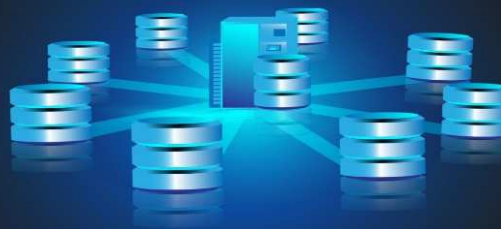


Data Warehouse

Dr. Zayneb TRABELSI



Plan

- BI et DW
- Modélisation Conceptuelle d'un DW
- Mise en œuvre du DW (Intégration des données)
- Modélisation Logique
- Analyse OLAP
- Langages de requêtes OLAP
- Reporting et Data Visualization
- Data Lake



Data Warehouse

BI et Data Warehouse



Contexte : BI (*Business Intelligence*)

- Besoin: prise de décisions stratégiques
- Pourquoi: besoin de réactivité
- Qui: les décideurs (non informaticiens)
- Comment: répondre aux demandes d'analyse des données, dégager des informations qualitatives nouvelles
 - Quels sont mes produits les plus vendus ?
 - Quels sont mes meilleurs magasins ?



Contexte: Problématique

- Données opérationnelles (de production)
 - Bases de données (Oracle, SQL Server)
 - Fichiers, etc.
- Caractéristiques de ces données:
 - Distribuées: systèmes éparpillés
 - Hétérogènes: systèmes et structures de données différents
 - Détaillées: organisation des données selon les processus fonctionnels
 - Volatiles: pas d'historisation
 - Peu/pas adaptées à l'analyse : les requêtes lourdes peuvent bloquer le système transactionnel



Système d'information dédié aux applications décisionnelles: un data warehouse



DW: Définition

- W. H. Inmon (1996): «Le data Warehouse est une collection de données **orientées sujet, intégrées, non volatiles et historisées**, organisées pour le support d'un processus d'aide à la décision»

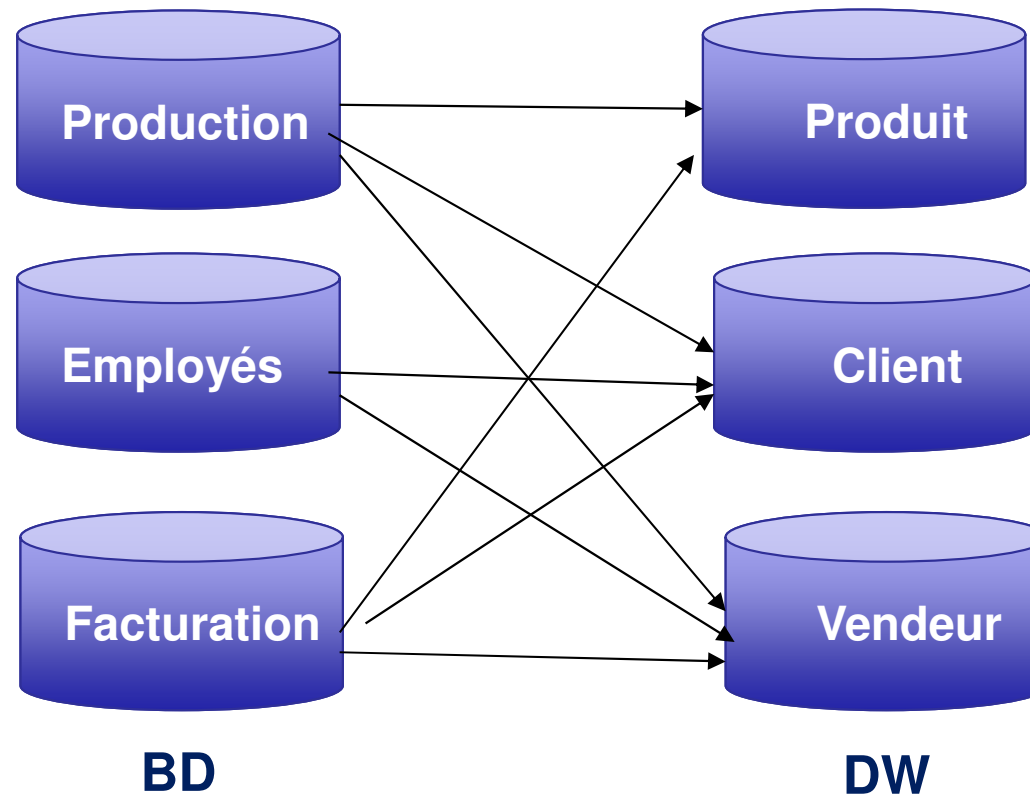


Mettre en place une base de données à **des fins d'analyse**



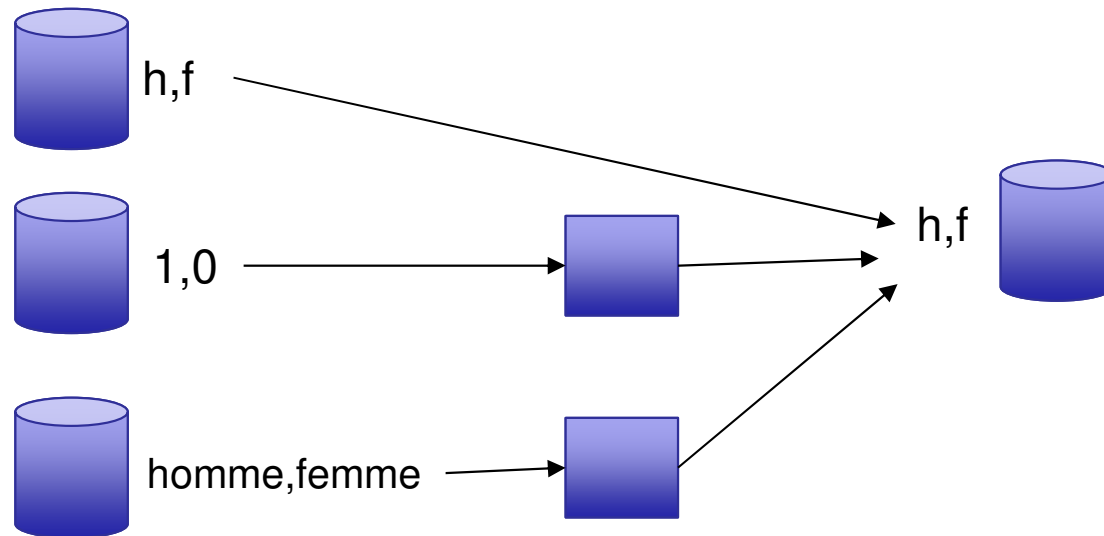
DW: Caractéristiques des données

- Orientées Sujet
 - Les données sont organisées par thème
 - Les données propres à un thème, les ventes par exemple, seront extraites des différentes bases de production et regroupées.



DW: Caractéristiques des données

- Intégrées
 - Les données proviennent de sources hétérogènes utilisant chacune un type de format
 - Elles sont intégrées avant d'être proposées à utilisation (référentiel unique)



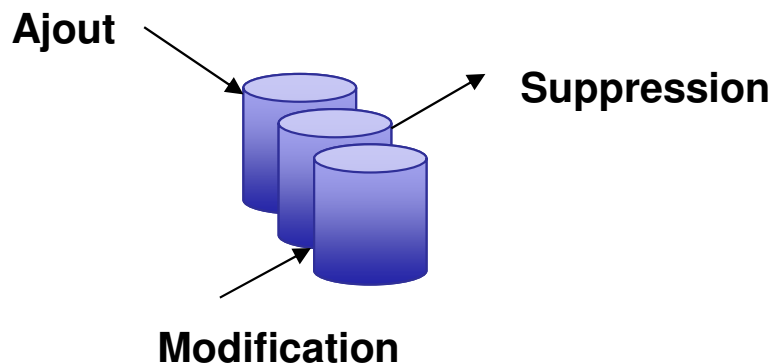
DW: Caractéristiques des données

- Historisées
 - Stockage de l'historique des données
 - Un référentiel temps doit être associé aux données



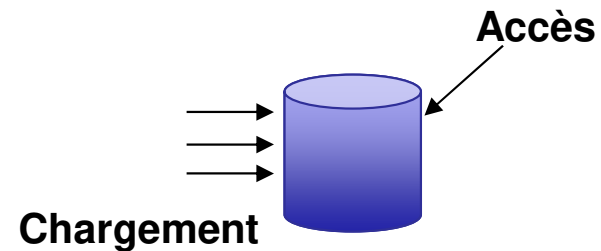
DW: Caractéristiques des données

- Non-volatiles
 - Conséquence de l'historisation
 - Une même requête effectuée à intervalle de temps, en précisant la date référence de l'information donnera le même résultat
 - Pas de mises à jour des données dans le DW



OLTP
(On-Line Transactional Processing)

BD



OLAP
(On-Line Analytical Processing)

DW



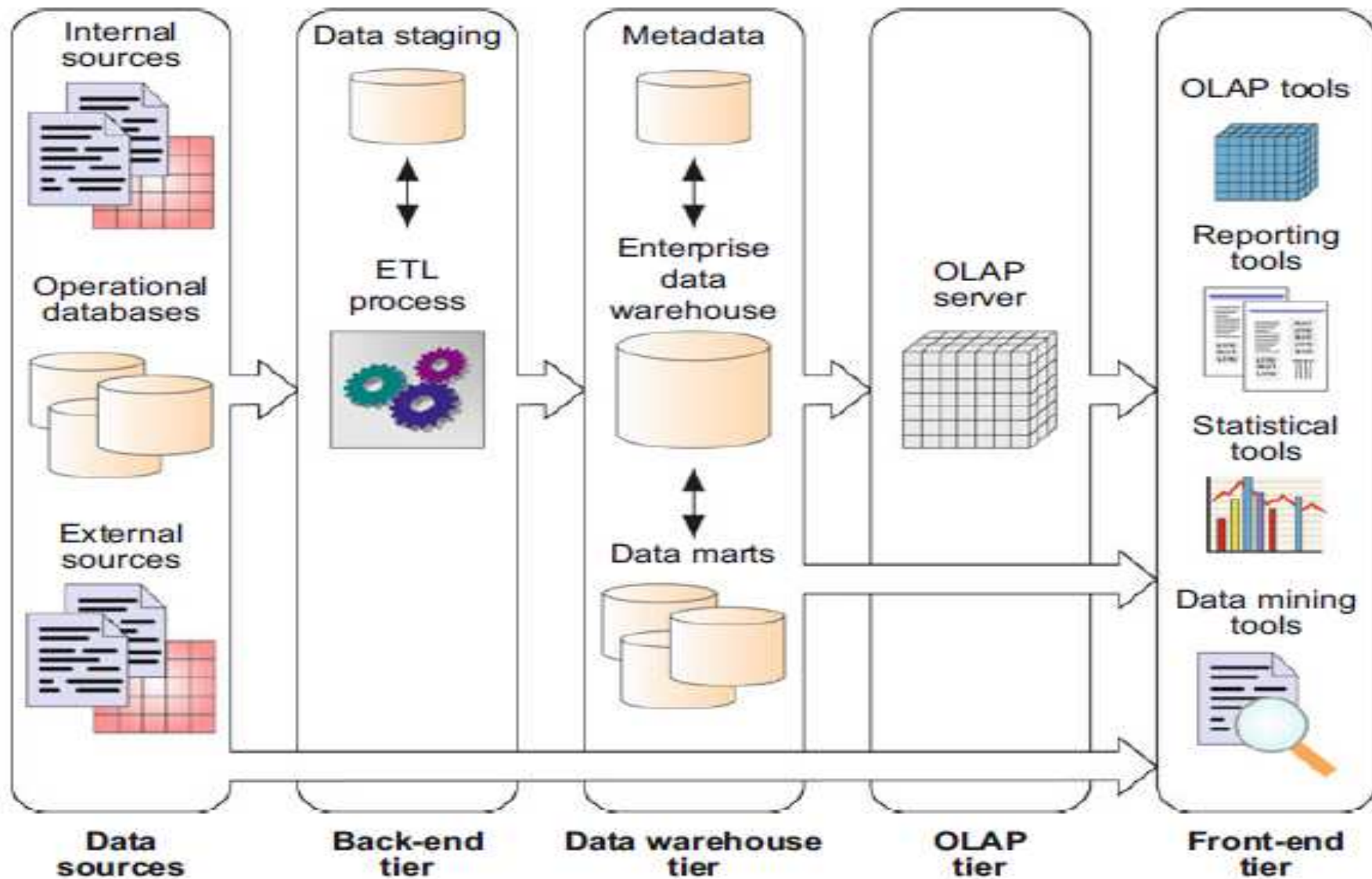
DW: OLTP vs OLAP

| Characteristic | Operational Database (OLTP) | Data Warehouse (OLAP) |
|---------------------|-----------------------------|---|
| Currency | Current | Historical |
| Details level | Individual | Individual and summary |
| Orientation | Process | Subject |
| Records per request | Few | Thousands |
| Update level | Highly volatile | Mostly refreshed (non volatile) |
| Data model | Relational | Relational (star schemas) and multidimensional (data cubes) |

11



DW: Architecture Générale

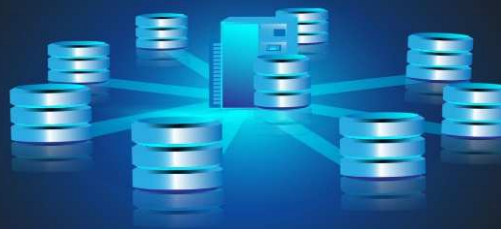


12



Data Warehouse

Modélisation Conceptuelle



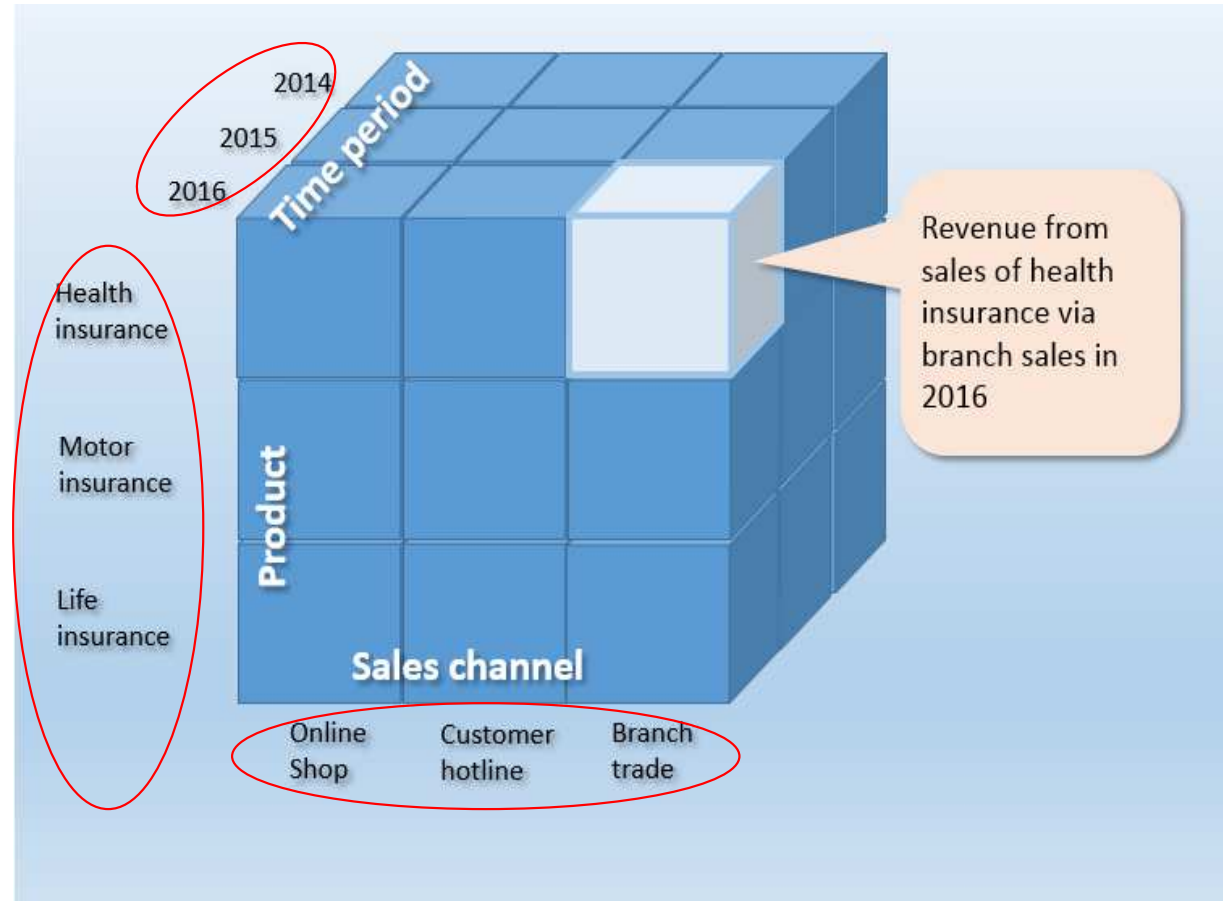
Modélisation Conceptuelle

- Nouvelle méthode de conception autour des concepts métiers
 - Ne pas normaliser au maximum
- Introduction de nouveaux types de table:
 - Table de faits
 - Table de dimensions
- Introduction de nouveaux modèles:
 - Modèle en étoile (star)
 - Modèle en flocon (snowflake)
 - Modèle en constellation (galaxy)



Modélisation Conceptuelle: Fait - Dimension

- Dimensions : axes d'analyse (date, produit, magasin, etc.)
- Fait: sujet d'analyse (mesures: revenu, quantités vendues, etc.)
 - la valeur d'une mesure, calculée ou mesurée, selon un membre de chacune des dimensions
- Exemple : **200 000 dinars** est un fait qui exprime la valeur de la mesure **Revenu** pour le membre **2016** de la dimension **Time** et le membre **Health** de la dimension **Product** et le membre **Branch** de la dimension **Sales Channel**



Modélisation Conceptuelle: Types de Faits

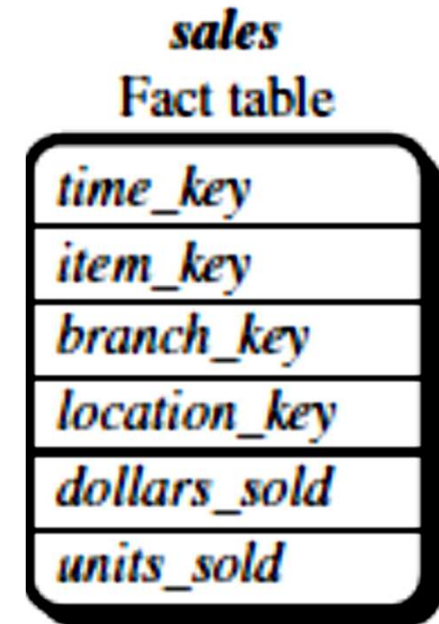
- Additif: additionnable suivant toutes les dimensions
 - Quantités vendues, chiffre d'affaire
 - Peut être le résultat d'un calcul:
 - Bénéfice = montant vente – coût
- Semi additif: additionnable suivant certaines dimensions
 - Solde d'un compte bancaire:
 - Inutile d'additionner sur les dates car cela représente des instantanés d'un niveau
 - Somme sur les comptes: tout ce que nous possédons en banque
- Non additif: fait non additionnable quelque soit la dimension
 - Prix unitaire: l'addition sur n'importe quelle dimension donne un nombre qui n'a pas de sens

16



Modélisation Conceptuelle: Table de fait

- Table principale du modèle dimensionnel
- Contient les données observables (les faits) sur le sujet étudié selon divers axes d'analyse (les dimensions)
- Attributs de la table des faits
 - **des clés étrangères formant une clé primaire**
 - des mesures associées à chaque clé primaire



Modélisation Conceptuelle: Table de dimension

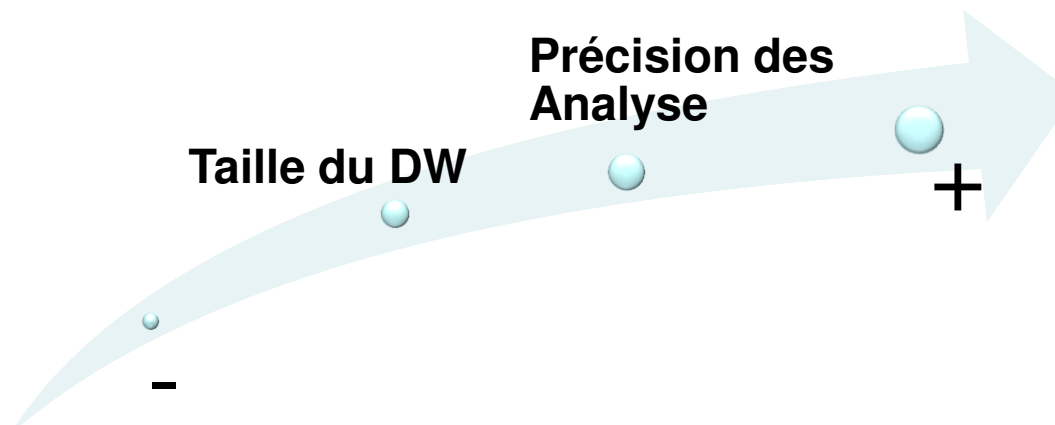
- Axe d'analyse selon lequel vont être étudiées les données observables (faits)
- Contient le détail sur les faits
- Dimension Temps
 - Commune à l'ensemble du DW
 - Reliée à toute table de faits

| <i>time</i> Dimension table | |
|--------------------------------|--|
| <i>time_key</i> | |
| <i>day</i> | |
| <i>day_of_the_week</i> | |
| <i>month</i> | |
| <i>quarter</i> | |
| <i>year</i> | |



Modélisation Conceptuelle: Granularité Dimension

- Une dimension contient des membres organisés en hiérarchie :
 - Chacun des membres appartient à un niveau hiérarchique (ou niveau de granularité) particulier
 - Granularité d'une dimension : nombre de niveaux hiérarchiques
 - Exemples :
 - Dimension temporelle : jour, mois, année
 - Dimension géographique : magasin, ville, région, pays
 - Dimension produit : produit, catégorie, marque



Modélisation Conceptuelle: Types de modèles

- Etoile (Star)
- Flocon (Snowflake)
- Constellation (Galaxy)

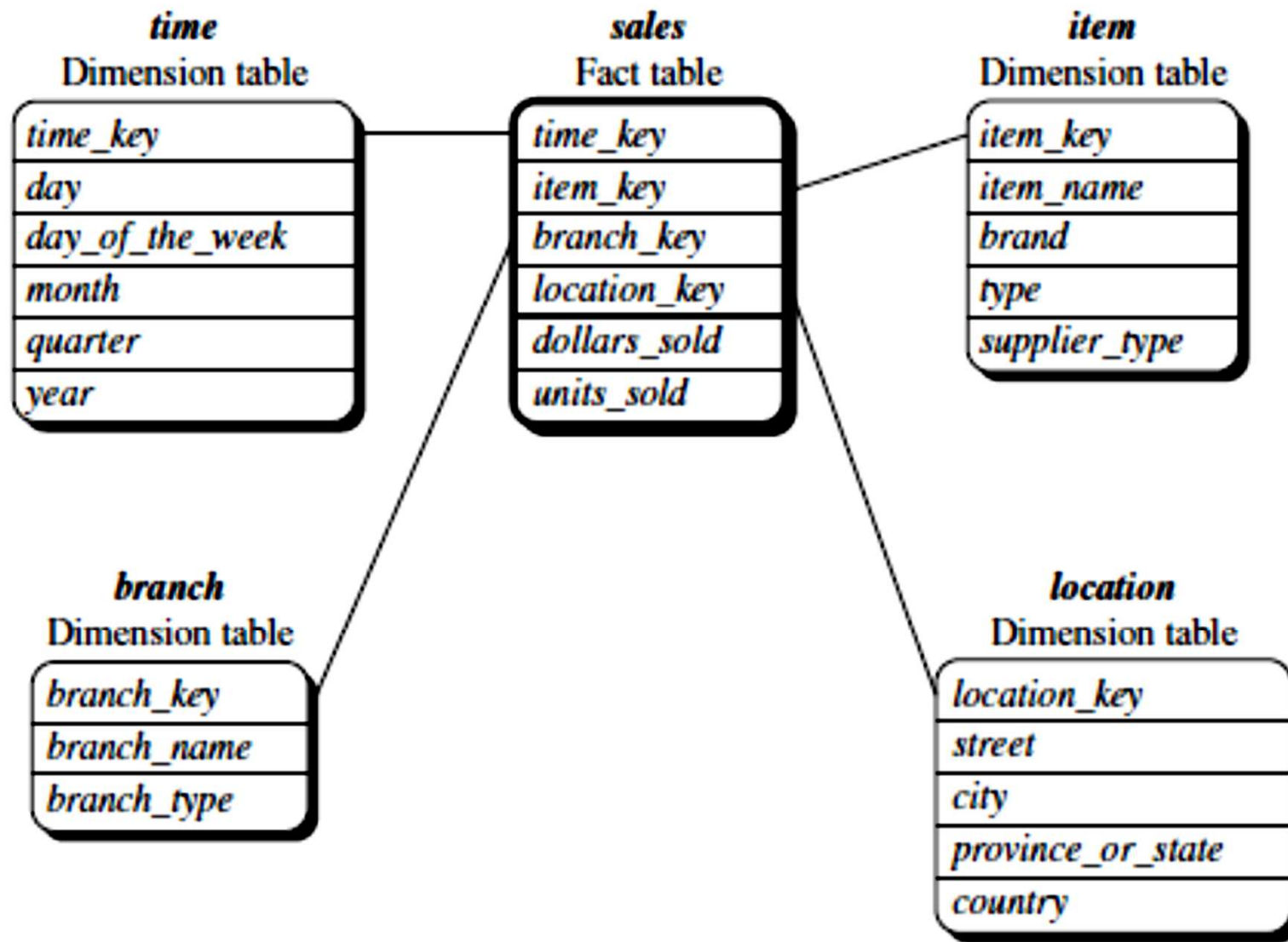


Modélisation Conceptuelle: Types de modèles - Etoile

- Une table de fait centrale et des dimensions
- Association de type *many-to-one* connectant les différentes dimensions aux faits
- Les dimensions n'ont pas de liaison entre elles
 - Avantages:
 - Facilité de navigation
 - Nombre de jointures limité
 - Inconvénients:
 - Redondance dans les dimensions



Modélisation Conceptuelle: Types de modèles - Etoile



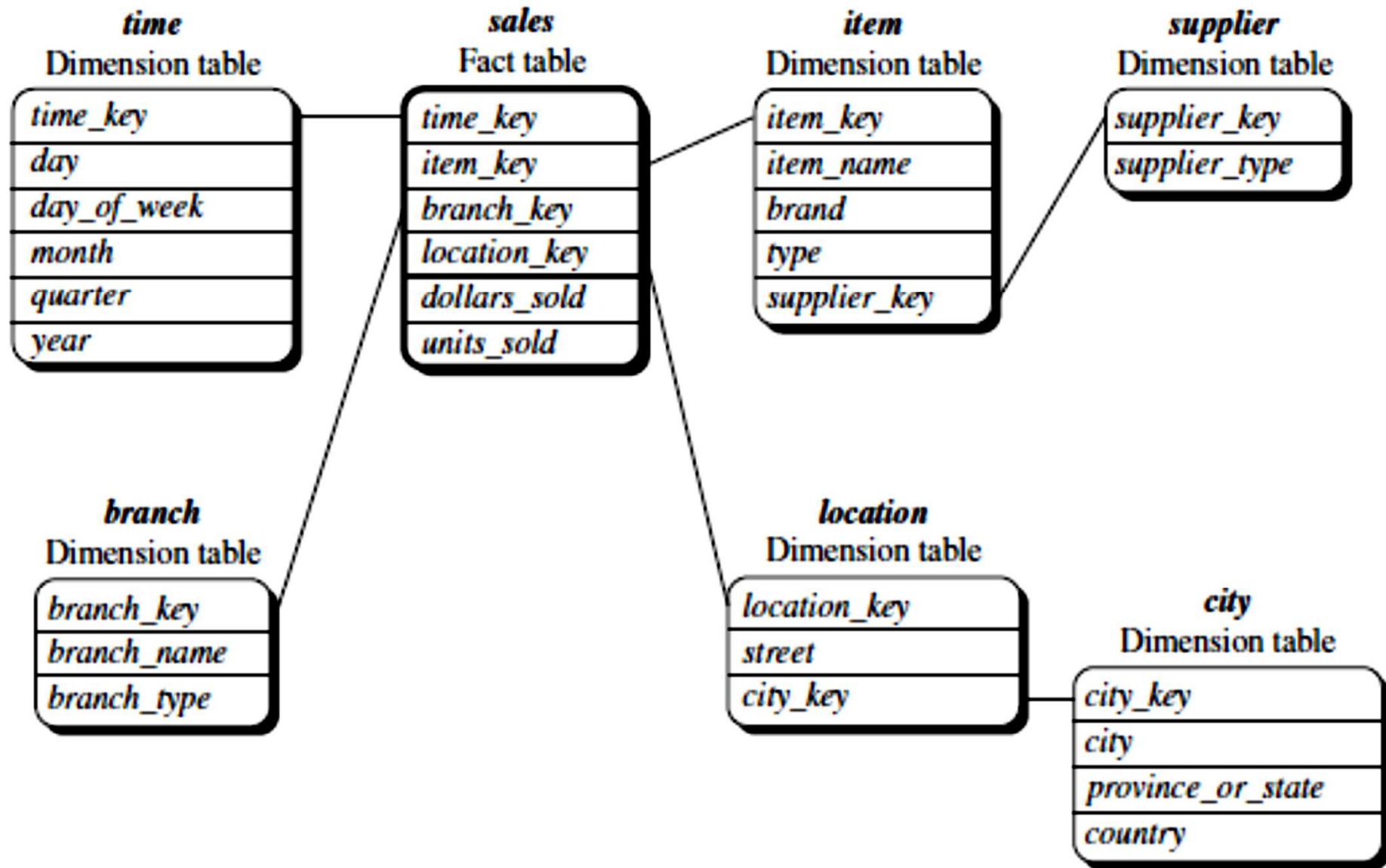
Modélisation Conceptuelle: Types de modèles - Flocon

- Evolution du modèle en étoile
- Une table de fait et des dimensions décomposées en sous-hiérarchies: en général, un seul niveau hiérarchique dans une table de dimension
- **La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine**
- Avantages:
 - Normalisation des dimensions
 - Économie d'espace disque
- Inconvénients:
 - Modèle plus complexe (jointures)
 - Requêtes moins performantes

23



Modélisation Conceptuelle: Types de modèles - Flocon



24



Modélisation Conceptuelle: Types de modèles - Exemple

Flocon

| Product_ID | Description | Brand | Prod_group_ID |
|------------|-------------|---------|---------------|
| 10 | E71 | Nokia | 4 |
| 11 | PS-42A | Samsung | 2 |
| 12 | 5800 | Nokia | 4 |
| | Bold | Berry | 4 |

| Prod_group_ID | Description | Prod_categ_ID |
|---------------|--------------|---------------|
| 2 | TV | 11 |
| 4 | Mobile Pho.. | 11 |

| Prod_categ_ID | Description |
|---------------|-------------|
| 11 | Electronics |

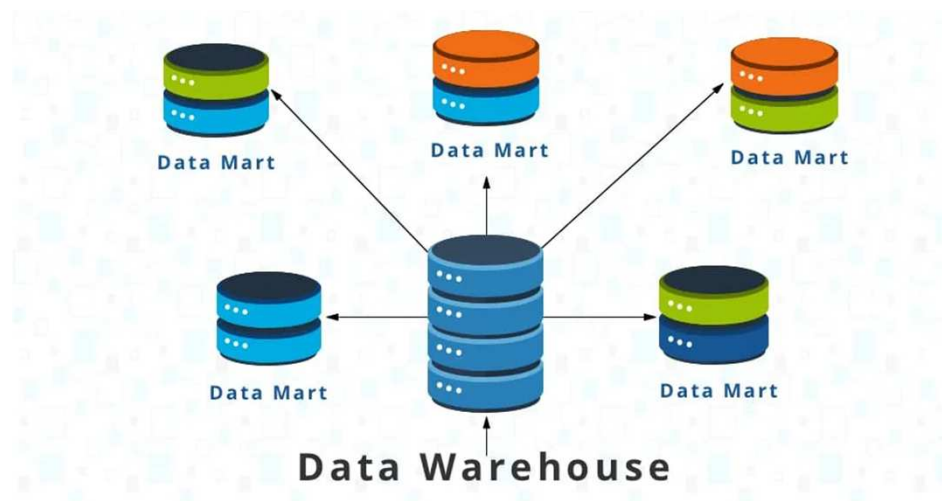
Etoile

| Product_ID | Description | ... | Prod. group | Prod. categ |
|------------|-------------|-----|-------------|-------------|
| 10 | E71 | ... | Mobile Ph.. | Electronics |
| 11 | PS-42A | ... | TV | Electronics |
| 12 | 5800 | | Mobile Ph.. | Electronics |
| 13 | Bold | | Mobile Ph.. | Electronics |



Modélisation Conceptuelle: Types de modèles - Constellation

- Ensemble de schémas en étoiles/flocons dans lesquels les tables de faits se partagent certaines tables de dimensions
- En général, on a:
 - un schéma de constellation de faits pour le **Data Warehouse**
 - Une étoile de la constellation pour un magasin de données **Data Mart**

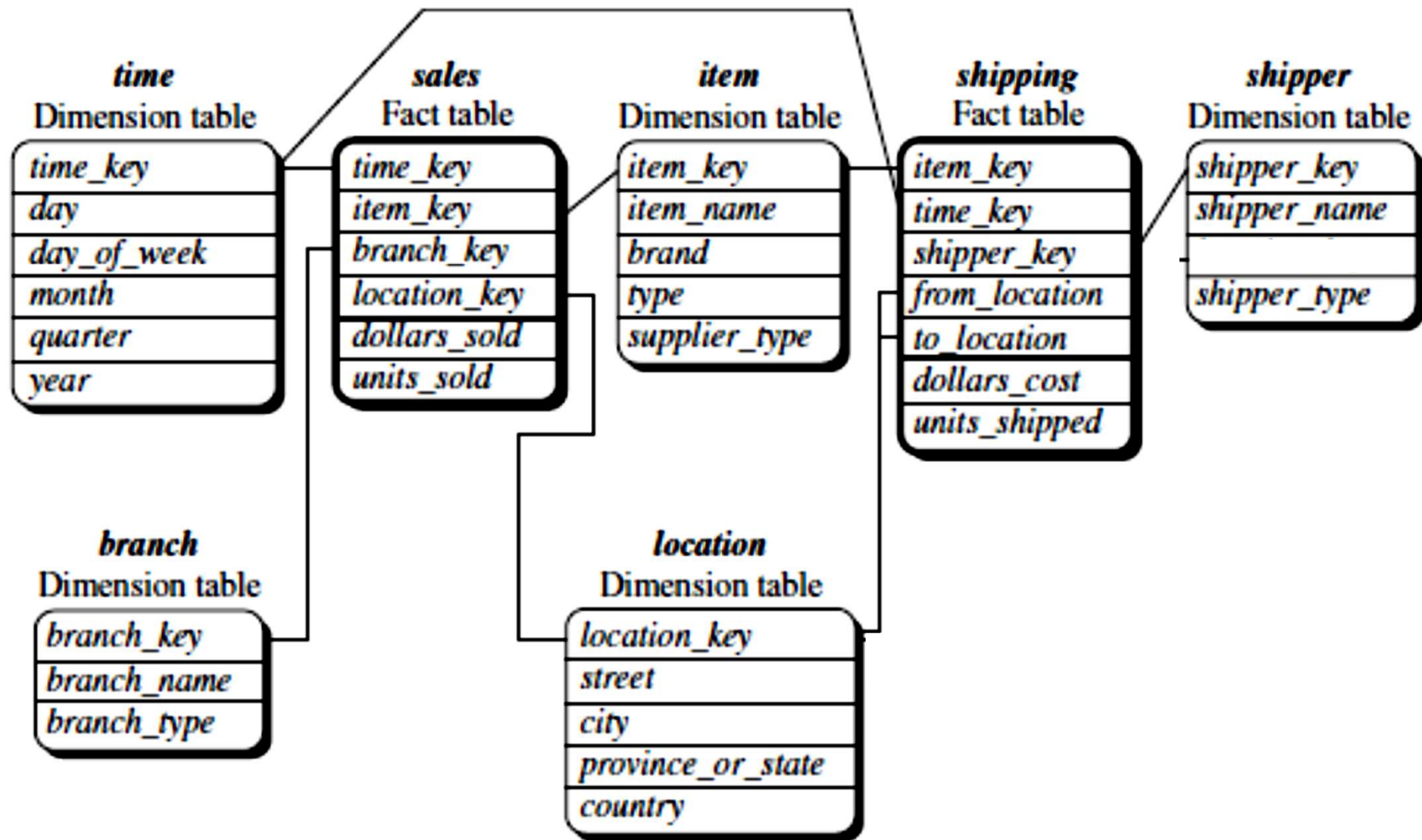


Source: Talend.com

26



Modélisation Conceptuelle: Types de modèles - Constellation



Modélisation Conceptuelle: Synthèse

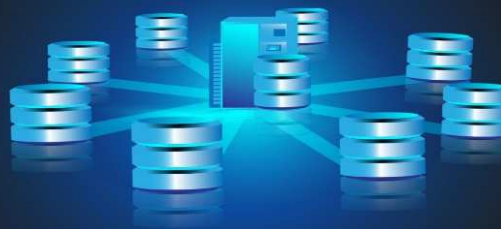
| Star Schema | Snowflake Schema |
|--|--|
| Hierarchies for the dimensions are stored in the dimensional table. | Hierarchies are divided into separate tables. |
| It contains a fact table surrounded by dimension tables. | One fact table surrounded by dimension table which are in turn surrounded by dimension table |
| In a star schema, only single join creates the relationship between the fact table and any dimension tables. | A snowflake schema requires many joins to fetch the data. |
| Simple DB Design. | Very Complex DB Design. |
| Denormalized Data structure and query also run faster. | Normalized Data Structure. |
| High level of Data redundancy | Very low-level data redundancy |
| Single Dimension table contains aggregated data. | Data Split into different Dimension Tables. |
| Cube processing is faster. | Cube processing might be slow because of the complex join. |

28



Data Warehouse

Intégration de Données



Choix de l'Architecture

- Architecture top-down (*Inmon*)
➡ Approche Data Warehouse
- Architecture bottom-up (*Kimball*)
➡ Approche Data Mart



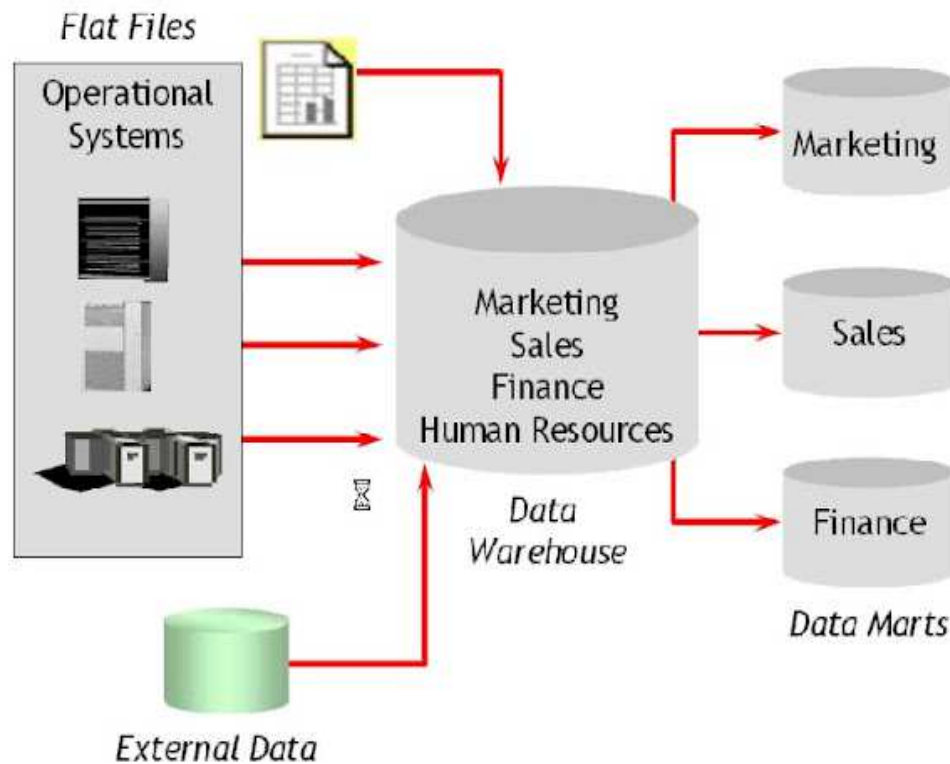
Choix de l'Architecture – DW vs Data Mart

| Caractéristiques | DW | DM |
|------------------------------|--|---|
| Echelle du modèle de données | Entreprise | Département |
| Champs applicatifs | Multi-sujet | Quelques sujets |
| Sources de données | Multiples (BD opérationnelle, BD externes, etc.) | Quelques unes (pour un besoin d'analyse spécifique) |
| Stockage | Plusieurs BD distribuées | Une BD |
| Taille | > 100 Go | 10 à 20 Go |

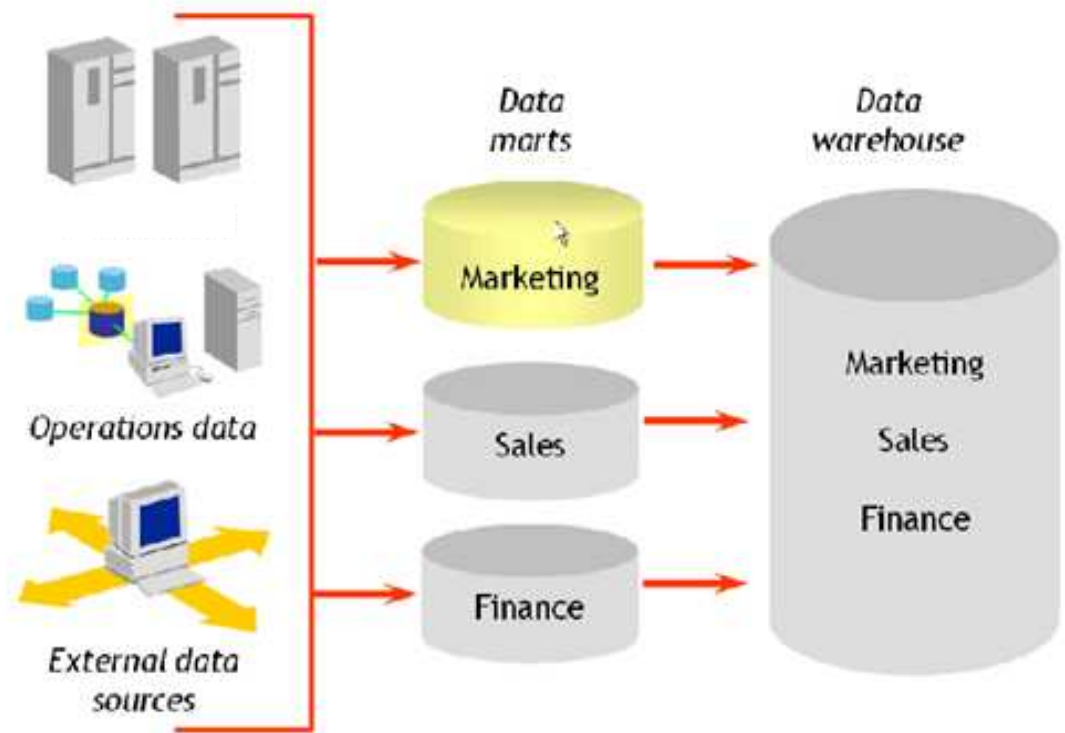


Choix de l'Architecture

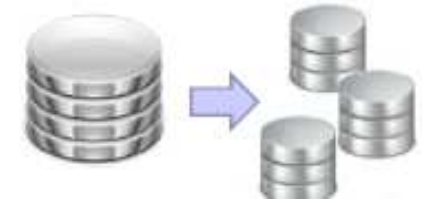
Top Down Approach



Bottom-Up Approach

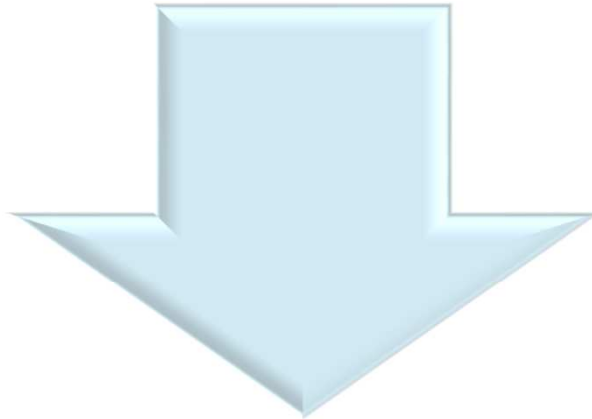


Choix de l'Architecture



Top-Down

- Conception intégrale du DW a priori
 - DM extraits du DW
- + Vision conceptuelle globale du DW
- + Normalisation des données, absence de redondance
- Difficulté de mise en oeuvre
- Manque d'évolutivité



Bottom-Up

- Construction incrémentale du DW
 - Le DW est une union de DMs
- + Simplicité de mise en oeuvre
- + Résultats rapides
- Difficulté d'intégration des DMs



Alimentation du DW

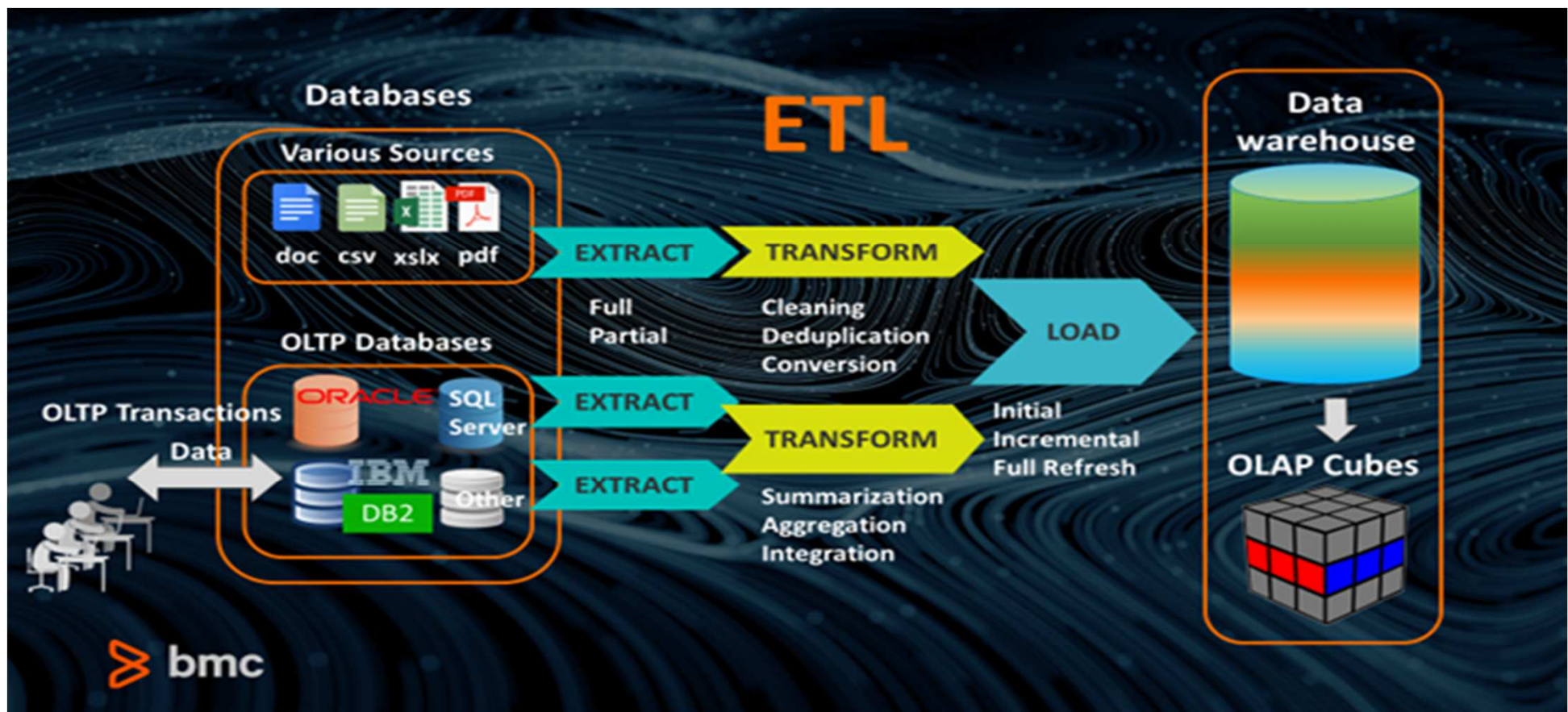
- Après avoir conçu le modèle de données du DW, on doit l'alimenter:
 1. Sélection des sources de données
 2. Extraction des données (*Extract*)
 3. Transformation (*Transform*)
 4. Chargement (*Load*)



Alimentation du DW

- Processus d'alimentation (*ETL Pipeline*)

Extract ➡ Transform ➡ Load

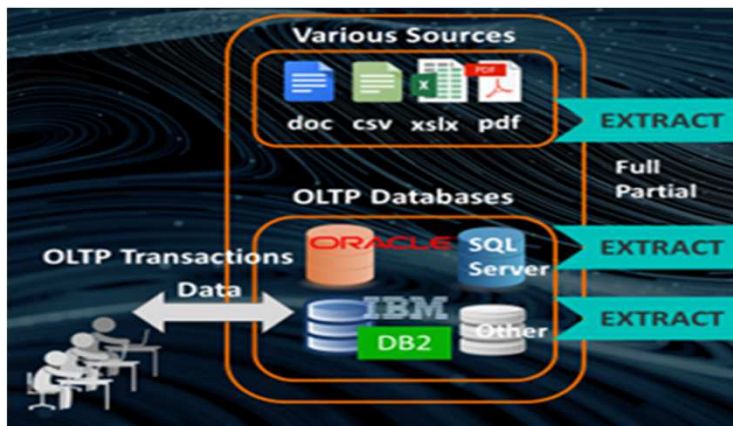


35



Alimentation du DW - Extract

- Quelles sont les données de production à sélectionner pour alimenter le DW ?
 - Toutes les données sources ne sont pas forcément utiles
- Rafraichissement du DW
 - Push: déclencheurs dans les sources de données
 - Pull: requêtage des sources
 - Contraintes: ne pas perturber les opérations OLTP



Sources de données variées

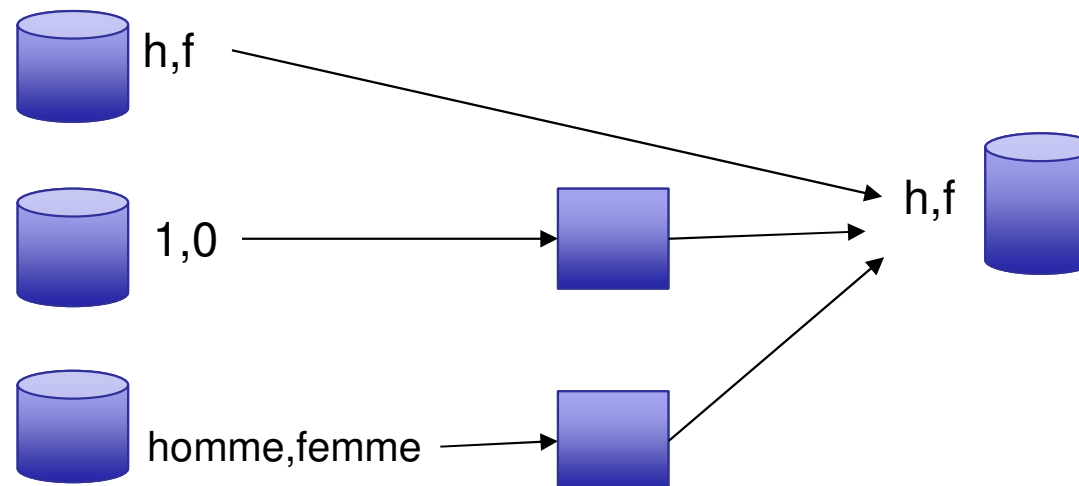


Homogénéiser ces données³⁶



Alimentation du DW - Transform

- Suite d'opérations ayant pour but de rendre les données cibles **homogènes** pour être traitées de façon **cohérente**
- Unification des données
 - Noms des attributs
 - Types (exemple: précision numérique)
 - Formats (exemple: dates)
 - Unités de mesures (conversion)



Alimentation du DW - Transform

- Nettoyage des données
 - Suppression des doublons
 - Traitement des valeurs manquantes (Ex: mapping des valeurs Null à 0 ou à la moyenne ou à la médiane de la colonne)
 - Détection des valeurs erronées ou incohérentes (Utilisation d'expressions régulières)



Alimentation du DW – Transform: Exemple

| Parsed data | |
|-------------|-------------------|
| First name | Aimee |
| Middle name | Christina |
| Last name | Parker |
| Job title | Prod. Mgr. |
| Firm | Microsoft |
| Street | One Microsoft Way |
| City | Redmond |
| State | WA |



| Corrected data | |
|----------------|-------------------|
| First name | Aimee |
| Middle name | Christina |
| Last name | Parker |
| Job title | Prod. Mgr. |
| Firm | Microsoft |
| Street | 15580 NE 31st St. |
| City | Redmond |
| State | WA |
| Postal Code | 98052 |
| Country | USA |



Alimentation du DW – Transform: Exemple

| Corrected data | |
|----------------|-------------------|
| First name | Aimee |
| Middle name | Christina |
| Last name | Parker |
| Job title | Prod. Mgr. |
| Firm | Microsoft |
| Street | 15580 NE 31st St. |
| City | Redmond |
| State | WA |
| Postal Code | 98052 |
| Country | USA |



| Standardized data | |
|-------------------|-----------------------|
| First name | Aimee |
| Middle name | Christina |
| Last name | Parker |
| Job title | Product Manager |
| Firm | Microsoft Corporation |
| Street | 15580 NE 31st Street |
| City | Redmond |
| State | WA |
| Postal Code | 98052 |
| Country | USA |



Alimentation du DW - Load

- Charger les données nettoyées et préparées dans le DW
- C'est une phase plutôt mécanique et la moins complexe
- C'est une opération assez longue: il faut définir des politiques de **chargement** et de **rafraichissement**
- Politiques de chargement/rafraichissement
 - Complet (*Full*) / incrémental (*Incremental*)



Alimentation du DW - Load

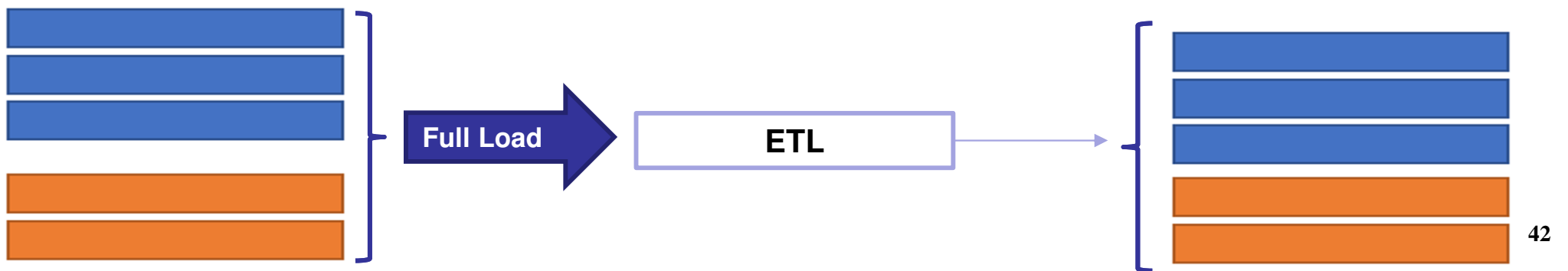
Initial Load



Incremental Load



Full Load















Outils d'Intégration de Données

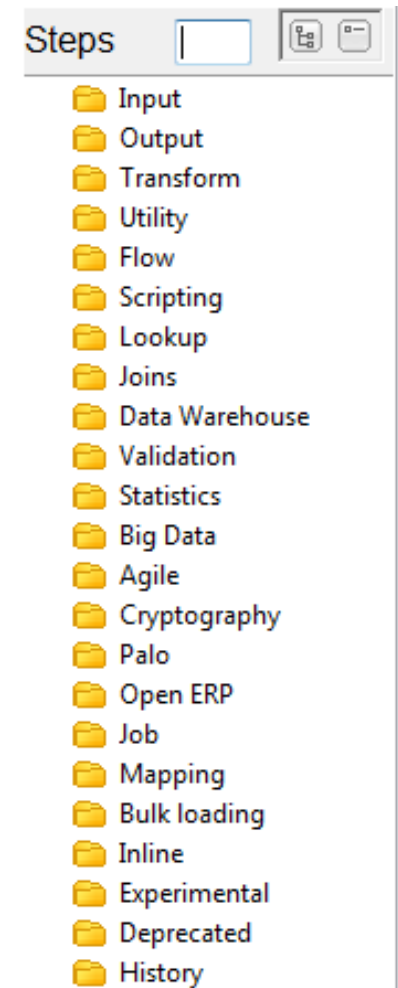
- Produits des Vendeurs Traditionnels
 - Vendeurs de BD: Oracle, IBM, Microsoft
 - Autres: SAP, Informatica, SAS, Information Builders
- Open source
 - Pentaho Data Integration
 - Talend Open Studio for Data Integration



Outils d'Intégration - Pentaho Data Integration

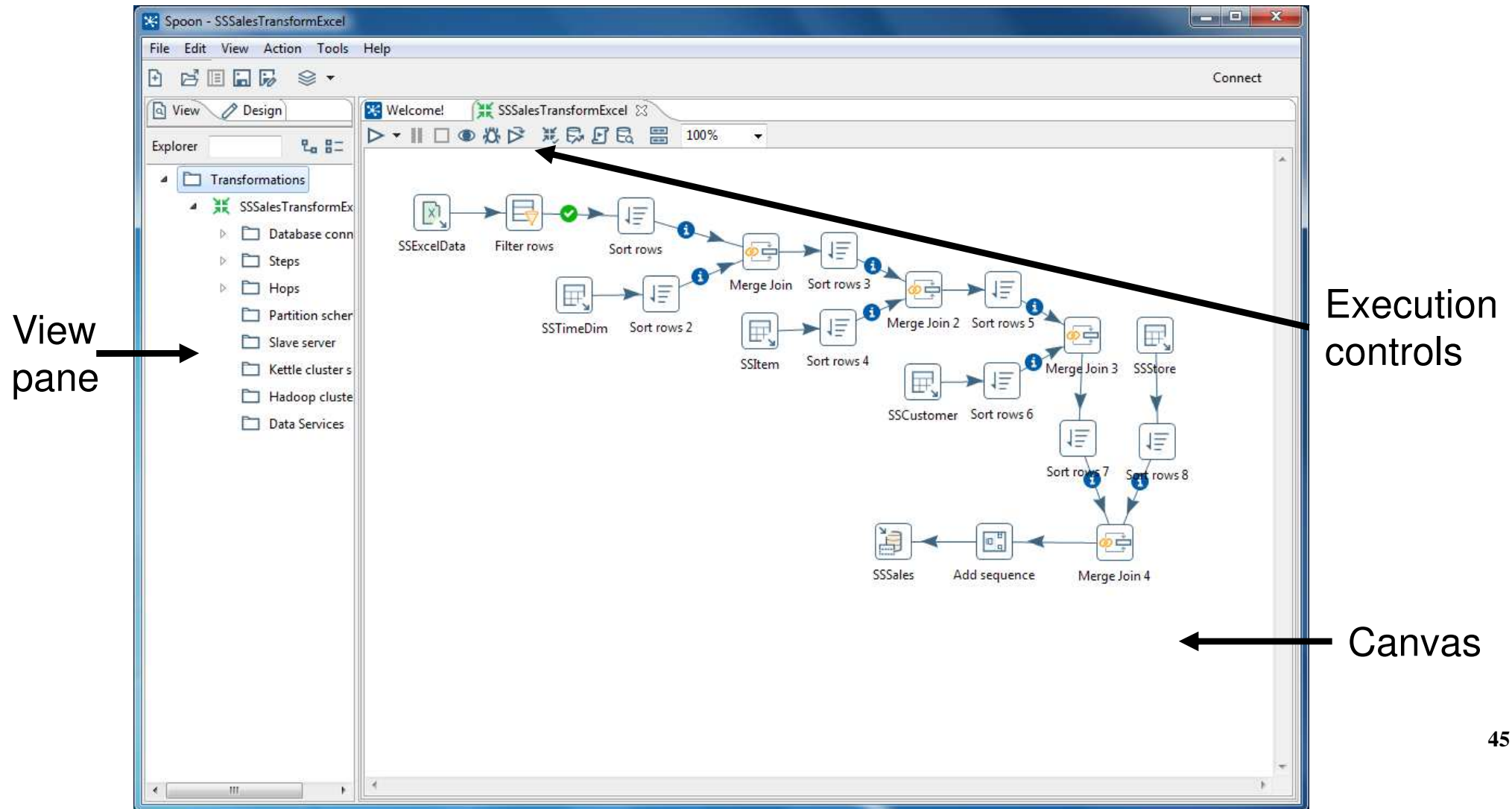
Transformations

- Step: process in a data flow
 - Input/Output  
 - Transform: sort, split, concatenate, ...  
 - Flow: filter rows  
 - Lookup: existence of rows, tables, files, ...  
 - Join: merge join, multiway merge, ...  
 - Validation: credit card, mail, data  
- Hop: directed connection between steps
- Database connections
- Distributed processing: partition, cluster, ...



Outils d'Intégration - Pentaho Data Integration

Spoon IDE

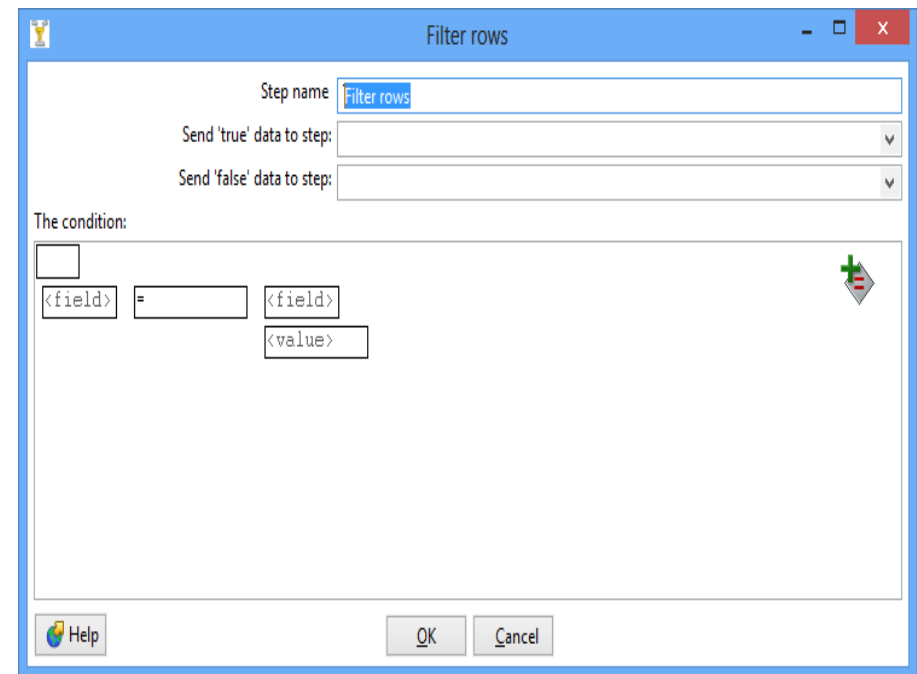
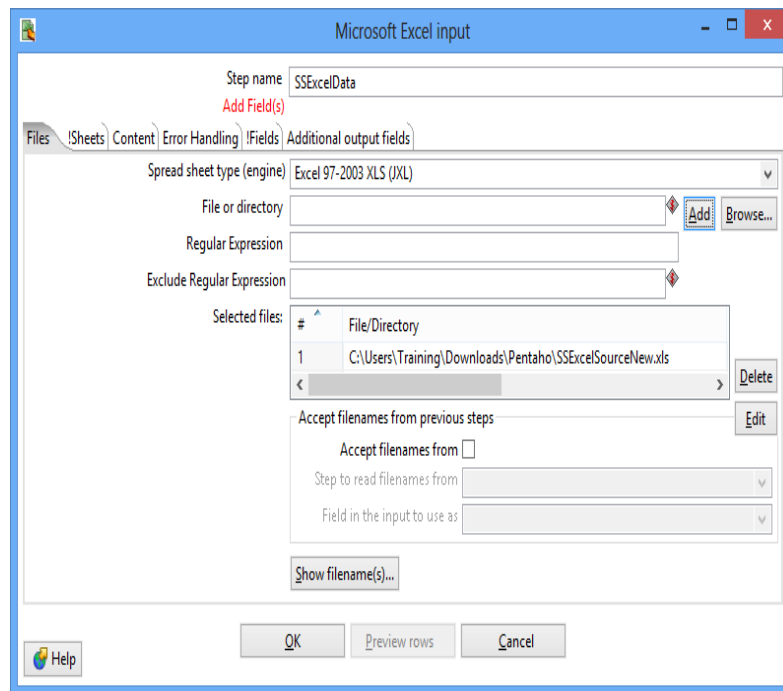


45



Outils d'Intégration - Pentaho Data Integration

Transformations Example

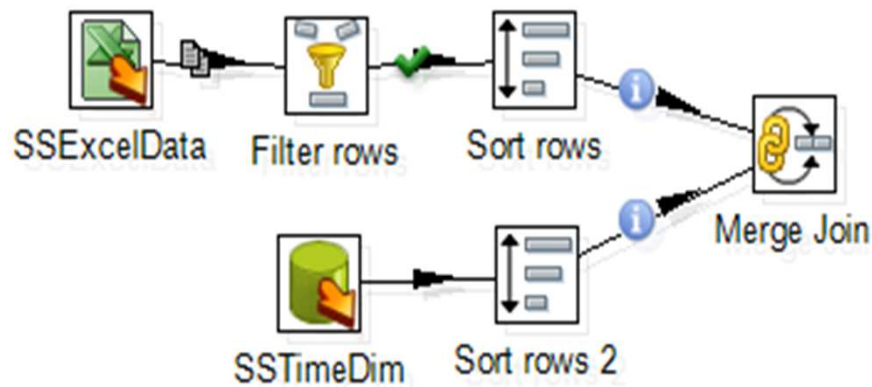


46



Outils d'Intégration - Pentaho Data Integration

Transformations Example



Merge Join

Step name: Merge Join

First Step: Sort rows

Second Step: Sort rows 2

Join Type: INNER

Keys for 1st step:

| # | Key field |
|---|-----------|
| 1 | Day |
| 2 | Month |
| 3 | Year |

Get key fields

Keys for 2nd step:

| # | Key field |
|---|------------|
| 1 | TIMEDAY |
| 2 | TIMEMON... |
| 3 | TIMEYEAR |

Get key fields

Help OK Cancel

