Université de Sherbrooke

Bases de données dimensionnelles

Modèle dimensionnel

UdeS:BDD_00

Christina KHNAISSER *(christina.khnaisser@usherbrooke.ca)*
Luc LAVOIE *(luc.lavoie@usherbrooke.ca)*
(les auteurs sont cités en ordre alphabétique nominal)

—

CoFELI/Scriptorum/BDD_00-Modele, version 1.2.1.b, en date du 2025-11-12
— 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
2025-11-12	LL	Mise à jour et corrections mineures.
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 de la théorie relationnelle, de la modélisation logique et des bases de données relationnelles.

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 pratique 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 [Adamson2010a, Adamson 2017a] et Jiang [Jiang2015a].

1. Mise en contexte

Un système d'information peut être construit pour le soutien à l'exécution de processus (système transactionnel, opérationnel) ou pour l'évaluation des processus (système analytique). Cette distinction oriente certains choix de modélisation et de mise en oeuvre.

Le maintien d'une image fidèle de l'état courant requis par le système opérationnel nécessite une réactivité importante et uniforme en même temps qu'un maintien rigoureux de la cohérence et de l'intégrité des informations. La synthèse d'informations par le système analytique nécessite le rassemblement (l'agrégation) de données provenant de plusieurs sources et accumulées au fil du temps.

Exemple

Le 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 de façon concurrente tout en conservant l'intégrité de la BD assurant le soutien au processus (en particulier les informations « en temps réel » sur l'état du processus).

Son système analytique vise principalement la synthèse d'informations à partir de celles obtenues en cours d'exécution des processus durant une période donnée et leur comparaison avec d'autres informations provenant de sources externes (services délocalisés, ministères, organismes subventionnaires, fondations, etc.). Par exemple, l'évolution du taux de recrutement au 1er cycle par département au cours des dix dernières années par rapport aux autres universités, les sources de financement des cent derniers doctorants, la variation du nombre d'inscriptions aux activités au cours des cinq dernières années en fonction des données de la santé publique et du ministère l'enseignement supérieur.

Modèle transactionnel

- Objectifs
  - Soutien « en temps réel » aux processus
- Besoins
  - Refléter l'état courant de chacune des entités informationnelles
- Problèmes
  - Maintien de l'intégrité au fil des modifications

    - Multiplicité des intervenants
    - Interactions concurrentes

Modèle analytique

- Objectifs
    - Analyse des processus selon différentes perspectives et sur différentes échelles de temps
- Besoins
    - Agréger des données provenant de plusieurs sources
    - Synthétiser des informations relatives à plusieurs processus
- Problèmes
    - Multiplicité des sources requises
    - Hétérogénéité consécutive
        - des modèles de connaissances
        - des modèles conceptuels
        - des modèles logiques
        - des technologies
        - des règles légales
        - des règles de gouvernance
        - des règles éthiques
    - Indépendance des évolutions des sources

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 (maintien intégrité et cohérence)	Recherche (agrégation et évaluation)
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 usuellement appliqués	Normalisation 1FN, FNBC *voire parfois 5FN*	Normalisation 1FN, 3FN *voire parfois 6FN*

1.2. Entrepôt de données

Vision classique

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.

L'expression « entrepôt de données » est entrée dans l'usage, mais elle est inappropriée. En effet, une base de données analytique ne peut se résumer à un entrepôt (lieu, bâtiment ou dispositif où l'on dépose *temporairement* des marchandises, des données) puisqu'elle a pour mission de conserver et rendre accessible les données historiques sur le long terme (de façon analogue à un service d'archive).

Vision 2020

Un entrepôt de données est un service d'archivage (collecte, sécurisation, conservation) des données d'une organisation offrant des fonctionnalités d'accès à celles-ci selon différents modèles d'interprétation.

La base de données dimensionnelle est l'un de ces modèles. Elle est conçue dans le but de faciliter l'étude des processus de l'organisation en regroupant les données pertinentes à chacun des évènements des processus. Le niveau de détail des évènements peut varier (activité, tâche, etc.), mais tous les évènements d'une même table doivent partager la même ligne de temps.

Les évènements sont décrits dans une table comprenant les attributs caractérisant les évènements (appelées « dimensions ») et les attributs mesurant les évènements (appelés faits). La clé est généralement formée par l'ensemble des dimensions.

Le temps, le lieu, les acteurs, les intrants, les extrants sont les dimensions les plus fréquentes. Le cout, la durée, l'effort, l'énergie, la qualité sont parmi les mesures les plus fréquentes.

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).

Chaine 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.

L'expression « aide à la décision » (dont la portée s'étend à toute décision individuelle, collective ou sociétale) n'est bien sûr pas équivalente à *business intelligence* (dont la

> portée est limitée au secteur commercial).
>
> Dans ce contexte, on peut également s'étonner du couplage des mots *business* (action accomplie au seul profit de l'acteur, indépendamment des répercussions sur autrui ou sur la société) et *intelligence*.

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*)
  - Liang (2015 - *data warehouse*)
- Lac de données (étang de données, *data lake*)
  - James Dixion (2010)
- Entrepôts de lac de données (marais de données, *data lakehouse*)
  - Databricks (2023)

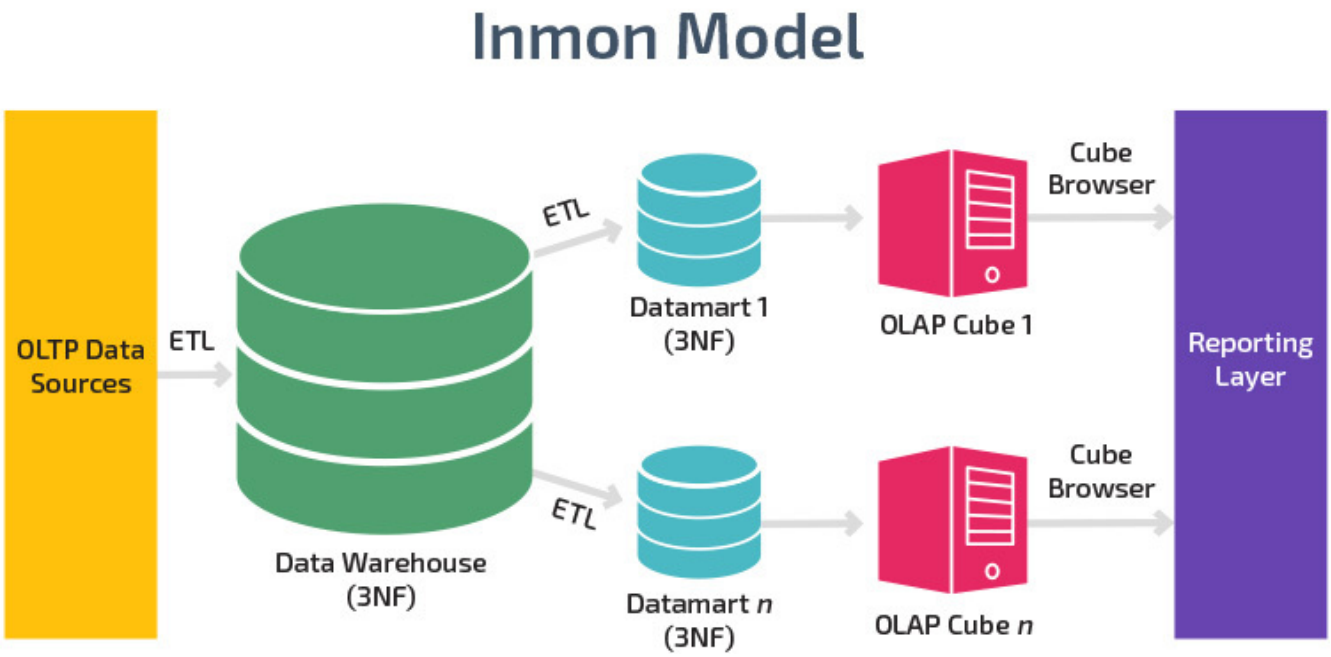


Figure 1. Entrepôt de données - Inmon [Panoply2024a]

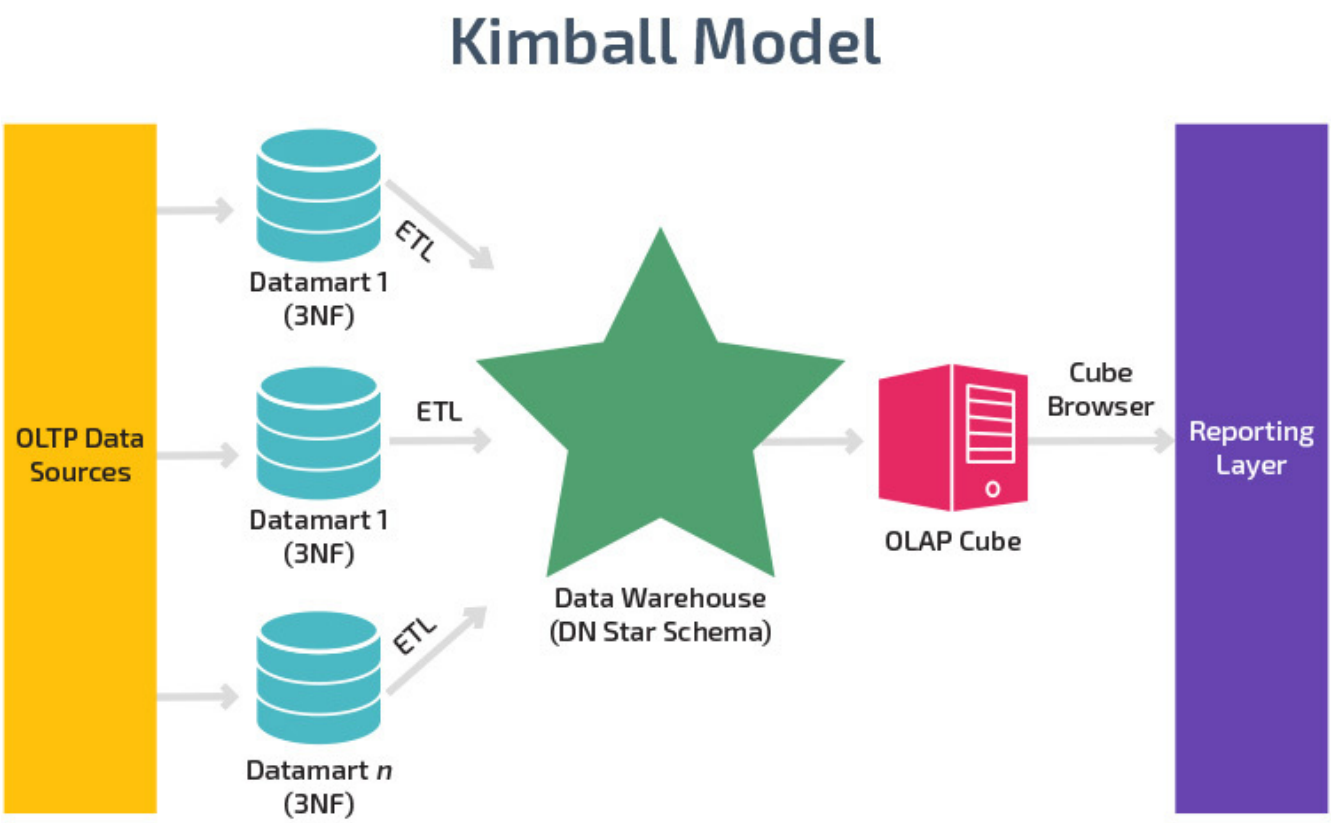


Figure 2. Entrepôt de données - Kimbal [Panoply2024a]

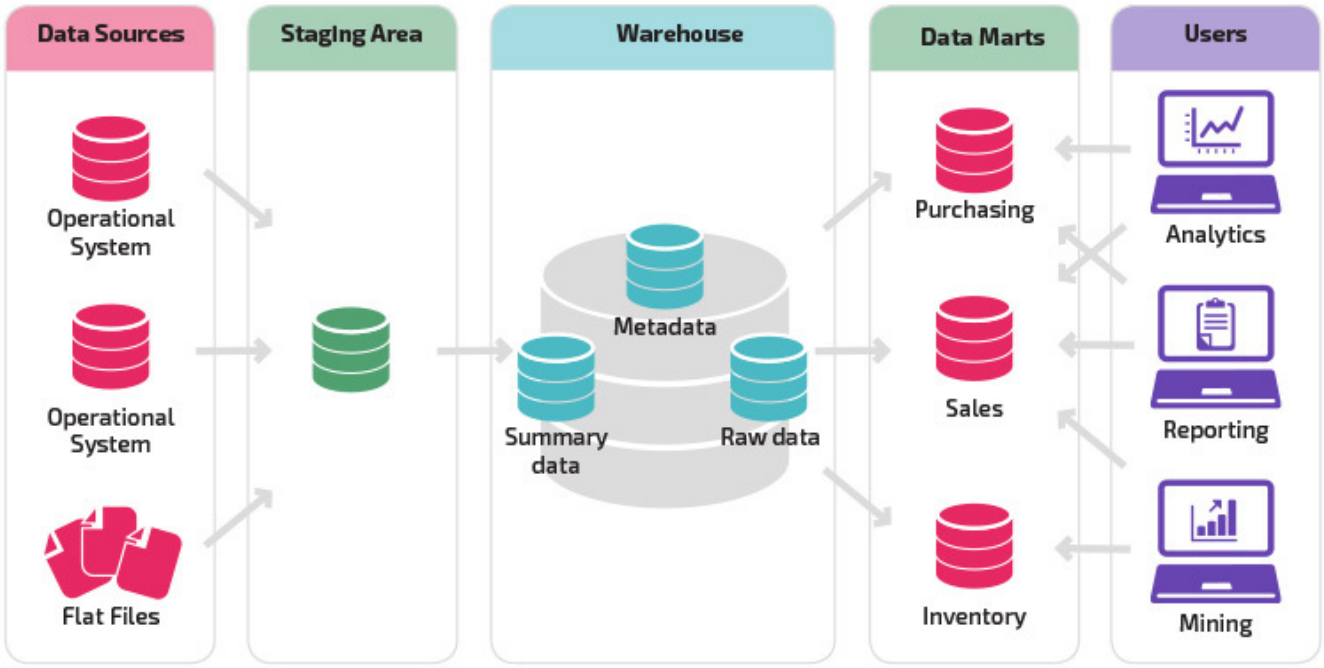


Figure 3. Entrepôt de données - Linstedt [Panoply2024a]

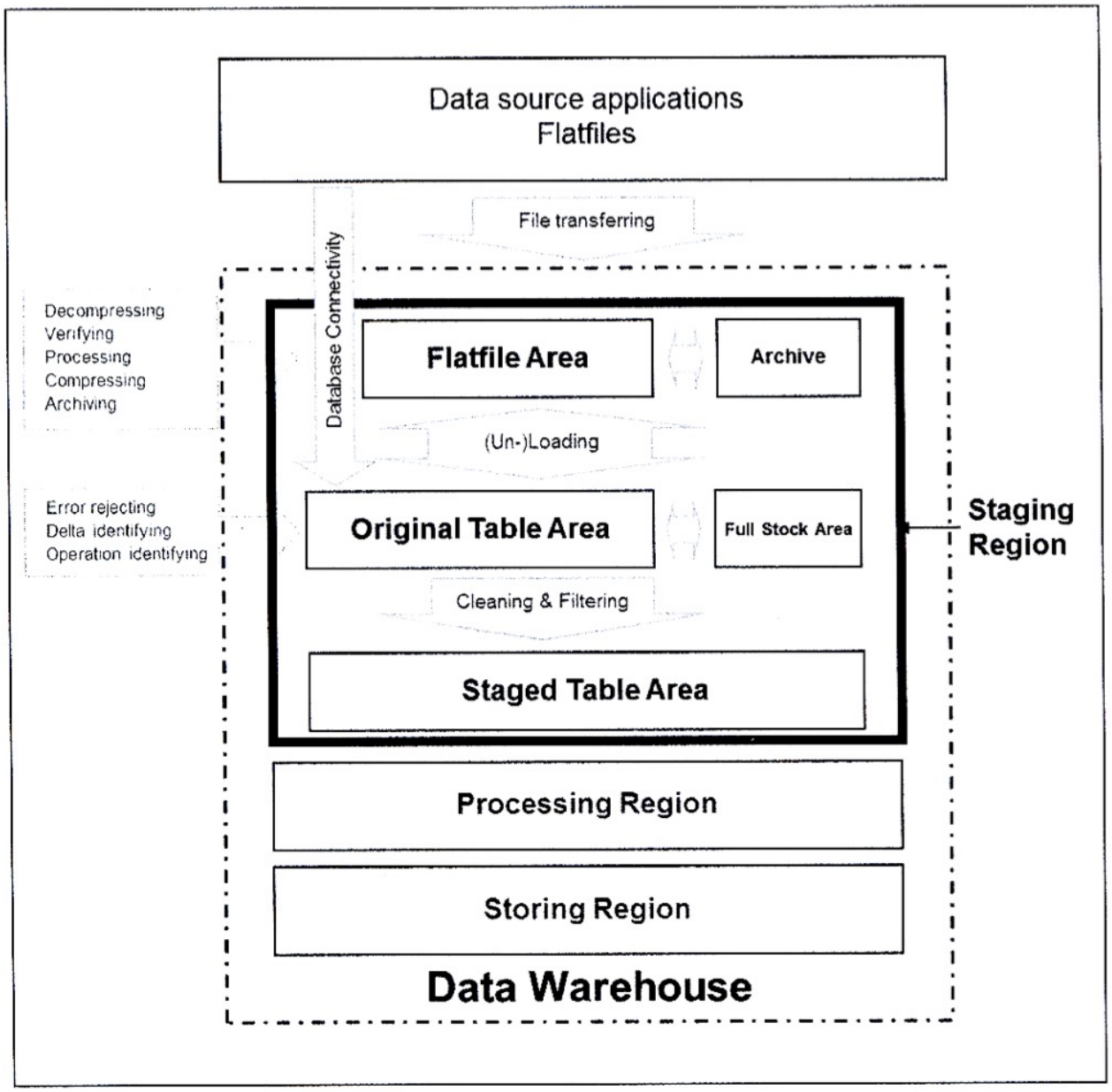


Figure 3.1: Staging Region

Figure 4. Staging region - Jiang [Jiang2015a]

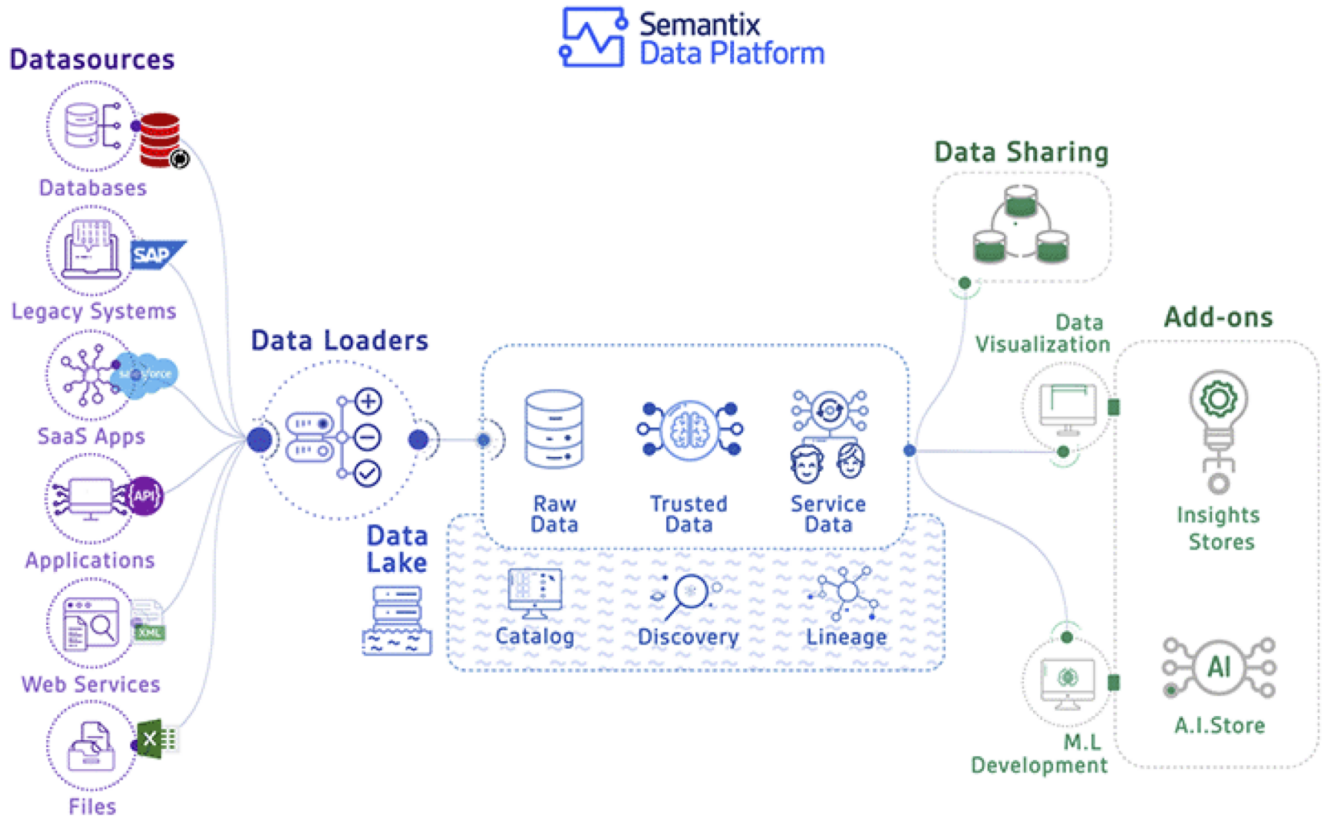


Figure 5. Lac de données (https://semantix.com.br/data-platform/)

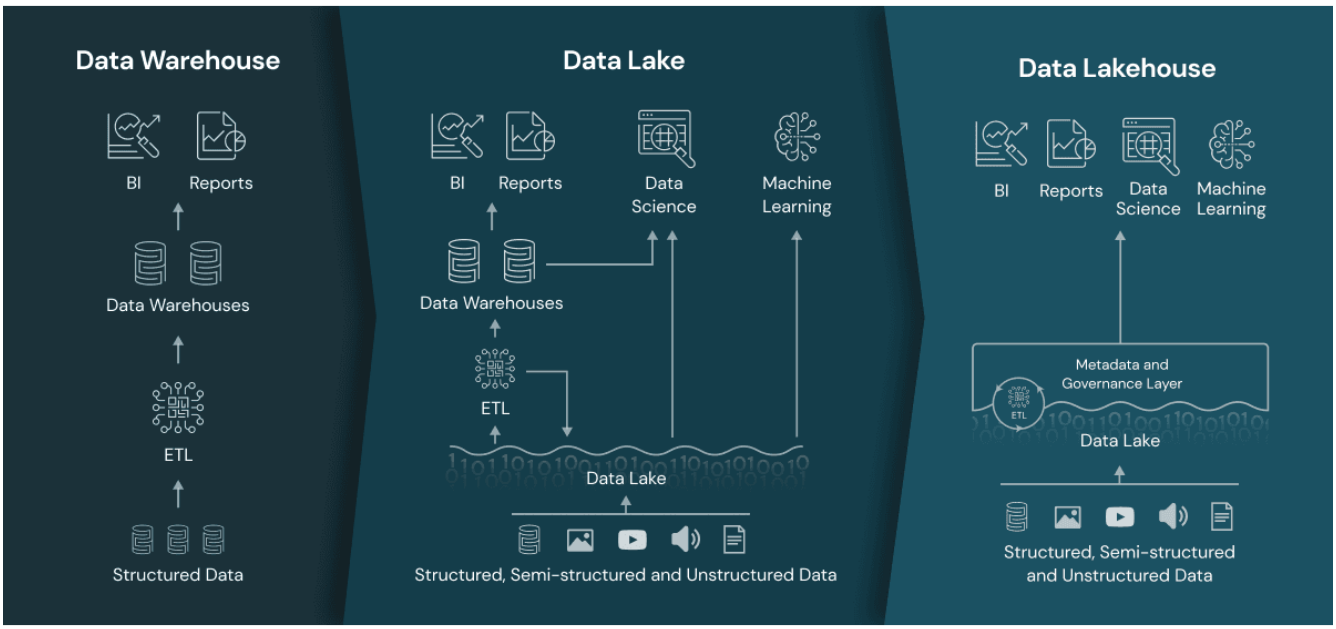


Figure 6. Lac d'entrepôts de données (https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html)

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 7. Gartner Magic Quadrant for Data Integration Tools 2022-2023

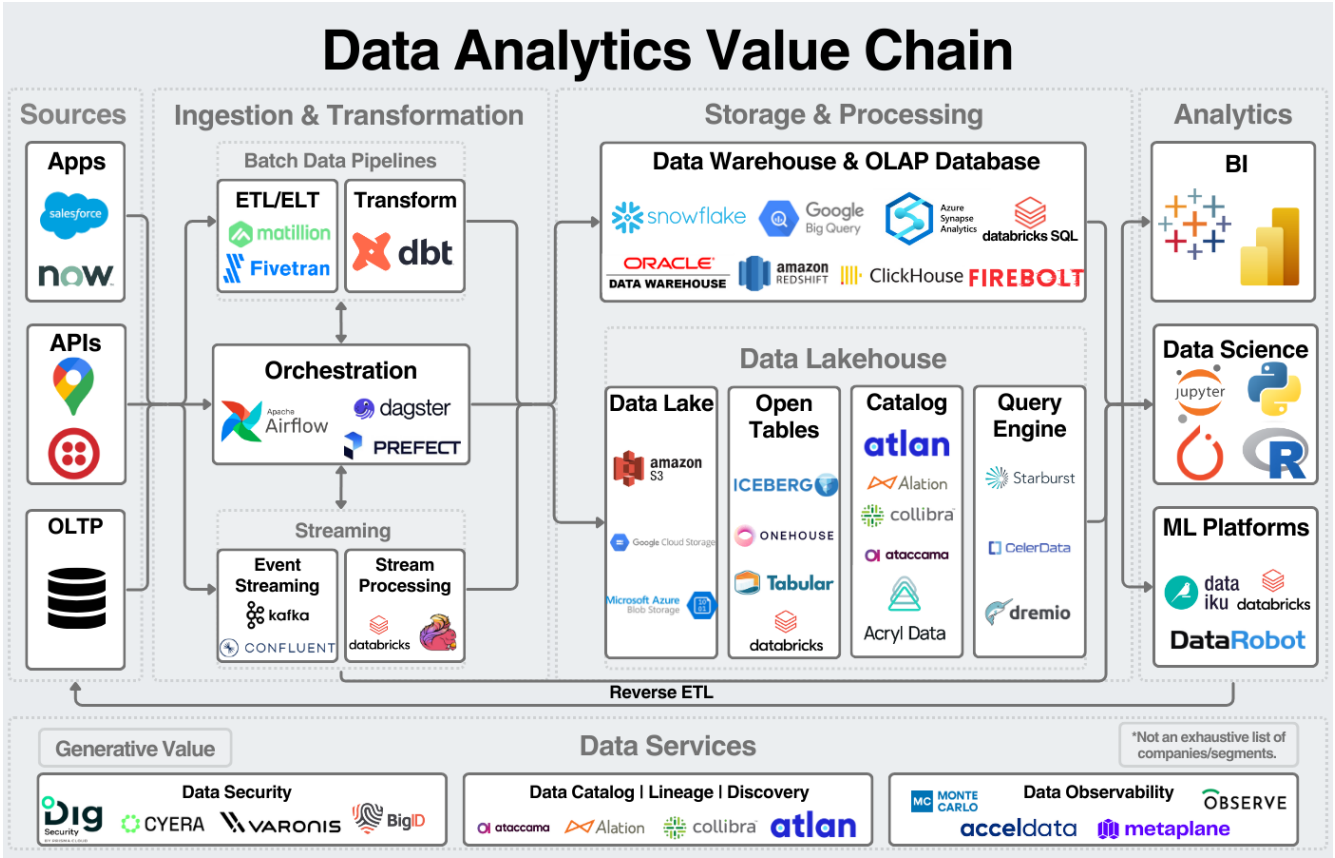


Figure 8. Technologies analytiques 2024 (https://www.generativevalue.com/p/a-primer-on-data-warehouses)

2. Définition du modèle

La modélisation dimensionnelle vise à définir comment les processus sont mesurés. Un modèle dimensionnel repose principalement sur deux types d'entités : les faits et les dimensions.

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).

Plusieurs auteurs distinguent les clés externes comme étant « naturelles » et les clés internes comment étant « artificielles ».

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.

Clé artificielle (interne)

Une clé artificielle est un attribut qui identifie 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 sont difficilement opérationnalisables en pratique : qui est capable de prédire l'avenir sans faute ?

Dans un contexte où :

- il existe une multiplicité de sources (dont les gouvernances sont susceptibles d'être différentes, voire indépendantes),
- d'autres sources pourraient être ajoutées dans le futur,

il est généralement téméraire de postuler qu'une clé externe préserve sa propriété de clé au sein de l'entrepôt et la conservera à long terme.

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 en prescrire la sémantique

ni même la garantir. Or, la seule façon d'être certain qu'une clé n'ait pas de sémantique externe est d'en définir une interne et de ne pas la diffuser à l'externe. D'où l'importance, voire l'obligation, de créer une clé sous le seul contrôle du SGBD et donc, une clé interne.

Pour cette raison, une règle de pratique mise de l'avant par plusieurs auteurs (dont Adamson, Kimball et Jiang), dicte de **toujours** définir une clé interne pour une dimension (et d'y répertorier les différentes clés externes selon les sources).

C'est notamment pour cette raison que les étiquettes « interne » et « externe » sont préférées aux appellations plus « naturelles » et « artificielles » utilisées par certains auteurs. Une clé « artificielle » externe n'est pas un bon choix.

Dans une table dimensionnelle, toutes les clés issues de l'alimentation ont, par définition, 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 des sources et les variations qui sont susceptibles de s'ensuivre.

Par ailleurs, une dimension est une entité qui **caractérise** nécessairement (**au moins un**) un fait d'un processus **mesuré** :

- **caractérise** : sans cette dimension, le processus (en fait, un évènement du processus) ne peut à coup sûr être distingué d'un autre ; pour être plus précis, il serait préférable d'utiliser l'expression « qui participe à l'identification de l'évènement ») ;
- **au moins un** : si la dimension ne caractérise aucun fait d'aucun processus (même indirectement dans le cas d'un flocon)… pourquoi est-elle là ?
- **mesuré** : ici, il est possible d'hésiter 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 interne 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).

(*Fin de la note éditoriale*)

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.

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 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.

Note éditoriale

Le lecteur aura compris que ces catégories ne sont pas consensuellement définies et que ce ne sont pas les seules à être proposées. Quand on aura ajouté à cela qu'elles ne sont généralement pas définies de façon aristotélicienne, ni même simplement arbitrables logiquement, on aura compris qu'il faudra être très circonspect au moment d'adopter les règles de pratique qui les utilisent. (*Fin de la note éditoriale*)

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 consensus.

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*)

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.

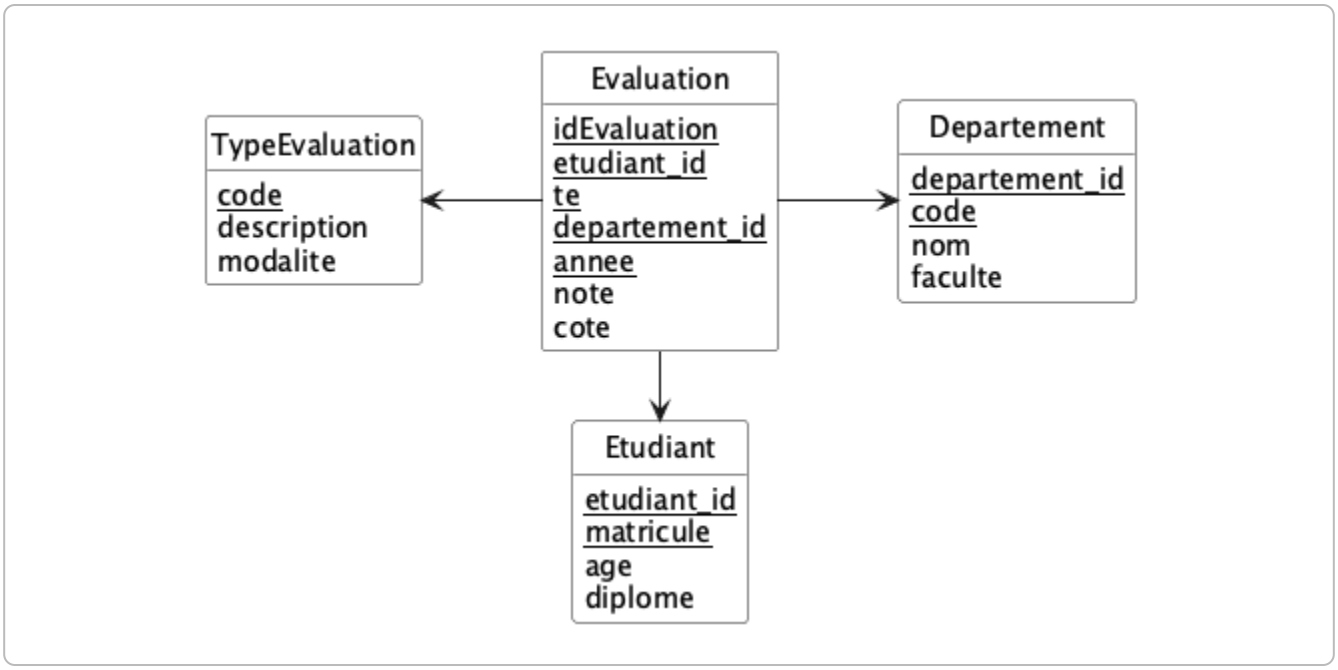
3.1. Schéma en étoile

- 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 [Adamson2010a]

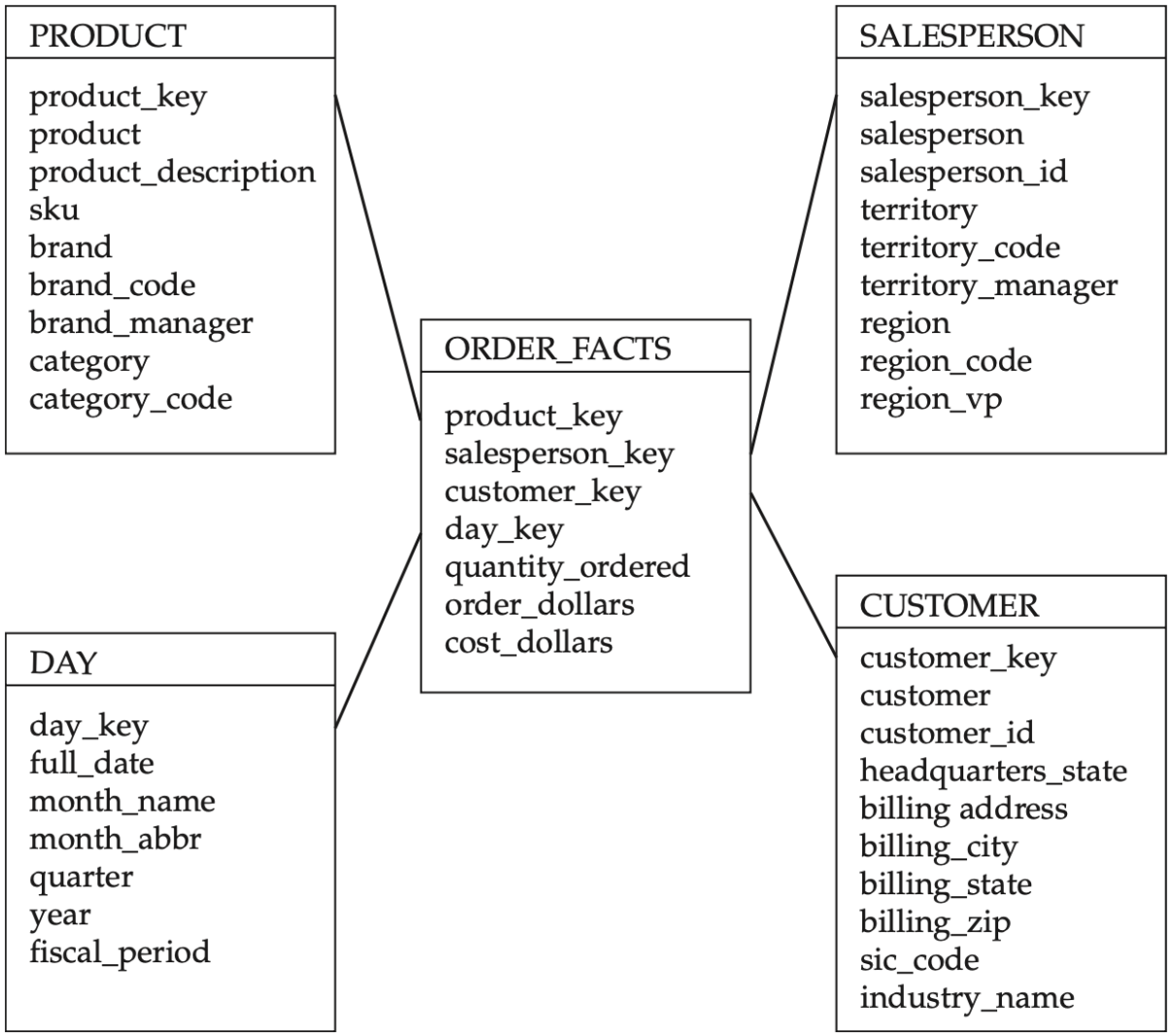


Figure 1-5 A simple star schema for the orders process

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

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

Une dimension peut être normalisée de deux façons :

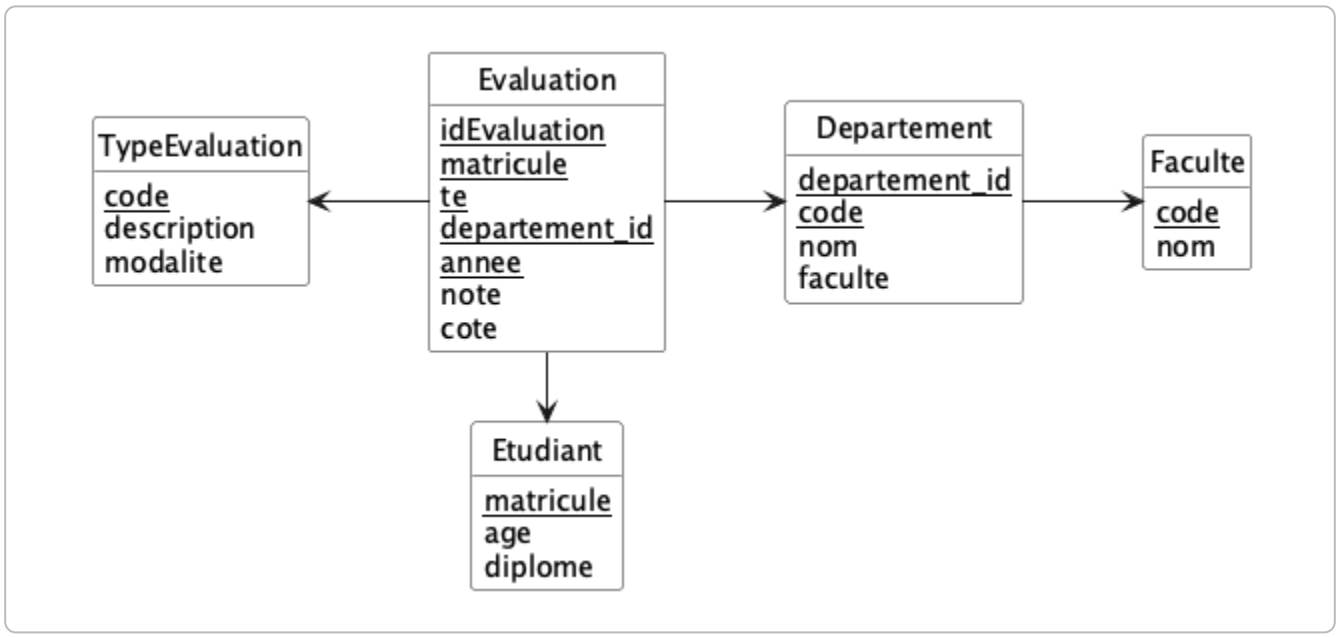
- normalisation en première forme normale, ce qui consiste à créer une relation pour chaque attribut multivalué (un attribut qui peut avoir plusieurs valeurs) ou annulable.
- normalisation hiérarchique, ce qui consiste à créer une relation par niveau de détail du plus granulaire au moins granulaire. 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 ou des contradictions. 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 ? Pourquoi en est-il autrement dans les bases de données normalisées utilisées en exploitation ?

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



Exemple 4. Schéma en flocon d'un processus de commande [Adamson2010a]

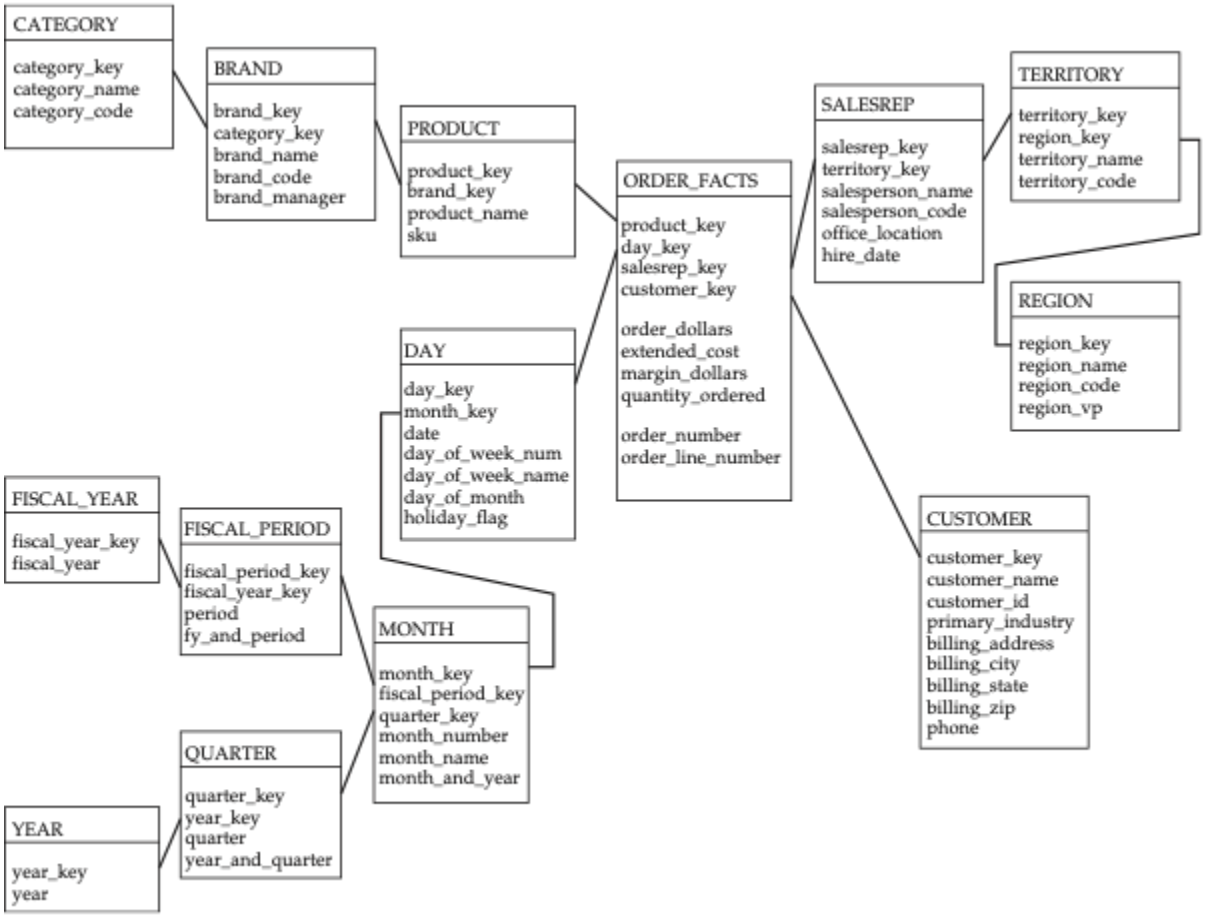


Figure 7-5 A snowflake schema

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

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.

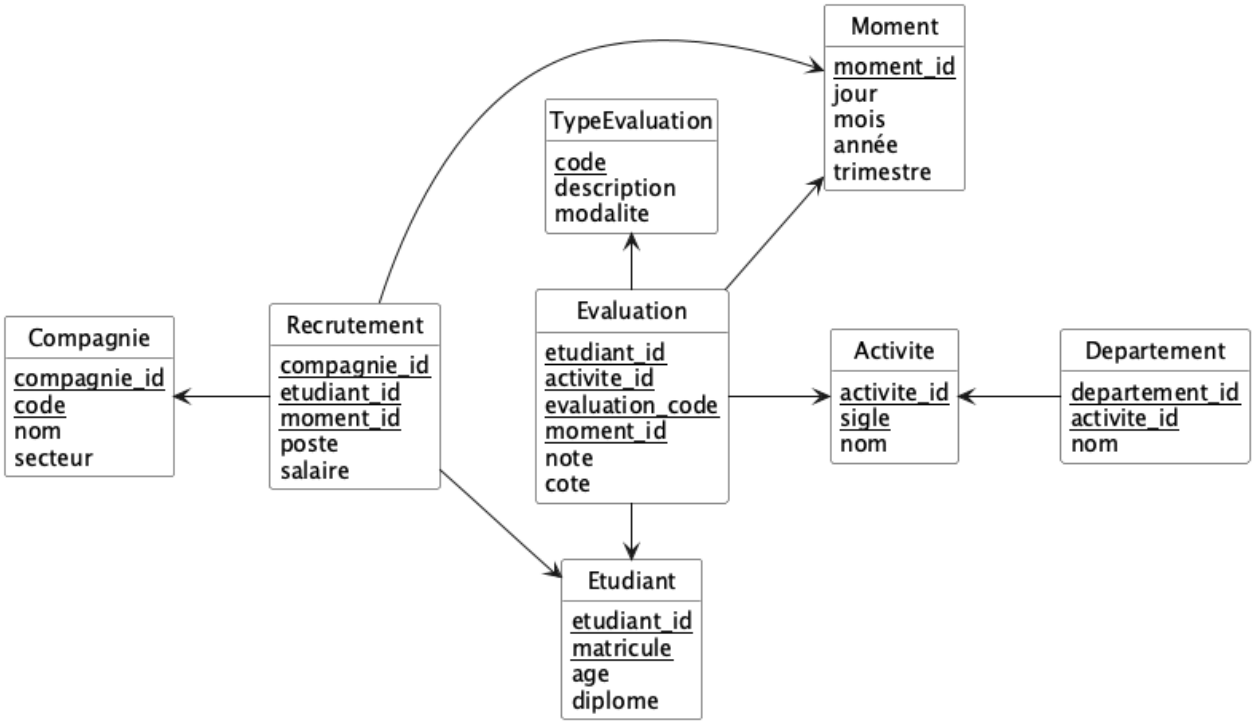
Généralement, dans un domaine (ou une organisation), il y a plusieurs processus qui ont lieu à différents moments. Dans des situations où vous évaluez individuellement les processus, une relation factuelle par processus est nécessaire. Pour choisir la bonne modélisation, deux questions se posent. Soit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



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

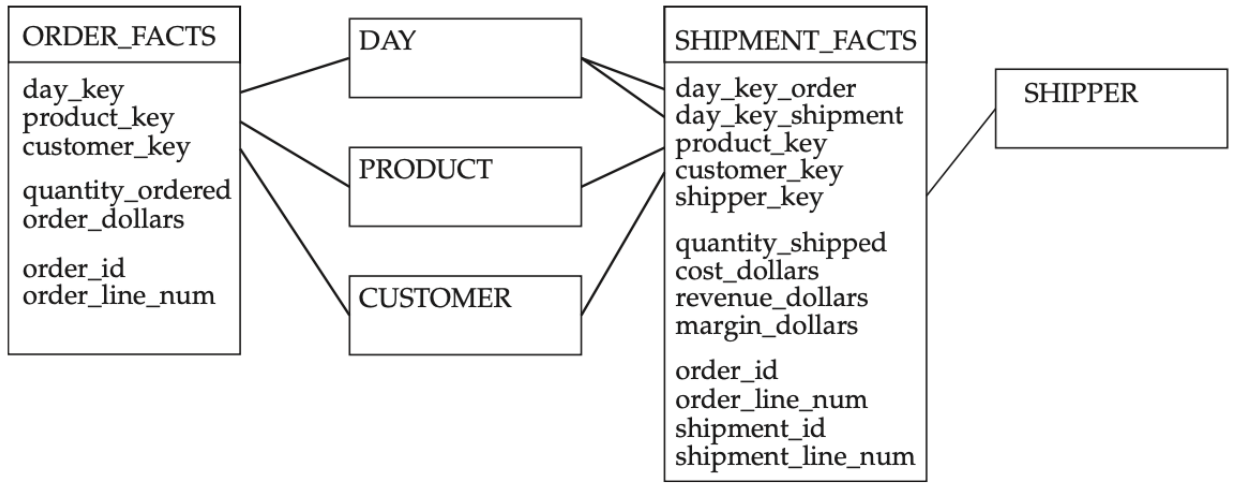


Figure 4-8 Separating the two processes into separate fact tables

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 numériques. 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. La notion de clé s'applique tout aussi bien aux relations, aux classes et aux ensembles d'entités. Une clé sera qualifiée de *déterminante* si ses attributs appartiennent à la même relation que les attributs déterminés, et *référentielle* s'ils appartiennent à une même autre relation (dite *référée*). 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** un 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, mais incapable de conjuguer deux règles de façon priorisée !

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.

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 !

procédé

1. Méthode employée pour produire un effet déterminé ou parvenir à un certain résultat.
2. Plus spécifiquement, en génie logiciel, méthode d'organisation des processus pour produire un ensemble d'artéfacts livrables.

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 2025-11-13 10:46:40 UTC

Université de Sherbrooke