Lundi 29/03

Professeur : Hervé TONDEUR, herve.tondeur@yahoo.fr

La donnée... L'or noir du 21e siècle

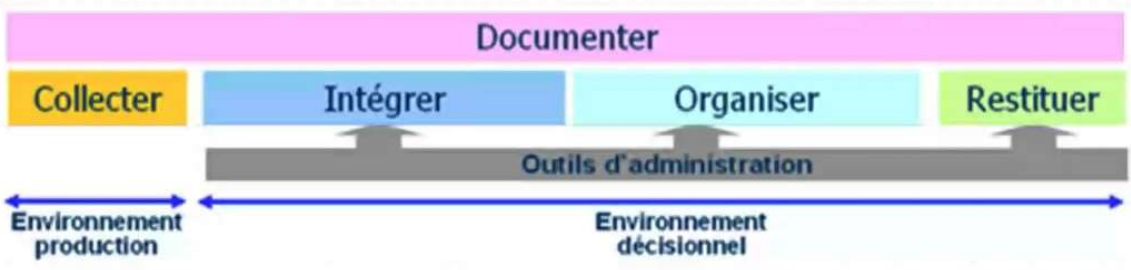
- Que veut réellement dire cette expression ?
- Comment une (E) peut réellement transformer ses données en richesse ?
- Quelles sont les données dont on parle ?
- Où se trouvent les gisements dont l'(E) peut prétendre explorer ?
- Quels sont les coûts pour extraire la valeur des données ?
- Comment, avec quels moyens et en combien de temps les (E) pourront profiter de cette richesse ?

Introduction

Nous allons essayer de comprendre :

- Ce qu'est une donnée et toutes les idées qui tournent autour d'une donnée
- Comment stocker, manipuler, interpréter ces données
- Comment construire des outils de stockage et restitution des données
- Comment comprendre le fonctionnement des différents types d'entrepôts
- Quelle est la ≠ entre les entrepôts de pilotage
- Les entrepôts décisionnels
- Les entrepôts d'études et recherches
- Comment peut-on exploiter ces entrepôts, les principes

Place d'un entrepôt de données dans un SI, notion d'urbanisation



SI : "Véhicule" des entités de l'organisation. Sa structure est constituée de l'ensemble des ressources (le personnel, le matériel, les logiciels, les procédures) **organisées pour : • Collecter / La collecte**

L'origine de l'info peut être interne (*comptes, stocks*) ou externe (*infos sur le concurrent, disposition nouvelle d'ordre fiscale ou sociale, météo, etc.*) Pour les infos d'origine externe, il est nécessaire d'y être **tout particulièrement attentif,** il convient donc d'**organiser des veilles technos**

- **Stocker ou Intégrer**

- **Traiter ou Organiser**

La phase de traitement commence avec le choix du support utilisé car il faut trouver un construction formalisée pour traiter l'info:

→ soit la centralisation (réalisée à un seul endroit dans l'E

→ soit la décentralisation (permet à chaque poste de travail d'échanger des infos et travailler en autonomie)

→ soit la distribution (permet un traitement au niveau d'un site unique ; la saisie et la diffusion s'effectuent grâce à des terminaux)

- **Communiquer ou Restituer** la donnée aux utilisateurs, aux managers et autres chefs de décision qui vont comprendre et appréhender l'info pour agir sur la collecte des données Elle doit répondre à 4 critères :

→ Origine et destination → Forme ?

→ Délai

→ Diffusion large ou restreinte ?

⇒ **il s'agit (les 4 étapes précédentes) de la documentation du SI** → La qualité coûte cher

mais apporte ..

Le SI coordonne les activités de l'organisation et lui permet ainsi, d'atteindre ses objectifs, grâce **à la structuration des échanges**

Un SI se construit à partir de l'analyse des processus métier de l'organisation et de leurs interractions/interrelations, et non seulement autour de solutions informatiques plus ou moins standardisées par le marché

⇒ Le SI est un **système sociotechnique** composé de 2 sous-systèmes : l'un social (consitué de personnes et de relations entre ces personnes) et l'autre technique Hardware et Software)

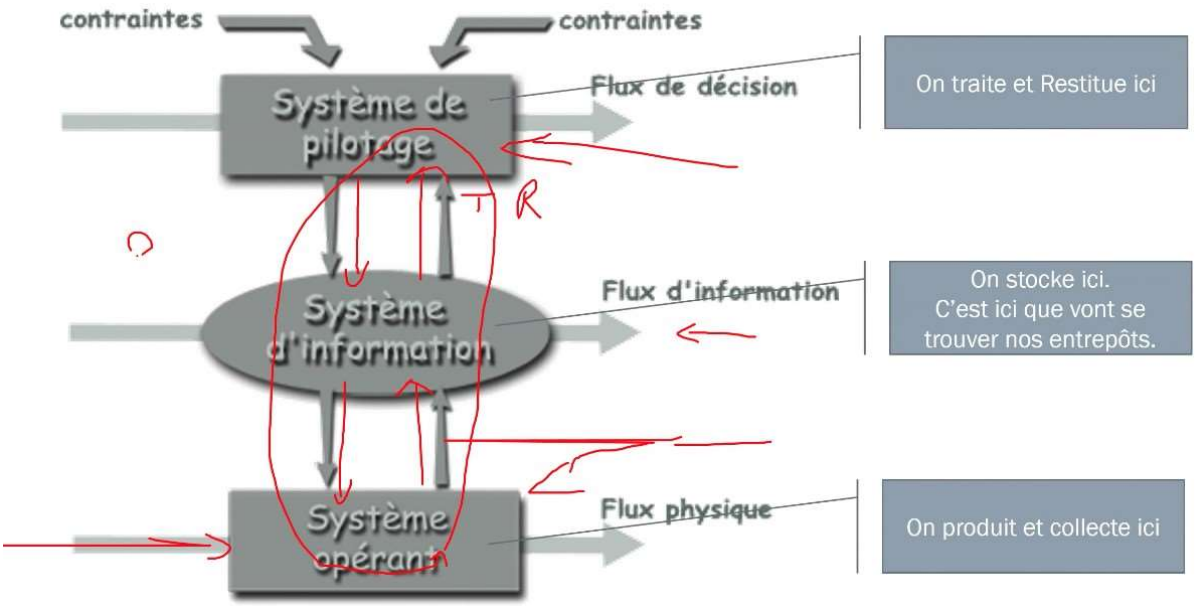
2 grandes catégories de systèmes selon les types d'application informatique :

- **les systèmes de conception** : fonctionnent selon des techniques temps réel ; ici il faut restituer et traiter au plus vite.
- **les SI de gestion** : qui emploient des techniques de gestions (majoritaire)

Ce qui ≠ ces 2 catégories c'est **la notion du temps**

Organisation : On peut tj décomposer une organisation en 3 sous-systèmes : le système opérant : on

- **produit** et **collecte** l'info
- le système d'information : on **stocke** l'info ici ; l**es entrepôts se trouvent ici** système de pilotage : on
- **traite** et **restitue** l'info



Vision des entrepôts de données selon le type d'utilisateur

Vision des entrepôts de données selon la perspective utilisateurs / DSI / Managériale /technicien informatique

Du point de vue utilisateurs, un entrepôt est un système qui permet d'avoir a disposition pour son quotidien d'informations pour des prises de décisions métiers rapides, de lui permettre de compléter son environnement informatif dans sont domaine opérationnel.

Du point de vue du Directeur du système d'information, un entrepôt est un système qui va permettre de rationaliser son SI et de permettra de pérenniser les informations et consolider le SI.

Du point de vue Managériale, les entrepôts de données sont des outils de pilotages, qui permettent de piloter l'entreprise, d'améliorer la qualité, d'ouvrir des marchés de donner un instantané de l'état de l'entreprise (du point de vue financier (dépenses/recettes, marketing, qualité, ressources humaines).

Du point de vue du technicien est un système informatique composé de SGBD, d'outils d'ingestions de données, d'outils de restitutions des données, de traitement et affichage des données pour les utilisateurs et manageurs.

Difficile de discuter d'entrepôt de données avec ≠ personnes car ≠ points de vue

Quelques définitions

Entrepôt : toute solution technique permettant de conserver des données de toutes provenance selon certaines règles :

les données doivent être :

→ ordonnées : l'entrepôt est organisé est structuré

→ intégrées : les données proviennent de sources hétérogènes utilisant chacune un type de format, elles sont intégrées avant utilisation

→ orientées sujet : dans un entrepôt, les données sont organisées par thème

→ historisées : les données non volatiles sont horodatées → non volatiles :

Données : toute info simple ou complexe qui permet de qualifier un info du monde réel ou imaginaire. Elle doit absolument représenter une information, attribut d'un objet ou d'une entité

Sources de données :

Les sources de données ont de multiples domaines de productions. Tous les domaines métiers produisent de la donnée de tout type par la saisie d'opérateurs

- par la capture de mesures
- par des captures
- par le travail et mesure de robots de productions
- par le calcul agrégé de données

Les sources de données internes :

- les données internes sont l'ensemble des elt d'information produites et stockées au sein d'une organisation

Les sources de données externes :

- les données externes sont toutes les données non générées par l'organisation et accessible via des sources extérieures

Types de données

Données non structurées : Un peu de tout, données à la fois exploitable par l'être humain mais difficilement exploitable par la machine

Données semi-structurées : On ajoute aux données non structurées des métadonnées

Métadonnées : ce sont des données qui décrivent des données

Données structurées : ce sont les données qui permettent la mise en place des entrepôts de données

Typage des données et représentation des données métiers et métadonnées

Qualité des données :

- Homogénéité
- Contexte d'acquisition
- Actualisation / Fraicheur des données
- Cohérence
- Inter opérabilité (plus pour le futur)
- Unicité / Gestion des doublons
- Validité scientifique (exemple : température d'un humain comprise entre 28 et 39 degrés)
- Complétude
- Précision
- Traçabilité

Modèle de données

Modèle qui décrit comment sont représentées les données dans une organisation métier, un système d'information ou une base de données (exemple UML).

Il est composé généralement de :

- Structure de données
- Intégrité de données
- Manipulation des données
- Recherche des données

Interopérabilité = cruciale aujourd'hui

Lundi 12/04 anis

2 manières d'appréhender la construction de datawarehouse

- **Ralph Kimball (bottom-up) :** Chaque datamart s'occupe des secteurs d'activités qui lui est dédié (comptabilité, vente, marketing…) Le datawarehouse est ainsi la combinaison des différents data marts qui facilite le reporting et l'analyse.

- **Bill Inmon (top-down) :** Le datawarehouse est un dépôt centralisé de toutes les données de l'entreprise (approche massive, on met tout dans un conteneur). Cela implique qu'il faut normaliser, organiser et modéliser cet entrepôt.

Définitions :

- **ETL = extraire les data**
-

Architecture 3 tiers :

Un entrepôt de donnée est divisé en 3 parties / tiers :

- **L'accès aux données** : contient le serveur de base de données utilisé pour extraire les données de nombreuses sources différentes
- **Le traitement** : contient différents serveurs et notamment un serveur OLAP, qui transforme les données en une structure mieux adaptée à l'analyse et aux requêtes complexes. Le serveur OLAP peut fonctionner de 2 manières : soit comme un système étendu de gestion de base de données soit
- **La présentation** : correspond à la couche client. Ce tiers contient les outils utilisés pour l'analyse de données de haut niveau, le reporting et le data mining (exploration des données)

Entrepôt de données

⇒ base de données regroupant une partie ou l'ensemble des données fonctionnelles d'une entreprise.

But : fournir un ensemble de données servant de référence unique, utilisée pour la prise de décision dans l'entreprise par le biais de statistiques et rapports via des outils de reporting. **Techniquement parlant**, il sert surtout à "**délester**" les bases de données opérationnelles des requêtes pouvant nuire à leur performance.

Infocentre : ancêtre des datawarehouses

Nouvelles solutions : Data Lake (lacs de données), ils permettent de stocker des données structurés et non structurés sans pré-traitements (notamment utile pour la BI et le machine Learning).

Base de données relationnelle / Modèle relationnel

Outil par défaut des entrepôts de données.

- Nombreux avantages (exemples propriétés ACID) (voir slide p 174)
- Désavantage : stockage vertical (Quand j'ai une base de données sur serveur, cette base de données va grossir = limite en taille, limite en temps de mémoires)
- Base de données No-SQL = stockage horizontal (distribué sur plusieurs machines)

Succès du modèle relationnel = slide 176

Formes normales

Règles de normalisation (normaliser une relation) : hiérarchique (d'abord niveau 1 ensuite niveau 2 etc. etc.) = p. 179

Liste formes normales : p.130

Infocentre

Avantages et désavantages p.146

Datawerouse

Facteurs à considérer lors de leurs constructions :

- L'interdépendance informationnelle entre les unités de l'entreprise (la bonne intégration)
- Les sources de données (1 vs plusieurs sources)
- La quantité de données
- La latence des données (mise à jour temps réelle ? ou quotidienne)
- L'urgence d'obtenir une solution fonctionnelle (il me faut cet entrepôt fonctionnel rapidement (datamarts) ou moins rapidement (EDW))
- Nombre d'utilisateurs
- Nature des tâches des utilisateurs finaux (simples rapports VS fouilles de données)
- Contraintes de ressources (financières, technologique, main d'œuvre)

- Objectif du projet (stratégique VS opérationnel)

Datamart vs Dataware house (slide 159)

Un datamart (magasin de données) se concentre sur 1 sujet d'analyse (ex les ventes OU les livraisons mais pas les deux)

Sert à faire des analyse simples et spécialisées comme la fluctuation de ventes par catégories de produits.

Nombre de sources limitées.

Processus d'extractions des données relativement simple

Même processus de conception que les entrepôts de données mais demande moins de ressources.

Magasins de données Datamarts vs EDW

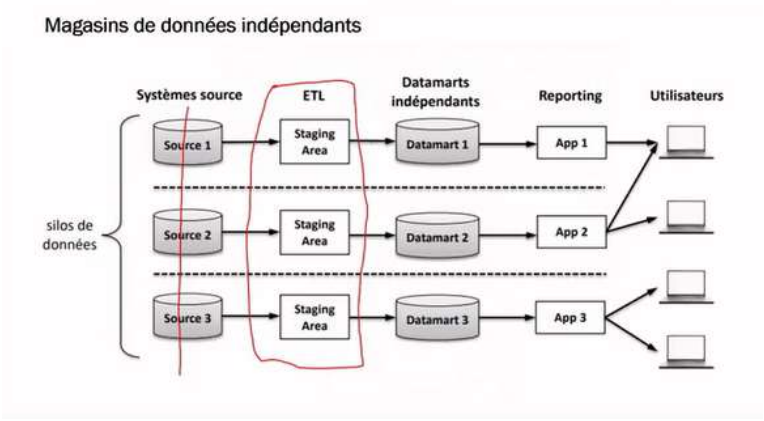
Caractéristique	Magasin de données	Entrepôt de données (EDW)
Portée	Un domaine d'analyse	Plusieurs domaines d'analyse
Temps de développement	Mois	Années
Coûts de développement	$ $	$ $ $ $
Complexité de développement	Faible à moyenne	Grande
Taille des données	Mb à plusieurs Gb	Gb jusqu'à plusieurs Pb
Horizon des données	Courantes et historiques	La plupart du temps historiques
Transformation des données	Faible à moyenne	Importante
Fréquence des mises-à-jour	Horaire, journalier ou hebdomadaire	Peut aller jusqu'à mensuel
Nombre d'utilisateurs simultanés	Dizaines	Centaines à milliers
Types d'utilisateur	Analystes dans le domaine spécifique et gestionnaires	Analyste d'entreprise et cadres seniors
Objectifs	Optimisation des activités dans le domaine spécifique	Optimisation inter-fonctionnelle et support à la décision

Architectures d'entrepôts de données

En voici 5 très classiques mais on peut en avoir d'autres :

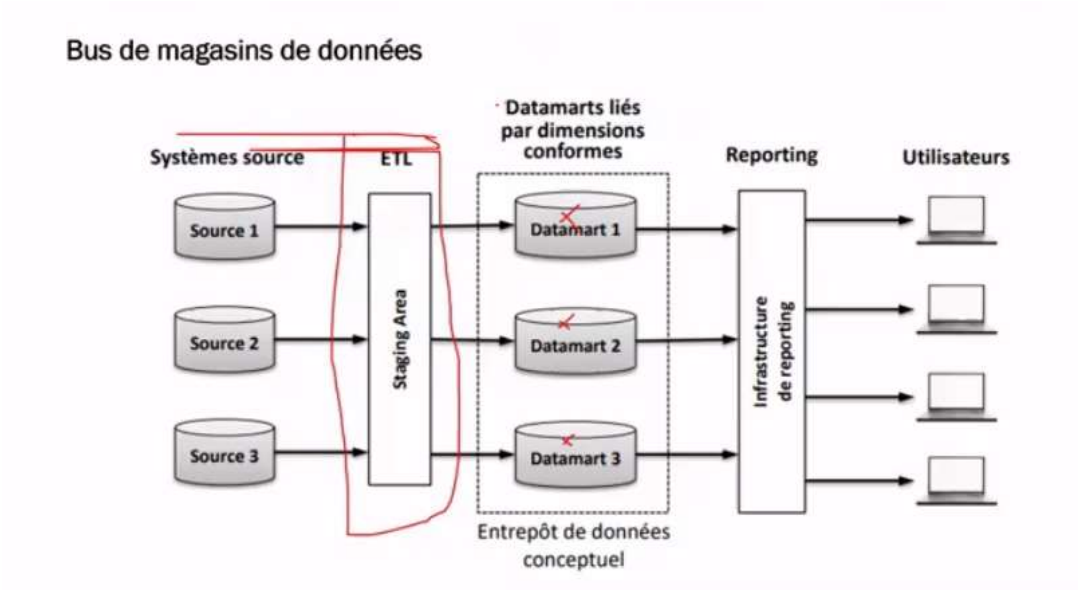
Magasins de données indépendants

Slide 163



Pour chacune des sources, on crée un datamart indépendant et créer des outils de reporting pour chacun de ces datamarts.

Architecture en bus de magasins de données

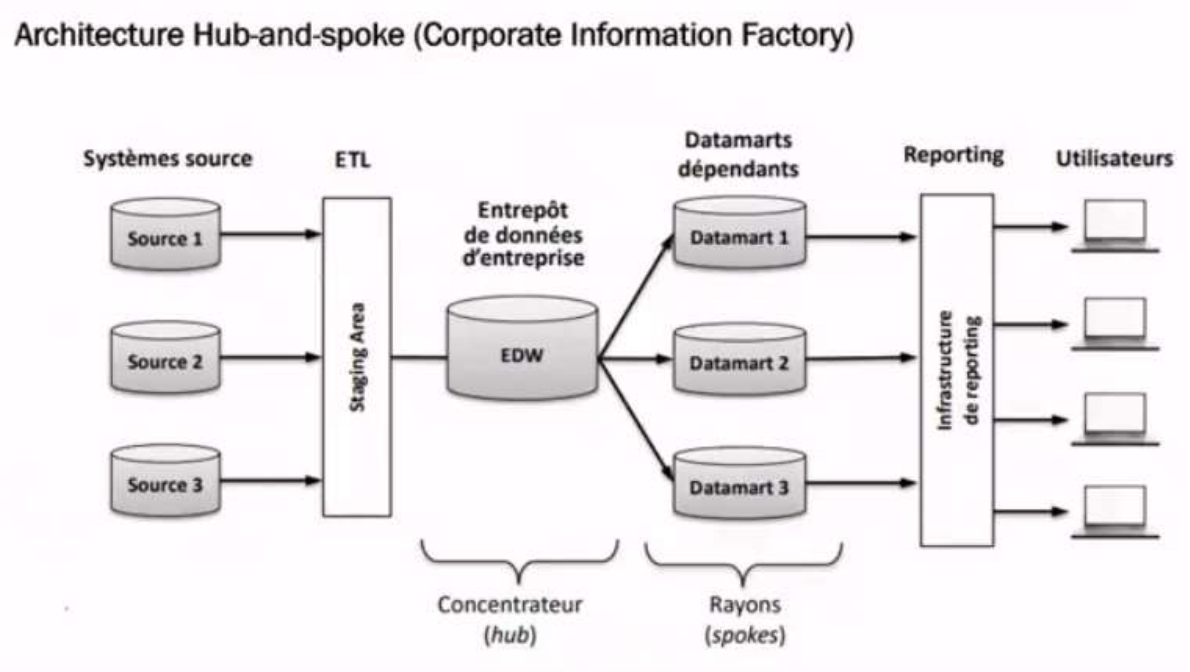


Chacune des sources est relié à un même ETL qui va s'occuper de l'extraction des données depuis ces même sources. Chaque datamart est indépendant mais peut être lié à l'aide de clés primaires (exemple : dans un hôpital, la dimension conforme sera le patient qui relie tous les data marts en entre eux. Il peut y avoir plusieurs dimensions conformes).

Nous avons un seul outil de reporting à parti de notre bus de magasins de données.

Avantages et inconvénients (slide 166)

Architecture Hub-and-Spoke



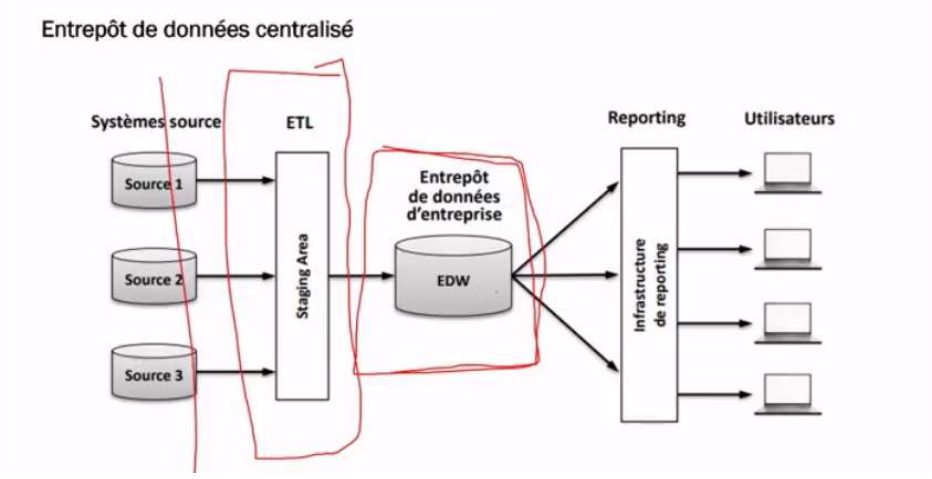
L'une des architectures les plus populaires utilisée par les gros systèmes (S-Data Hub).

Chacune des sources est relié à un même ETL qui va s'occuper de l'extraction des données depuis ces même sources. Ces données seront stockées dans un EDW (Entrepôt de données d'entreprise / un gros entrepôt) qu'on appelle un concentrateur ou un hub de données.

Ces datas vont venir alimenter des datamarts de manière automatique ou à la demande qui vont permettent une analyse fine sur ces données grâce à un outil de reporting unique. Hub and Spoke (collecte et mise en rayon)

Avantages et inconvénients p 168 = très bonne solution très utilisée

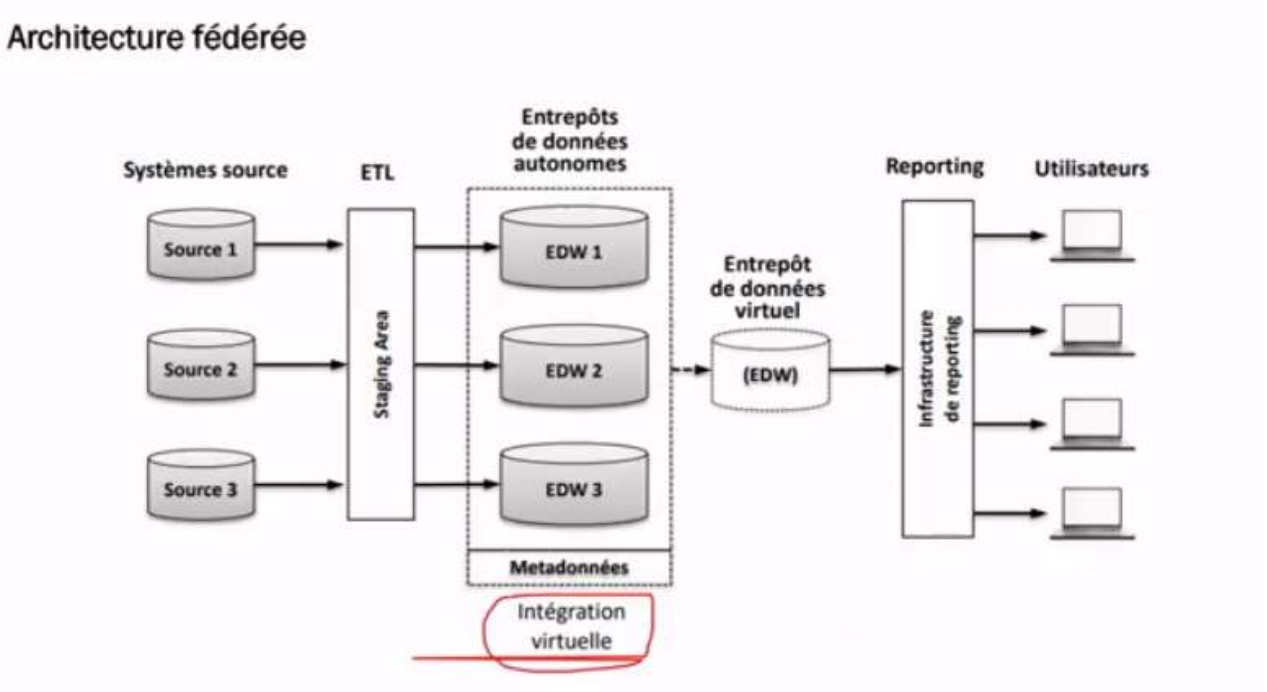
Entrepôt de données centralisé



Même chose que précédemment sauf qu'il n'y a pas de data marts.

Avantages et inconvénients p 170 = Bonne solution mais couteuse en termes d'évolution

Architecture fédérée

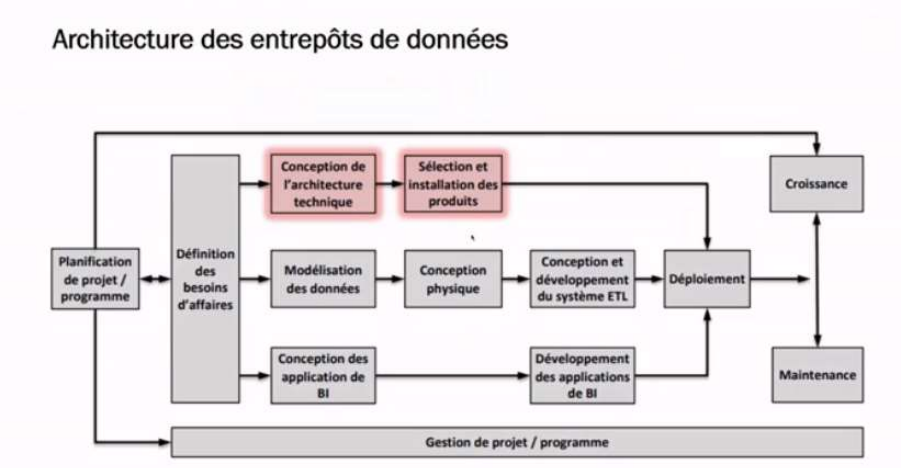


L'architecture de pros (quand on a plusieurs sources de données etc. etc.)

Cours 13 Avril

Différence entre le modèle en étoile et flocon de neige pour la conception des entrepôts de données

Bonnes pratiques pour mettre en place un entrepôt de données



Pour concevoir un data Warehouse, on utilise trois modèles :

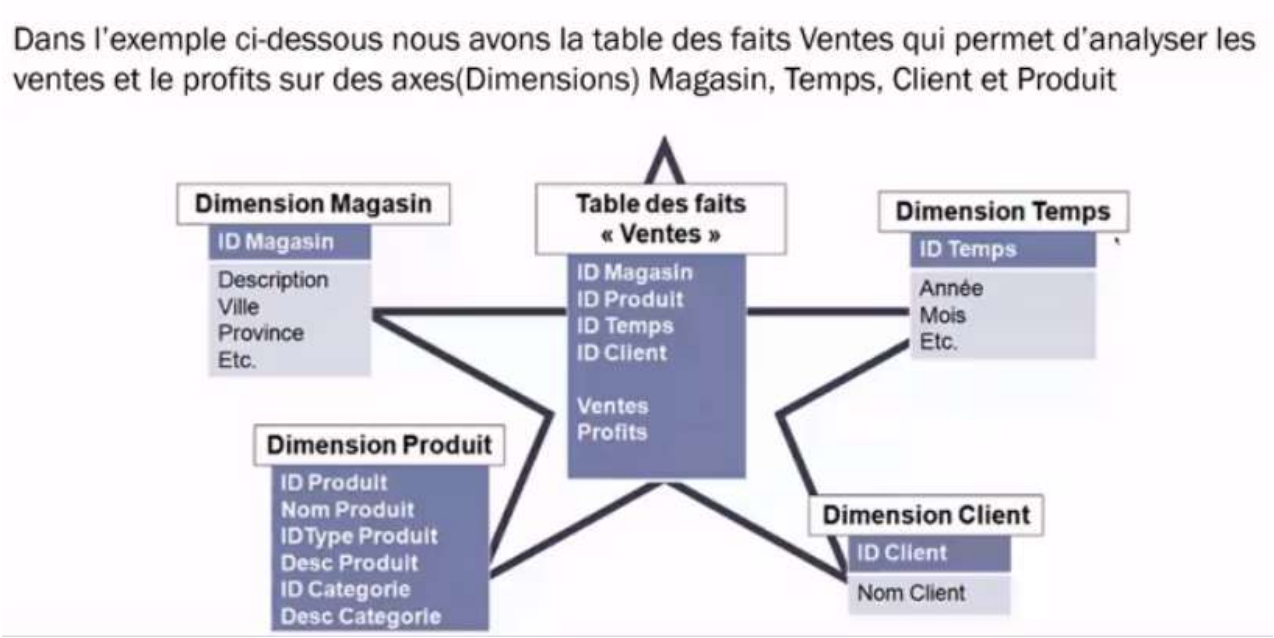
- Le modèle en étoile

- Modèle flocon de neige
- Le modèle galaxie ou constellation

Dimensions = axes sur lesquels on veut faire l'analyse
Faits = sur quoi je porte l'analyse / ce que nous voulons analyser (ex : table des faits pour la vente qui permet d'évaluer le chiffre d'affaire net)

Modèle en étoile slide 177

Modèle en étoile : table de fait centrale qui est liée par les tables de dimensions dénormalisées. Les dimensions ne sont jamais liées entre elles.



Exemple questions que ce schéma permet de réponde :

- Quelles sont les ventes par magasins ? Par ville ?
- Quelles sont les ventes par produits ? Quelles sont les produits les plus vendus ?
- Nombre de ventes par mois ? Par année ?
- Les ventes par type de client ?

Idéal pour les requêtes en jointure.

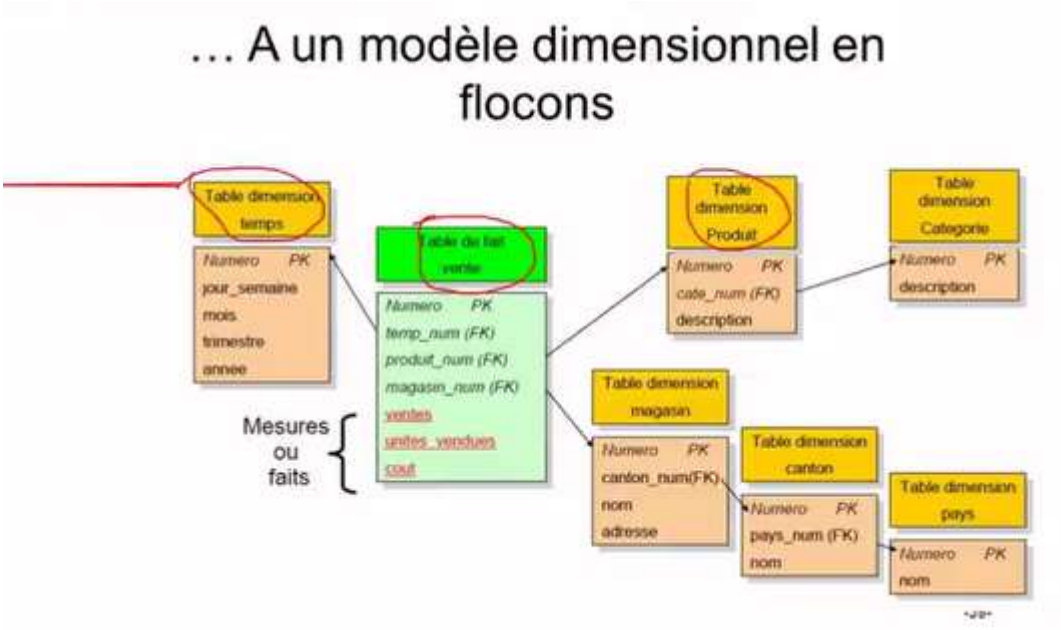
Indexation = enjeu majeur pr améliorer vitesse des requêtes.

Modèle en flocon (p 181)

Modèle intéressant si gros entrepôt de données.

Un modèle en flocon = modèle en étoile vu précédemment avec la normalisation des dimensions.

- Chaque table de dimension est renormalisée pour faire apparaitre une hiérarchie sous-jacente.
- Intérêt de ce modèle : gain d'espace de stockage de l'ordre de 5 à 10%
- Pas d'amélioration de la qualité des analyses



Ici on a subdivisé la table produit et normaliser pour extraire la table catégorie.
On a bien à faire un modèle en étoile étendu.

Modèle en constellation ou en galaxie

C'est un ensemble d'étoiles ou de flocons dans lequel les tables de faits se partagent certaines tables de dimensions forme un modèle « en constellation » ou dit en « galaxie ».

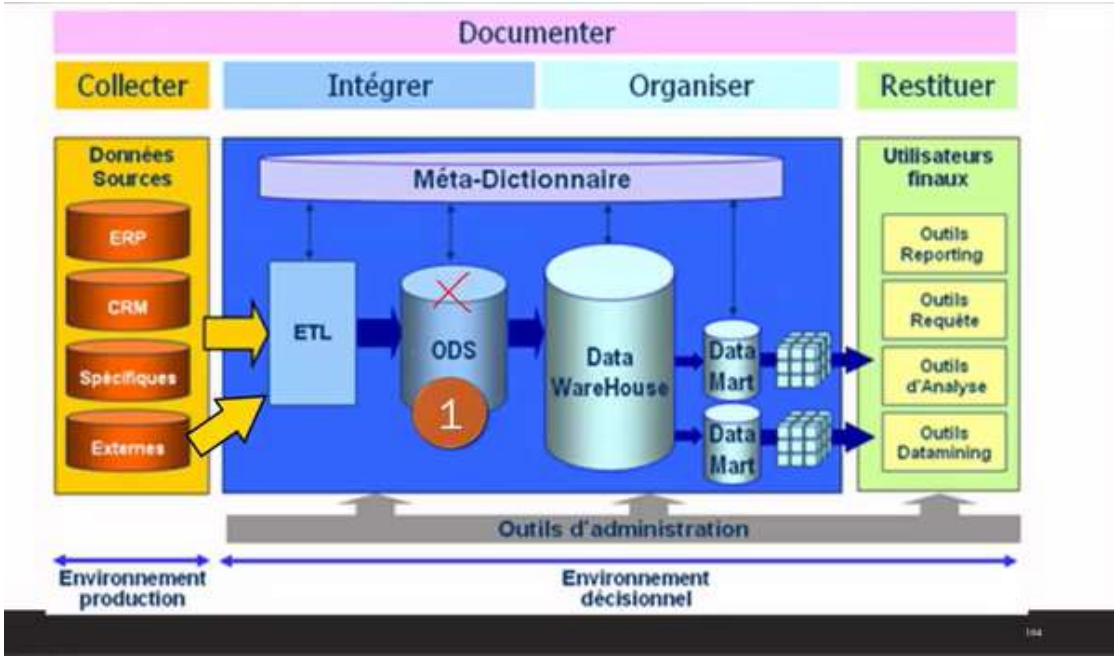
Les ODS / Comptoirs de données opérationnel

Je construis un entrepôt de donnée consultable à n'importe quelle moment (phase opérationnel et non analytique)

Il permet de pouvoir travailler en temps réel :

- Similaire à un datamart ou datawarehouse
- Intègre et consolide des données de sources hétérogènes
- Peuvent servir de sources à des entrepôts de données ou à des systèmes opérationnels
- Contiennent rarement des données historiques / Pas d'historisation
- Mettent à jour les données au lieu de les ajouter

- Effectue les changements en temps réel plutôt que de les faire en lot
- NE REMPLACE PAS LES ENTREPOTS DE DONNEES



Types d'hébergement des datawarehouse ou entrepôt de données

On premise (sur place)

Etape collecter est généralement toujours On Premise

Pour la partie Intégrer et organiser on peut les mettre on premise mais mettre un entrepôt de donnée chez soi demande beaucoup de compétences et de couts (serveurs) et nécessite de l'expertise et une maitrise de l'infrastructure, une maitrise des notions BDD, une maitrise du développement et une maitrise des données sources.

Entrepôts de données hébergés dans le cloud (slide 188)

Cela peut être :

- Des pures infrastructures mise à disposition
- Des plateformes mise à disposition
- Mise à disposition de services complets

Moins rentable à long terme

Plateformes d'exemple (PaaS et SaaS) : AWS, IBM et Microsoft (Azure avec PowerBI pour le retraitement et la mise à disposition de la donnée)

Méta Dictionnaire (p. 192)

Important d'avoir de la qualité pour les informations qui circulent au sein du SI.
Ce méta dictionnaire doit garantir la qualité des données.

Quelles données ? Toutes les données ne méritent pas d'être contrôlées. On doit privilégier les données importantes sur lesquels ont ne peut pas se tromper (données de référence)

MDM : Master Data Management

Le concept de MDM regroupe tous les moyens et l'ensemble des méthodes pour construire un référentiel de qualité.

Cela comprend :

- Le nettoyage des données
- La mise en cohérence
- La consolidation
- La mise à jour
- L'élimination des doublons
- L'établissement des descriptifs des données de référence de l'entreprise

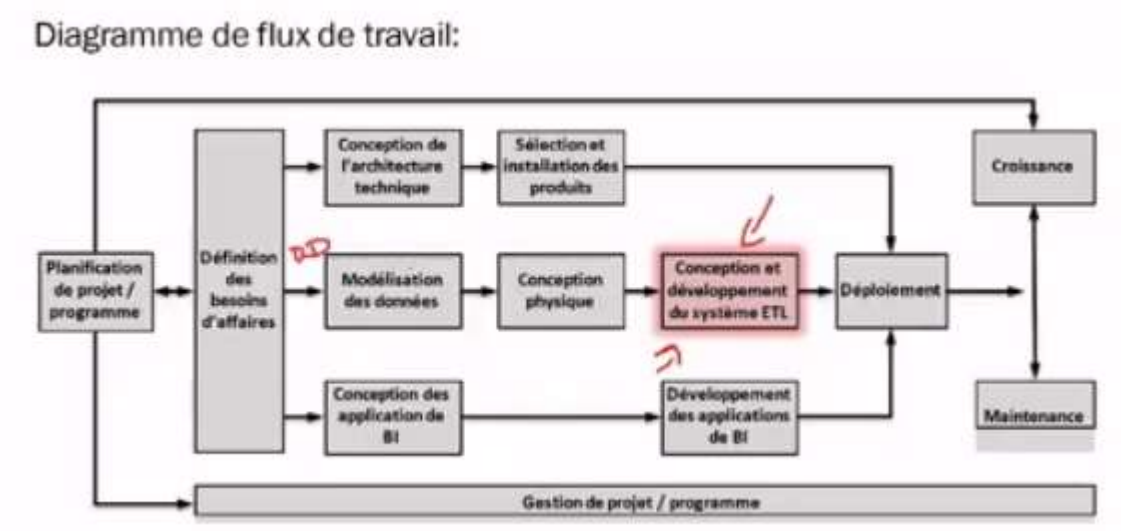
Tous les ETL n'ont pas de MDM et sont difficile à gérer notamment niveau temps/

ETL : Service ingestion ou Service Back-Room

Il s'agit de la couche de préparation des données. Il permet l'extraction des données sources et le chargement vers les entrepôts (ODS ou Datawarehouse ou un mix de tout ça).

Il possède des services de management (ETL Management Services) et des Data Stores (ETL Data Stores).

Fonctionnement ETL (slide 201) :



Une fois que notre modèle de données / dictionnaire des données a été défini, on va pouvoir concevoir et développer notre système d'extraction.

- Extraction : Un ETL se connecte à des système sources (ERP / CRM / BDD Opérationnel / ODS / Externe etc. etc.) et extrait les données.
- Consolidation : Il va ensuite contrôler ces données à l'aide du dictionnaire de données.
- Livraison : ce qu'on va insérer ces données au sein de notre entrepôt.

Approches d'intégration (p. 207) :

- ETL : 90-95% du temps
- EII : très rarement utilisé (fédère toutes les données de l'entreprise de plusieurs sources et fournit un accès en temps réel aux données = je consomme et je consolide)
- EAI : beaucoup utilisé domaine santé et bancaire (échange de message via bus commun / et route grâce à l'IA) (bien pour alimenter un ODS)

Comparaison entre les approches d'intégration (p 217)

Principales étapes de développement ETL :

p. 221

Staging area = zone préparation de données