Université de Sherbrooke

Bases de données dimensionnelles

Modèle dimensionnel

UdeS:BDD_00

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

CoFELI/Scriptorum/BDD_00-Modele, version 1.2.1.a, en date du 2025-11-04

— 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 de la théorie relationnelle, de la modélisation logique et des bases de données relationnelles.

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'informations 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 d'entités la description du contexte de chaque mesure

1.3. Applications

- Aide à la décision (*business intelligence*)
  - Modélisation
  - Définition et calcul des mesures
  - Définition et évaluation de *scenarrii*
- 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, 2e, 3e 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

1.4.1. Architectures

- Entrepôts de données (*data warehouse*)
  - Bill Inmon (1970 — *corporate information factory*, *data mart*)
  - 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)

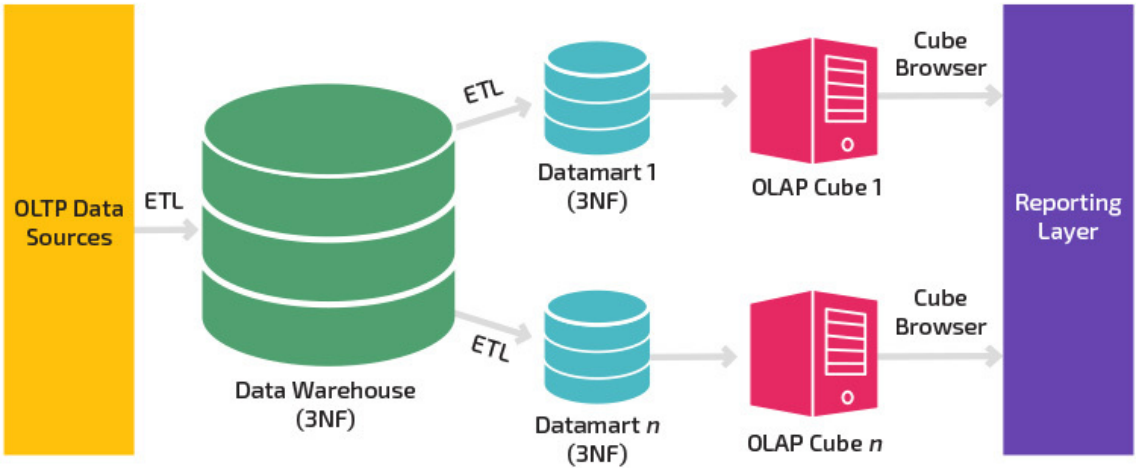


Figure 1. Entrepôt de données - Inmon [Panoply2024a]

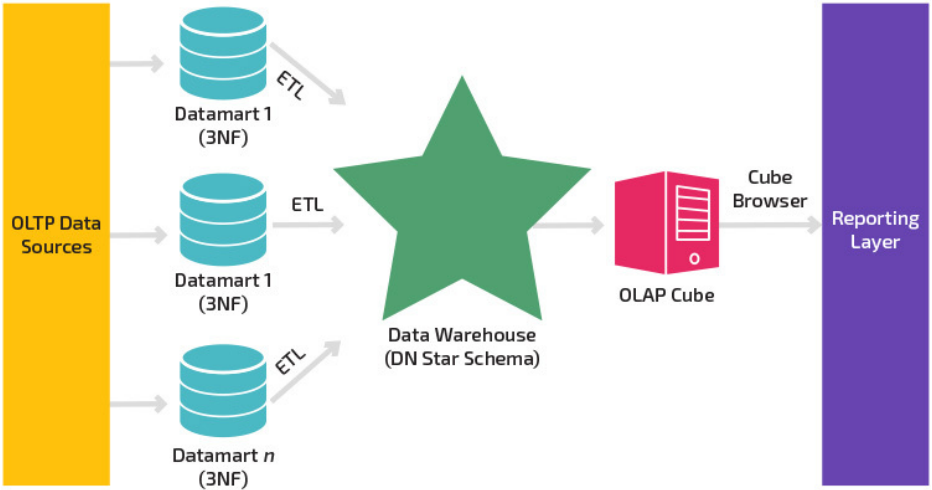


Figure 2. Entrepôt de données - Kimbal [Panoply2024a]

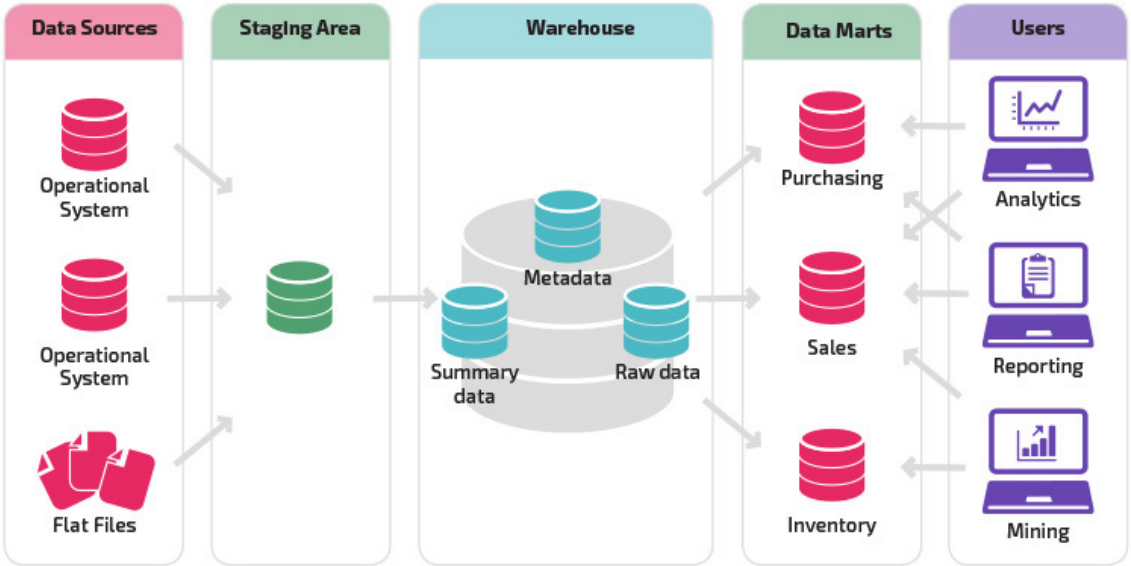


Figure 3. Entrepôt de données - Linstedt [Panoply2024a]

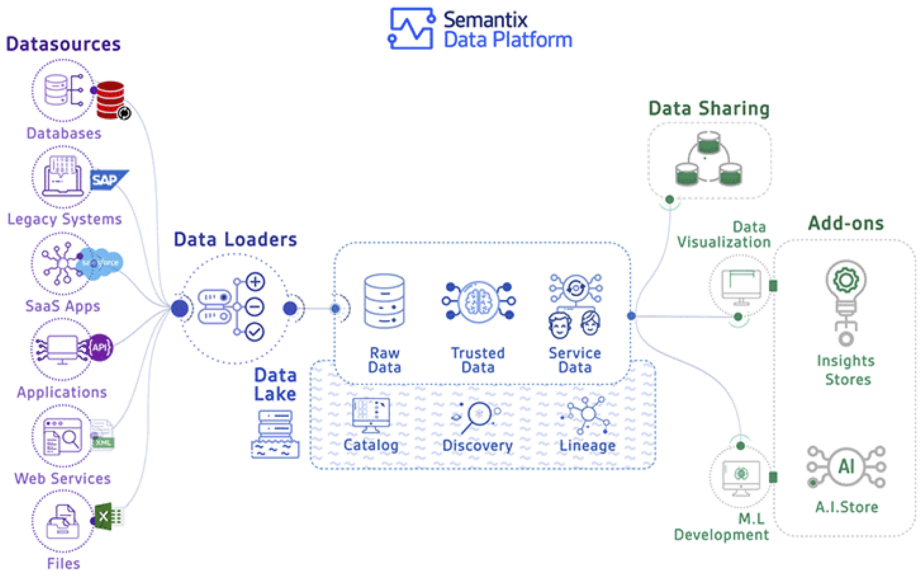


Figure 4. Lac de données (https://semantix.com.br/data-platform/)

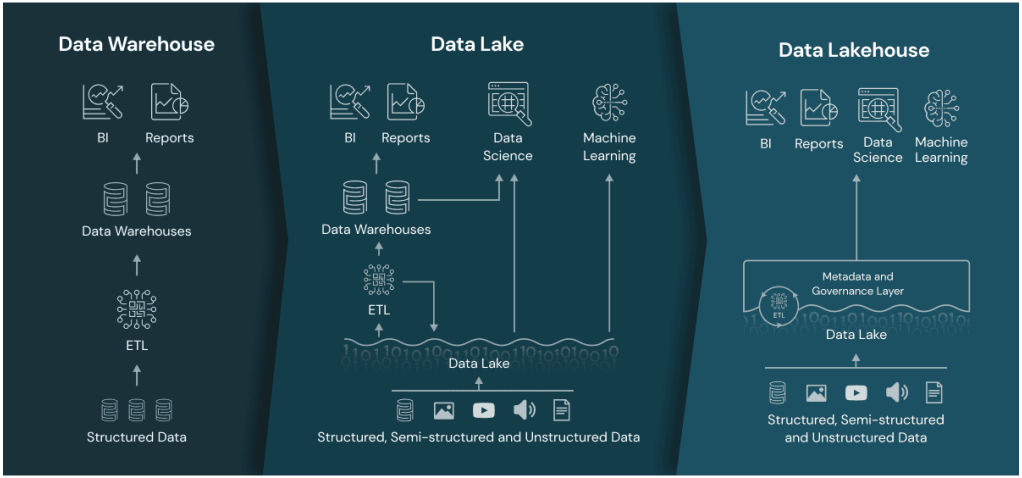


Figure 5. Lac d'entrepôts de données (https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html)

Toutes les figures précédentes sont sujettes à caution. Certaines sont fausses. Elles sont cependant représentatives des opinions les plus répandues.

Voir les critiques suivantes :

- La non-prise en compte des règles de modélisation et de conception découlant des principes d'analyse des processus, voir [Adamson2010a] pour les règles de pratique la prise en compte.
- La non-prise en compte de l'évolutivité des modèles, des technologies et des usages, voir [Ambler2006a] pour les structures et les méthodes permettant la prise en compte.
- La non-prise en compte de la dynamique des données des sources, voir [Jiang2015a] pour la prise en compte de cette dynamique, la nécessité de l'abandon de l'ETL au profit de l'ELT et la nécessaire structuration du *staging area* en archive historicisée.

1.4.2. Technologies

Figure 6. Gartner Magic Quadrant for Data Integration Tools 2022-2023

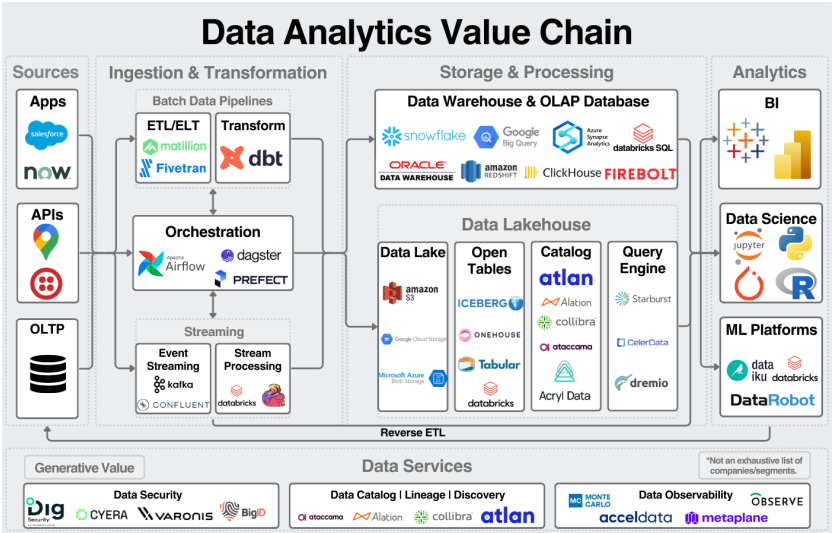


Figure 7. Technologies analytiques 2024 (https://www.generativevalue.com/p/a-primer-on-data-warehouses)

2. Définition du modèle

Concepts

- Dimension : caractéristique d'un évènement d'un processus.
- Fait : mesure d'un évènement 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 une suite (chronologique) d'évènements, chaque évènement :

- est caractérisé par un ensemble de propriétés (dimensions).
- est sujet à un ensemble de mesures (faits)

Voici des exemples

Université

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

Entrepôt

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

Rappel

processus

Ensemble d'activités logiquement interreliées permettant d'élaborer un résultat (ensemble d'artefacts déterminés).

- L'enchainement des activités au sein d'un processus répond généralement aux prescriptions d'un procédé.
- En anglais, procédé et processus se disent tous deux « process » d'où, parfois, une certaine confusion !

2.2. Dimension

Une dimension est une entité qui caractérise nécessairement un évènement (appartenant au processus ciblé) à être mesuré.

Une dimension est représentée par une (variable de) relation, fréquemment nommée relation dimensionnelle. En pratique, elle est composée d'une clé primaire interne et de plusieurs attributs provenant de sources externes. Ces attributs ont pour fonction de caractériser, de documenter, la dimension. Des sous-ensembles de ces attributs peuvent former des clés externes relativement aux sources, parfois même relativement à l'entrepôt (donc à la relation dimensionnelle elle-même).

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 descriptifs* :
  représente une portion du dictionnaire de données sous la forme d'une paire d'attributs, le premier attribut représente un code et le deuxième la description de l'entité (de la catégorie d'entité) associée au code.
  Exemple :
  code catégorie, nom catégorie : 1, Nourriture ;
  code type évaluation, nom type d'évaluation : TP, travail pratique.
- *Attribut composé* :
  représente un attribut qui contient plusieurs parties.
  Exemple :
  +1 514 123-4567 (code du pays, code régional, code local)

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

2.3. Fait

Un fait représente un évènement ciblé (et donc mesuré) d'un processus d'un processus soumis à l'analyse.

Des (instances de) faits ayant la même définition, les mêmes mesures, la même synchronicité et la même granularité pour un même processus sont regroupés dans une (variable de) relation, fréquemment nommée relation factuelle.

Une relation factuelle est définie par

- une clé primaire formée de l'ensemble des clés déterminantes des dimensions la caractérisant (ces dernières devenant ainsi des clés référentielles vers leurs dimensions respectives) ;
- l'ensemble des attributs représentant les mesures retenues en regard du processus analysé.

La communauté de pratique distingue plusieurs catégories d'attributs :

- *Attribut agrégable* :
  les valeurs de l'attribut doivent agrégables ; certains auteurs prescrivent que tous les attributs non-clé d'un fait doivent être agrégable (un sous-ensemble de ceux-ci les limitant, à tort, au cas additif).
  Exemple :
  sommes des notes, moyenne de la classe.
- *Attribut calculé (dérivé)* :
  sa valeur est le résultat d'une fonction dont les paramètres sont fournis par d'autres attributs du fait ; certains auteurs permettent également les paramètres fournis par des attributs des dimensions référées par le fait.
  Exemple :
  côte d'un cours, âge, etc.

2.4. La question du temps

- Le temps est-il une dimension comme une autre ?
- Une relation factuelle doit-elle nécessairement avoir une dimension temporelle ?

2.4.1. Quelques propositions fréquentes

Proposition 1

Créer une dimension temporelle unique à laquelle toutes les tables de faits réfèrent. Les tables des autres dimensions peuvent y référer aussi ; l'inclusion de l'attribut référentiel à leur clé primaire ne fait pas concensus.

Proposition 2

Ajouter une estampille temporelle à toutes les tables et l'ajouter à la clé primaire.

Proposition 3

Ajouter un intervalle temporel à toutes les tables et l'ajouter à la clé primaire. Dans ce cas, lesquelles des contraintes suivantes doivent être respectées : non-contradiction, non-redondance, non-circonlocution ?

Constat

Sauf la proposition 3 lorsqu'elle inclut les trois contraintes, toutes les propositions repoussent le problème de la modélisation temporelle à chaque requête analytique. Il en découle un risque certain d'incohérences entre les requêtes n'adoptant pas le même modèle.

2.4.2. Quelques problèmes communs

À quelle horloge les estampilles temporelles réfèrent-elles ?

- Celle du SGBD de l'entrepôt ?
  - Dans ce cas, suppose-t-on que les horloges des sources y sont toutes synchronisées ?
- Celle du SGBD de la source ?
  - Comment rendre compte du temps d'attributs de sources différentes au sein d'un même tuple ?
  - Comment rendre compte des contradictions temporelles entre les sources ?

À quoi fait référence l'estampille temporelle ?

- Pour les dimensions
  - Au temps de validation à la source ?
  - Au temps de transaction à la source ?
  - Au temps d'exportation à la source ?
  - Au temps d'importation à l'entrepôt ?
  - Au temps d'intégration à l'entrepôt ?
  - etc.
- Pour les faits
  - Au moment du calcul des attributs calculés ?
  - À un moment calculé à partir des estampilles des dimensions contributives ?
  - etc.

3. Mise en oeuvre du modèle

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

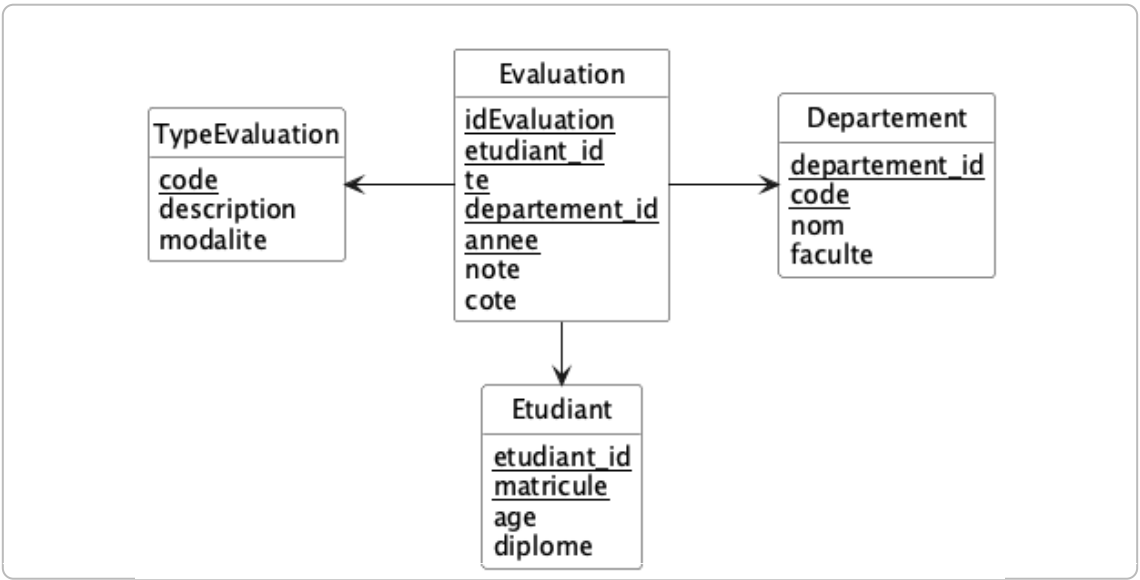
Le schéma étoile est propre à un processus. Lorsqu'il y a plusieurs processus, cela forme une constellation (avec plusieurs relations factuelles).

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 [Adamson2010a]

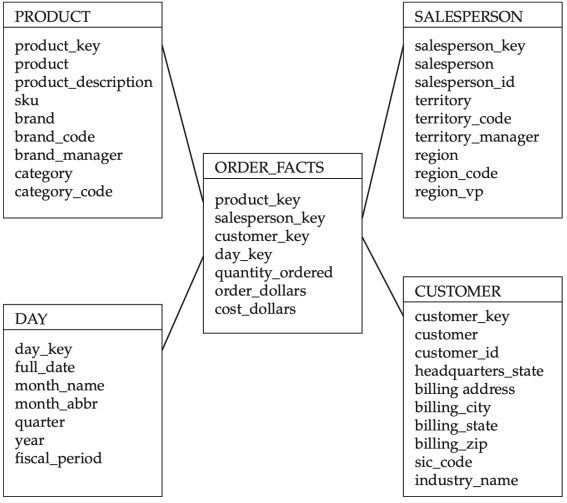


Figure 1-5 A simple star schema for the orders process

Dimension temporelle

Par la définition même du processus, la dimension temporelle est toujours présente dans un schéma dimensionnel.

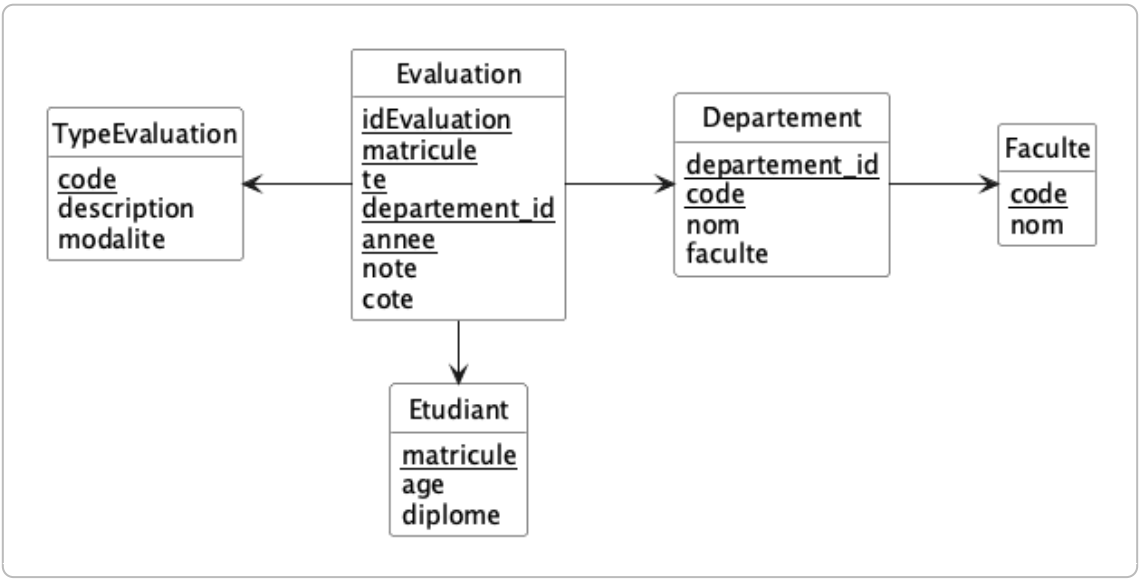
Il arrive fréquemment toutefois qu'il n'y ait guère d'autres attributs à lui être associée que le point temporel lui-même. Il est donc fréquent que cette dimension soit « dégénérée », c'est-à-dire incluse directement comme attribut dans les tables de faits.

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 [Adamson2010a]

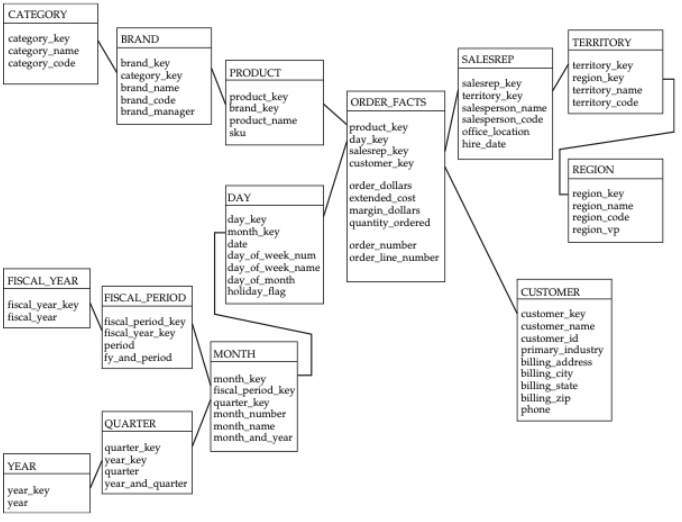


Figure 7-5 A snowflake schema

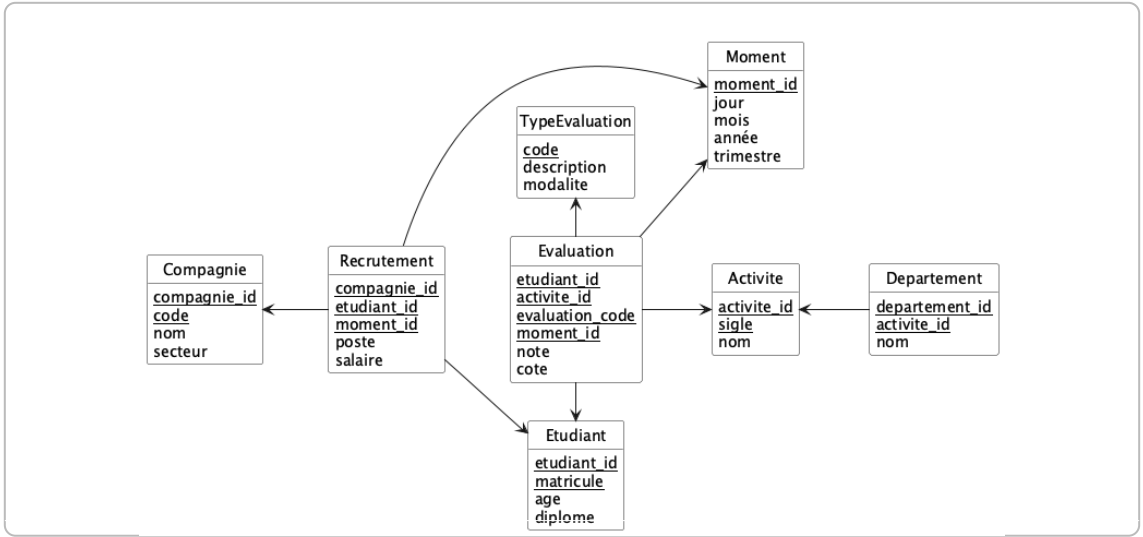
3.3. Schéma en constellation

- 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 de recrutement de personnes étudiantes



Exemple 6. Schéma en constellation, processus de commande et processus de livraison [Adamson2010a]

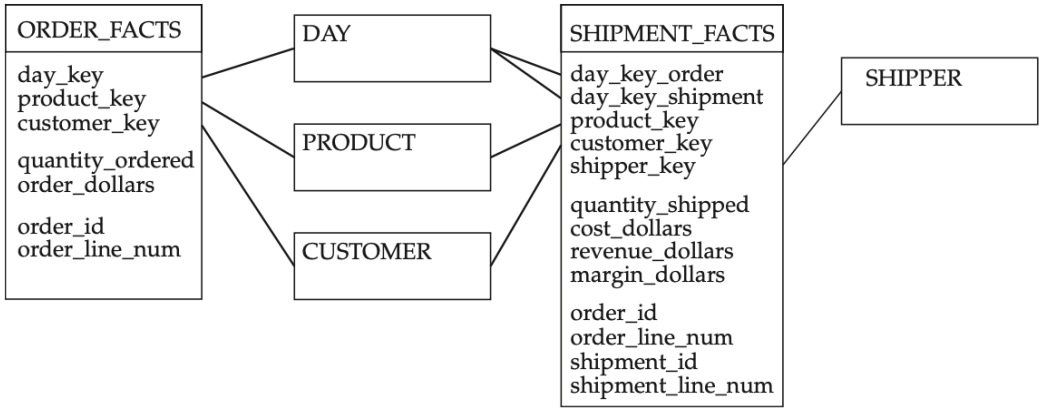


Figure 4-8 Separating the two processes into separate fact tables

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.

[Adamson2017a]
Christopher ADAMSON ;
Chris Adamson's Blog (2007-2017) ;
https://blog.chrisadamson.com ;
dernière consultation 2025-10-31

[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.

[Kimball2013a]
Ralph KIMBALL, Margy ROSS ;
The data warehouse toolkit: the definitive guide to dimensional modeling ;
3rd Edition, John Wiley, 2013 ;
ISBN 978-1118530801.

[Panoply2024a]
Panoply ;
Data Warehouse Guide ;
https://panoply.io/data-warehouse-guide/ ;
dernière consultation 2025-10-31

Produit le 2025-11-05 13:28:00 UTC

Université de Sherbrooke