KAPEI

Indicateur de performance

www.kapei-conseil.com

Clément LEFAURE
Co-gérant / Directeur technique

clefaure@kapei-conseil.com
30, rue Pré-Gaudry 69007 Lyon

Modélisation Datavault

Aix*Marseille université

Master Informatique
Science et Ingénierie des Données (SID)

05

Objectifs de la présentation

- Rappeler les modélisations existantes
- Présenter le concept de datavault
- Etudier l'éventuel intérêt de ce nouveau paradigme
- Donner les références pour en savoir plus

Modélisation « datavault »

- **1 – Présentation**
- **2 – Les modèles existants**
  - Relationnel et « formes normales »
  - Le multi-dimensionnel
  - Objectifs du datavault
- **3 – Le datavault en quelques mots**
  - Les composants du datavault modeling
  - Architecture du datavault
  - Etapes de chargement
- **4 – Conclusion**
  - Est-ce vraiment nouveau ?
  - Avantages / Inconvénients
  - Quand et pour quoi l'utiliser ?

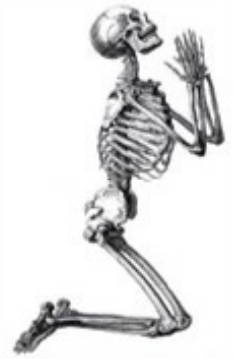
1 – PRESENTATION DU DATAVAULT

1

Qu'est-ce que le « Datavault » ?

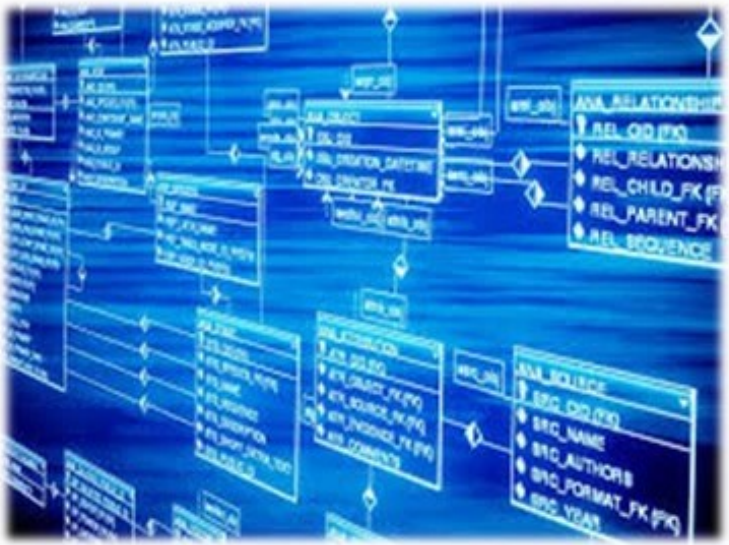
- Un modèle qui nous vient des Pays-Bas :
  - Dan Linstedt a conçu la modélisation Data Vault en 1990.
  - Il l'a placé dans le domaine public en 2000
- Il est utilisé par de grands groupes partout dans le monde (Volvo, IBM, HP, Deloitte, Dutsch police, Logica, Microsoft, Informatica, US universities… la liste est longue)
- Il est conçu pour gour gérer les **datawarehouses entreprises** (i.e. Enterprise Datawarehouse ou EDW)

Dan Linstedt,
le grand gourou



Pauvre mortel

2 – MODELES EXISTANTS



2

Le relationnel et les formes normales

- La plupart des bases relationnelles sont modélisées à l'aide des *formes normales*:
  - 1ère forme normale (1NF) :
    - Les tables ont des **clés uniques**
    - Elles contiennent des **valeurs non répétitives** (le cas contraire consiste à mettre une liste dans un seul attribut).
    - Les attributs sont **constants dans le temps** (utiliser par exemple la date de naissance plutôt que l'âge).
  - 2ieme forme normale (2NF) :
    - Tout attribut ne composant pas un identifiant **dépend d'un identifiant**.
  - 3ième forme normale (3NF) :
    - Tout attribut ne composant pas un identifiant **dépend directement d'un identifiant**.
- Les autres formes normales ne sont guère utilisées

Le relationnel et les formes normales: 1NF

- Table non normalisée (type "fichier Excel"):

Produit	Fournisseur
téléviseur	VIDEO SA, HITEK LTD

- Première forme normale (1NF)

Produit	Fournisseur
téléviseur	VIDEO SA
téléviseur	HITEK LTD

Le relationnel et les formes normales: 2NF et 3NF

Deuxième forme normale (2NF):

Produit	Fournisseur	Adresse fournisseur
téléviseur	VIDEO SA	13 rue du cherche-midi
écran plat	VIDEO SA	13 rue du cherche-midi
téléviseur	HITEK LTD	25 Bond Street

Produit	Fournisseur
téléviseur	VIDEO SA
téléviseur	HITEK LTD
écran plat	VIDEO SA

Fournisseur	Adresse fournisseur
VIDEO SA	13 rue du cherche-midi
HITEK LTD	25 Bond Street

Troisième forme normale (3NF):

Fournisseur	Adresse fournisseur	Ville	Pays
VIDEO SA	13 rue du cherche-midi	PARIS	FRANCE
HITEK LTD	25 Bond Street	LONDON	ENGLAND

Fournisseur	Adresse fournisseur	Ville
VIDEO SA	13 rue du cherche-midi	PARIS
HITEK LTD	25 Bond Street	LONDON

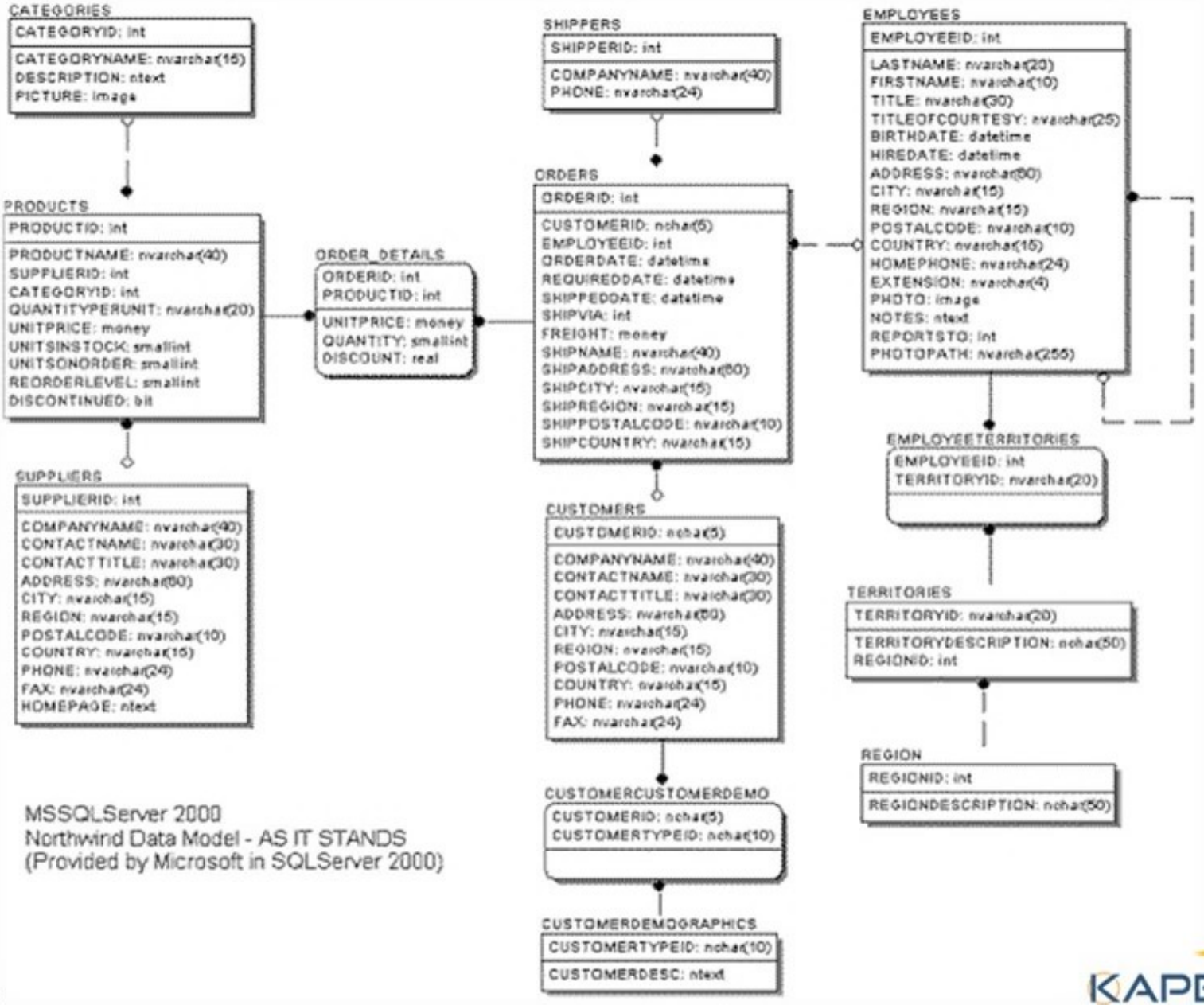
Ville	Pays
PARIS	FRANCE
LONDON	ENGLAND

Schéma normalisé 3NF

Exemple classique

(Commandes /

lignes de commande)



KAPEI

Modèle relationnel PROS / CONS

- Avantages:
  - Normalisé (3NF)
  - Pas de duplication de données
  - Utilisé pour les applications transactionnelles depuis 30 ans

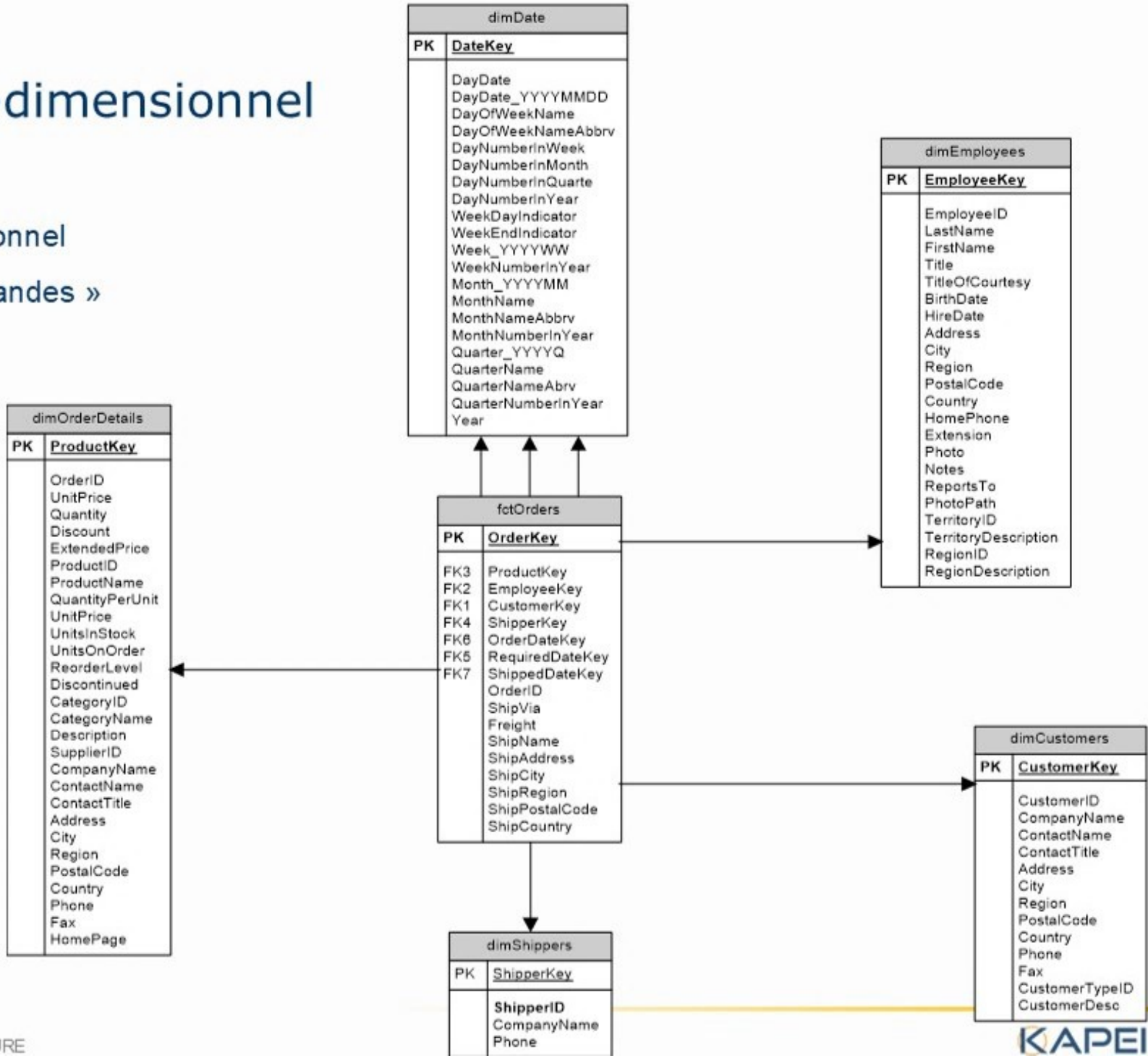
- Inconvénients:
  - La structure de la donnée n'est pas dédiée à l'analyse
  - Il manque un historique du contexte
  - Les jointures multiples consomment beaucoup de ressources

Modèle multi-dimensionnel

Modèle multi-dimensionnel

Table de fait « Commandes »

(Ralph Kimball)



KAPEI

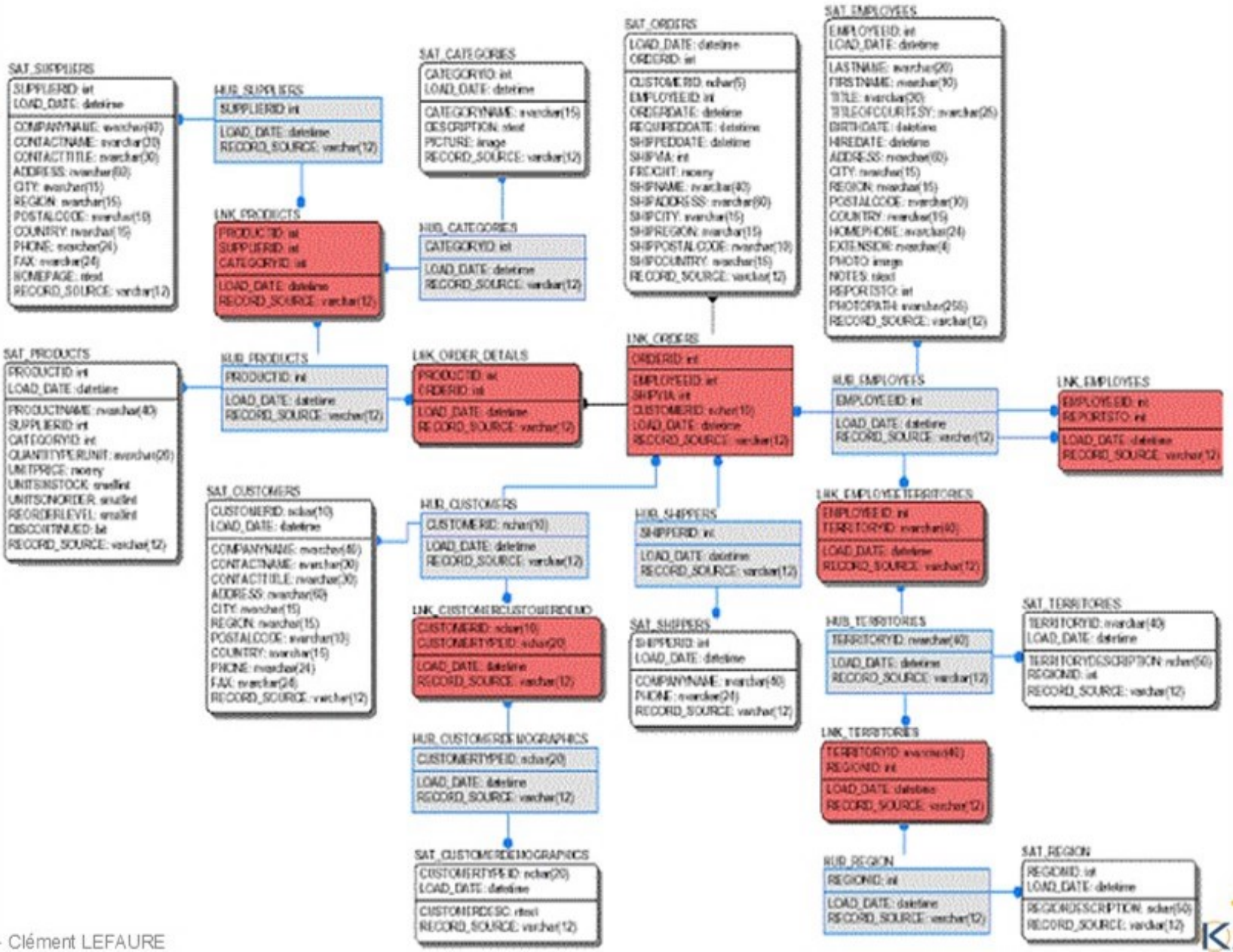
Modèle multi-dimensionnel PROS / CONS

- Avantage:
  - Modèle orienté analyse pour :
    - Comprendre des tendances
    - Prédire les futurs comportements et futurs besoins
  - Possibilités d'analyse d'une grande profondeur d'historique
  - Temps de réponse intéressants (OLAP)
- Inconvénients:
  - La dénormalisation et la duplication de données est autorisée: risque de problème de performance avec les chargements de données **temps réel** (ou pseudo temps réel)
  - A l'origine conçu pour analyser en se focalisant sur un seul domaine fonctionnel ("star schema" + logique de "datamart")
  - Difficultés avec les tables de faits évolutives ou à granularité variable

Datavault: un mix entre le relationnel et le dimensionnel

Datavault

« Commandes »



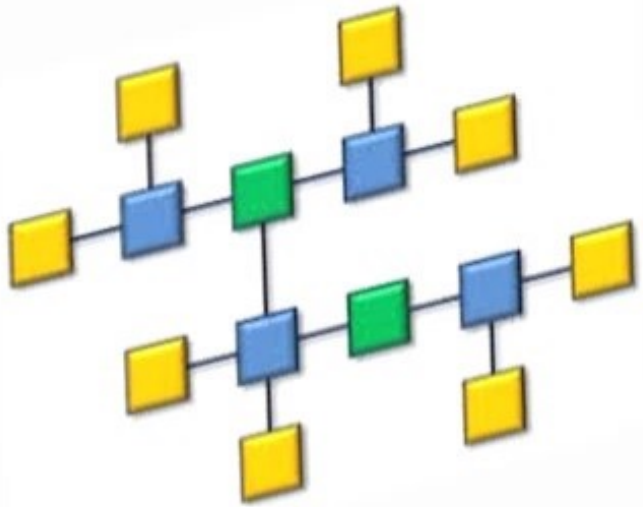
KAPEI

Objectifs du Datavault

*"Data Vault is a **detail oriented**, **historical tracking** and uniquely linked set of **normalized tables** that support one or more functional areas of business."*

- Approche hybride 3NF / « Star schema »
- Modèle normalisée
- Avec historisation des données
- Couvrant de multiples domaines fonctionnels de l'entreprise
- Venant de sources de données opérationnelles multiples (et tracking)
- Avec gestion du temps réel (ou pseudo temps-réel)

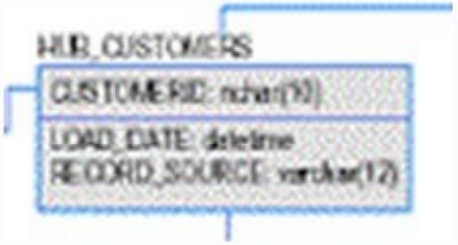
3 – LE DATAVAULT EN QUELQUES MOTS



3

Composants du Datavault: **Hubs**

- Les « hubs »
  - Une simple table contenant une liste de **clés business**
  - Ces clés sont celles utilisées dans la source tous les jours:
    - EX : numéro de facture, identifiant d'employé etc...
  - Attributs types :
    - Business key
    - Surrogate key
    - Load Date time stamp
    - Record source



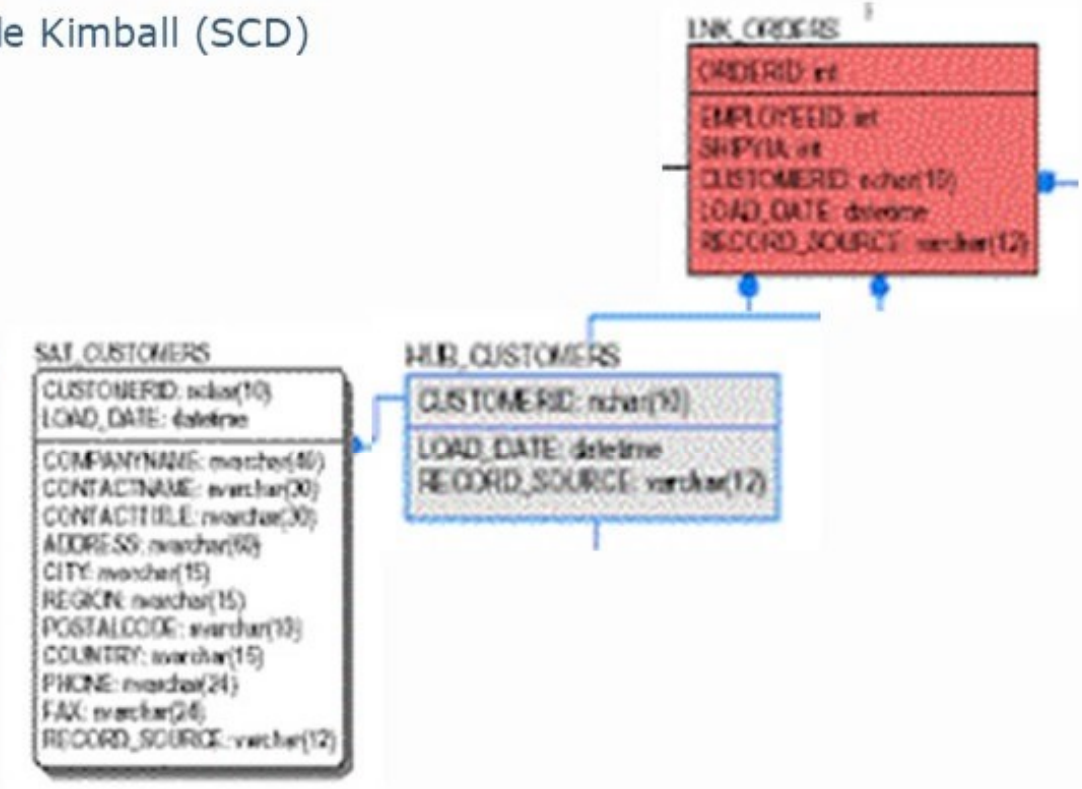
Composants du Datavault: **Liens**

- Les « liens »
  - Ils représentent la relation (la transaction) entre plusieurs composants business
    - Ils contiennent des clés business
  - Cela ressemble un peu aux "tables de faits" de Kimball
  - Attributs types :
    - Surrogate Key
    - Hub 1 Key to Hub N Key
    - Load Date Time Stamp
    - Record Source

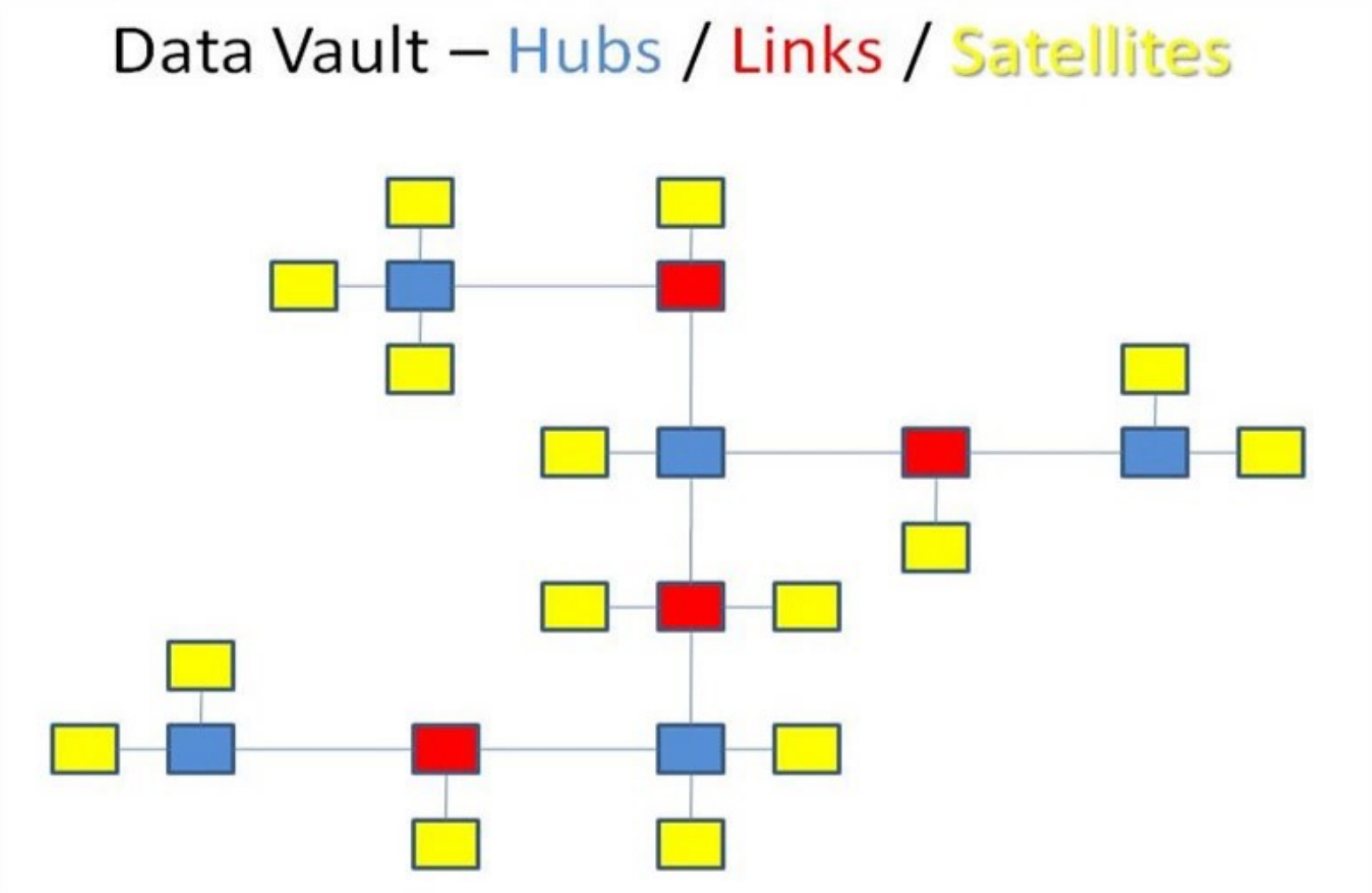
KAPEI

Composants du Datavault: **Satellites**

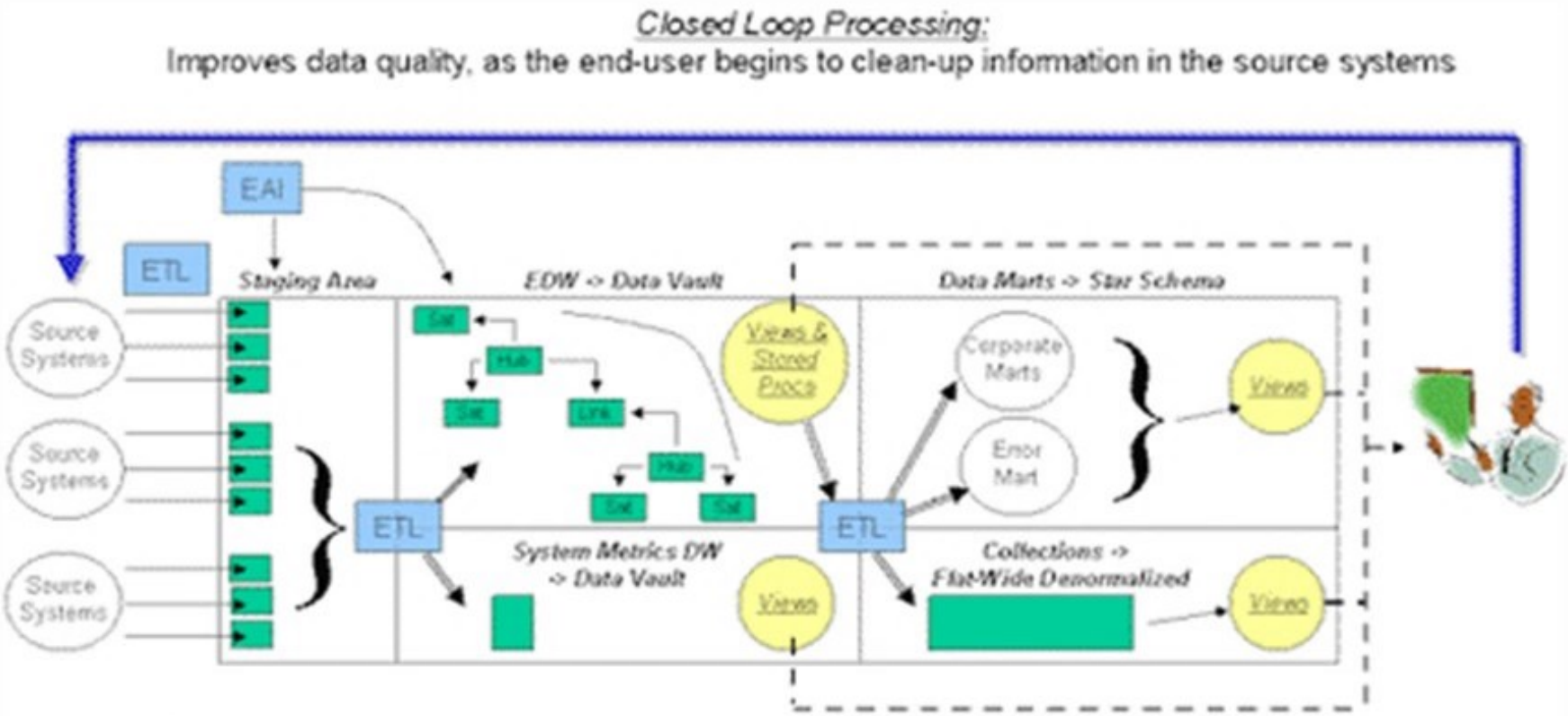
- Les « satellites »
  - Les satellites sont des attributs descriptifs des "Hubs"
  - Toute cette information peut changer dans le temps
  - Ressemble aux dimension de Type 2 de Kimball (SCD)
  - Attributs types :
    - Satellite Primary Key:
      - Hub Or Link PK
      - and Load Date Time Stamp
      - (Begin Date)
    - Satellite Optional PK:
      - Seq. Surrogate Num
    - End Date (Optional)
    - Record Source



Vue macro sur un modèle Datavault



Architecture du Datavault



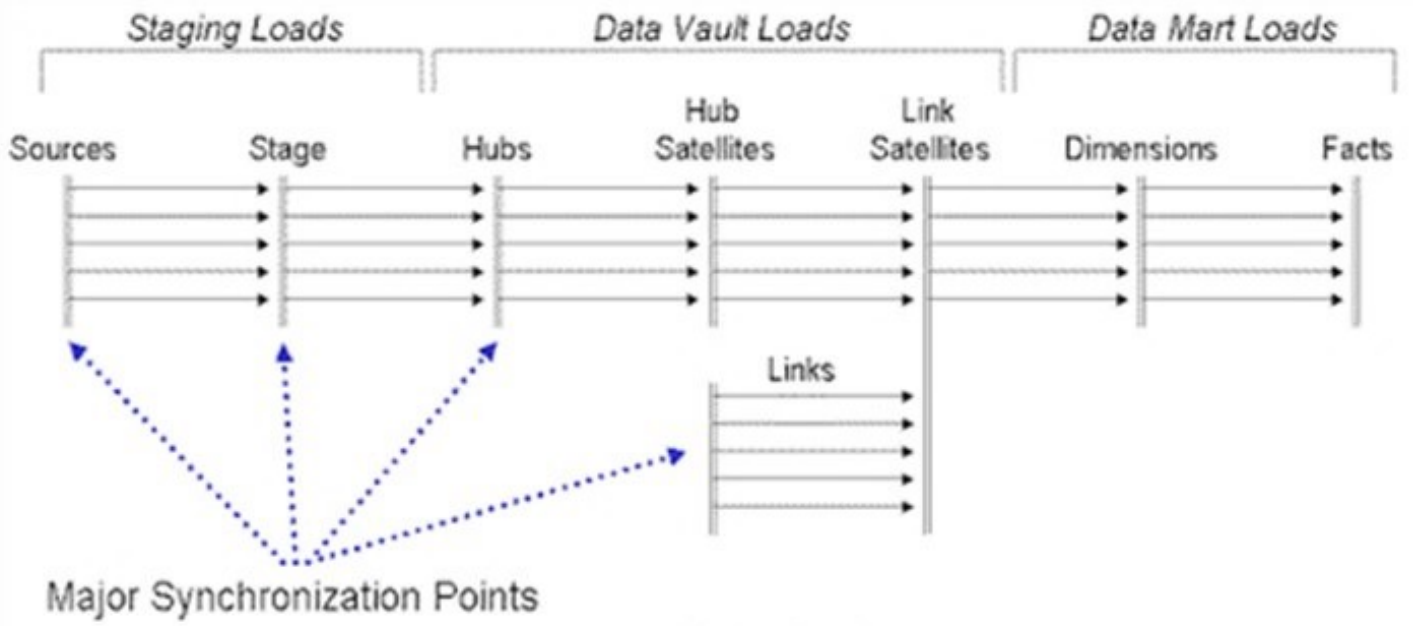
OPTIMIZE
- Parallel Processing
- Maximize Resources
- Optimize Throughput And Scalability
- Flexibility
- Information and Process Tracability

DW Team Owns
- Indexes
- Partitioning
- Restartability
- Parallelism
- Volume
- Data Processing

Why Views?
- Ability to control access
- Query Monitoring and Maintenance
- Isolate Reports from Architecture Changes
- Fast Re-use of Multiple DB Buffers
- Layer of Abstraction
- Logical Data Marts

Business User Owns
- Business Rules
- Accessibility
- Data Formatting
- Quality
- Information (the data)

Etapes de chargement

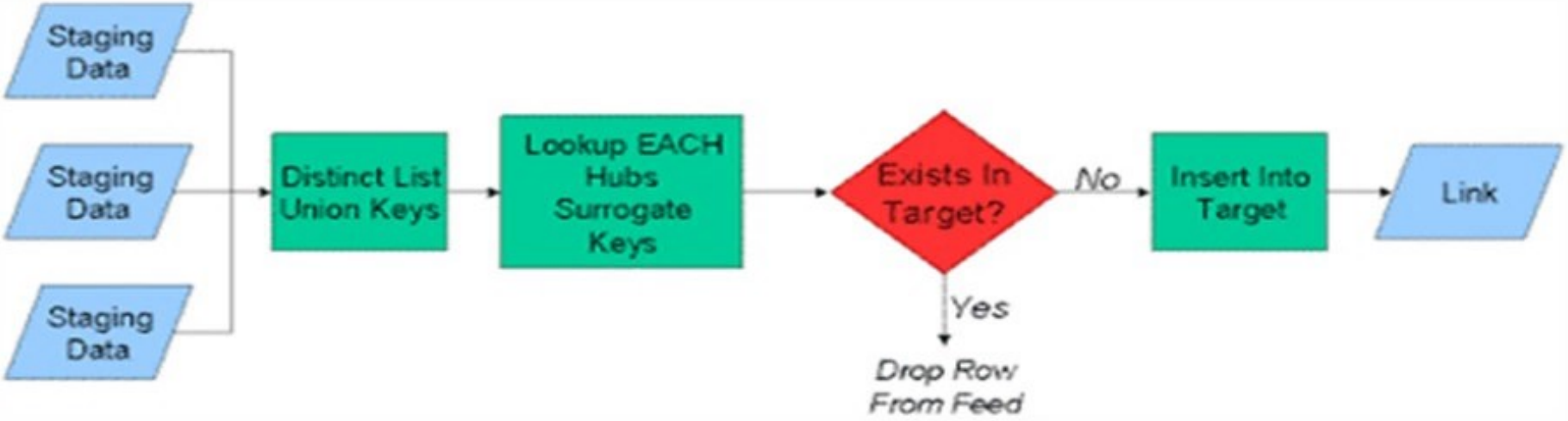


Processing:

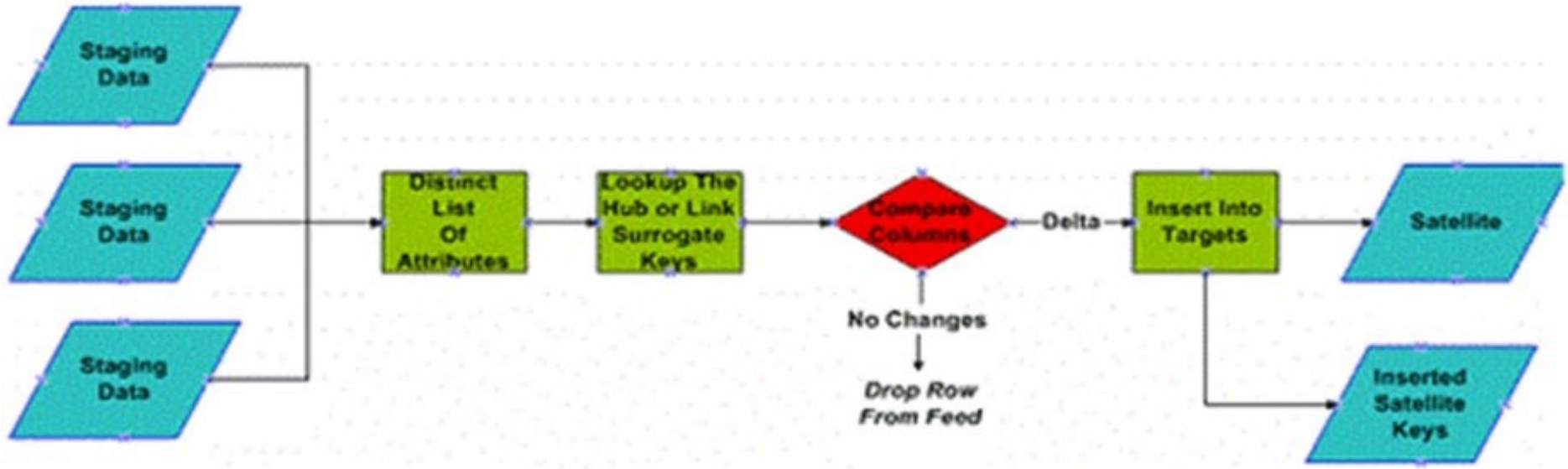
- All loads are done in parallel
- Sets of processes "wait" for the previous set to complete
- Processes are run as soon as data is ready.
- No other "waiting" time is required.
- Load Dependencies are greatly reduced.

Chargement des Hubs

Chargement des Liens



Chargement des satellites



4 – CONCLUSION

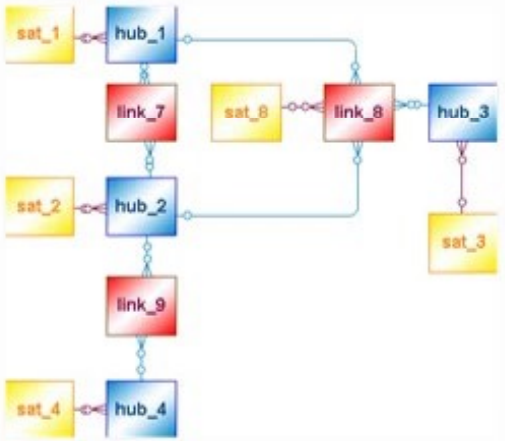
4

Le Datavault **n**'est **pas**

- Un système opérationnel ou un ERP
  - Le datavault n'est pas conçu pour gérer du transactionnel (peu d'updates)
- Un cube OLAP
  - En revanche des cubes ou des datamarts peuvent être clients du datavault
- Un MDM
  - Le datavault ne permet pas de gérer les données référentielles
  - Il permet juste de gérer les sources multiples
  - Mais il peut aussi se situer en parallèle d'un MDM

En résumé…

- C'est une sorte d'ODS amélioré (« ODS++ »)
- Proche de l'ancienne notion d' « infocentre »
- Ou d'une base de données temporelle :
  - Les données sont « time-stampées »
  - Et tracées par source
- Ce n'est pas si différent de ce que l'on fait déjà



Avantages du Datavault

- Datavault permet de construire un EDW (enterprise datawarehouse) d'une grande souplesse
- Différents sources et types de données peuvent être intégrés facilement
- Avec une traçabilité complète des données
- Ses techniques de modélisation permettent de supporter vraiment du "temps réel"
- Permet d'intégrer facilement des domaines fonctionnels très divers
- Son architecture évolutive garantie

En doublant les capacité hardware on diminue par deux le temps de restitution !

Inconvénients du Datavault

- Plus difficile à maintenir qu'un schéma en étoile
  - Il est possible de développer un « framework » pour automatiser la maintenance (voir le site pour plus d'info)
- La création du modèle est assez couteuse en temps et énergie et requiert de bonnes compétences en modélisation
- Relativement peu d'experts à l'heure actuelle comparé aux autres méthodologies

Quand et pour quoi utiliser le Datavault ?

- Pour construire un datawarehouse entreprise centralisé:
  - Besoin de temps réel
  - Multiples sources de données contradictoires
  - Forte évolutivité des données
- Pour s'appuyer sur une méthodologie et éviter "réinventer la roue"
  - Si les problématiques ont déjà été étudiées, autant réutiliser l'existant
  - Possibilité de piocher les bonnes idées "à la carte"
- Pour s'en servir de caution méthodologique :
  - Pour rassurer client et consultant

Bibliographie et crédits

- Le site officiel :
  - http://danlinstedt.com/

- Wikipédia:
  - http://en.wikipedia.org/wiki/Data_Vault_Modeling
- Datavault Academy :
  - https://www.youtube.com/user/DataVaultAcademy

Quizz

- Que veut dire le « vault » de datavault ?
- Le datavault est une base relationnelle. Vrai ou faux ?
- Une modélisation en datavault permet-il de minimiser le volume d'une base de données ?
- Modéliser en datavault permet de réduire le temps d'analyse en phase de conception. Vrai ou faux ?
- Le datavault est-il apparut avant le décisionnel ?
- Doit-on développer un datavault par domaine métier?
- BONUS: Quand doit-on passer en datavault ?

A retrouver sur Kahoot.com