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 —

Sommaire

Introduction aux bases de données dimensionnelles (entrepôts de données).

Mise en garde

Le présent document est en cours d'élaboration ; en conséquence, il est incomplet et peut contenir des erreurs.

Historique

diffusion	resp.	description
2025-04-05	CK	Mise à jour et corrections mineures.
2024-10-31	CK	Mise à jour des sections.
2024-08-16	CK	Ébauche initiale à partir de documents antérieurs produits entre 2004 et 2023.

Table des matières

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.

Contenu des sections

- La section 1 expose la différence entre les bases de données transactionnelles et les bases de données analytiques.
- La section 2 présente le modèle dimensionnel.
- La section 3 présente les principes de mise en oeuvre d'un modèle dimensionnel.
- La section 4 présente les règles de pratiques pour la mise en oeuvre d'une base de données dimensionnelles.

Évolution du document

La première version du document a été établie sur la base des travaux publiés par Adamson.

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

Un système d'information peut être construit pour l'exécution de processus (système transactionnel) ou pour l'évaluation des processus (système analytique). Cette distinction dirige certains choix de modélisation et de mise en ouvre. Par exemple, un système transactionnel d'une Université enregistre des informations sur les personnes étudiantes, sur les facultés et sur les activités pédagogiques offertes par département. Cela nécessite la spécification d'interactions spécifiques entre une base de données et plusieurs intervenants pour ajouter, mettre à jour ou supprimer des données d'une façon concurrente. Ces interactions permettent de garder l'intégrité de l'état courant des entités modélisées assurant le bon fonctionnement des processus. Or, avec un système analytique, on vise principalement la synthèse d'information à partir des informations obtenues en cours d'exécution des processus. Par exemple, le taux de recrutement au 1er cycle par département, le taux de réussite au doctorat, la variation du nombre

d'inscriptions dans les 5 dernières années. La synthèse d'information nécessite le rassemblement (l'agrégation) de plusieurs données accumulées au fil du temps.

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

Un entrepôt de données (*data warehouse*) fournit une vue unifiée des données provenant de différents systèmes (sources de données) pour augmenter la couverture et l'efficacité de la prise de décisions stratégiques. Une prise de décision stratégique est une action entreprise par les décideurs afin d'améliorer la performance de l'organisation.

Le modèle d'un entrepôt de données est le plus souvent construit selon la méthode de modélisation dimensionnelle. Un modèle dimensionnel vise à modéliser les mesures des processus d'un domaine. Le modèle dimensionnel est constitué d'un ensemble de mesures permettant d'analyser et d'évaluer un processus tout en permettant la documentation du 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, 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

- 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)
- ...

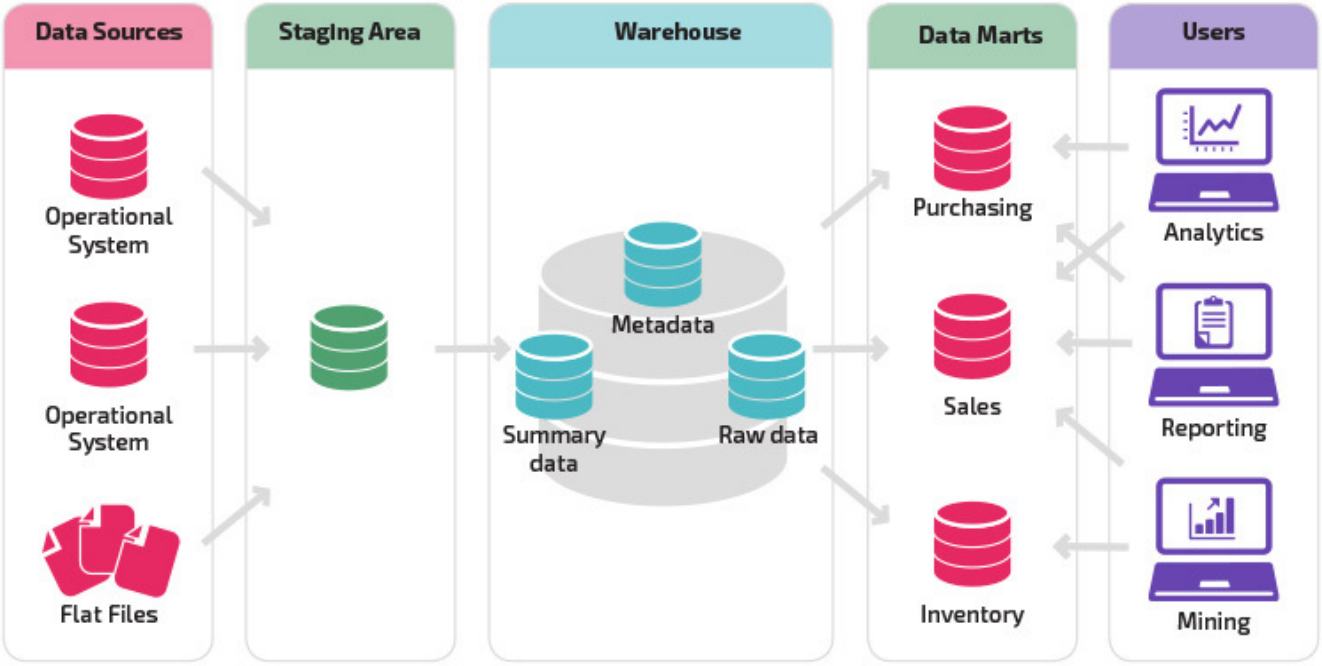
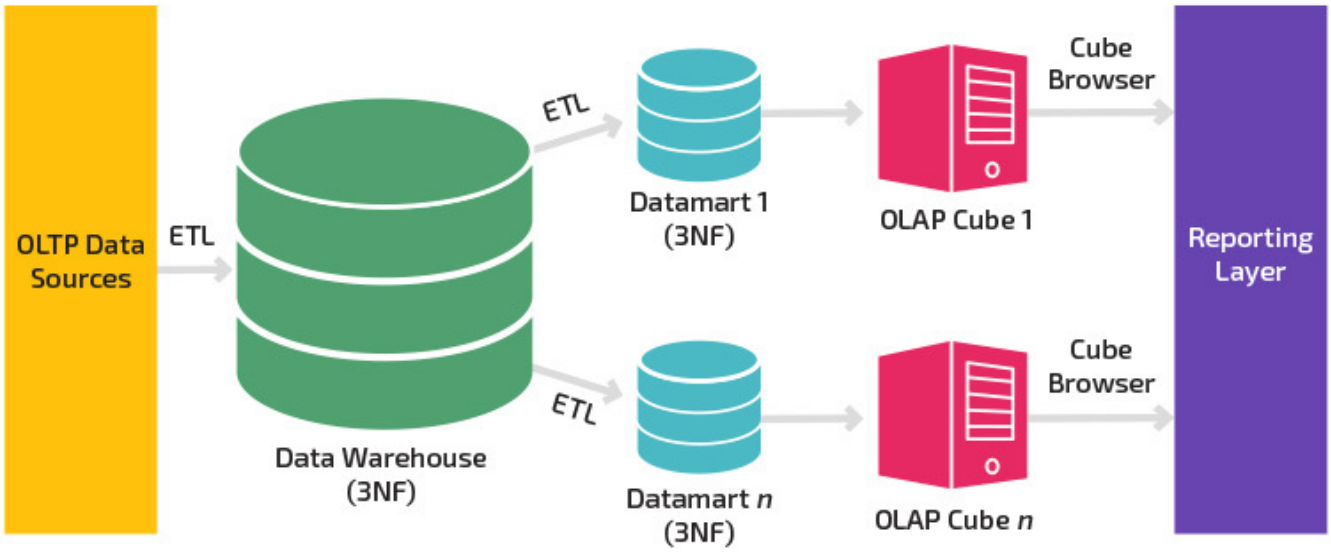
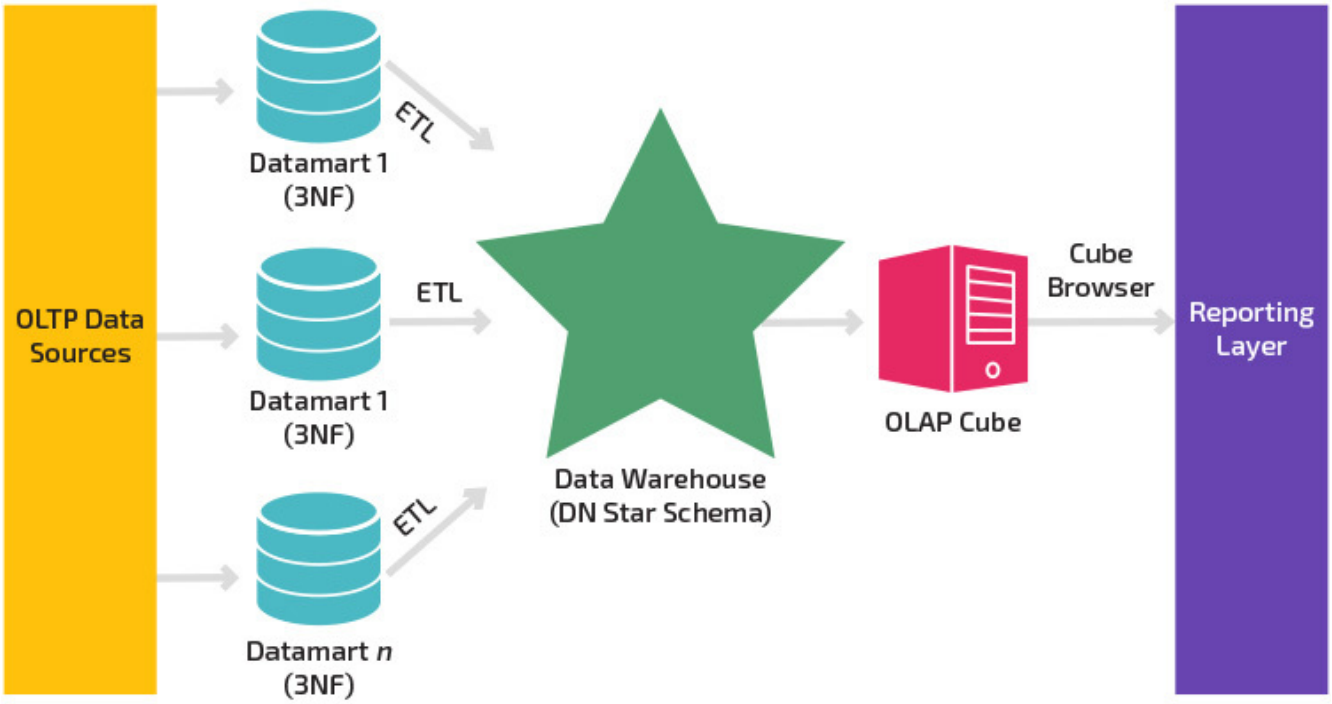


Figure 1. Entrepôt de données (https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/)

Inmon Model



Kimball Model



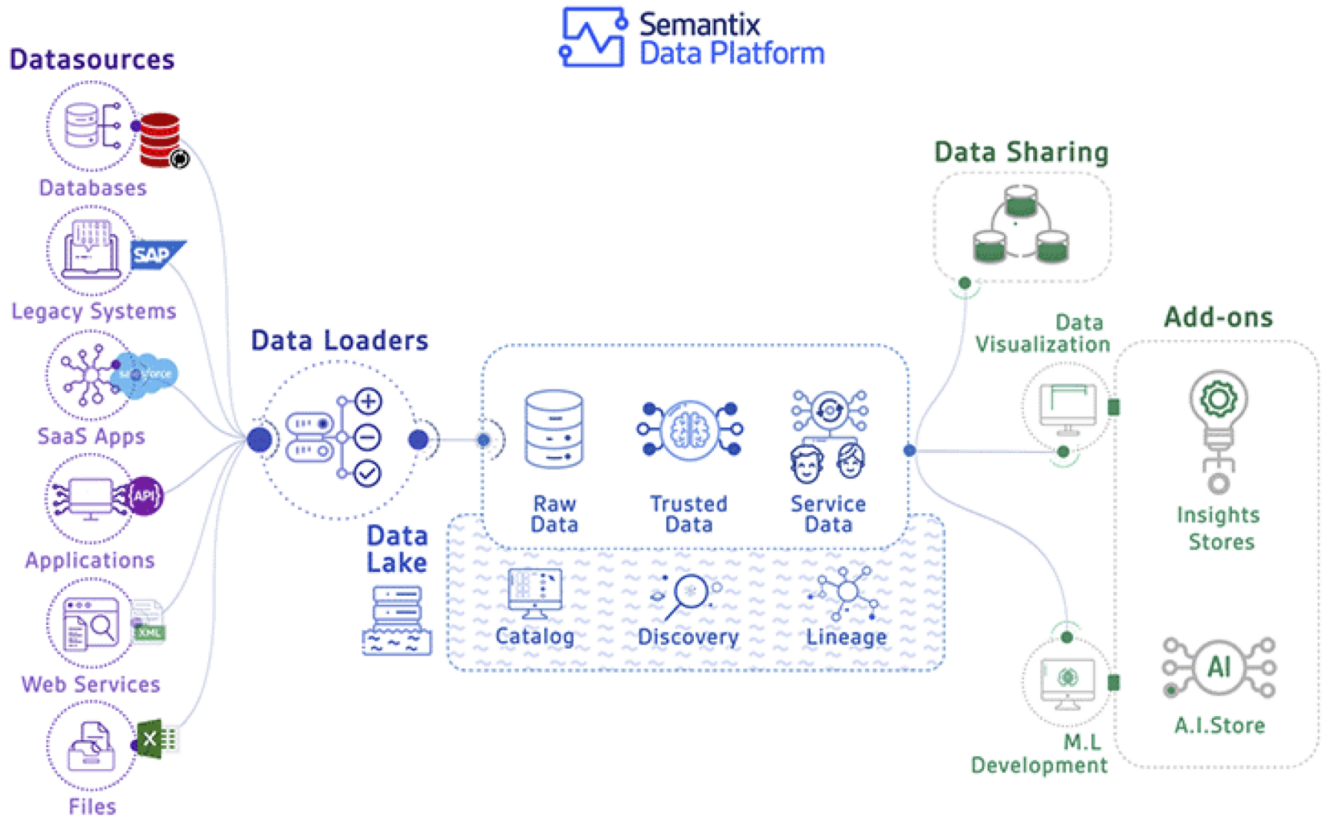


Figure 2. Lac de données (https://semantix.com.br/data-platform/)

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

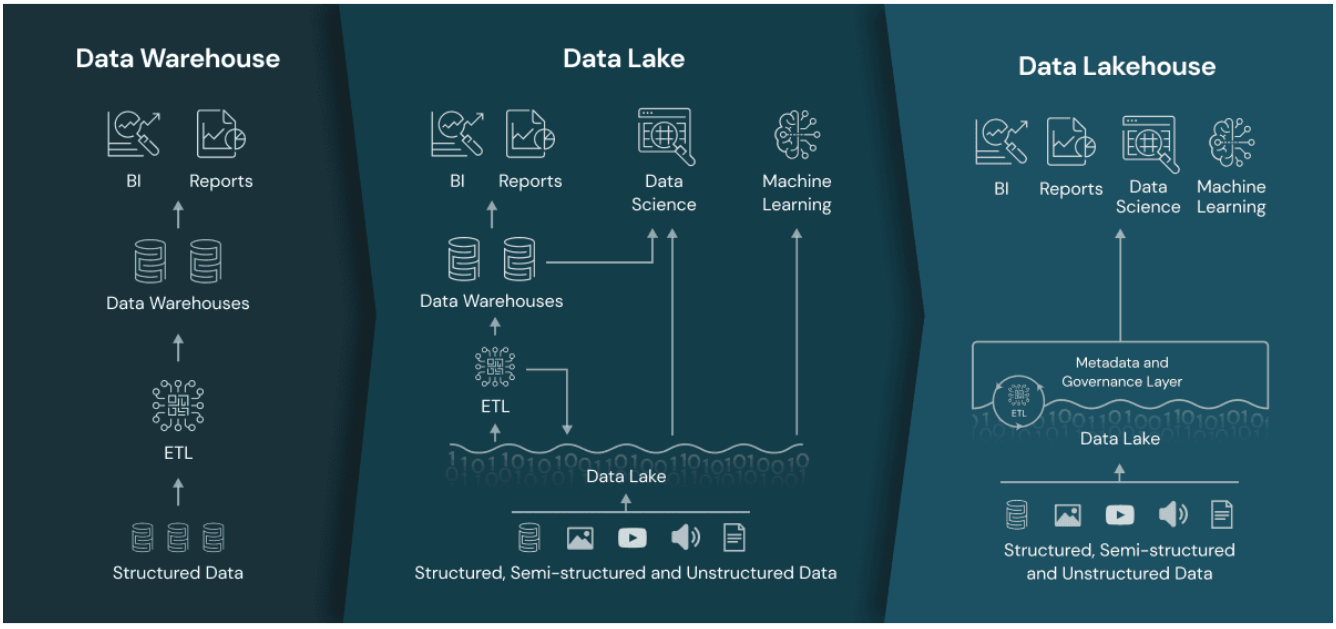


Figure 4. Lac d'entrepôts de données (https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html)

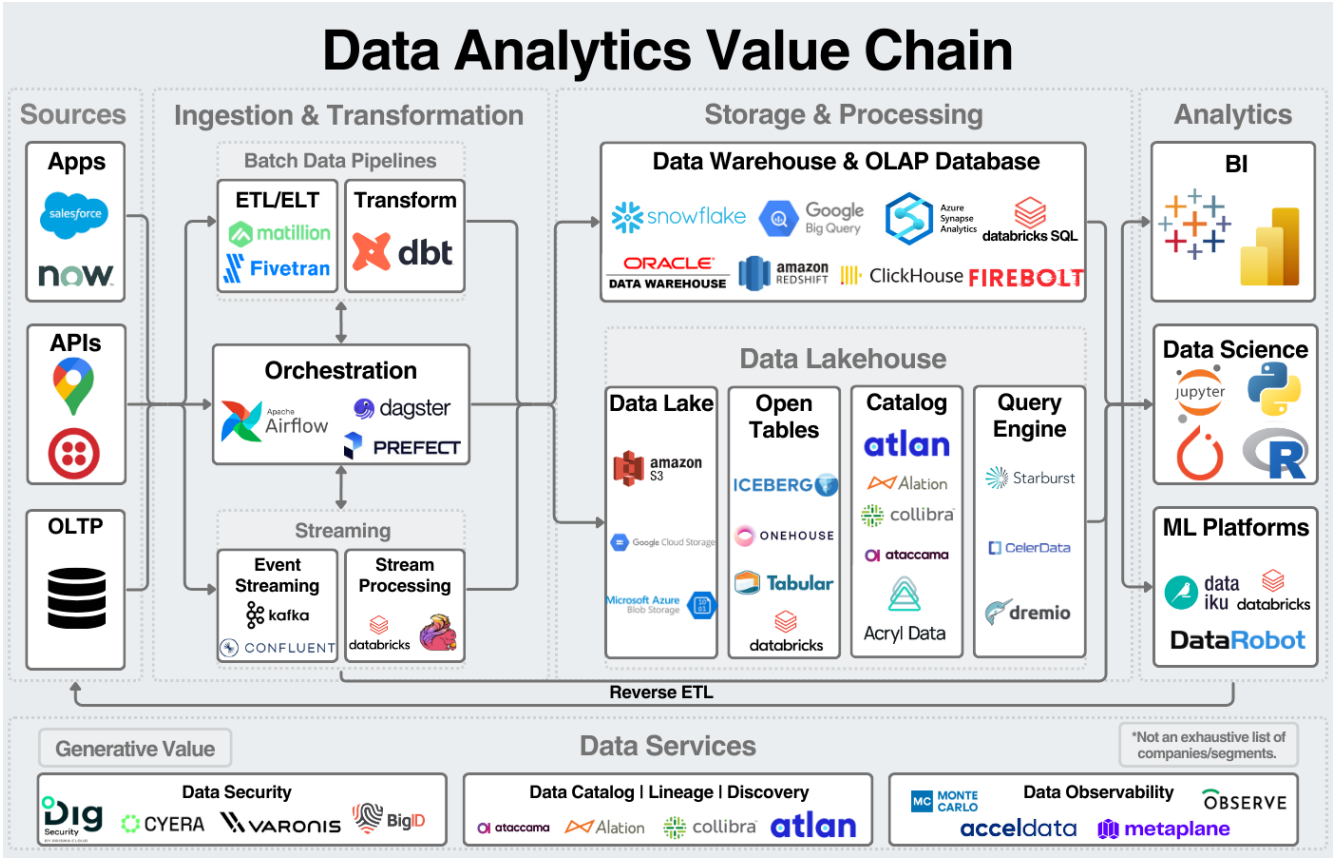


Figure 5. Technologies analytiques 2024 (https://www.generativevalue.com/p/a-primer-on-data-warehouses)

2. Définition du modèle

Le modèle dimensionnel est formé de deux composants : fait et dimension. La modélisation dimensionnelle vise à définir comment les processus sont mesurés.

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.

Note éditoriale

Ces définitions de clé «naturelle» (par opposition à «artifielle») sont difficilement opérationalisables. Il est plus juste (et arbitrable) de fonder la définition sur la présence (ou l'absence) de sémantique externe associée à la clé. Dès qu'il y a une sémantique externe au SGBD, celui-ci ne peut la contrôler. D'où l'importance, parfois, de créer une clé sous le seul contrôle du SGBD.

C'est notamment pour cette raison que j'ai préféré utiliser les étiquettes «interne» et «externe» aux appellations plus usuelles.

Dans une table dimensionnelle, toutes les clés issues de l'alimentation ont une sémantique externe, que le SGBD de l'entrepôt ne contrôle donc pas. Or, il lui faut contrôler au moins une clé afin d'arbitrer la multiplicité et les variations des sources.

Par ailleurs, une dimension est une entité qui **caractérise** nécessairement (**au moins un**) un processus **mesuré** : * caractérise : sans cette dimension, le processus (j'aimerais mieux dire l'évènement) ne peut à coup sûr être distingué d'un autre ; pour être plus spécifique, il serait préférable d'utiliser l'expression «qui participe à l'identification de l'évènement) ; * au moins un : si la dimension ne caractérise aucun processus (possiblement indirectement dans le cas d'un flocon)… pourquoi est-elle là ? * mesuré : ici, j'hésite à en faire une obligation — faute d'attribut de mesure (de «fait») on aurait quand même la confirmation de l'existence de l'évènement (mais comment l'existence pourrait-elle être constatée sans mesure?)

La clé artificielle ne doit pas avoir une sémantique associée. Elle est souvent représentée par un attribut de type UUID (Universally Unique IDentifier) ou un entier généré lors du processus d'alimentation (en utilisant les mécanismes appropriés mis à disposition par le SGBD, par exemple, INTEGER GENERATED ALWAYS AS IDENTITY en SQL).

Finalement, cette définition est différente de la définition traditionnelle en ce qu'elle fait référence à la source de données. Or, c'est bien embêtant. Car la source change, évolue. Utiliser la dépendance fonctionnelle et la sémantique est un fondement plus solide.

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.

Cette catégorisation est empirique, non formalisée et contestée. Elle permet toutefois d'énoncer des règles de pratiques en utilisant un vocabulaire commun.

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
- Les catégories d'attributs ne sont pas exclusives.
- Un attribut calculé qui ne peut pas être utilisé pour faire des sommations ne doit pas faire partie de la relation factuelle. (ex. pourcentage de vente)
- Une bonne règle de pratique me semble être la suivante : tout attribut non-clé doit être agrégeable, il faut éviter les attributs dérivés et synthétisés sous peine d'incohérence :
  - attribut dérivé, lors de la mise à jour du fait (problème bénin, une discipline stricte peut pallier ce problème) ;
  - attribut synthétisé, lors de la mise à jour des tuples des dimensions concernées (problème important, car il est vraisemblablement trop couteux de déployer tous les automatismes requis).

3. Mise en oeuvre du modèle

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

Un modèle dimensionnel est mis en oeuvre dans une base de données relationnelle selon différentes formes : étoile, flocon ou constellation.

Un schéma dimensionnel est composé de deux types de relation (table) : relation factuelle et relation dimensionnelle. Ce schéma est optimisé en vue de la consultation (R), de la recherche d'informations et de la synthèse (agrégations) de mesure. Il est en général convenu que ces schémas ne sont pas modifiables en cours d'exploitation et que les mises à jour (U) demeurent exceptionnelles, le plus souvent limitées à des ajouts (C) et les suppressions totalement exclues.

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.

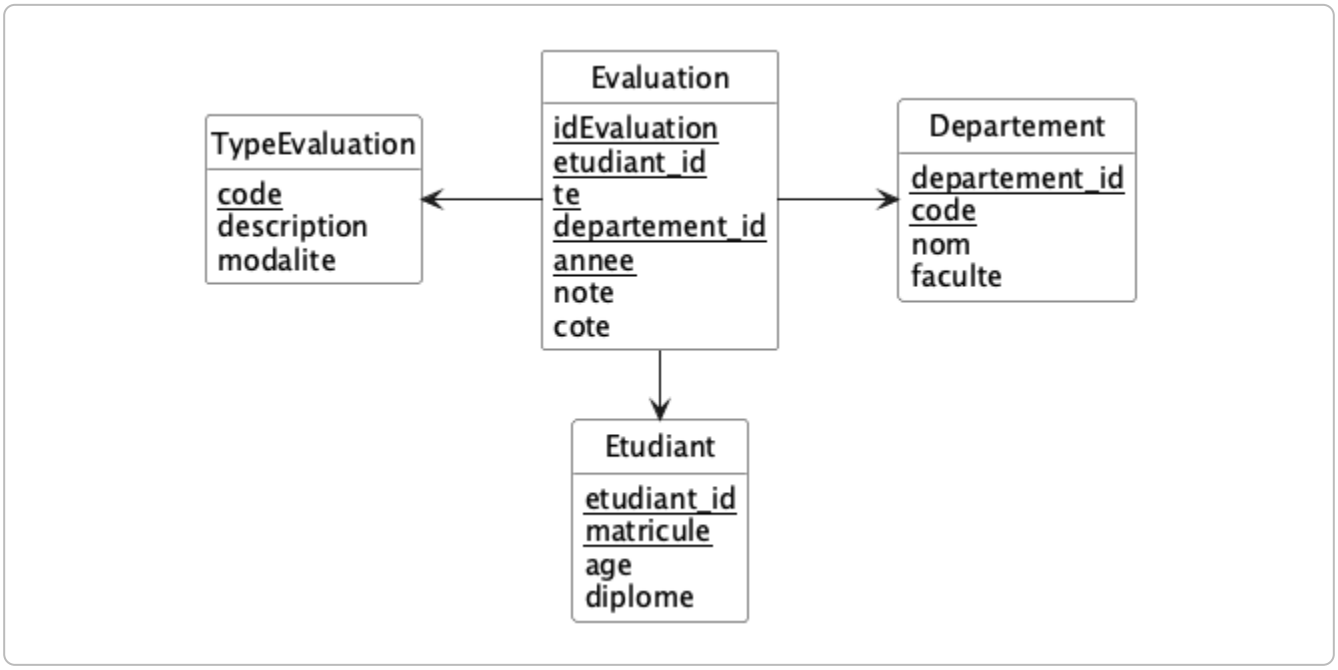
3.1. Schéma en étoile

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

Un schéma en étoile est formé d'une relation factuelle qui représente un processus et des relations dimensionnelles directement liées qui décrivent un fait.

Notez que les relations sont rarement en troisième forme normale (parfois même pas en 1re forme normale), pour faciliter la formulation de requêtes !! (pas sûr, mais bon…)

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



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

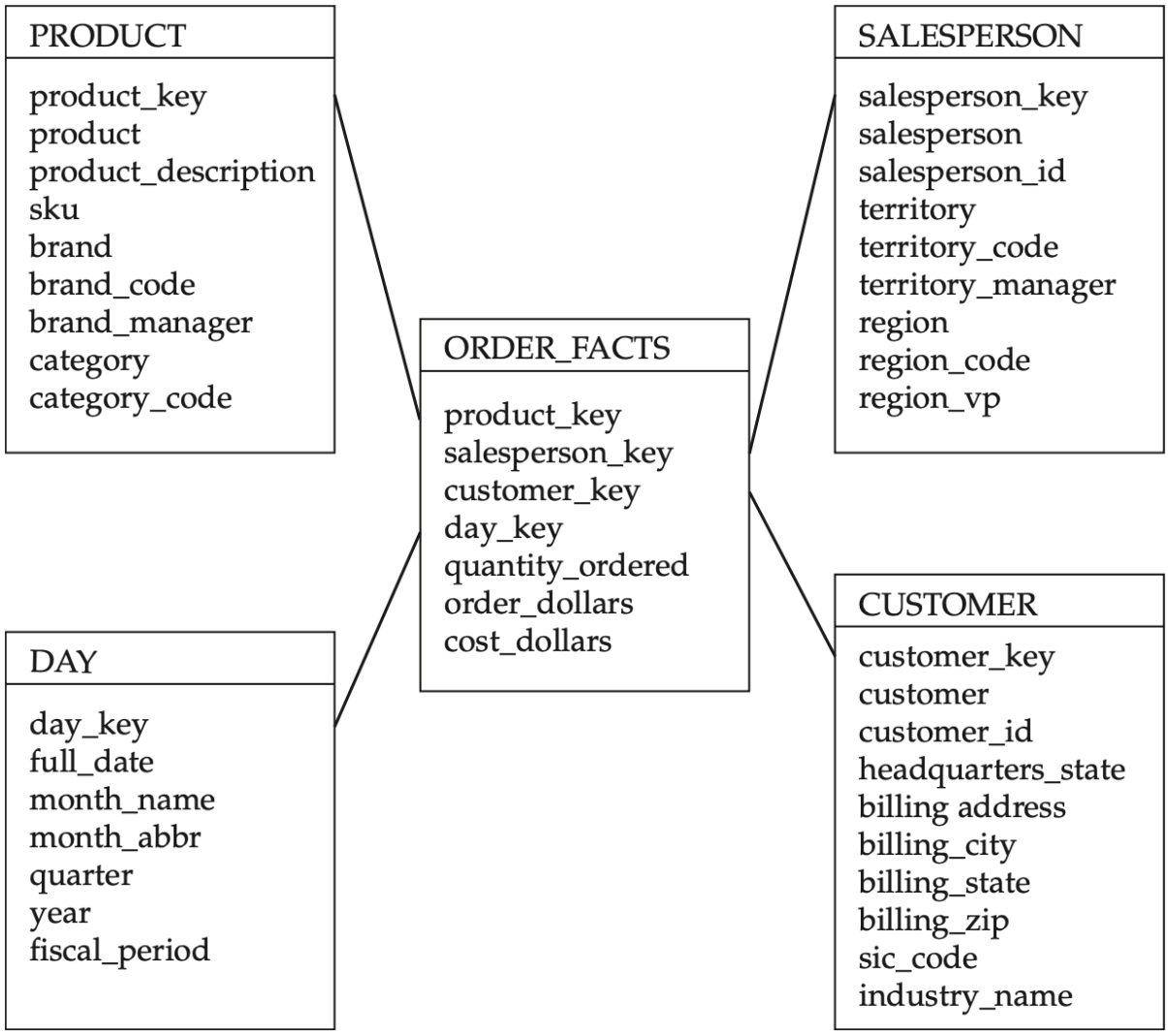


Figure 1-5 A simple star schema for the orders process

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

Un schéma en flocon est un schéma en étoile où les dimensions sont normalisées.

Une dimension peut être normalisée de deux façons :

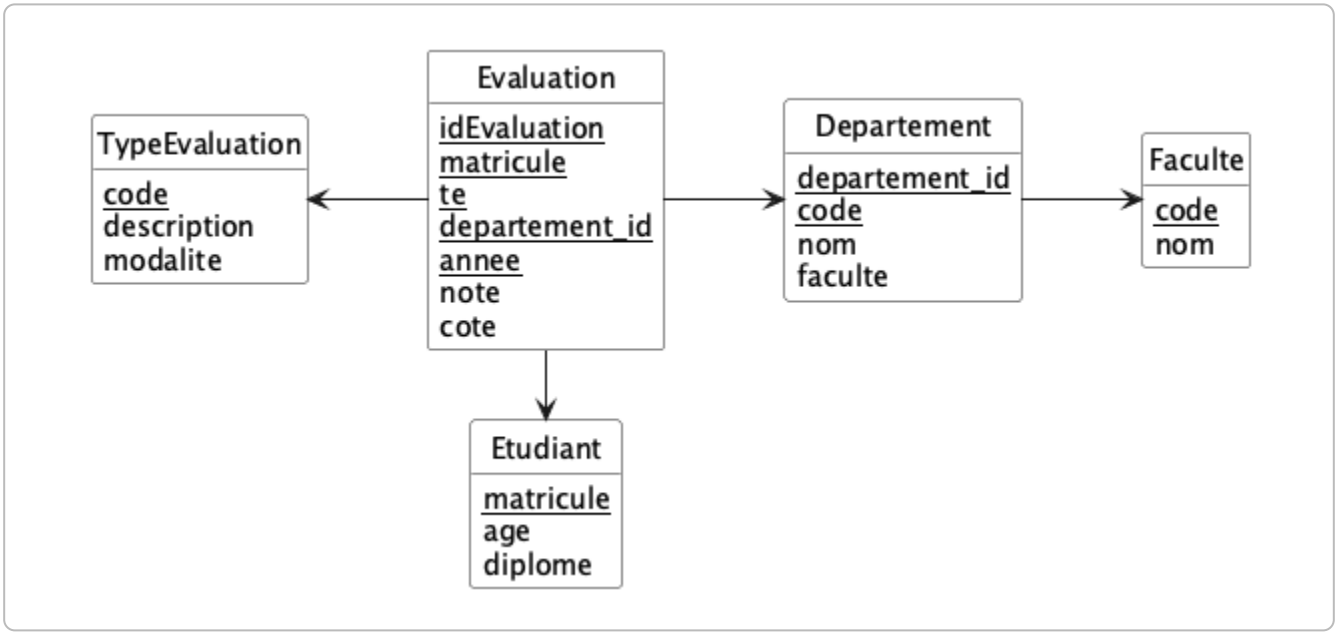
- normalisation en première forme normale, ce qui consiste à créer une relation pour chaque attribut multivalué (un attribut qui peut avoir plusieurs valeurs) ou annulable.
- normalisation hiérarchique, ce qui consiste à créer une relation par niveau de détail du plus granulaire au moins granulaire. La première relation représente le niveau le plus granulaire et le dernier niveau est le niveau le moins granulaire. Chaque relation est définie par une clé artificielle et les attributs propres du niveau de granularité. La relation du niveau le plus granulaire est reliée à la relation du niveau supérieur par une clé référentielle.

Cette pratique n'est pas recommandée dans le monde des entrepôts de données, même si la dénormalisation engendre des doublons. La normalisation dans le contexte d'un entrepôt de données compliquerait l'alimentation et ralentirait l'exécution des requêtes. L'intégrité des données est réputée être garantie par le processus d'alimentation.

Question

Quelles sont les études validées par les pairs ainsi que les méta-analyses corroborant ces trois affirmations ?

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



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

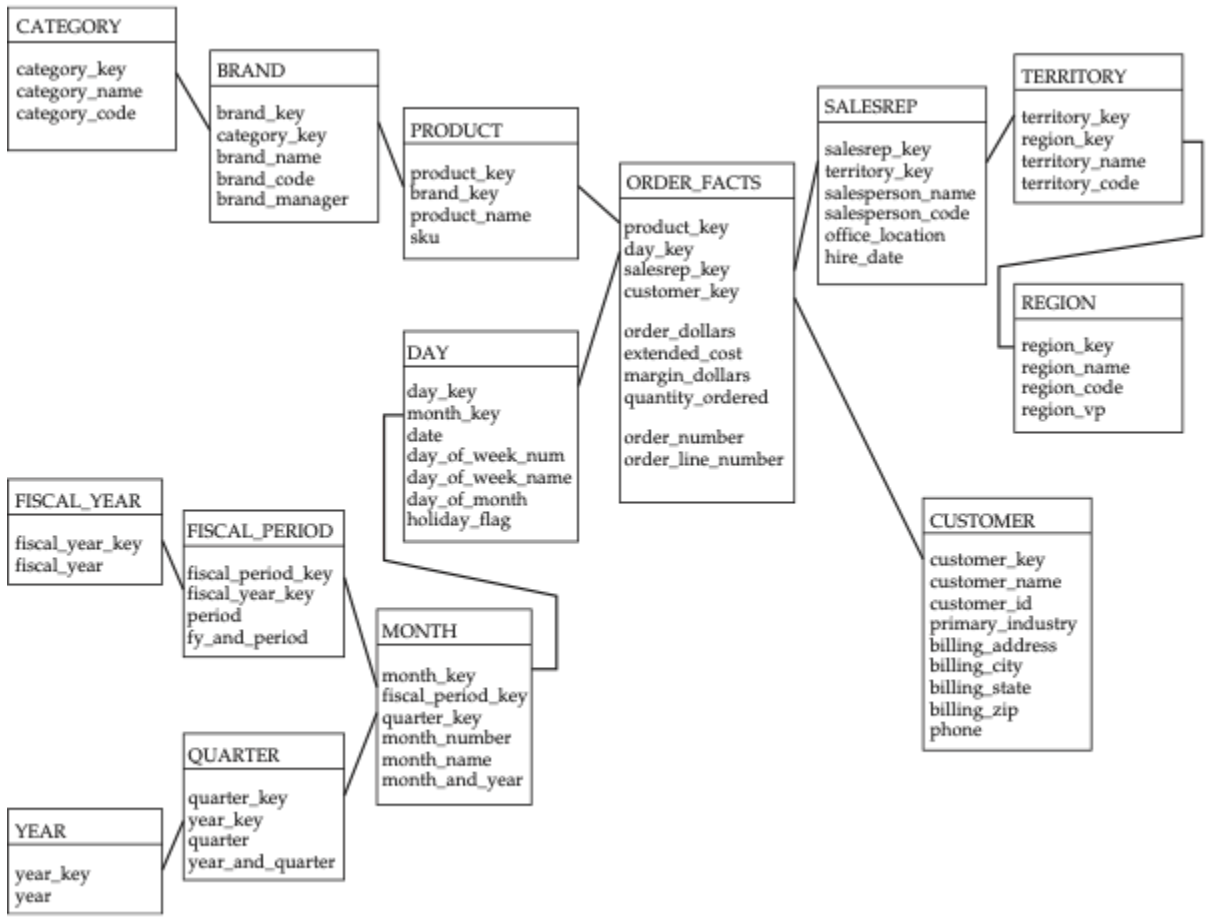


Figure 7-5 A snowflake schema

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

Un schéma en constellation est composé de plusieurs schémas d'étoiles qui partagent des dimensions. Conséquemment, il n'y a pas de relation de clés référentielles entre les diverses relations factuelles, mais uniquement entre les relations factuelles et les relations dimensionnelles.

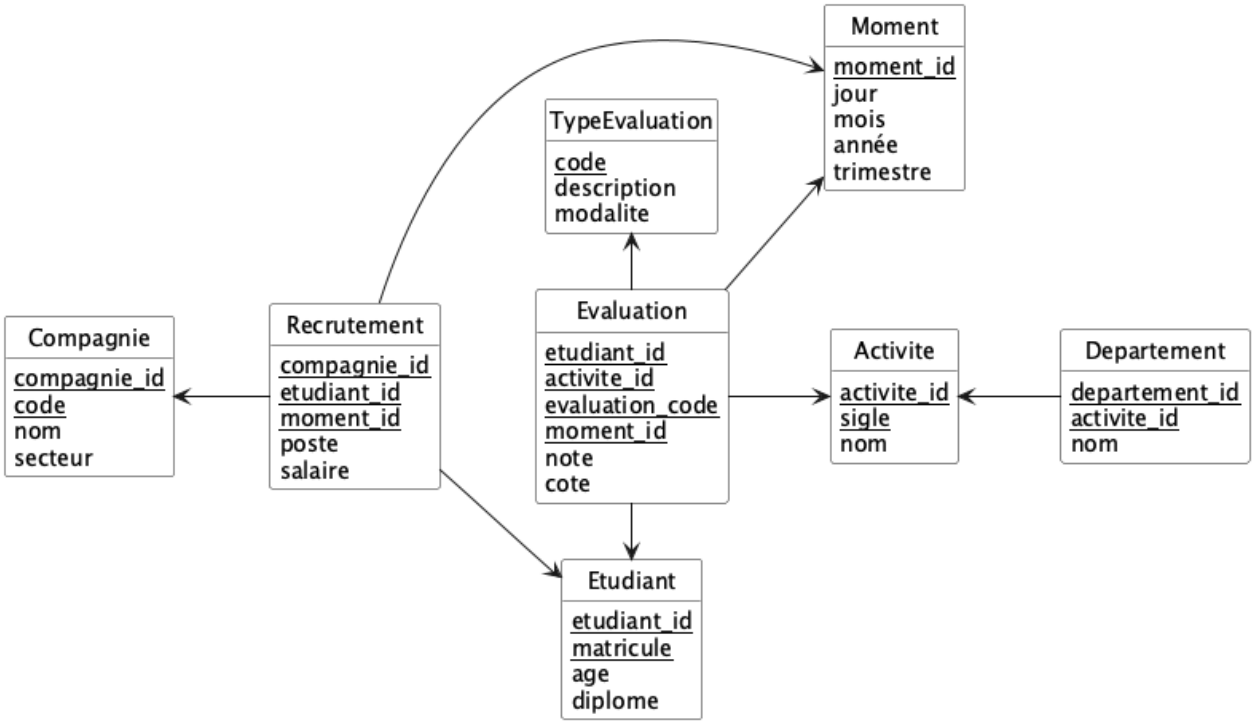
Généralement, dans un domaine (ou une organisation), il y a plusieurs processus qui ont lieu à différents moments. Dans des situations où vous évaluez individuellement les processus, une relation factuelle par processus est nécessaire. Pour choisir la bonne modélisation, deux questions se posent. Soit deux faits :

1. Ces faits se produisent-ils simultanément (moment différent)?
2. Ces faits sont-ils disponibles au même niveau de détail (granularité différente)?

Si la réponse à l'une de ces questions est « non », les faits représentent des processus différents.

Lorsque deux faits décrivent des évènements qui se produisent à différents moments, à différents niveaux de granularité ou nécessitent des dimensions différentes, cela représente deux processus.

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

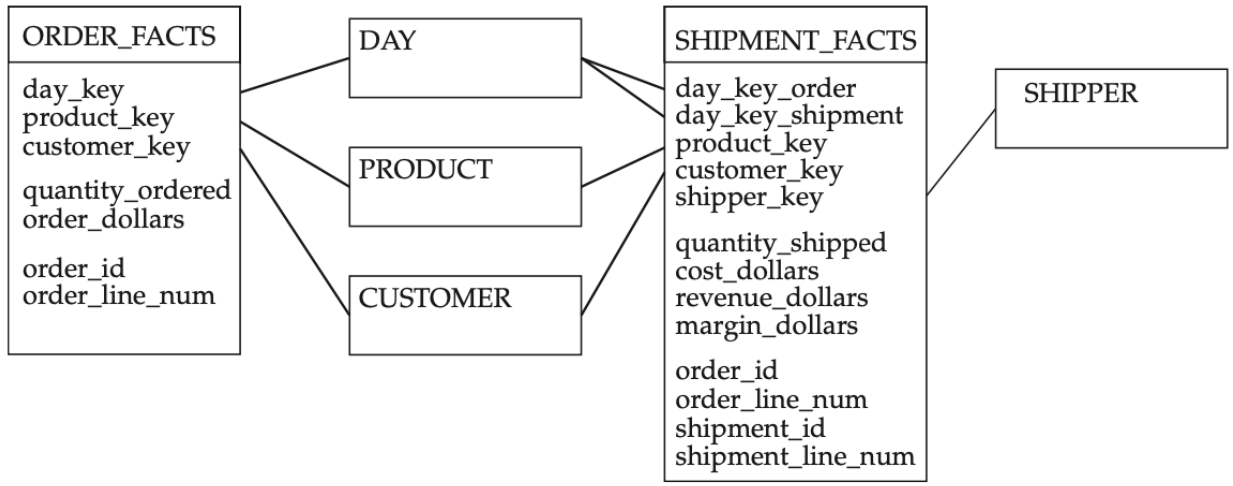


Figure 4-8 Separating the two processes into separate fact tables

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.

Glossaire

Clé

En regard d'une relation, une clé est un ensemble d'attributs qui détermine fonctionnellement tous les autres attributs de la relation. Ainsi, deux tuples d'une même relation ne peuvent avoir la même (valeur de) clé. Une clé est dite *stricte* si aucun de ses attributs ne peut en être retiré sans qu'elle perde la propriété de clé. Cette notion s'applique tout aussi bien aux classes et aux ensembles d'entités. Une clé stricte est parfois appelée *clé candidate*, un calque de *candidate key* en langue états-unienne.
Orthographe le nom *clé* est fréquemment utilisé en apposition, par exemple *un attribut clé*. Au pluriel, conformément à la règle régissant l'apposition, il reste invariable, *des attributs clé*. Il est aussi utilisé pour caractériser une entité qui n'est pas une clé, *une non-clé, des attributs non-clé*, **avec** trait d'union comme le prescrit la règle régissant la négation des noms. Votre correcteur orthographique n'est pas de cet avis ? Alors, il s'agit vraisemblablement d'un outil utilisant l'IA, donc incapable de conjuguer deux règles à la fois !

Clé externe (naturelle)

Une clé externe est une clé stricte non interne (voir clé interne).

Clé interne (clé artificielle, *surrogate key*)

Une clé interne est une clé stricte déterminée indépendamment du domaine d'application (donc non fondée sur le modèle de connaissances, le modèle ontologique, le modèle conceptuel, la sémantique de la source de données, la pratique du métier, le contexte d'utilisation).
Corolaire : Toute clé stricte est soit interne, soit externe.
Note 1 : Une clé externe ayant une sémantique dépendante du domaine d'application est susceptible de devoir être modifiée afin de refléter adéquatement les caractéristiques de l'entité qu'elle détermine (dépendance fonctionnelle cachée). À défaut de quoi, elle pourrait induire une interprétation incorrecte ou perdre sa propriété de clé.
Note 2 : Souvent, cette clé est insuffisante dans un contexte historique parce que sa valeur, comme sa sémantique, est susceptible d'évoluer dans le temps. D'où l'importance, parfois, de créer une clé sous le seul contrôle du MLD par l'entremise du SGBD.
Note 3 : Une clé interne prend souvent la forme d'une suite de symboles générés séquentiellement, chronologiquement ou pseudo-aléatoirement.

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