base de données relationnelle.

1. Mise en contexte

- Besoins transactionnels versus besoins analytiques
  - Des données d'une entité spécifique
  - Des données agrégées d'une ou de plusieurs entités
- Multiplicité des intervenants versus multiplicité des sources et des modèles
  - Interaction concurrentielle
  - Recherche d'information à partir de plusieurs sources
  - Hétérogénéité des
    - modèles de connaissances
    - modèles conceptuels
    - modèles logiques
    - technologies

    - règles légales
    - règles de gouvernance
    - règles éthiques
- Indépendance des évolutions
- Modèles fondés sur les processus
    - Exécution des processus
    - Évaluation des processus

1.1. Contexte

Tableau 1. Notation des opérateurs logiques

	Transactionnel	**Analytique**
Objectif	Soutenir l'exécution des processus	Analyser et évaluer des processus
Fonctions	CRUD/ÉMIR	R(ucd)/ÉmirA
Optimisation	Mise à jour (concurrence)	Recherche (performance)
Portée	Transaction	Lot de transactions
Nature des requêtes prédéfinie et stable ad hoc et variable	 plus de 90 % moins de 10 %	 plus de 50 % moins de 50 %
Temporalité	Courante	Historique
Recherche d'information	On-Line Transaction Processing (OLTP)	On-Line Analytic Processing (OLAP)
Principes de conception couramment appliqués	Normalisation : 1FN, FNBC *voire parfois 5FN*	Normalisation 1FN, 3FN *voire parfois 6FN*

1.2. Entrepôt de données

- Vue unifiée de plusieurs sources de données
- Modélisation dimensionnelle
  - Ensemble de mesures permettant d'évaluer un processus
  - Ensemble des entités qui décrit le contexte de chaque mesure

1.3. Applications

- Systèmes d'aide à la décision et intelligence d'affaires
- Forage de données (*data mining*)
  - Prédiction
  - Classification
  - Inférence

Voici des exemples

Université

- Évaluation du rendement des personnes étudiantes des différentes facultés.
- Évaluation de l'opportunité d'obtention d'un emploi après l'obtention d'un diplôme (1er, 2^{e}, 3^{e} cycle).

Chaîne de distribution et de vente au détail

- Évaluation du profit des ventes dans les différentes succursales du pays.
- Évaluation de la satisfaction des clients par rapport à la qualité et la diversité des produits.

1.4. Architectures et technologies

- Entrepôts de données (*data warehouse*)
  - Bill Inmon (1970 - *Corporate Information Factory*)
  - Ralph Kimball(1990 - *Dimensional data model*)
  - Dan Linstedt (2000 - *Data vault*)
- Lac de données (étang de données, *data lake*)
  - James Dixion (2010)
- Entrepôts de lac de données (marais de données, *data lakehouse*)
  - Databricks (2023)
- …

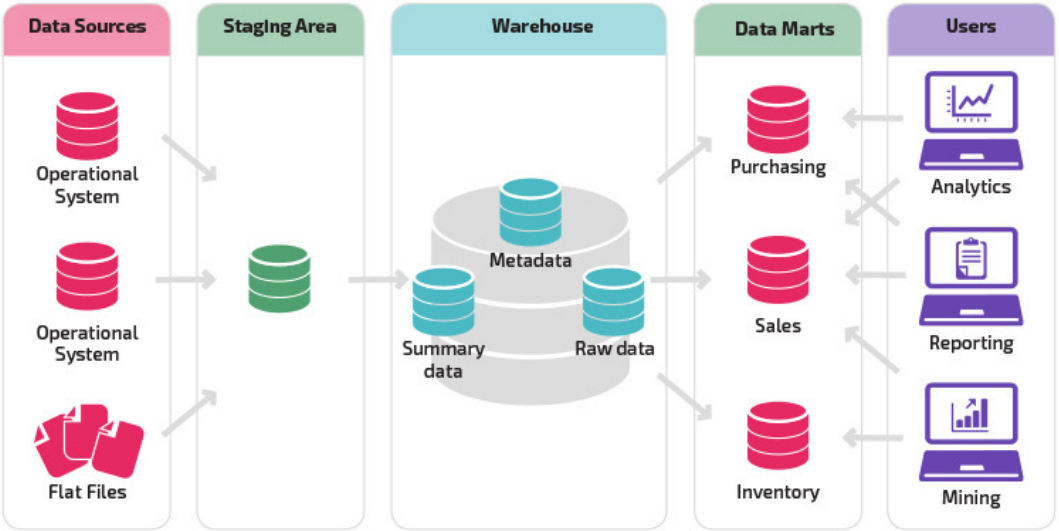
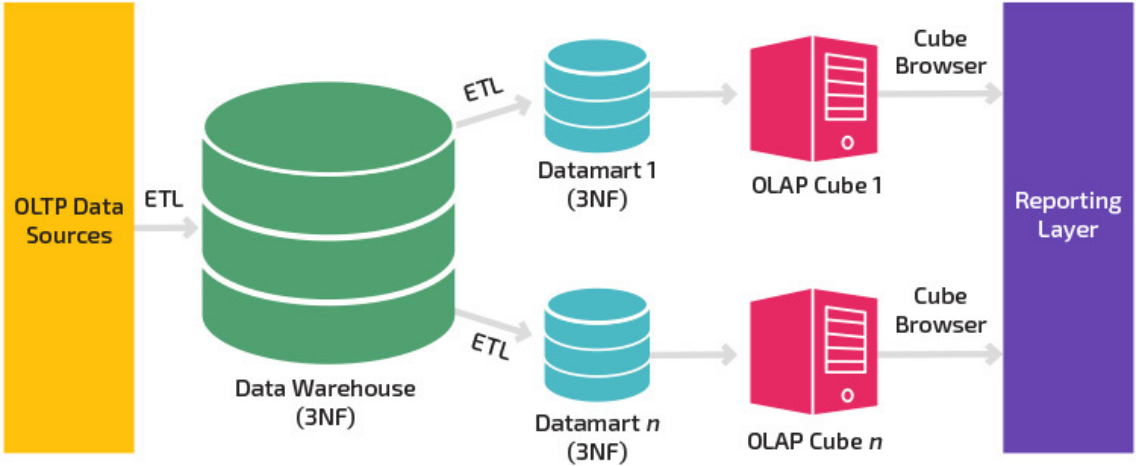
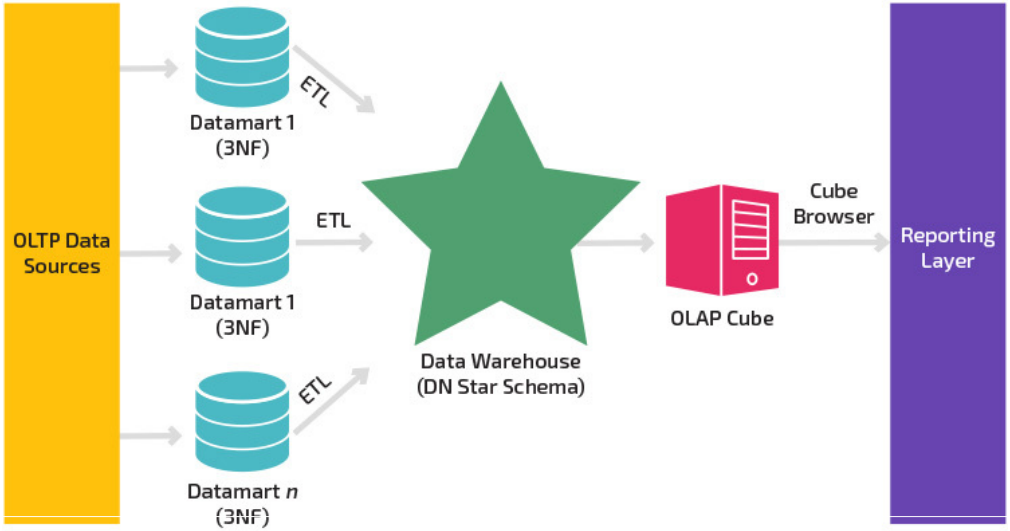


Figure 1. Entrepôt de données (https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/)

Inmon Model



Kimball Model



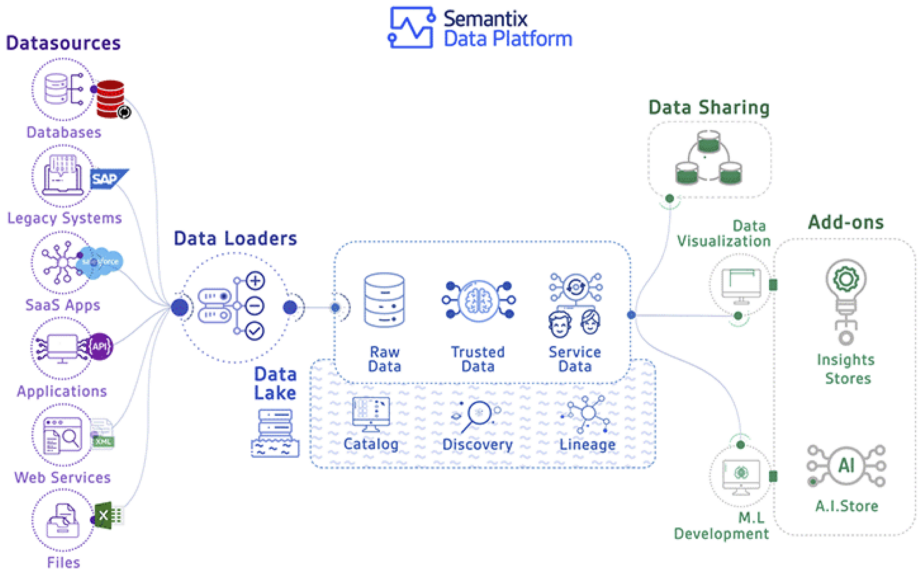


Figure 2. Lac de données (https://semantix.com.br/data-platform/)

Toutes les figures précédentes sont sujettes à caution. Certaines sont carrément fausses. Ce sont malheureusement les plus répandues.

Voir les critiques suivantes :

- [Adamson2010a] chap. 2.
- [Ambler2006a] chap. 1, 2 et 3
- [Jiang2015a]

Figure 3. Gartner Magic Quadrant for Data Integration Tools 2022-2023

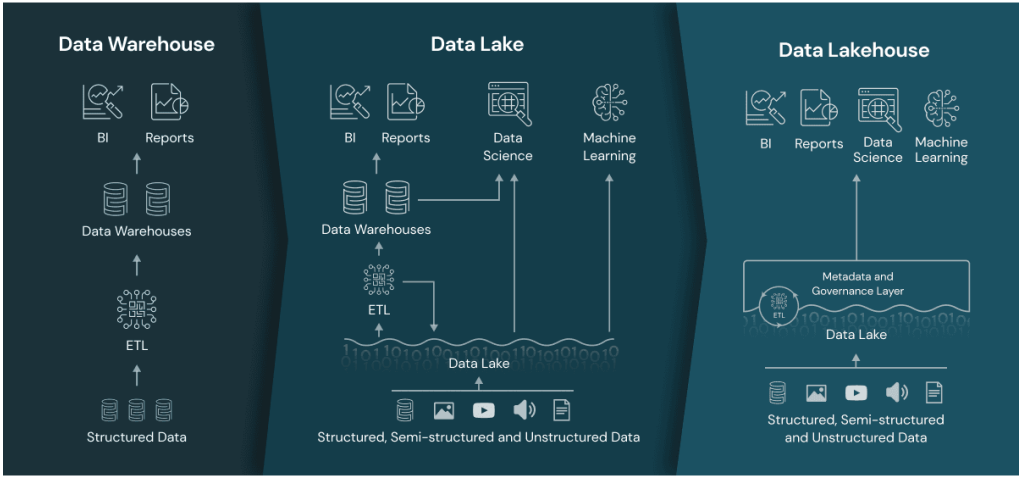


Figure 4. Lac d'entrepôts de données (https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html)

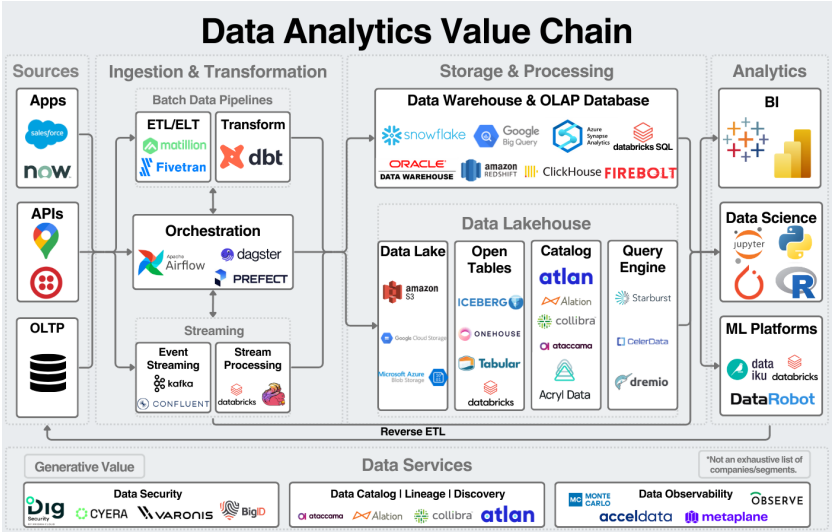


Figure 5. Technologies analytiques 2024 (https://www.generativevalue.com/p/a-primer-on-data-warehouses)

2. Définition du modèle

Concepts

- Dimension : caractéristique d'un processus.
- Fait : mesure d'un processus.

Objectifs

- Historicisation des données d'un processus.
- Synthétisation des mesures d'un processus.

2.1. Processus

Un processus y est décrit par un ensemble d'évènements, chaque évènement possède des mesures (faits) et se caractérise par un ensemble de propriétés (dimensions).

Voici des exemples

Université

- L’inscription des personnes étudiantes à une activité pédagogique.
- L’évaluation d’une personne étudiante par différents types d’évaluation.

Entrepôt

- La commande de produits par un client.
- La livraison de produits commandés par un fournisseur à un client.

2.2. Dimension

Une dimension est une entité qui caractérise nécessairement un processus mesuré.

Une dimension est représentée par une relation dimensionnelle.

Une relation dimensionnelle est définie par une clé artificielle, des clés naturelles, des attributs .

Clé naturelle (externe)

Une clé naturelle est un ensemble d'attributs qui identifie d'une façon unique une entité dans un domaine. Parfois cette clé peut être spécifique à une source de données ou non accessible. Dans ce contexte, plusieurs clés naturelles peuvent permettre d'identifier une même entité.

La clé naturelle doit être définie seulement si les sources de données qui participent à l'entrepôt de données utilisent la **même** clé naturelle. Sinon, il faut créer une table de correspondance entre la clé artificielle et chaque clé naturelle de chaque source.

Clé artificielle (interne)

Une clé artificielle est un attribut qui identifier d'une façon unique une entité dans l'entrepôt de données. Comme les attributs qui composent la dimension proviennent souvent de plusieurs sources externes, en général, aucune des clés externes ne peut être garantie. Pour cette raison, il est d'usage d'ajouter systématiquement une clé interne aux attributs de la dimension.

Attributs non-clé

Les attributs non-clé sont généralement descriptifs. Ils sont souvent utilisés pour définir les agrégations et les conditions de restrictions ou pour l'ordonnancement des faits de la relation factuelle.

La communauté de pratique distingue plusieurs catégories d'attributs, telles que :

- Attributs code-description : représente un dictionnaire de données. C'est une paire d'attributs, le premier représente un code et le deuxième la description du code.
  Exemple : code catégorie, nom catégorie : 1, Nourriture ;
  code type évaluation, nom type d'évaluation : TP, travail pratique
- Attribut composition : représente un attribut qui contient plusieurs parties.
  Exemple : +1 514-1234-1923 (code du pays, code de la région, code du téléphone)

- Attribut calculé : dérivé à partir d'une fonction sur un attribut dans la même dimension.
  Exemple : date de naissance ⇒ âge.
- Attribut agrégeable : peut être utilisé par une fonction d'agrégation pour calculé un fait.
  Exemple : note ⇒ moyenne des notes.

2.3. Fait

Un fait représente un évènement produisant des mesures du processus qu'on veut évaluer.

Des faits ayant les mêmes mesures, la même synchronicité et la même granularité pour un même processus sont regroupés dans une relation, nommée relation factuelle.

Une relation factuelle est définie par

- l'ensemble des clés artificielles des dimensions la caractérisant ;
- l'ensemble des attributs représentant les mesures du processus modélisé.

Les clés artificielles des dimensions permettent de lier un fait aux informations des dimensions par les clés référentielles.

Nous distinguons plusieurs catégories d'attributs:

- Attribut agrégeable: il faut qu'une fonction d'agréation puisse lui être appliquée Beaucoup d'auteurs prescrivent que tous les attributs non-clé d'un fait doivent être agrégeable (certains, à tort, les limitant au cas additif). Exemple : sommes des notes, moyenne de la classe;
- Attribut calculé (Attribut dérivé): l'attribut dérivé est calculé à partir d'autres attributs du fait que ce soit par une expression logique, arithmétique, rationnelle ou autre.+ Exemple : côte d'un cours, âge, cout total, pourcentage de vent

3. Mise en oeuvre du modèle

- Étoile (*star*)
- Flocon (*snowflake*)
- Constellation (*starflake*)

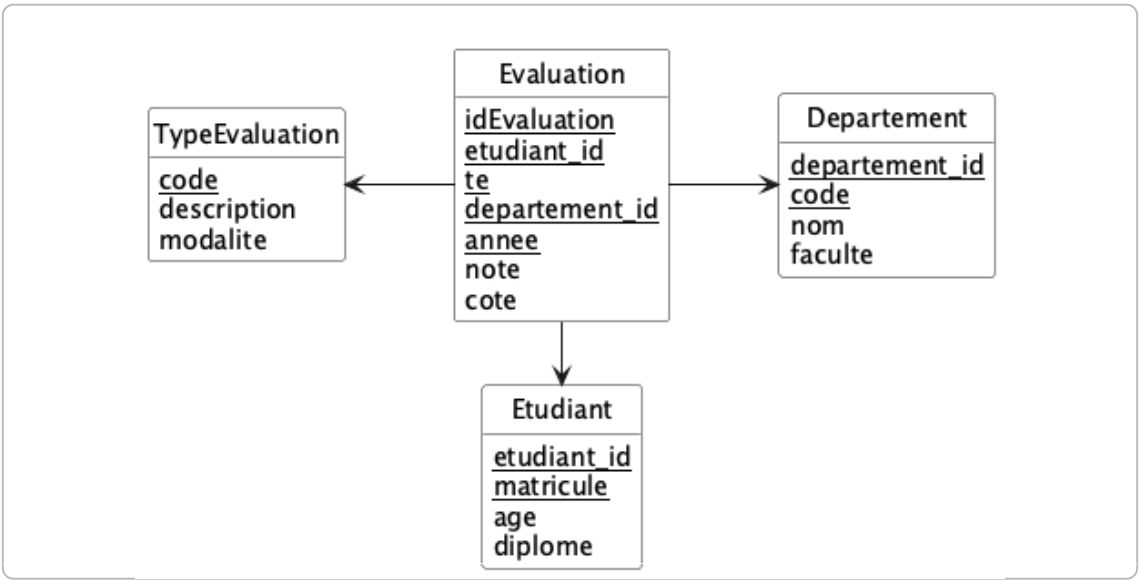
Le schéma étoile est propre à un processus. Lorsqu'il y a plusieurs processus, cela forme une constellation (avec plusieurs relations factuelles).

En général, une même relation dimensionnelle peut être référée par plusieurs relations factuelles. Un schéma avec des relations dimensionnelles hiérarchiques forme un flocon.

3.1. Schéma en étoile

- Relation factuelle (table de faits)
- Relations dimensionnelles (tables des dimensions)

Exemple 1. Schéma en étoile d'un processus d'évaluation de personnes diplômées



Exemple 2. Schéma en étoile d'un processus de commande

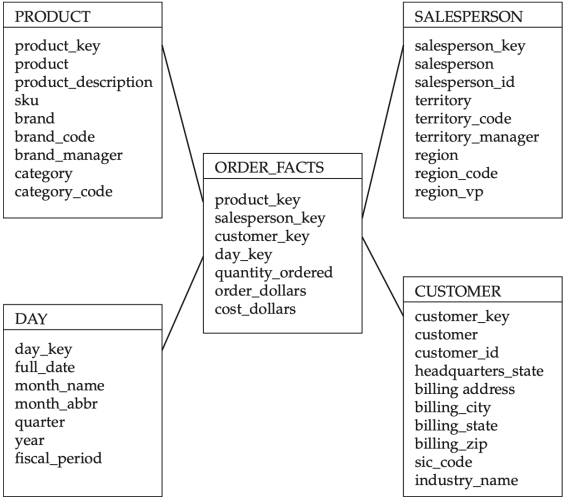


Figure 1-5 A simple star schema for the orders process

Dimension temporelle

Par la définition même du processus, la dimension temporelle (la dimension DAY dans l'exemple) est toujours présente dans un schéma dimensionnel.

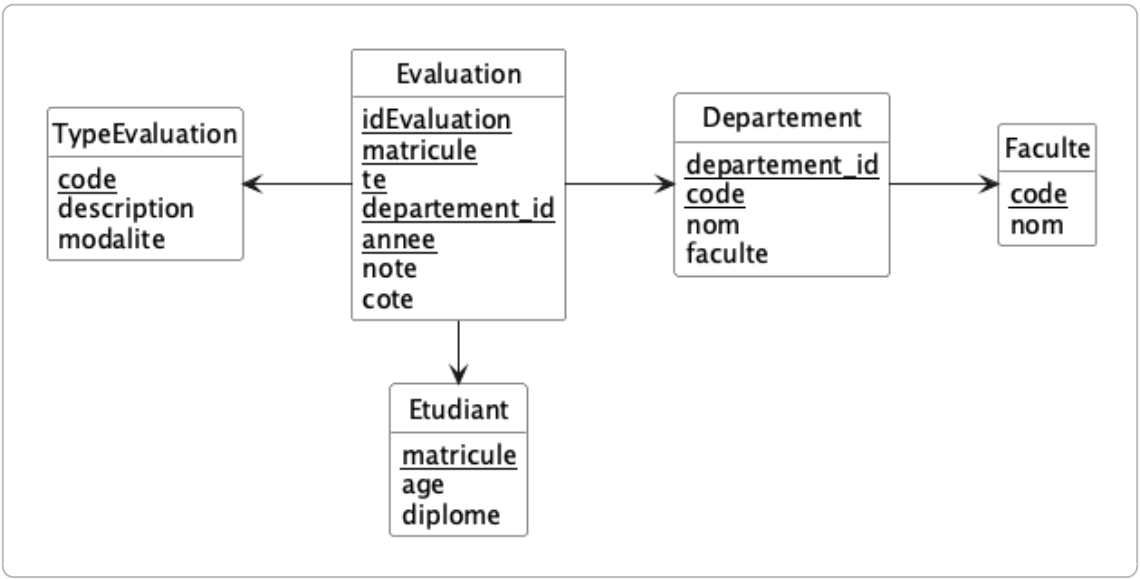
Il arrive fréquemment toutefois qu'il n'y ait guère d'autres attributs à lui être associée que le point temporel lui-même. Il est donc fréquent que cette dimension soit « dégénérée », et incluse directement comme attribut dans les tables de faits.

Par contre, plusieurs auteurs déconseillent cette pratique, car elle induit souvent la perte de données requises pour la nécessaire mise en correspondance des temps des différents processus.

3.2. Schéma en flocon

- Relation factuelle
- Relations dimensionnelles hiérarchiques
  - Quartier ← Ville ← Région
  - Jour ← Mois ← Trimestre ← Année
  - Produit ← Marque ← Catégorie

Exemple 3. Schéma en flocon d'un processus d'évaluation de personnes diplômées



Exemple 4. Schéma en flocon d'un processus de commande

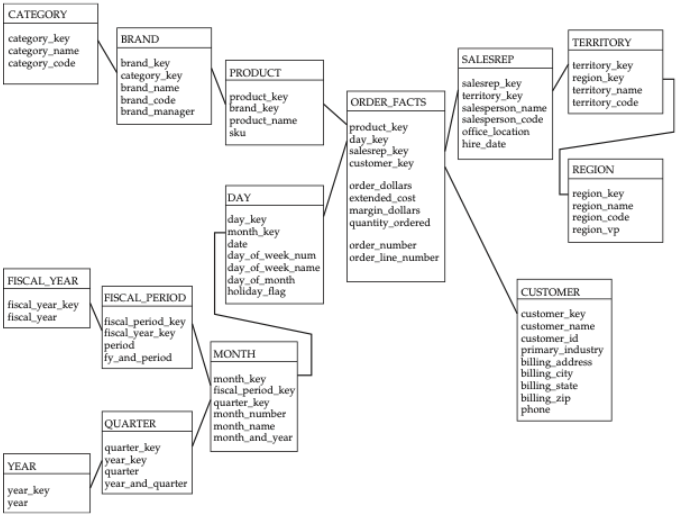


Figure 7-5 A snowflake schema

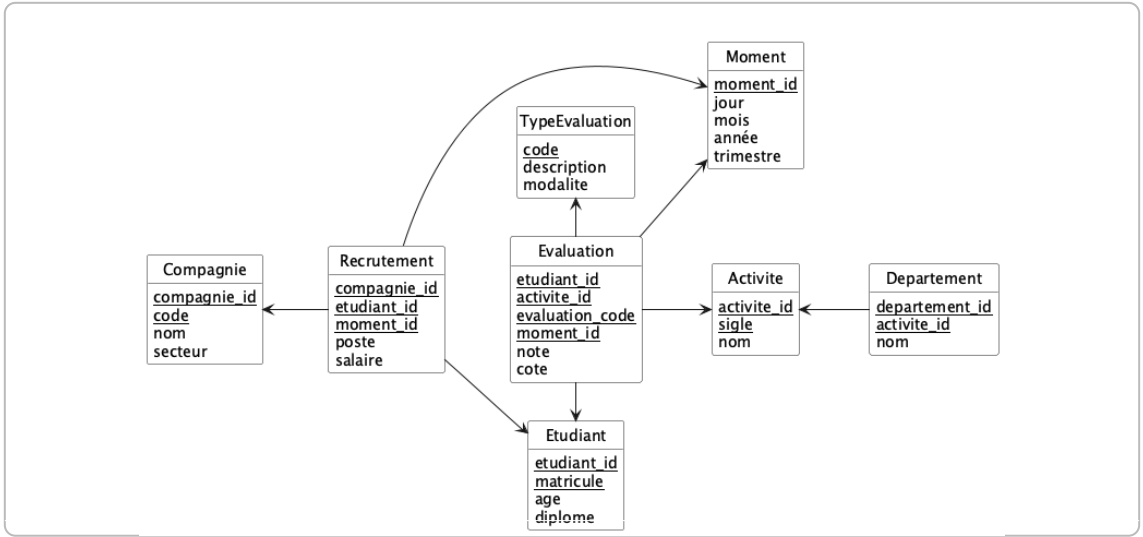
3.3. Schéma en constellation

- Relations factuelles
- Relations dimensionnelles

Exemple.

- Analyser la quantité commandée par jour, client et produit
- Analyser la quantité expédiée par jour, client, produit et expéditeur

Exemple 5. Schéma en constellation d'un processus d'évaluation et recrutement de personnes étudiantes



Exemple 6. Schéma en constellation d'un processus de commande et de livraison

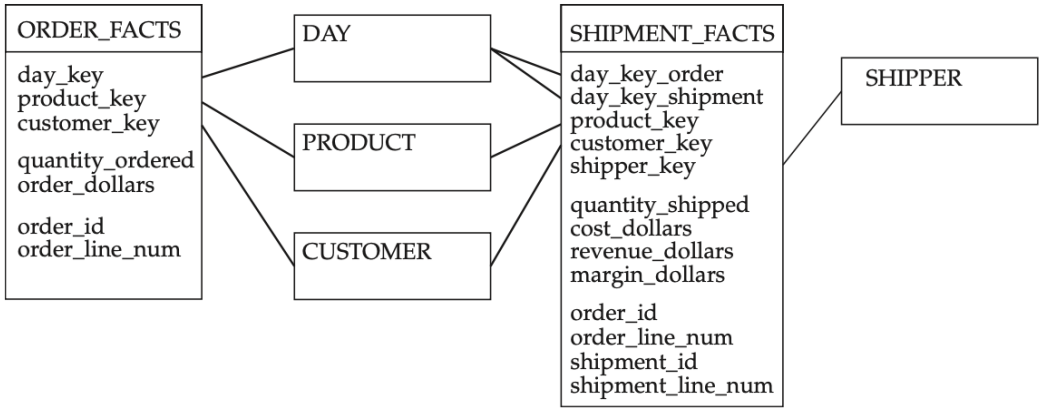


Figure 4-8 Separating the two processes into separate fact tables

4. Règles de pratique

- Attribuer une clé artificielle à chaque relation de dimension. Cet attribut sera utilisé pour identifier de manière unique chaque tuple de la relation.
- Fournir un ensemble complet d'attributs de dimension. Chaque nouvel attribut augmente considérablement le nombre de possibilités d'analyse.
- Prêter une attention particulière à l'utilisation des attributs numérique. Les attributs utilisés pour filtrer les requêtes, ordonner les données, définir l'agrégation ou gérer les relations hiérarchiques.
- Définir une relation factuelle par processus pour permettre d'évaluer les processus individuellement. Lorsque deux ou plusieurs faits ne se produisent pas simultanément ou utilisent des dimensions différentes, ils représentent des processus différents. Les placer dans une seule relation factuelle entravera l'analyse des processus individuels.

Références

[Adamson2010a]
Christopher ADAMSON ;
The complete reference star schema ;
McGraw-Hill, New York (NY, US), 2010 ;
ISBN 978-0-07-174432-4.

- *Chapitres 1, 3, 4.*

[Ambler2006a]
Scott W. AAMBLER, Pramod J. SADALAGE ;
Refactoring Databases ;
Addison-Wesley, Upper Sadle River (NJ, US), 2006 ;
ISBN 978-0-321-77451-4.

[Jiang2015a]
Bin JIANG ;
Constructing Data Warehouses with Metadata-Driven Generic Operators, and More
Architecture, Methodology, and Paradigm, Concepts, Algorithms, and Operators, Principles, Recommendations, and Exercises ;
2nd edition, DBJ Publishing, 2015 ;
ISBN 978-15086873-13.

Produit le 2025-10-31 12:22:36 UTC

Université de Sherbrooke