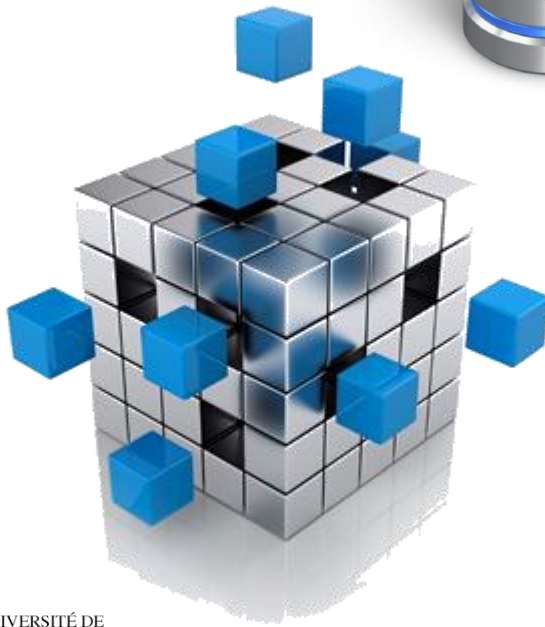


INF 735

Entrepôt et Forage de Données



Bloc 2
Modélisation
par Robert J. Laurin

Plan du cours – Les blocs

(Bloc1)

Introduction: Le besoin, concepts et définitions

(Bloc 4)

ETC:
Acquisition de
données

(Bloc 2)

Modélisation
(Entrepôt)

(Bloc 3)

Outils de
présentation,
OLAP et Forage

(Bloc 5) Architecture et Méta données

(Bloc 6) Définition des besoins et gestion de projet

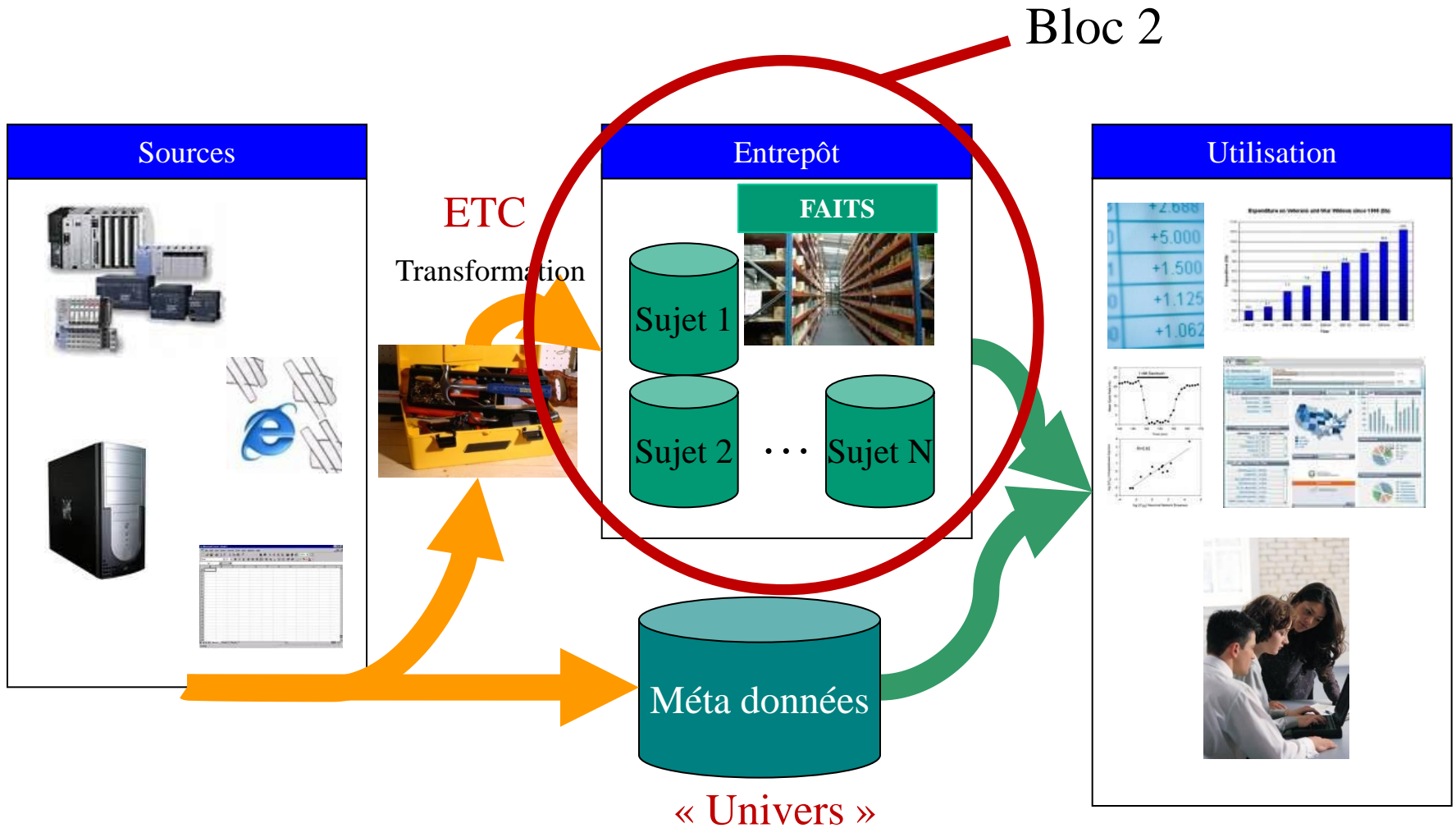
(Bloc 7) Techniques de réalisation et opération



- Suggéré:
Data Warehousing Fundamentals, A Comprehensive Guide for IT Professionals,
Paulraj Ponniah
 - Chapitres 10 et 11
- Annexes:
 - Dimensional model design checklist
 - Logical table design
 - Physical database design
 - Dimensional model document
 - Derived fact worksheet
 - INF735 – Bloc 2 – Modélisation étoile - Exercices

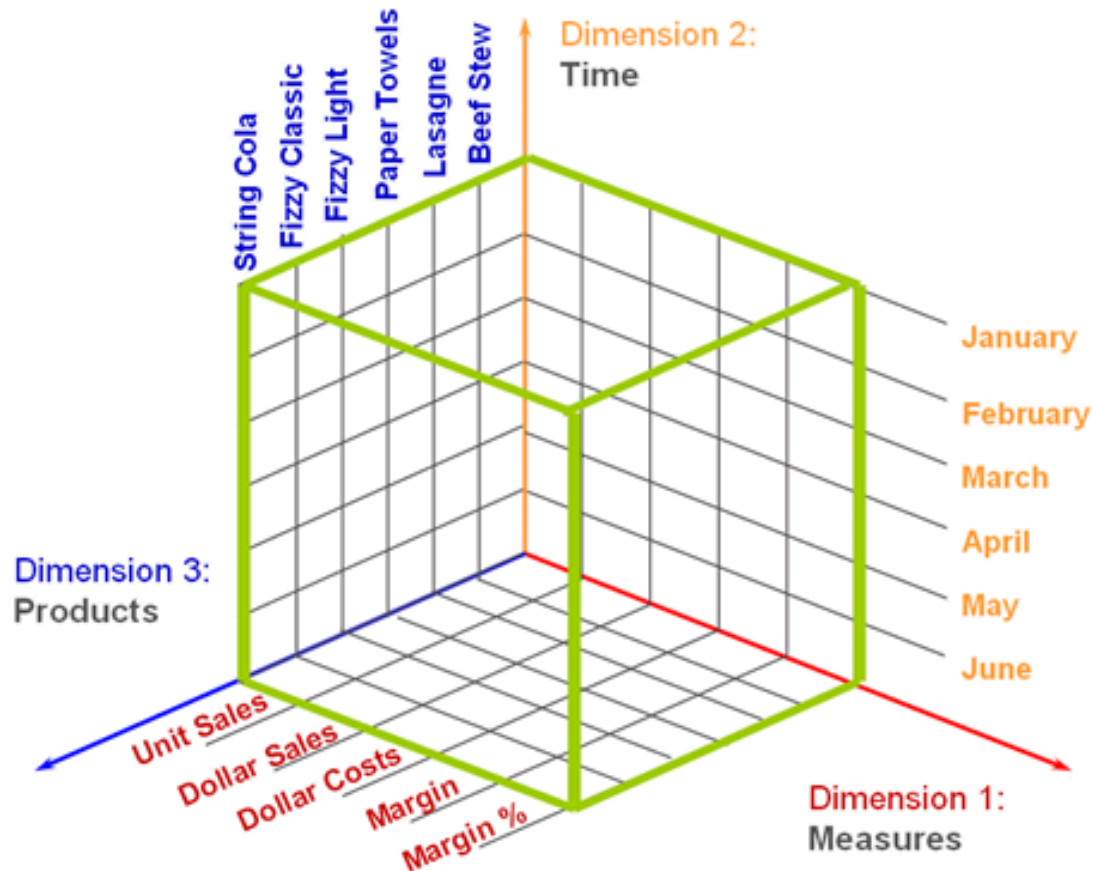
- 3 dimensions (et +) = 1 cube
- Pas de 3^{ième} forme normale
- Le modèle vient du besoin informationnel
- L'entrepôt grandira plus vite en espace disque que les systèmes opérationnels puisque les doublons y sont non seulement permis mais encouragés.

Rappel de la structure générale



MODÈLE MULTIDIMENSIONNEL

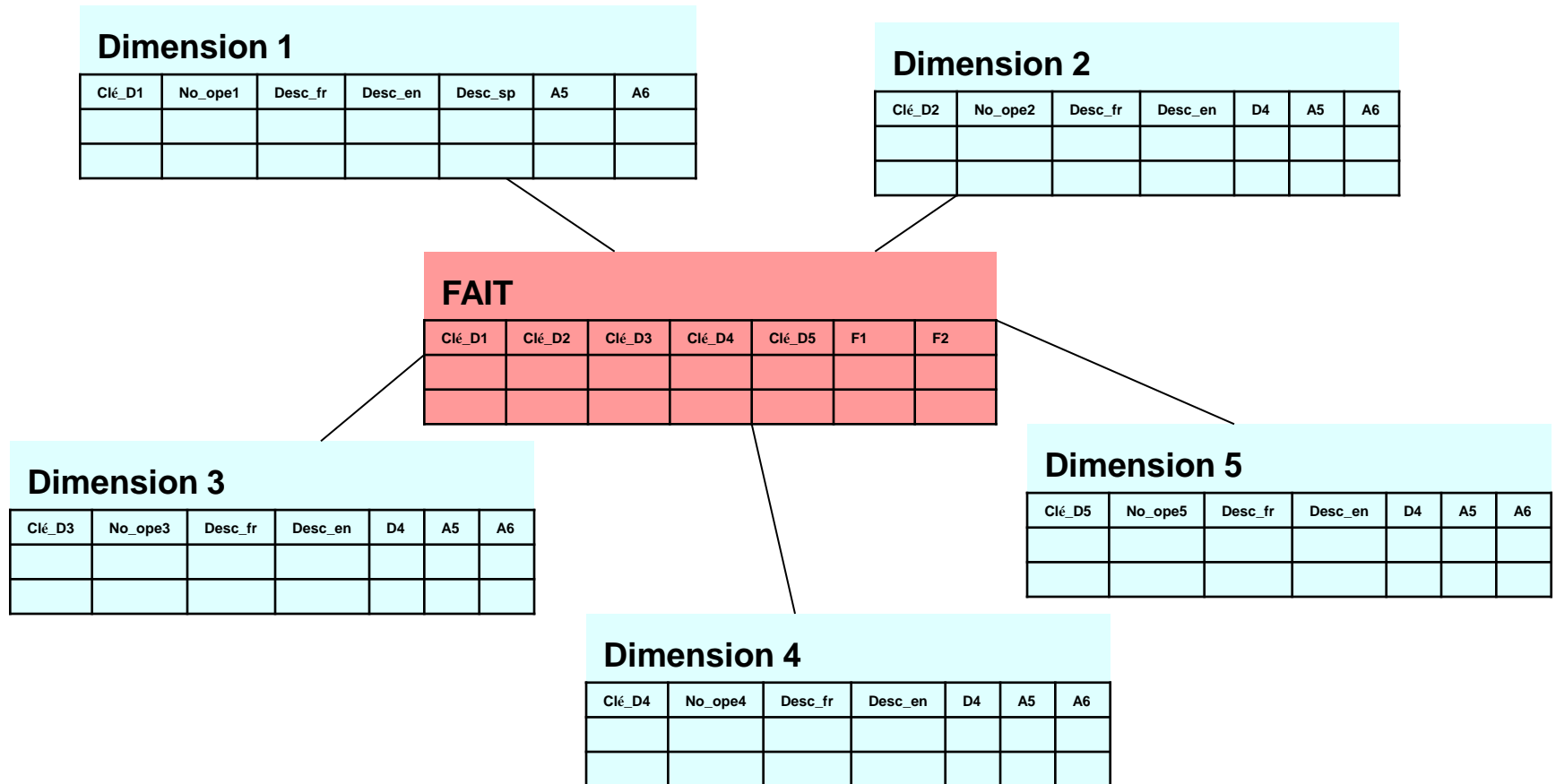
Cube



Source: Microsoft

1- ÉTOILE

1. Étoile



Modélisation dimensionnelle .vs. Opérationnel (INF732)

- Opérationnel (Entité Relation/Association):
 - Tables et jointures en chaines
 - Normalisées (3 FN)
- Étoile:
 - Faits et dimensions
 - Pas normalisé
 - Jointures simples

Technique de conception logique permettant de structurer les données de manière à les rendre intuitives aux utilisateur d'affaires et offrir une bonne performance aux requêtes.

- Divise les données en faits et dimensions;
- Les faits (mesures) sont généralement des valeurs numériques provenant des processus d'affaires;
- Les dimensions fournissent le contexte (qui, quoi, quand, où, pourquoi et comment) des faits;
- Schéma en étoile: une table de faits entourée de plusieurs tables de dimension.

- Ce que c'est:
 - C'est **la** technique viable pour l'entrepôt de données
 - Série de dimensions (éléments d'analyse) sur la données (fait selon une granularité)
 - Chaque dimension représente un choix pour l'utilisateur.
(Le modèle Entité-Relation ne permet pas à l'utilisateur de naviguer)
- À quel niveau de détail (faits)?
 - Selon le besoin exprimé
 - Idéalement le plus de détail possible
 - Selon les dimensions créées.

Modèle en étoile - Exemple

Créer pour BD commande herbe à puce

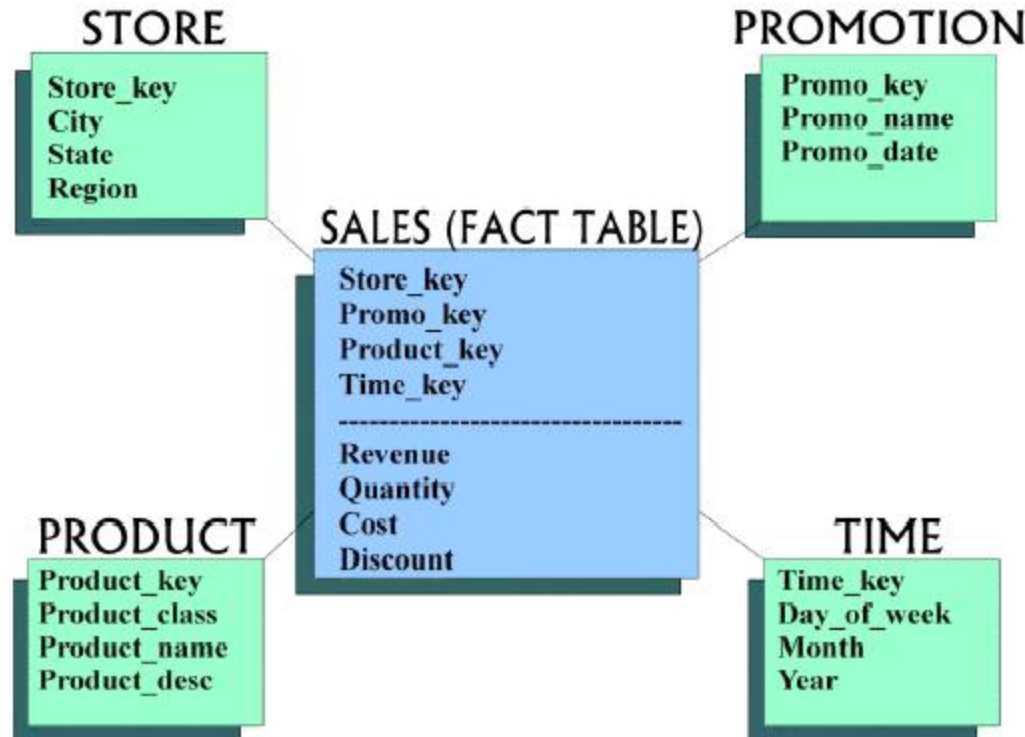
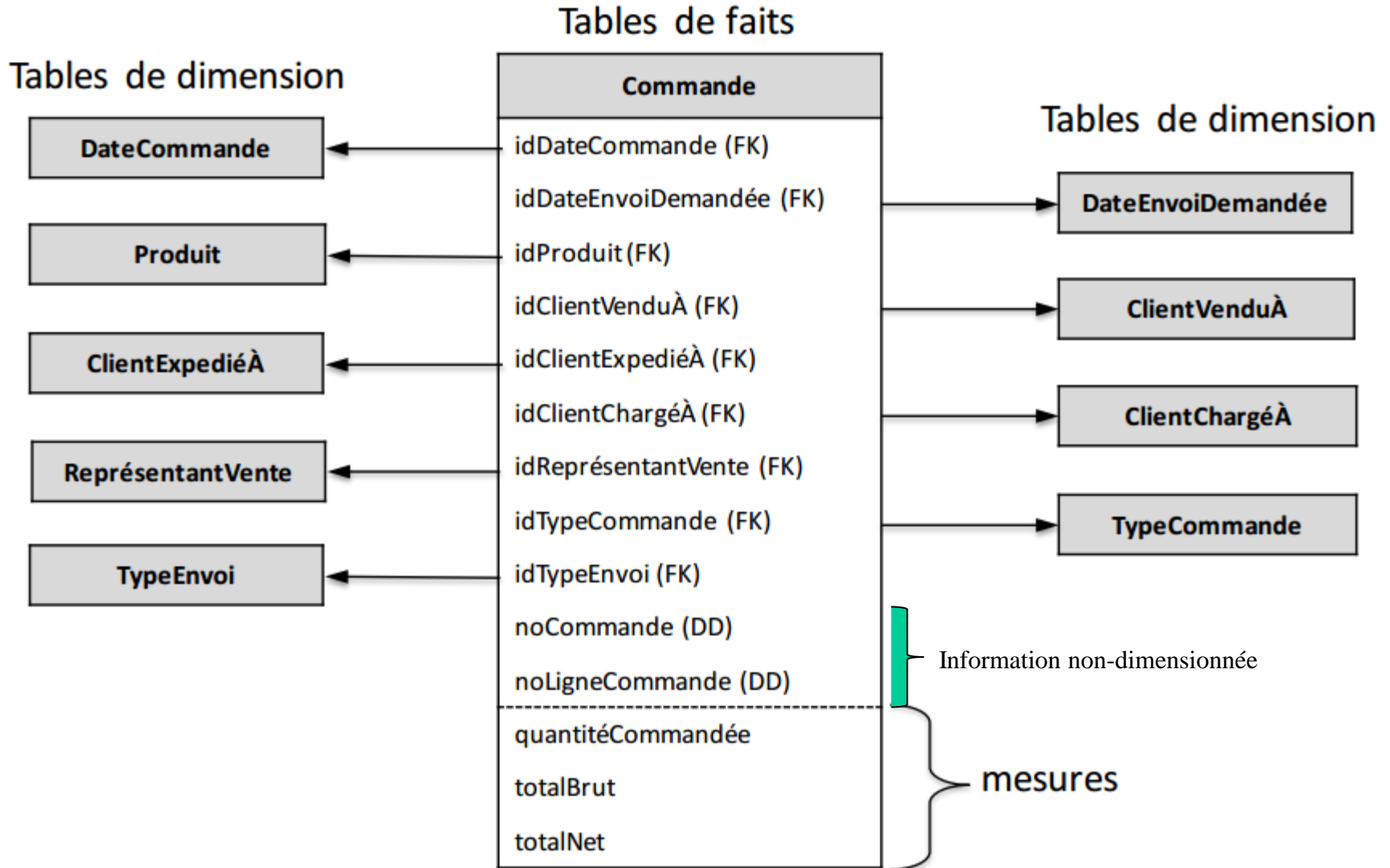


Figure 1. Star Schema Example

by Cheryl Grandy



Exemple (modèle logique)

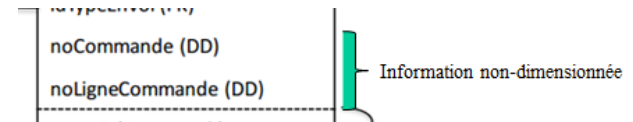


- Avantages
 - Relationnel avec des liens 1 à plusieurs
 - pas normalisée
 - Naturel pour l'utilisateur
 - Optimise la navigation
 - Aide le traitement des requêtes informationnelles
 - Tremplin pour d'autres schémas propriétaires.

« Un entrepôt utile est basé sur la granularité (donnée de base) ! »

Provient de la transaction ou évènement, pas du contexte

- Mesures / Quantifiable / calculé
 - Additives (Qté vendue, \$ vendu)
 - Semi-additives (ex: solde client selon le mois)
 - Non-additives (pourcentage d'escompte)
- Non Quantifiable
 - Ex: présence et absence (*mettre 1*)
- Information à occurrence unique par transaction/évènement
- Booléen indicateur



- **Tables de dimensions**

- ➔ Qualitatifs, circonstances et éléments déterminants du fait

- Artificielle (pas la clé naturelle, i.e. celle de la source opérationnelle)

- + Permet les modifications de type 2 et 3
 - + Ne changera pas dans le temps
 - + Permet d'identifier la même entité même si a 2 clés naturelles différentes

- **Table(s) de faits**

- ➔ Éléments quantifiables, mesurables, calculés

- Ensemble des clés étrangères

- + identifiant (ex: no de facture)

- ou séquentiel au besoin*

Pas de clé nulle

Pas de clé étrangère nulle

Modèle en étoile – Exemple de requête

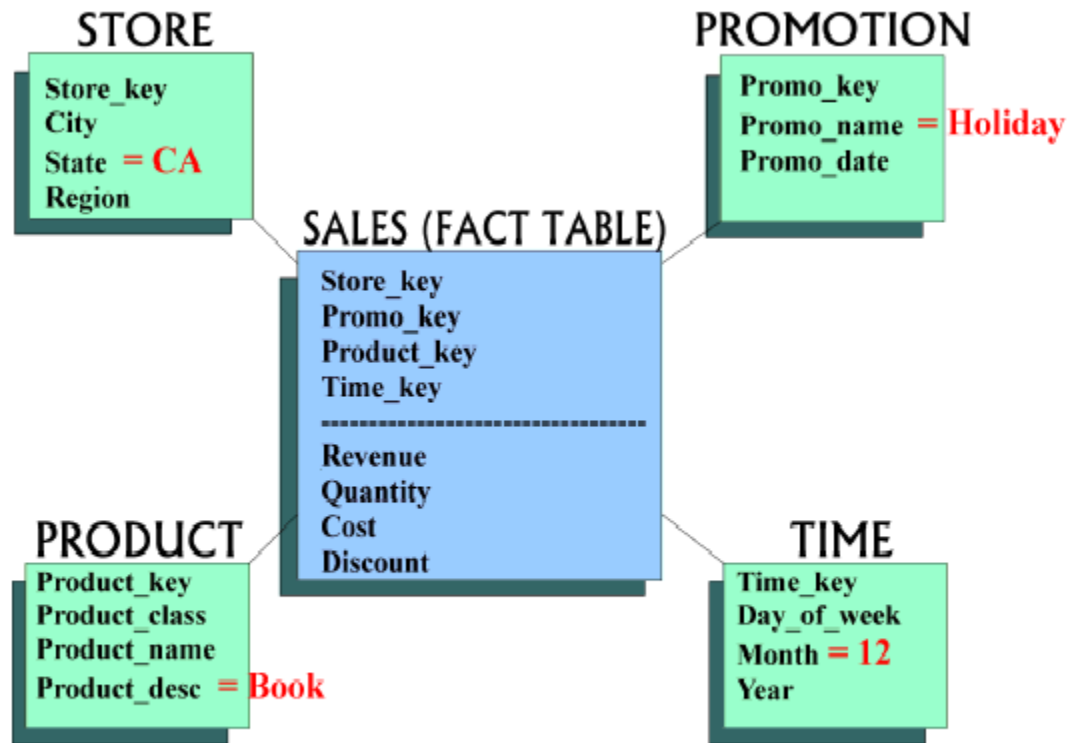


Figure 4. Multidimensional Query Example

by Cheryl Grandy

DISC
Dynamic Information Systems Corporation

Annexe

2- FLOCON

2. Flocon

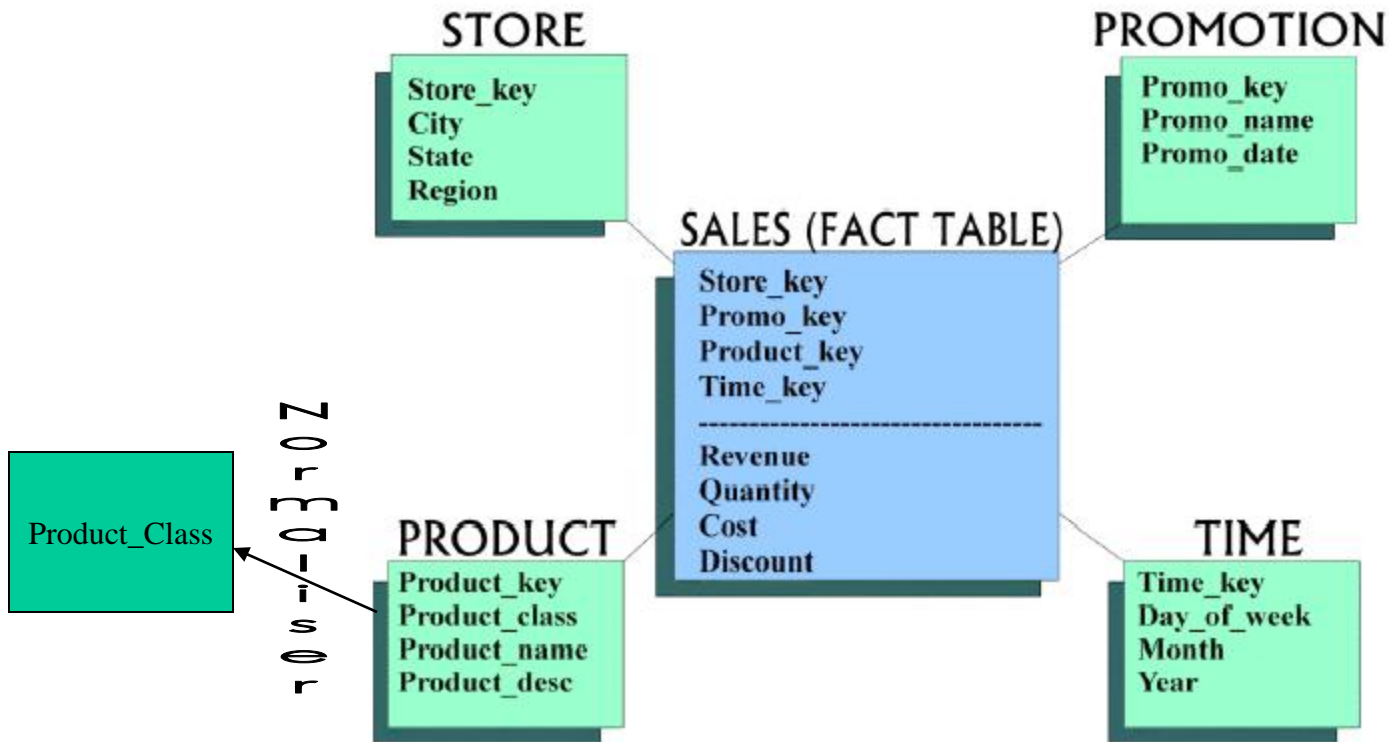




Figure 1. Star Schema Example

Hiérarchie des dimensions

- Un ensemble d'attributs ayant une relation hiérarchique (ex: catégorie et produit)
- Définissent les chemins d'accès dans les données (drill-down paths);
- Solutions:
 - Dimensions séparées liées directement au fait (**)
 - Dans 1 seule table de dimension
 - Flocon

** Hiérarchie peut être établie dans la métadonnée du cube ou « univers »

Quand ?

- Peu de données par ligne de dimension ou liste de possibilités d'un attribut trop longue pour un query
-  Facilité de répliquer la normalisation du system opérationnel
- Le flocon reflète la façon de penser des utilisateurs
- L'outil trop simple est plus performant de cette façon (exemple: COGNOS Powerplay)
-  Ajout à une dimension mal planifié.

Normalisation - ATTENTION

Bien que les flocons existent, il est préférable de ne pas normaliser.

Il est d'avis général que:

- OK de normaliser pour l'approvisionnement en données (« Staging »)
- Pas normaliser pour le « query » utilisateur
 - Difficile à comprendre par un utilisateur
 - Pas performant pour l'informationnel
- Pour présentation à un utilisateur, la données doit être dimensionnelle!

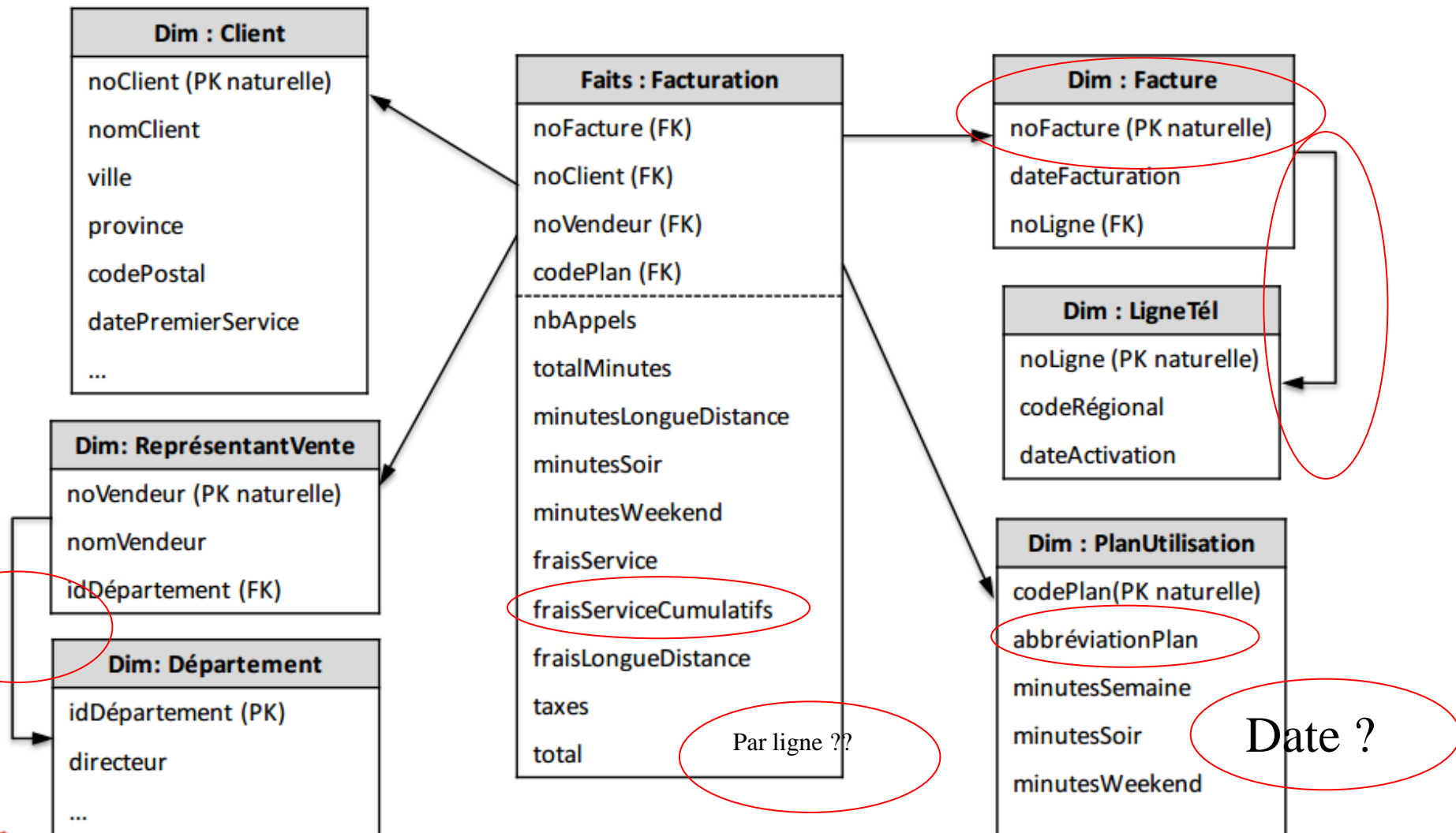
Santé et Bien-être Canada considère que le danger de la normalisation dans la présentation informationnelle croît avec l'usage. Éviter de normaliser.

Désavantages du flocon:

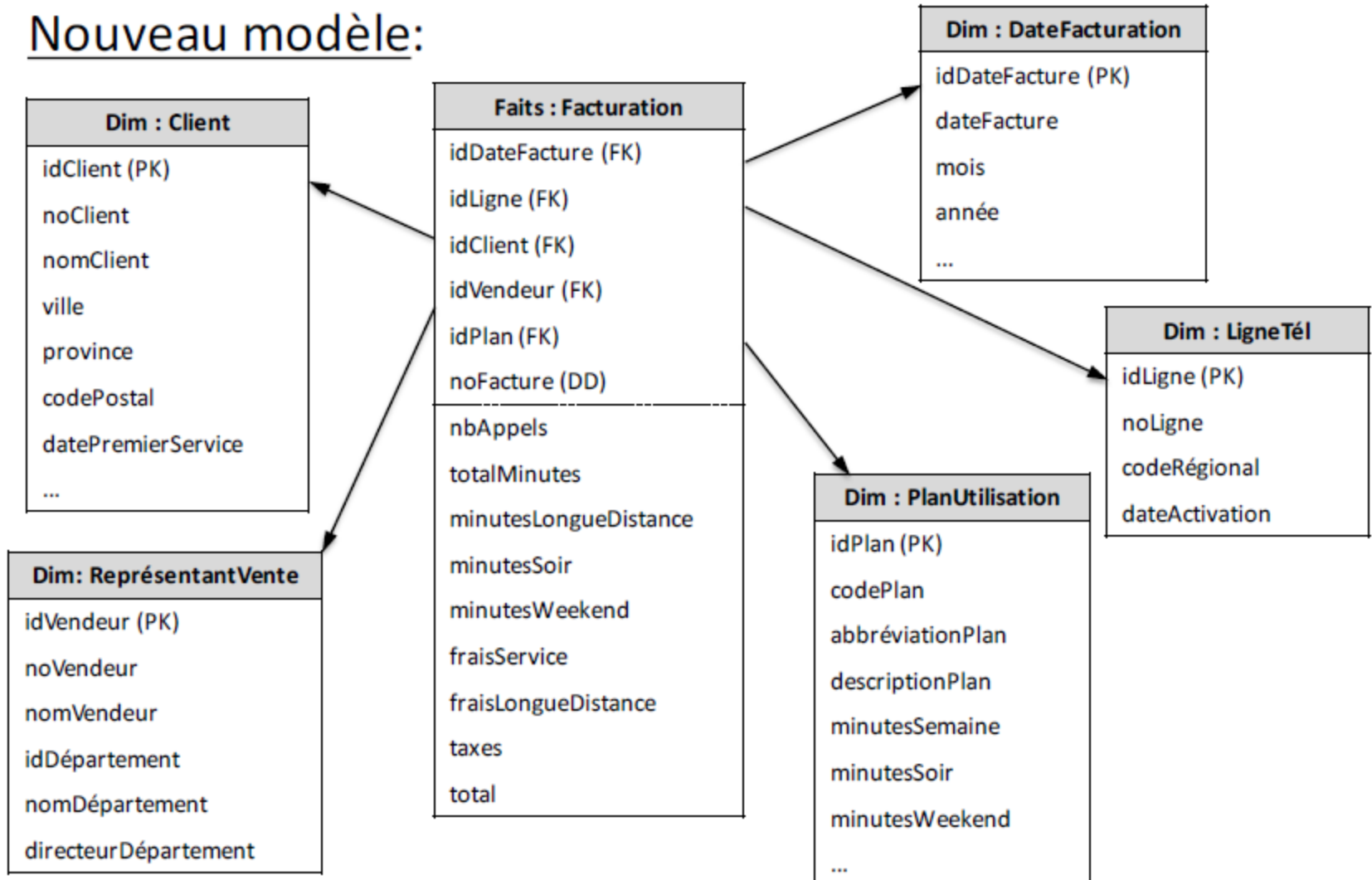
- Plus il y a de niveaux, plus ce sera difficile à comprendre pour un utilisateur
- Complexité au chargement
- Espace disque gagné minime et ne doit pas être le facteur déterminant
- Ralentit la capacité de navigation et la capacité à se déplacer d'un niveau de sommarisation à l'autre
- Empêche l'utilisation des index bitmap.

EXEMPLES

Un exemple avec erreurs

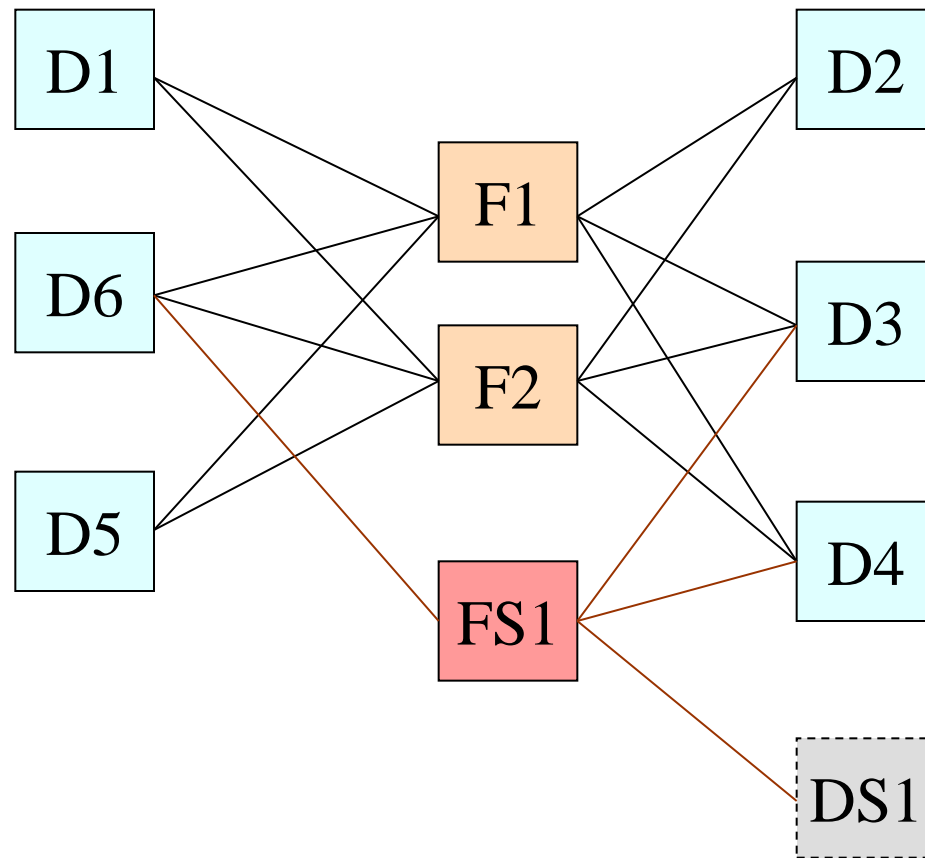


- Nouveau modèle:



AGRÉGATIONS

3. Étoile avec Agrégation



Sujet A

- Chercher le 10 pour 1...
- Attention à l'effet ETC
- Outil doit être adapté

« certains faits ne peuvent pas être additionnés ! »

Agrégation – Meilleur design

- Les agrégats doivent résider dans une table de faits séparée des données atomiques
- Chaque niveau d'agrégation doit avoir sa table de fait
- Créer une famille de schémas de façon à trouver les agrégats et le détail
- Par défaut, toute interrogation, SQL ou outil doit pointer sur la table de faits de détails et ses dimensions.
- Attention, tous les faits ne s'additionnent pas...

IMAGES OU « SNAPSHOTS »

Une image (« snapshot »)

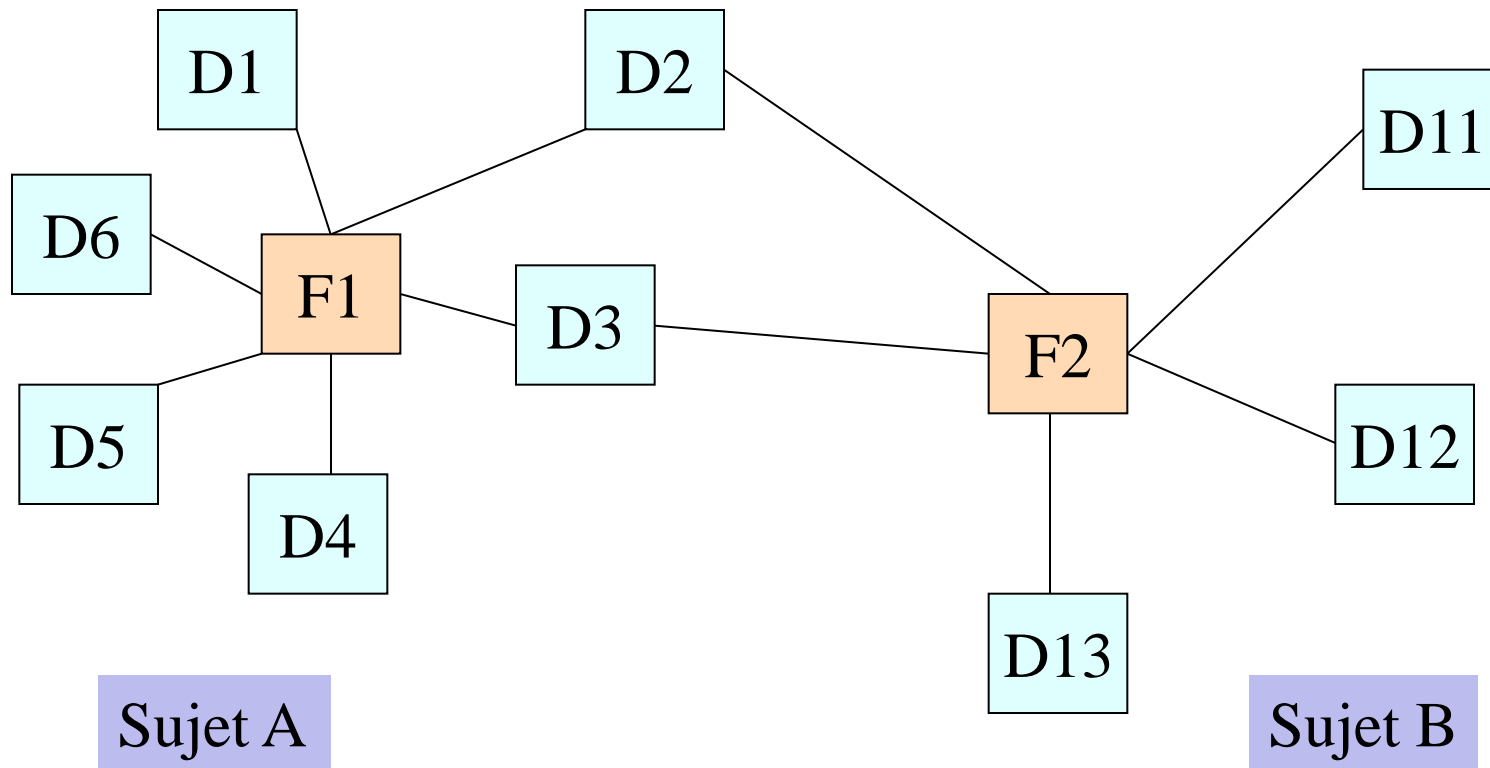
- Pour entreposer un moment précis
- Pour pouvoir comparer des intervalles

Ex: L'inventaire de matière première à midi tous les jours
 Positions des bateaux de la flotte

Exercices

CONSTELLATION

4. Famille d'étoiles ou Constellation



Modèle en Étoile – pré-requis

- Les dimensions représentent la même chose pour tous
- Les dimensions sont standardisées, décrites et publiées – font parti de la « Méta Donnée »
- Les comptoirs (« Data marts ») utilisent une conformité stricte aux dimensions standardisées
- Si on ne peut adhérer à une dimension standardisé, Donner un autre nom et documenter clairement son utilisation

« Un –Data Mart- est une façon élégante d’avaler l’entrepôt une bouché à la fois ! »

- Réutilisation d'une (ou plusieurs) dimension
- Lier plusieurs Faits
- Attention à la conformité
 - Définition commune
 - Représentation commune
 - Homonymes et synonymes
 - Même mesure
 - Même calculs

MODÉLISATION - CAS

Trop de dimensions - ATTENTION

Attention aux mille-pattes:

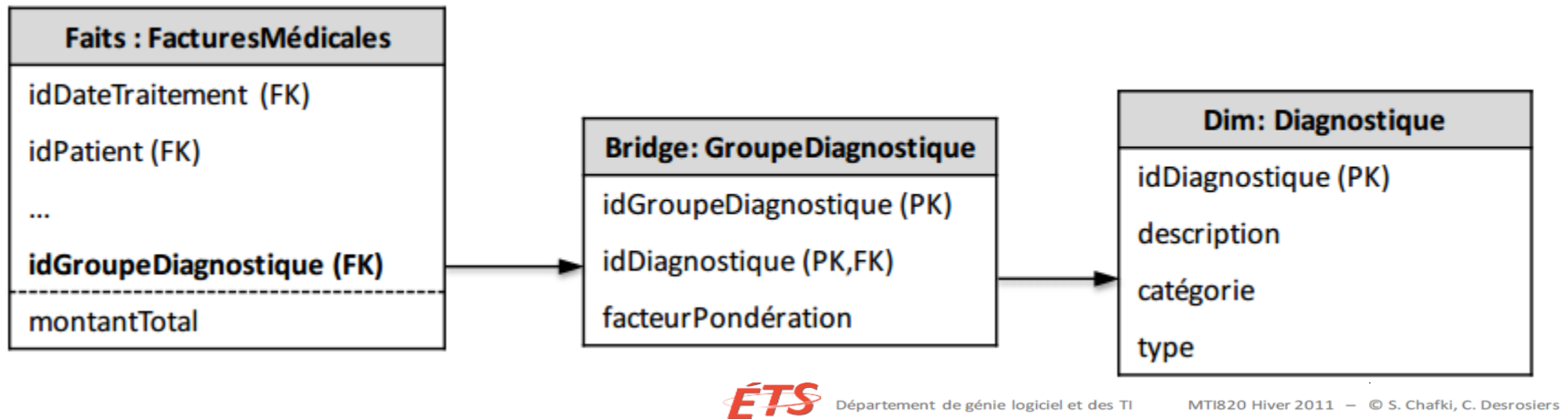
- Un nombre trop grand de dimensions indique généralement que des dimensions pourraient être combinées.
- Toutefois, les dimensions trop profondes ou trop larges peuvent poser problème.

Objectif: Meilleure navigation possible!

- Large
- Profonde

Table de pont (bridge table)

- Relie 1 à N de fait à dimensions



- Cas spéciaux seulement!
 - On peut généralement ajuster la table de fait...
 - Parfois signe d'un modèle déficient

Dimension DATE – plus qu’une simple date

- Date
- Jour de la semaine
- Julienne
- No. Semaine
- No. Mois
- Époque
- Jours depuis le début d’année fiscale
- Heure de début du jour
- Premier (dernier) jour de la semaine/mois/an/période
- Période comptable
- Quart
- Dates dans différents format
- Fête selon religion / pays
- Non de la fête
- Férié selon pays/religion
- LA date relative (0: aujourd’hui, -1: hier... normalement calculée mais parfois stockée)

- Les dates en valeurs absolues pour ce qu'elles représentent dans sa dimension
- Les dates en valeurs relatives calculées pour le fait.

- La JUNK – Dimensionner ou dans le fait
- Les tables de faits pour les éléments sans faits
- Enregistrer le fait (sans fait) qui n'a pas eu lieu

- Plusieurs tables faits pour 1 dimension (cardinalités différentes)
- Plusieurs attribut du fait du même domaine (ex: Dates)
 - Même dimension ou plusieurs dimensions ?

- Temps ou heure .vs. Date
 - Date dimensionnée
 - Intervalles dimensionnables
 - Heure précise au fait (trop d'entrées à la dimension)

- Hiérarchie (Flocon ou au Cube)
- Devises multiples

- Fuseau horaire
- Détail (transactions) .vs. Snapshot (portrait)

TYPES DE CHANGEMENTS

Changements dans les dimensions (lents)

- Type 1 : Une simple erreur
- Type 2: Impact sur l'historique
 - + Historique = date
- Type 3: Gestion de scénarios (modifications tentatives ou sans effet)

- Type 1 : Une simple erreur

Écraser l'ancienne valeur

Type 2 - Exemple

- Type 2: Impact sur l'historique

+ Historique = date

Exemple: Client_X change de province du Québec à l'Ontario le 19 mai 1980...

Dim_Client

| Clé_client | No_Client | Nom | Province | Actif | Date_changement |
|------------|-----------|----------|----------|-------|-----------------|
| ... | ... | ... | ... | ... | ... |
| 23456 | Cl234 | Client_X | Québec | N | 1971-06-20 |
| 36680 | Cl234 | Client_X | Ontario | Y | 1980-05-19 |
| ... | ... | ... | ... | ... | ... |

Fait_vente

| Clé_client | Vente_\$ | (date) |
|------------|----------|------------|
| ... | | |
| 23456 | 100 | 12/03/1979 |
| 23456 | 355 | 22/05/1979 |
| 23456 | 233 | 05/01/1980 |
| 36680 | 545 | 07/12/1985 |
| 36680 | 666 | 12/04/2001 |

Type 3 - Exemple

- Type 3: Gestion de scénarios (modifications tentatives ou sans effet)

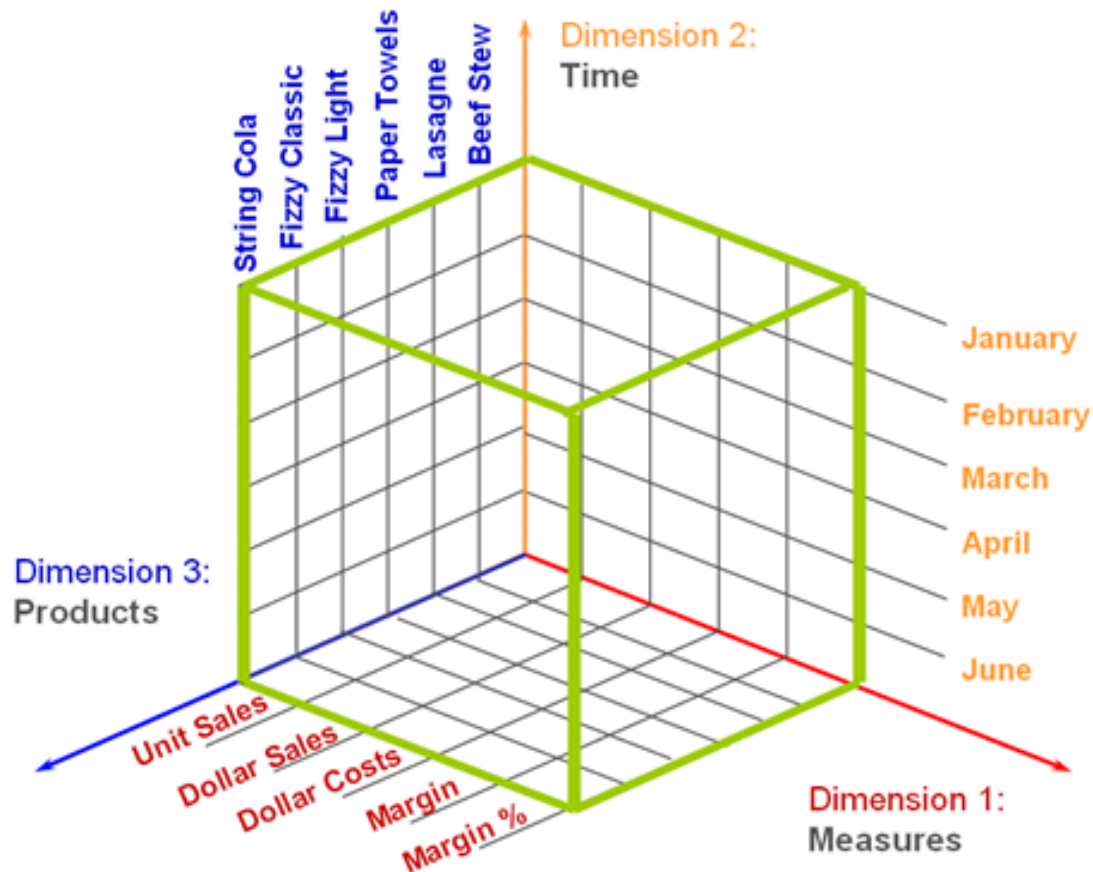
Changements dans les dimensions (Rapides)

Rapide = Plusieurs fois par semaine/mois

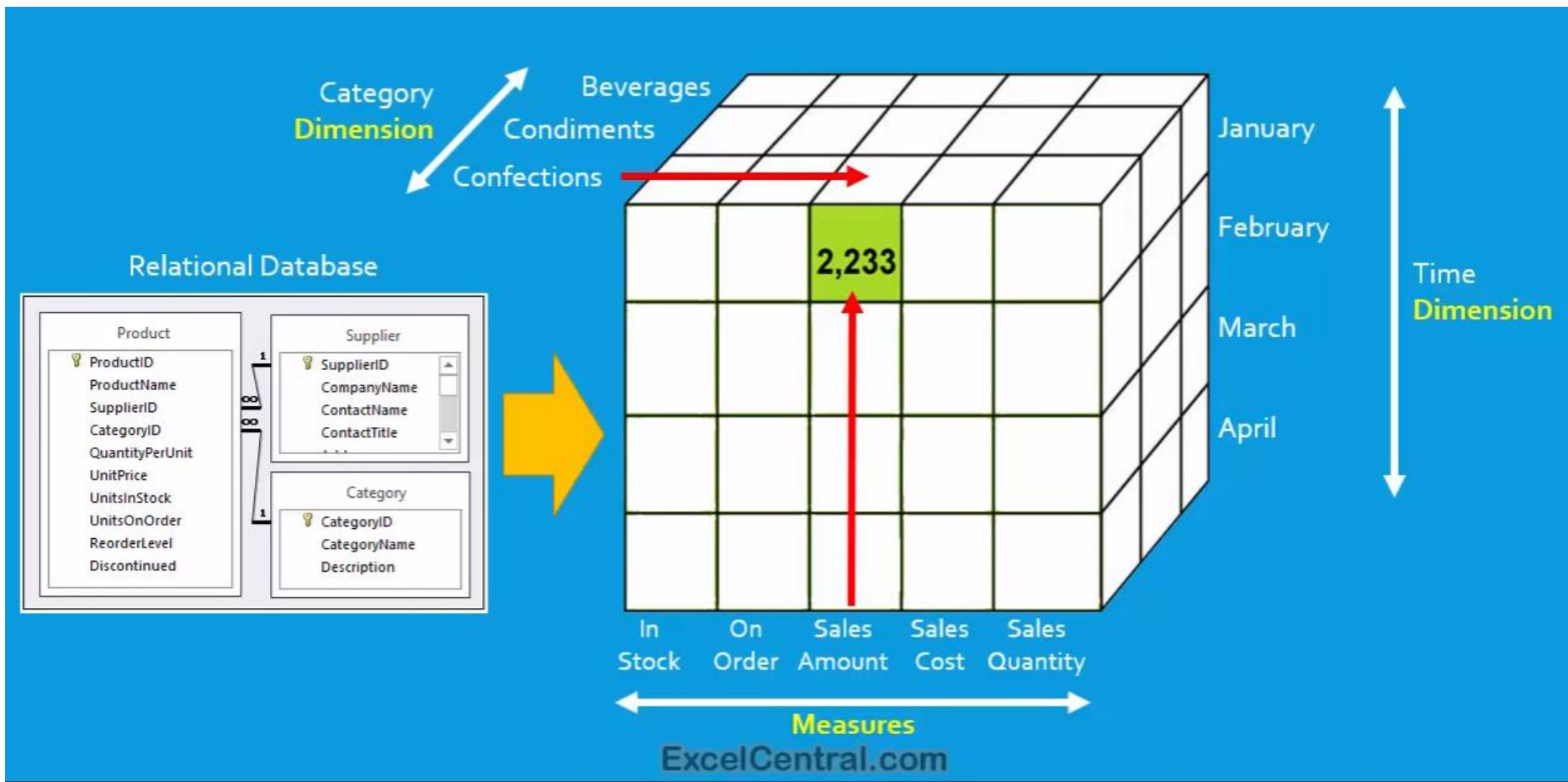
- Créer une nouvelle dimension avec les attributs qui changent
- Assigner une clé de plus au fait

3- CUBE

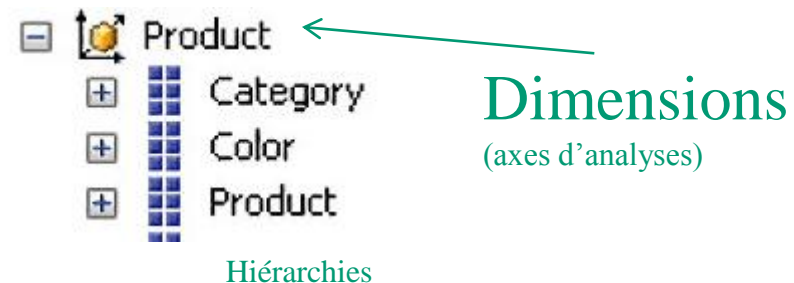
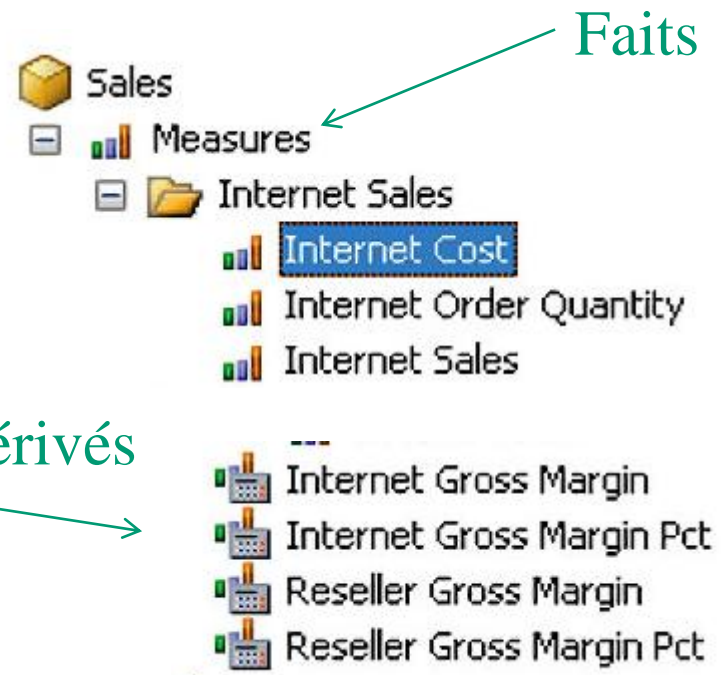
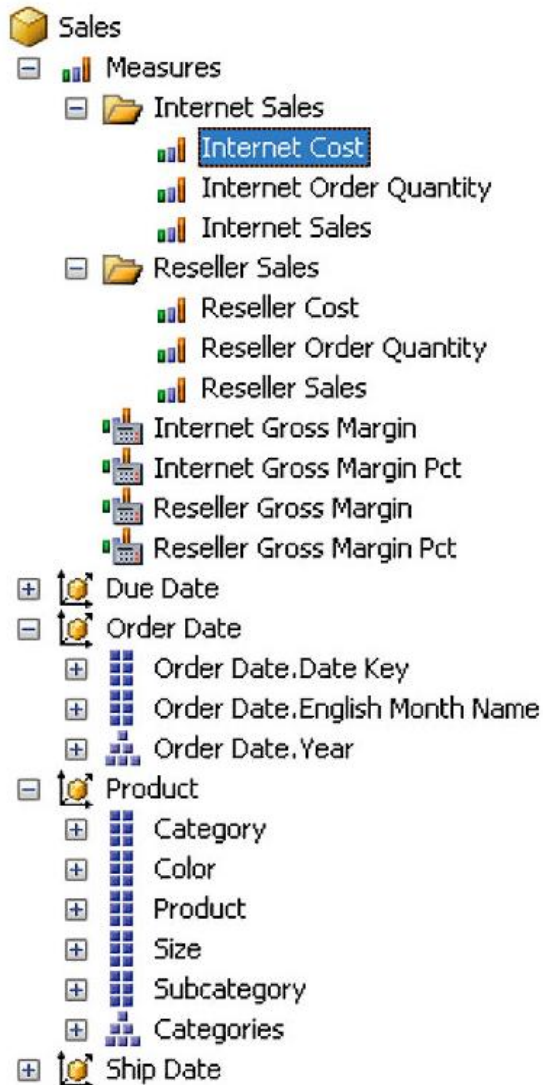
Cube



Source: Microsoft



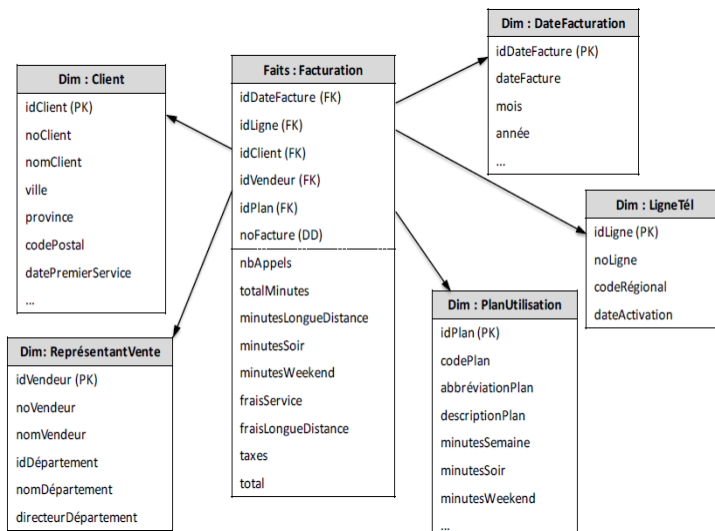
Le OLAP dans l'outil de présentation



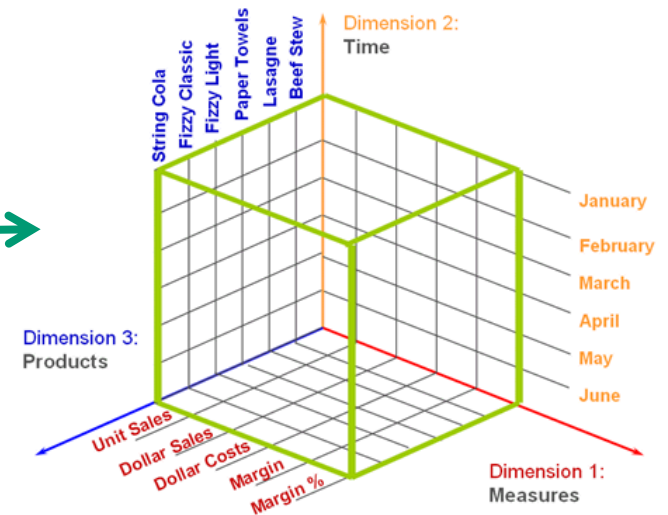
MEILLEURS PRATIQUES

Meilleurs pratiques multidimensionnelles

- Créer toujours vos modèles étoiles (relationnel)
- Créez des cubes (propriétaire) si besoin:
 - Performance
 - Intégration de la méta-information et hiérarchie
 - Présentation dans l'outil d'interrogation



Étoiles



Cubes

AUTRES CONSIDÉRATIONS

Quelques faits...

- 90% des requêtes utilisateurs sont de natures multidimensionnelles
- Les entrepôts doublent en volume chaque année
- La plupart des requêtes sont imprévisibles
- Les sommaires et les index consomment généralement plus d'espace disque que la donnée détaillée
- 99% des requêtes impliquent une agrégation (sommaire)
- La crédibilité repose sur la disponibilité de l'entrepôt et le rafraîchissement des données.

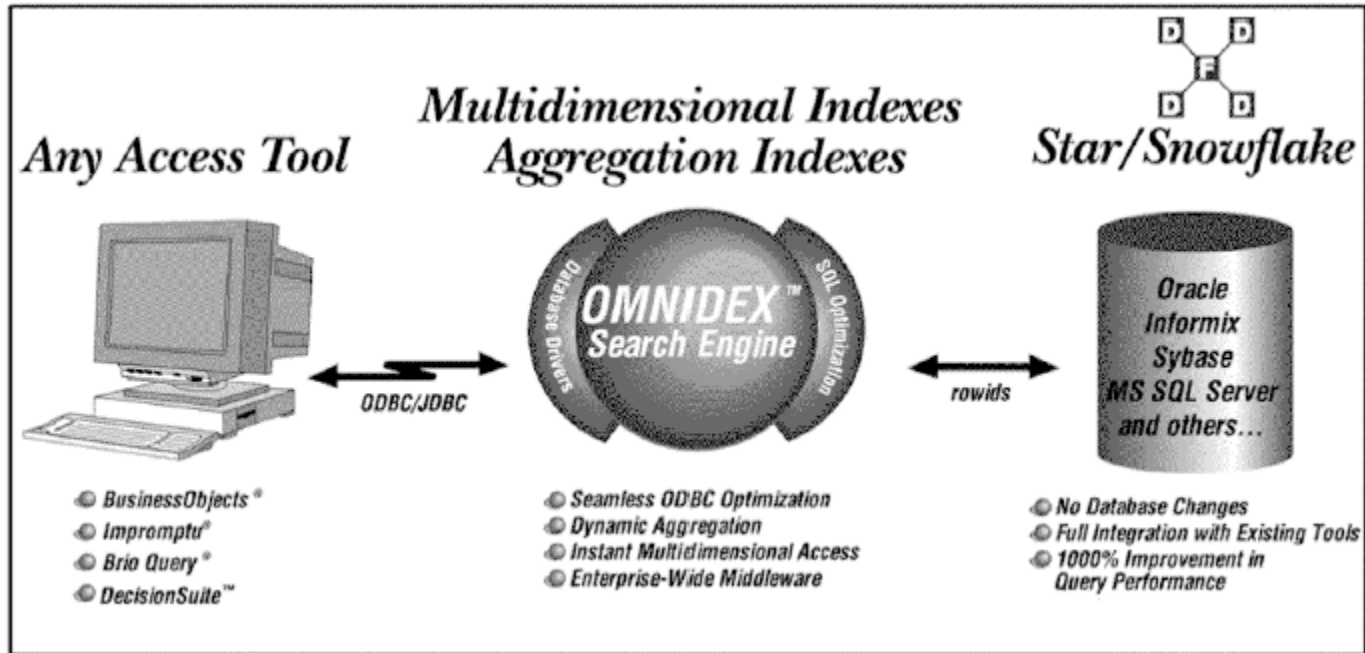
- Mythe #1: Bon que pour le DSS...
- Mythe #2: Personne ne comprend la modélisation dimensionnelle
- Mythe #3: Ne fonctionne qu'avec la vente au détail
- Mythe #4: Le Flocon remplace l'étoile
- Mythe #5: Difficile d'introduire une nouvelle dimension ou un nouveau type de donnée
- Mythe #6: Le Big Data remplace la modélisation étoile

- Lister les sujets/comptoirs potentiels
- Lister les dimensions
- Créer la matrice des deux (voir les recoupements)
- Obtenir la standardisation (politique)
- Modéliser:
 - Choisir le sujet/comptoir
 - Décider de la granularité (.vs. Dimensions)
 - Choisir les dimensions
 - Choisir les faits à présenter

- Les noms qui seront choisis à cette étape marqueront le projet et demeureront à jamais – bien choisir
- Un attribut ne devrait vivre que dans une dimension mais un fait peut être répété dans plusieurs tables de faits
- Si une dimension joue plusieurs rôles, nommer différemment et expliquer ce rôle

Modèle dimensionnel - Optimisation

- Des outils sont disponibles pour l'optimisation de l'indexation et agrégation





<http://www.rkimball.com/html/designtips.html>

Ralph

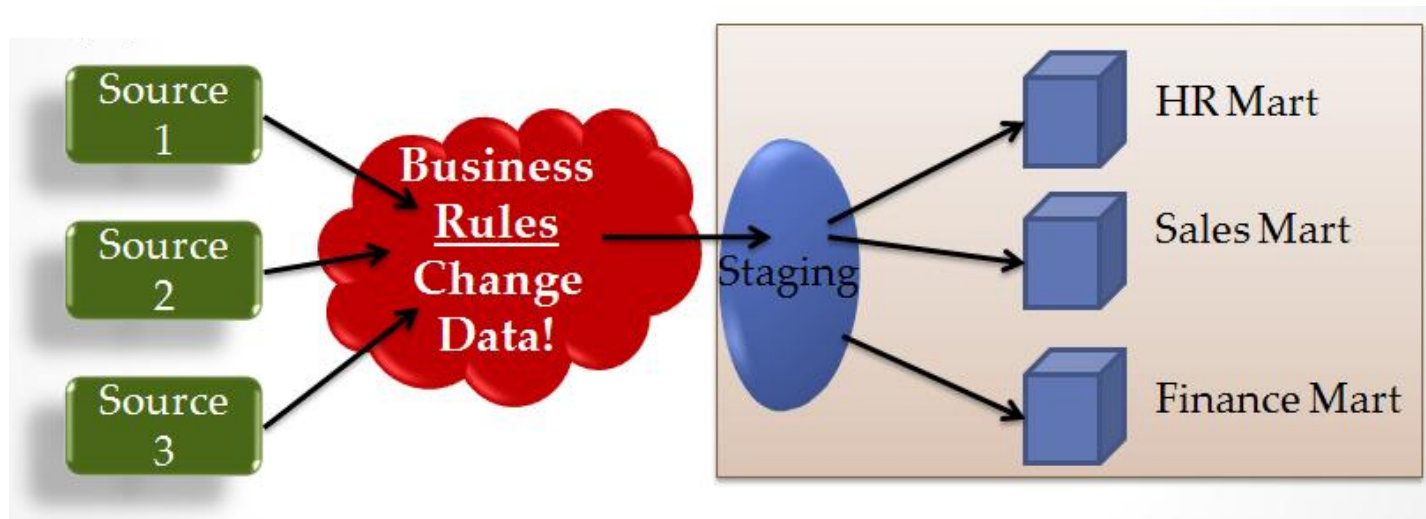
Gartner Magic Quadrant for Data Warehouse and Data Management solutions for Analytics



UNE PARENTHÈSE SUR LE DATA VAULT

Modélisation Entrepôts de données 1.0 - Comptoirs

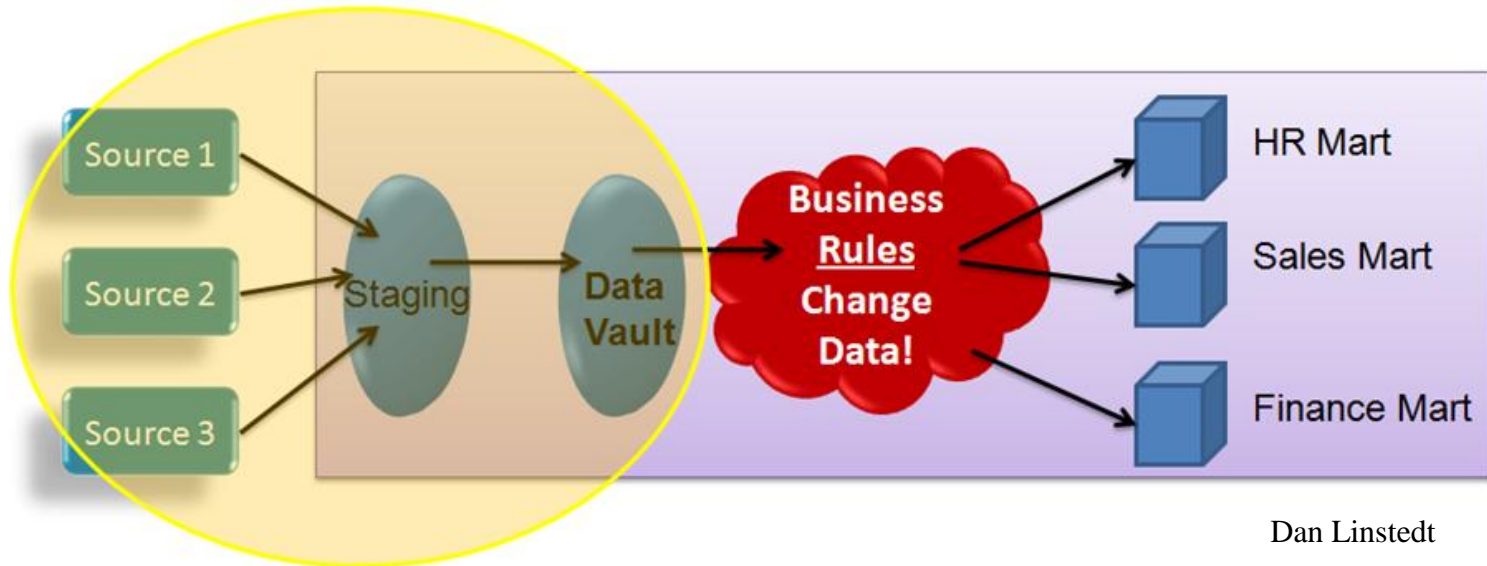
- Utilisation de Comptoirs (« Data marts ») privilégié
- Modélisation étoile idéale
- Alimentation SOURCE → ETC (Staging) → Comptoir



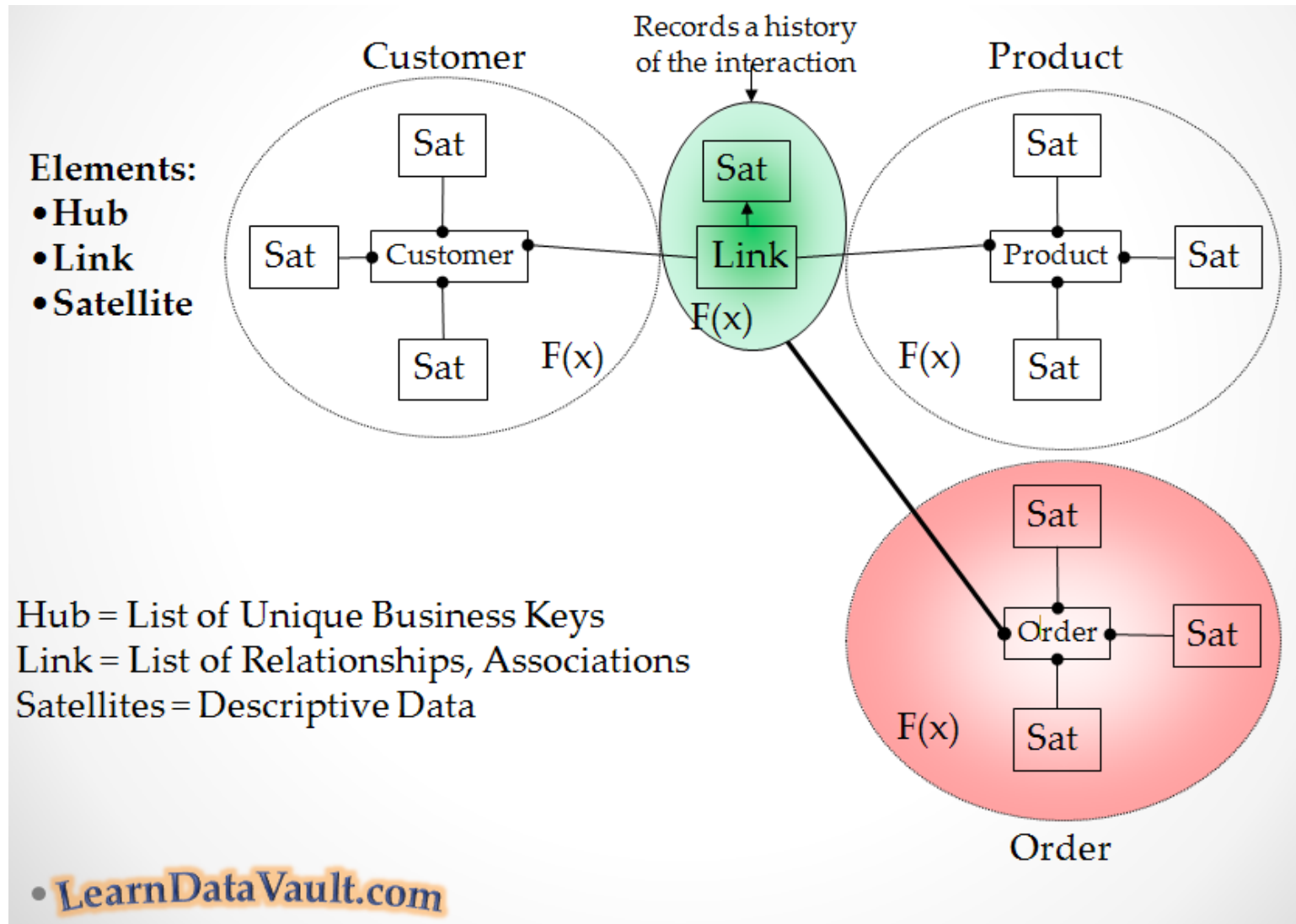
Modélisation Entrepôts de données 2.0 - Entreprise

- **Construction d'un Entrepôt de données d'Entreprise**
- Modélisation étoile idéale (pour OLAP) par sujet
- Alimentation

SOURCE → ETC (Staging) → DATA VAULT → Comptoir



« Data Vault » (par Dan Linstedt)



« Data Vault » (par Dan Linstedt)

Idéal pour:

- Entrepôt ENTREPRISE!
- Garder tous le détail transactionnel
- Prouver la traçabilité (SOX, vérifications)
- Multiples centres de données
- Énormément de sources de données
- ETC en « temps réel »

