Antisèche Business Intelligence (Informatique Décisionnelle)

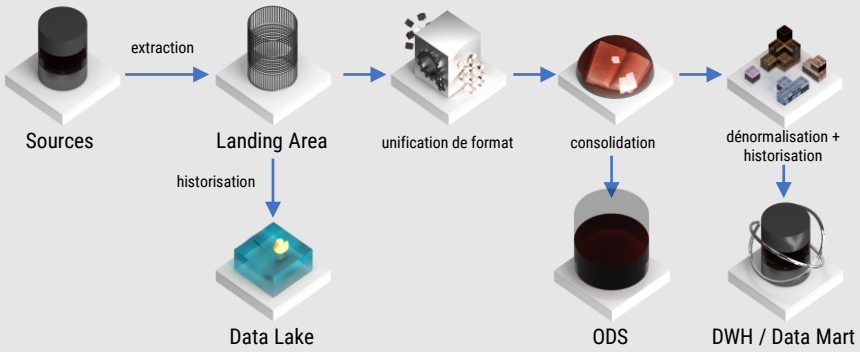
définitions & architecture

BI – approches et technologies d'analyse descriptive, prédictive et prescriptive de données cross-service et multi-système.

DWH, Data Warehouse / Entrepôt de Données – source de données consolidées et préparées pour l'analyse. Souvent DWH est historisé (i.e. préserve l'historique d'étapes de vie des objets métier).

Consolidation – recherche et élimination de multiples représentations des objets métier (doublons).
Consolidation se divise en **rapprochement** (identification de doublons, record linkage, duplicate detection) et **dédoublonnage** (construction de représentation unique, deduplication).

Le chargement de DWH est souvent fait avec les outils **ETL** ou **ELT**, dont les plus connus sont IBM DataStage, Talend, Informatica PowerCenter, Informatica Data Quality, Microsoft SSIS, Oracle ODI, etc. Les outils **CDC (Change Data Capture)** permettent d'optimiser l'extraction de données.



Les données durant le chargement passent par les étapes suivantes :

- unification de format,
- dénormalisation,
- consolidation,
- historisation.

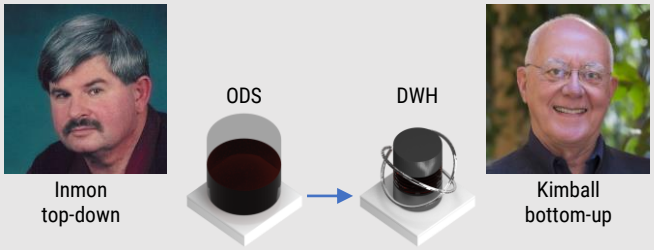
Les données telles qu'elles peuvent être sauvegardées dans **Landing Area. Landing Area** avec des données temporaires, tables de transcodification et des rejets fais partie de **Staging Area**.

Landing Area historisée et stockée dans Hadoop s'appelle **Data Lake.**

Les données cross-systèmes dans un format unifié s'appellent **pivot**, soit **data in canonical data model.** Souvent avec l'unification de format on exécute également la standardisation et la validation de données.

Les données après la consolidation, si elles sont sauvegardées dans une base de données, s'appellent **ODS (Operational Data Store)**, soit **CIF.** *Attention à l'utilisation incorrecte du terme en France et certains autres pays – parfois ce terme est utilisé dans le sens de Landing Area.*

Les données après la dénormalisation, préparation pour l'analyse et historisation s'appellent **DWH**. Si les données concernent qu'une seule domaine de business elles s'appellent **Data Mart.**



Si la construction de DWH se commence par la construction d'ODS c'est l'approche d'Inmon **(top-down)**, sinon si l'approche est plus « agile » et se commence directement par la construction de DWH par essaie et erreur (sans ODS) c'est l'approche de Kimball **(bottom-up)**.

Dans les projets, l'architecture mixte est parfois appliquée : top-down pour le référentiel et bottom-up pour le transactionnel.

Antisèche Business Intelligence (Informatique Décisionnelle)

qualité

Deux approches de base au **rapprochement** de données :

- Rapprochement **par clé fonctionnelle (business key)** – identification de doublons en utilisant une clé unique (code contrat, code TVA, etc).
- Rapprochement plus avancé – à la base de modèle statistique de **Newcombe, Fellegi et Sunter.** Cette approche permet d'ignorer des fautes de saisie et de traiter la majorité de données du référentiel.

Tactiques de **dédoublonnage** : valeurs plus récents, plus fréquents, validées, plus longues ou obtenues depuis SI prioritaire.

Unification de format et rapprochement sont liés avec la **standardisation** et la **validation de données.**

Standardisation – la transformation de données qui permet de diminuer la différence entre différentes représentations des mêmes données. Le but de la standardisation c'est d'améliorer la qualité de rapprochement et d'extraire l'information cachée dans les lignes de caractères.

Exemples de standardisation :

- Suppression d'accents / mise en majuscule (Mikaël -> MIKAEL).
- Parsing d'adresse (extraction de ville/code postal/nom de la rue/numéro de bâtiment, etc).
- Extraction de forme juridique de société.
- Unification de format des dates (03 avr 14 -> 03.04.2014).
- Transcodification de la nomenclature (FRA, FR, France -> FR).
- Extraction de code contrat depuis la description.

Validation – vérification des règles de gestion imposées sur les données. Par exemple : somme de contrôle (checksum) pour les codes TVA/EAN/NIR; ordre entre les dates de naissance de mère et de sa fille; taille de container et sa charge; règles de complétude d'adresse; intégrité des clés étrangères, etc.

Gestion d'erreurs : soit via **rejet** (suppression), soit via **auto-correction**, soit via utilisation de **statut/flag.**

modélisation

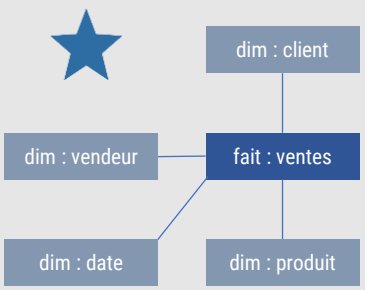
Dénormalisation – simplification de format via l'augmentation de la redondance dans les données (ex. jointure de plusieurs tables).

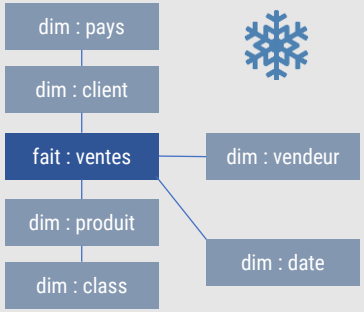
Normalisation – diminution de redondance dans un modèle de données.

La forme normale la plus connue, utilisée souvent dans les SI et dans l'ODS est **la 3éme forme normale (3NF)** : attribues de chaque objet (table) dépendent de la clé, la totalité de la clé et rien d'autre sauf la clé.

DWH souvent ne respecte pas la 3NF pour des raisons de simplicité de modèle et de la performance.

DWH est souvent **modélisé** de manière **dimensionnelle**, i.e. en forme **d'étoiles** ou des **flocons**. Ces modèles contiennent une table des transactions **(faits)** qui référencie des tables de référence **(dimensions).** Différence entre étoile et flocon est dans le nombre de niveaux de tables de dimensions :





Dimension **conforme** est une dimension utilisée par plusieurs tables de faits.

Modèle dimensionnel convient idéalement pour l'analyse de données financières, mais elle *peut ne pas convenir* dans les autres cas.

Pour la meilleure performance, DWH doit être implémenté avec une **base de données analytique** (ex. Vertica, SybaseIQ, DB2 BLU, Exasol, Greenplum, etc).

Antisèche Business Intelligence (Informatique Décisionnelle)

dimensions & faits

Tables de faits contiennent :

- **mesures**, i.e. les valeurs qui peuvent être utilisées comme les indicateurs de la valeur de la transaction,
- liens vers les tables de dimensions.

Les mesures peuvent être :

- **Additives,** i.e. telles que nous pouvons agréger sur tous les jeux de dimensions (ex. montant en Euro).
- **Semi-additives,** i.e. telles que nous pouvons agréger sur certaines dimensions, mais pas les autres (ex. quantité de produit vendu; montants en différentes devises).
- **Non-additives,** i.e. telles que nous ne pouvons pas agréger (différents ratios; pourcentage de marge; pourcentage d'impôt).

Dans le DWH classique les données dans les dimensions doivent être historisées. Il existe plusieurs types d'historisation (aussi connus comme Slowly Changing Dimension Types) :

- **Type 0.** Aucune historisation, aucun mise à jour de données.
- **Type 1.** Aucune historisation, les données sont mises à jour pour garder la dernière version.
- **Type 2.** Chaque mise à jour génère une nouvelle ligne horodatée.
- **Type 3.** Chaque mise à jour préserve la nouvelle donnée dans le nouveau champ (limité par nombre de champs prévus).
- **Type 4.** Variation sur le type 1 et type 2 : créer deux tables : la première préserve l'état actualisé et la deuxième trace toutes modifications.

…les autres types existent, mais ils sont plutôt exotiques et se basent sur la combinaison des approches 1, 2, 3 et 4.

Pour les dimensions du type 2, la clé étrangère stockée dans la table de faits est choisie au moment d'insertion/mise à jour du fait pour que la version de données dans le dimension corresponde aux données actuelles au moment de la transaction.

Assez souvent l'analyse BI ne nécessite pas de données historisées car l'identité d'objet de transaction est plus importante que son état.

analyse

Analyse de données :

- Rapports prédéfinies (Cognos, Microstrategy, etc).
- Ad-hoc reporting (self service) – outils sont entre les mains d'utilisateurs (Tableau, Spotfire, etc).
- Analyse avancée : data mining, machine learning, modélisation statistique (Python, R, Matlab, SAS, SPSS, etc).

Visualisations de base :

- Camembert ou graphique à barres empilées – comparaison de partie contre total (camembert étant inférieur pour la lisibilité).
- Barres – comparaison entre les classes ou visualisation de temps discrétisé (ex. par trimestres).
- Histogramme – distribution d'un paramètre.
- Nuage des points – distribution de deux paramètres, analyse de dépendance.
- Graphique linéaire – progression dans le temps continu.
- Arbre hiérarchique (tree map) – partie contre total dans la structure de plusieurs niveaux.
- Cartographie – pour les distributions géographiques si la position a de l'importance (jamais positionner les camemberts sur la carte).
- Graphique à bulles – à ne pas utiliser ou si nécessaire de visualiser la 3émé valeur, elle doit être proportionnelle à la surface de la bulle et pas au rayon.
- Treillis – comparaison par secteur/groupe/région.

Analyse avancée, tâches fréquentes :

- Régression – prédiction d'une valeur continue.
- Classification – prédiction d'un class prédéfini.
- Segmentation/clustering – division de données en classes de valeurs proches (class non-prédéfini).
- Réduction de la dimensionalité – projection de données sur l'espace plus petit (afin de les traiter ou de les visualiser).
- Détection d'anomalies.
- Extraction des règles d'association, i.e. corrélation d'évènements.
- Systèmes de recommandation.