Entreposage de données

Rémy Choquet

Processus ETL, Modélisation d'entrepôts, Restitution

Sommaire

- Généralités
- L'entreposage de données
- Les processus ETL
- L'analyse multidimensionnelle (OLAP)
- L'entreposage de données complexes et axes de recherche

Un peu d'histoire

- Au cours des dernières décennies, les entreprises (au sens large) passent à l'ère de l'information.
- Elles ont un défi majeur: faire évoluer leur SI à vocation de production vers un SI décisionnel dont la vocation est le pilotage et l'aide à la décision.
- Ils ont été rapidement reconnus et sont aujourd'hui largement utilisés pour l'analyse de données mais aussi en base pour l'analyse

Comment...

- Des données hétérogènes venant de sources multiples seront nécessaires pour construire un entrepôt, il devra être réactif.
- Un S.I.D. est un ensemble de données organisées de façon spécifique, facilement accessibles et destinées à la prise de décision.
- Les systèmes de gestion sont dédiés aux métiers; les systèmes décisionnels sont dédiés au pilotage de l'entreprise.

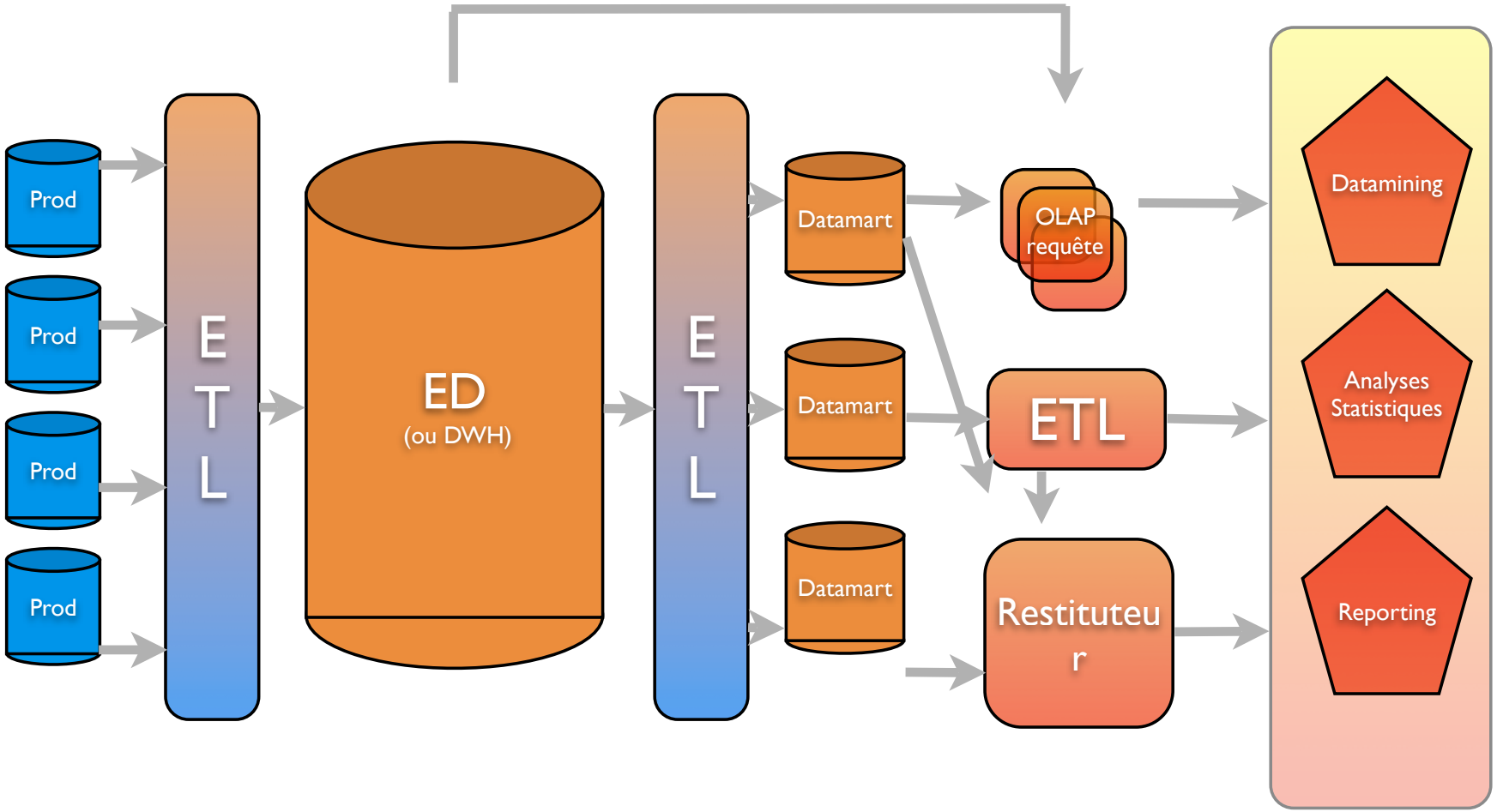
Qu'est ce qu'un Entrepôt de Données?

D'après BILL Inmon:

"Un ED est une collection de données thématiques, intégrées, non volatiles et historisées, organisées pour la prise de décision."

- Thématiques: thèmes par activités majeures;
- Intégrées: diverses sources de données;
- Non volatiles: non modifiables ou effaçables;
- Historisées: trace des données, suivre

Vision d'ensemble



Phases de construction d'un ED

- Conception (définition de la finalité de l'ED):
  - Piloter qu'elle activité, quels besoins d'analyse;
  - Déterminer et recenser les données à entreposer;
  - Définir les aspects techniques de la réalisation;
  - Le modèle de données;
  - Les démarches d'alimentation;

Phases de construction d'un ED

- Construction:
  - Extraction des données des différentes bases de production;
  - Nettoyage des données et mise en oeuvre des règles d'homogénéisation des données grâce aux métadonnées;
  - Voir ETL...
- Administration:
  - Assurer la qualité et la pérénité des données

Phases de construction d'un ED

- Restitution:
  - Mise en place des stratégies de restitution
  - Ordonnancement
  - Choix des outils de restitution
  - Reporting

Modélisation d'un ED: Modèle en Etoile

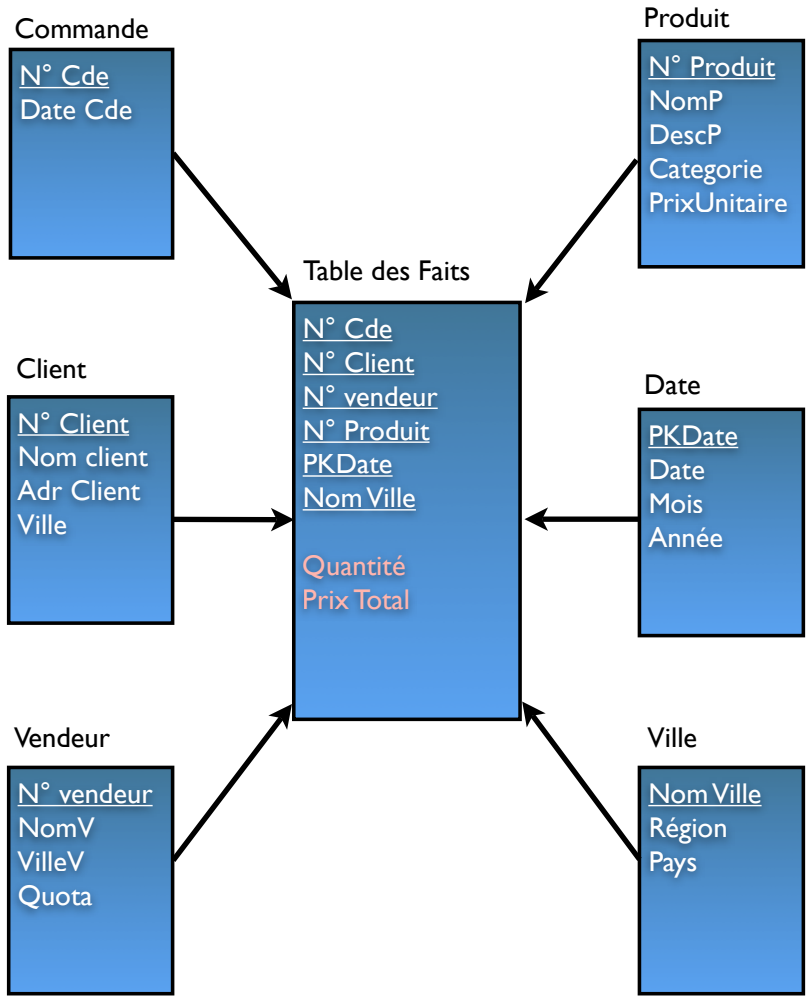
- Une ou plusieurs tables de faits, une ou plusieurs mesures
- Plusieurs tables de dimensions
- Granularité définie par les identifiants dans la table des faits

Avantages:

- Facilité de navigation
- Performances, nombre de jointures limitées
- Gestion des agrégats
- Fiabilité des résultats

Inconvénients:

- Toutes les dimensions ne concernent pas les mesures
- Redondances dans les dimensions
- Alimentation complexe



Propriétés des mesures et des dimensions

Mesures:

- Additivité: somme sur toutes les mesures (CA, quantité vendue)
- Semi-additivité: somme sur certaines dimensions (Nb contact client)
- Non-additivité: pas de somme, recalculer (encours moyen fin de mois)

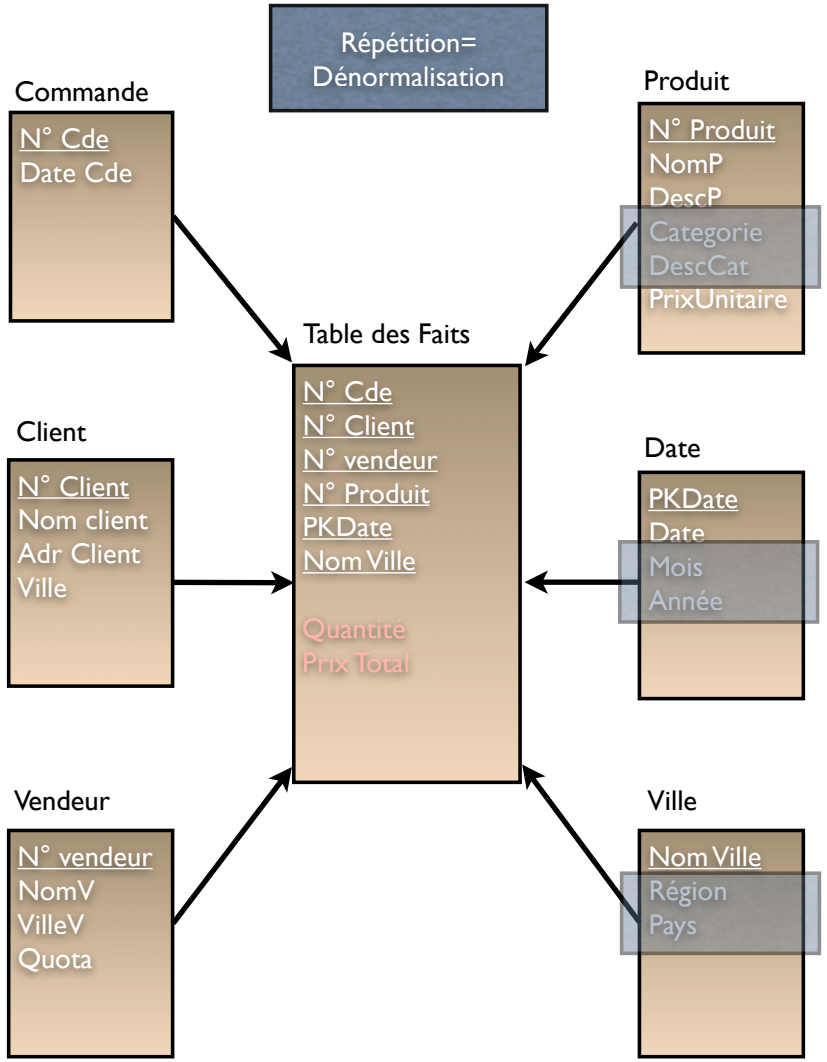
Dimensions:

- Elle est définie sur un thème ou un axe (attributs) selon lequel les données seront analysées
- Sont des points de vue depuis lesquels les mesures peuvent

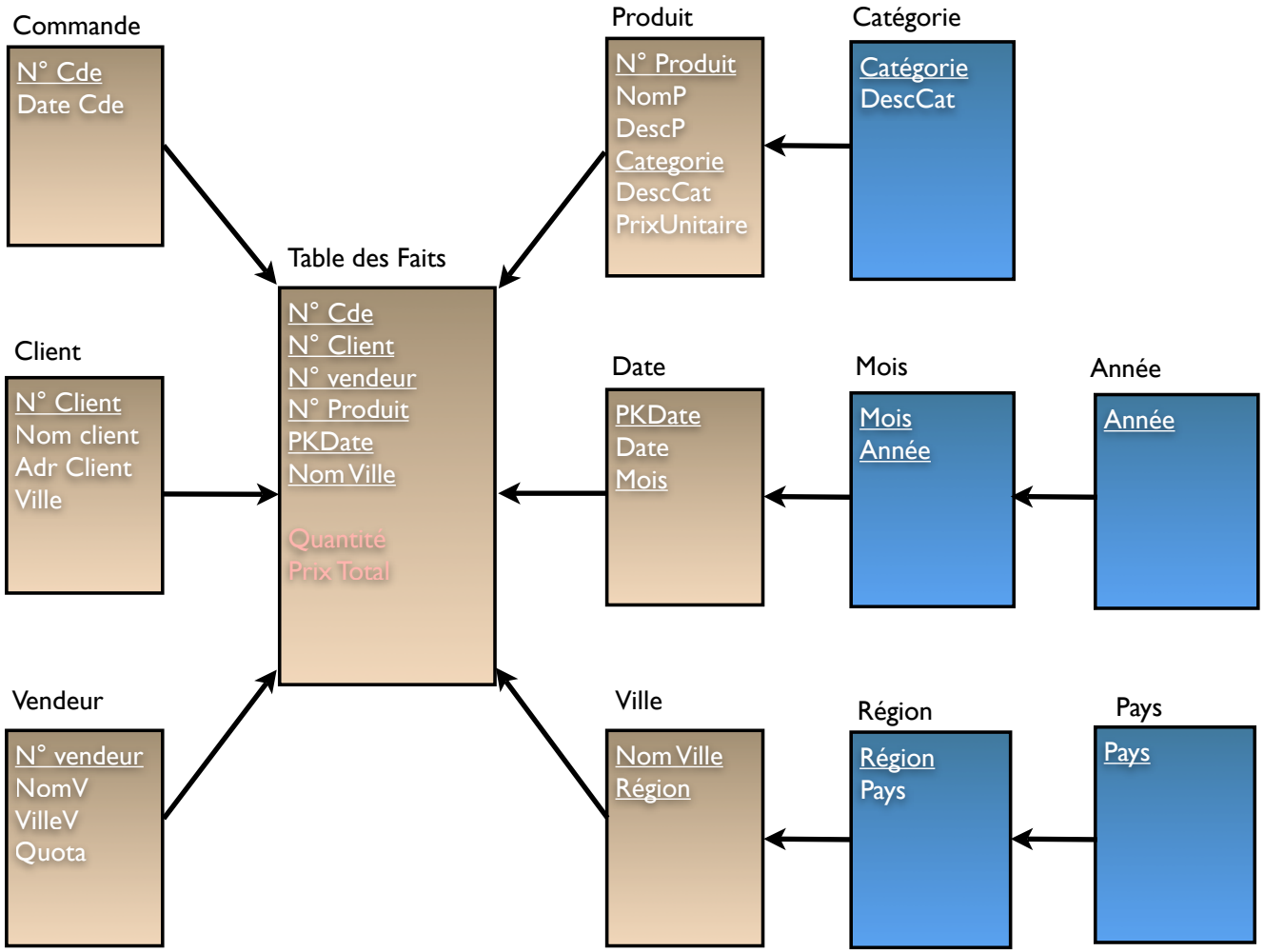
Modélisation en flocon de neige

- Le modèle de l'ED doit être simple à comprendre. On peut augmenter sa lisibilité en regroupant certaines dimensions. On définit ainsi des **hiérarchies**
- Ex:

Client	Commune	Département	Région	Pays	Continent



Modélisation en flocon de neige



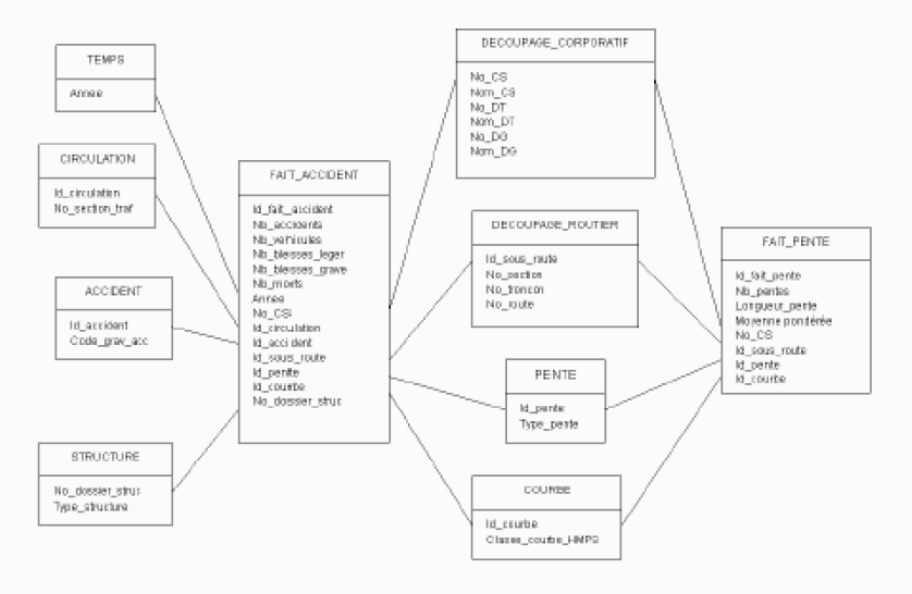
Modélisation en flocon de neige

- Modèle en flocons de neige = Modèle en étoile + Normalisation des dimensions

- ***Avantages:***
  - Réduction du volume
  - Permettre des analyses par pallier (drill down) sur la dimension hiérarchisée
- ***Inconvénients:***
  - navigation difficile
  - nombreuses jointures

Modèle en constellation

La modélisation en constellation consiste en la fusion de plusieurs modèles en étoile par une dimension commune
Il comprend donc plusieurs tables de faits et des tables de dimensions communes ou non à ces faits.



Estimations du volume d'un ED

- Exemples :

- Grande distribution :

CA annuel : 80 000 M$
Prix moyen d'un article d'un ticket : 5$
Nbre d'articles vendus pour un an : 80 * 109 / 5 = 16 * 109
Volume du DW :
16*109 *3 ans * 24 octets = 1,54 To (1,54*1012 = 1 540 Go)

- Téléphonie :

Nbre d'appels quotidiens : 100 millions
Historique : 3 ans * 365 jours= 1 095 jours
Volume du DW :
100 millions * 1 095 jours * 24 octets = 3,94 To

- Cartes de crédit :

Nbre de clients : 50 millions
Nbre moyen mensuel de transactions : 30
Volume :
50 millions * 26 mois * 30 transactions * 24 octets = 1,73 To

ETL: Principes

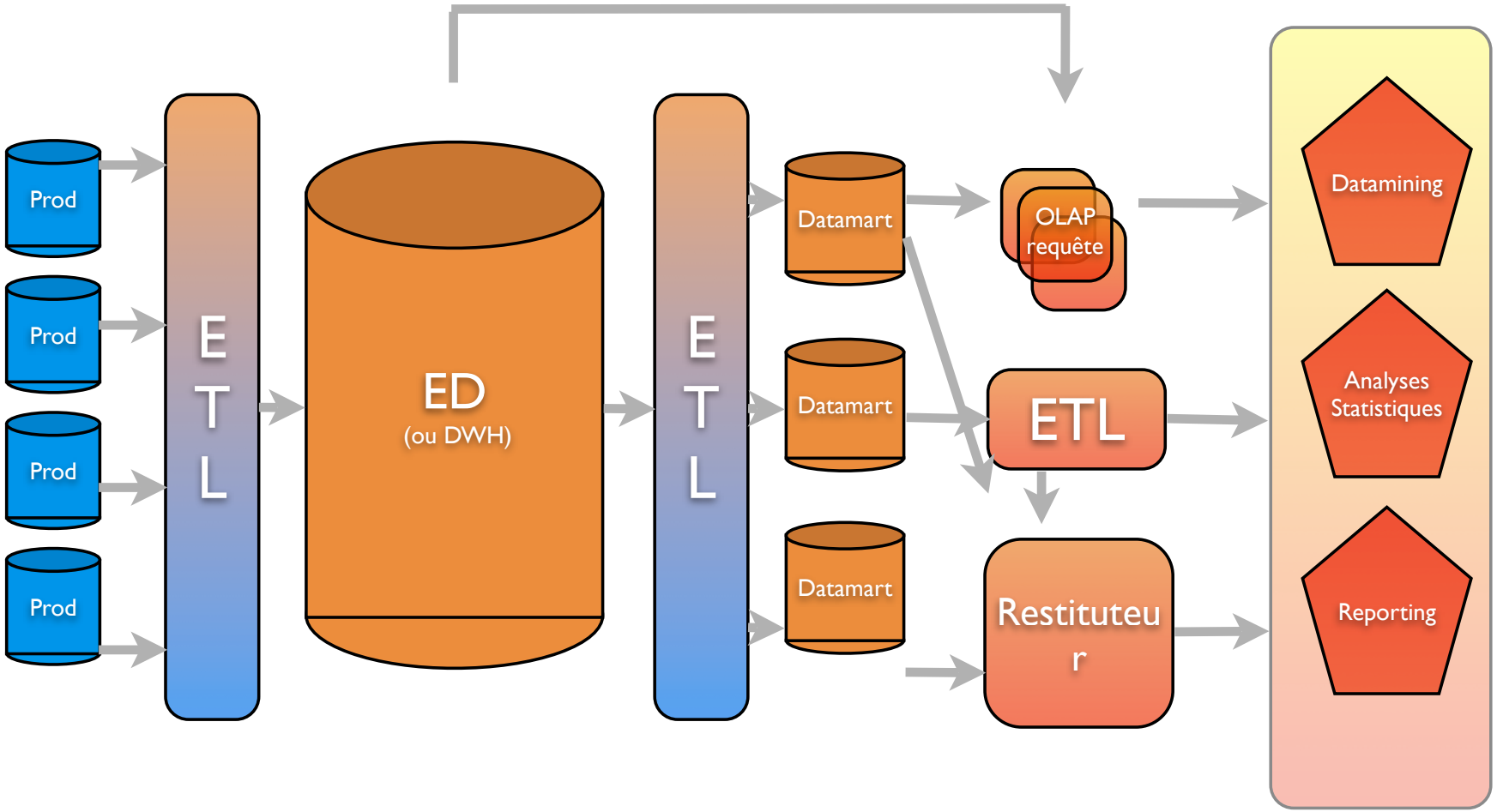
- « Extract-Transform-Load » est connu sous le terme ETL (ou parfois : datapumping). Il s'agit d'une technologie informatique intergicielle (comprendre middleware) permettant d'effectuer des synchronisations massives d'information d'une base de données vers une autre. Selon le contexte, on traduira par « alimentation », « extraction », « transformation », « constitution » ou « conversion », souvent combinés.

- Elle est basée sur des connecteurs servant à exporter ou importer les données dans les applications (Ex : connecteur Oracle ou SAP...), des transformateurs qui manipulent les données (agrégations, filtres, conversions...), et des mises en correspondance (mappages). Le but est l'intégration de l'entreprise par ses données.

- A l'origine, les solutions d'ETL sont apparues pour le chargement régulier de données agrégées dans les entrepôts de données (ou datawarehouse), avant de se diversifier vers les autres domaines logiciels. Ces solutions sont largement utilisées dans le monde bancaire et financier, ainsi que dans l'industrie, vu la multiplication des nombreuses interfaces.

- Des technologies complémentaires sont apparues par la suite : l'EAI (Intégration d'applications d'entreprise), puis l'ESB (Enterprise Service Bus).

Vision d'ensemble



Acheter ou développer?

Avantages d'une suite ETL:

- Développement simple, rapide et moins coûteux. Les coûts de l'outil seront amortis rapidement pour les projets sophistiqués ou de grandes envergures.
- Des ressources disposant de connaissances du domaine d'affaire et n'ayant pas de grandes compétences en programmation peuvent développer avec l'outil.
- La plupart des outils ETL intègrent des référentiels de gestion du métadata, tout en permettant de synchroniser le métadata avec les systèmes sources, les bases de données de l'entrepôt et autres outils BI.
- La plupart des outils ETL permettent la génération automatique du métadata à chaque étape du processus ETL et renforce la mise en place d'une méthodologie commune de gestion de métadata qui doit être respectée par tous les développeurs.
- La plupart des outils ETL dispose de programme intégré qui permet de faciliter la documentation, la création et la gestion de changement. L'outil ETL doit bien gérer les dépendances complexes et les erreurs qui peuvent surgir en cours d'exécution.
- Le référentiel de métadata de la plupart des outils ETL peut produire automatiquement des rapports de mise en correspondance des données (data lineage, looking backward) et d'analyse de dépendance de données (looking forward).
- Les outils ETL disposent de connecteurs intégrés pour la plupart des sources de données. Ils permettent aussi d'effectuer des conversions complexes de types de données (selon la source et la destination)
- Les outils ETL offrent des mécanismes de cryptage de compression en ligne de données.
- La plupart des outils ETL offre une très bonne performance même pour une grande quantité de données. Considérer donc d'acheter un outil ETL le volume de données est grand ou encore s'il va le devenir !
- Un outil ETL peut, le cas échéant, gérer des scénarios d'équilibrage de la charge entre les serveurs.

Acheter ou développer?

Avantages d'un ETL maison:

- Les outils de tests unitaires automatique sont disponibles seulement pour les outils développé maison. Par exemple Junit .
- Les techniques de programmation orientée objet permettent de rendre consistantes la gestion des erreurs, la validation et la mise à jour du métadata.
- Il est possible de gérer manuellement le métadata dans le code et de créer des interfaces pour la gestion de ce dernier.
- Disponibilité des programmeurs dans l'entreprise.
- Un outil ETL est limité aux capacités du fournisseur.
- Un outil ETL est limité à l'outil de scripting propriétaire.
- Un outil développé maison donne une grande flexibilité si le besoin se présente. Il est possible de faire tout.

Mon expérience dans le domaine nous a démontré que même pour les projets de petite envergure, il est conseillé de développer votre système ETL en utilisant une suite ETL. Nous résumons ici les avantages d'une telle solution :

- Définir une fois, appliquer plusieurs fois (partage et réutilisation)
- L'analyse d'impact
- Le référentiel de métadata.
- L'agrégation incrémentale
- La gestion des traitements par lot.
- Connectivité simplifiée
- Traitements parallèles et équilibrage de la charge
- L'expérience et le support du fournisseur.

Les 38 sous systèmes ETL (Kimball 2004) 1/5

1 - Système d'extraction – [EXTRACTION]

- Gestionnaire de connections : Connecteurs aux différentes sources de données (ODBC, Native...) et destinations de données.
- Mécanisme de filtre et de tri à la source : Ce module devrait permettre d'effectuer l'équivalent de l'utilisation de la clause Where dans un SQL. Formatage et conversion de données (Date, nombre...).Stockage de données dans l'environnement ETL (INSERT, UPDATE, DELETE)

2 - Système de détection des changements (Change data capture ou le CDC) CDC : [EXTRACTION]

- Lecteur des fichiers de log de bases de données. (Sorte de log miner);
- Comparateur d'enregistrements selon le CRC.

3 - Système d'analyse de données. [EXTRACTION, TRANSFORMATION]

- Analyse des propriétés des colonnes.
- Analyse de la structure des données incluant les clés étrangères, les clés primaires, les relations.
- Analyse des règles de gestion.

4 - Système de Nettoyage de données. [TRANSFORMATION]

- En général un système à base de dictionnaire pour analyse de noms et adresses des individus et organisations et possiblement les produits et les emplacement géographiques;
- Déduplication des données (Le même client peut provenir de plusieurs systèmes)
- Utilisation des techniques de logique floue (peut-être vrai, probablement vrai, peut-être faux, probablement faux)
- Utilisation des techniques de fusion (Merge).
- Gestion des clés d'affaires au niveau des systèmes sources.

5 - Système de validation de la conformité de données. [TRANSFORMATION]

- Identification et renforcement des attributs des dimensions conforme;
- Identification et renforcement des attributs des faits conformes;

6 - Gestionnaire de dimension d'audit. [TRANSFORMATION]

- assemblage du métadata concernant le chargement de chaque table de fait dans une dimension d'audit;
- Attachement de la dimension d'audit à la table de fait comme une dimension normale.

7 - Système de gestion de la qualité de données. [TRANSFORMATION]

- Appliquer des tests à la volée à tous les flux de données pour déceler des problèmes de qualité de données;

Les 38 sous systèmes ETL (Kimball 2004) 2/5

8 - Système de gestion des erreurs. [EXTRACTION, TRANSFORMATION, CHARGEMENT]

- Surveiller et détecter les erreurs en temps réel;
- Automatiser la reprise après erreur ;
- Traiter les messages reçus du système de gestion de la qualité de données.

9 - Système de gestion des clés de substitution (surrogate key). [TRANSFORMATION]

- Produire et gérer d'une façon centralisée les clés de substitution (dimension et fait) ;
- Être indépendant de la base de données.

10 - Gestionnaire de Slowly Changing Dimension (SCD). [TRANSFORMATION]

- Gérer les trois types de SCD (Type1 : Écraser, Type 2 : Nouvel enregistrement, Type 3 : Nouvelle colonne)

11 - Gestionnaire des dimensions arrivant en retard. [TRANSFORMATION]

- Insérer et mettre à jour des données associées (fait ou dimension) à une dimension que l'on reçoit en retard. Par dimension, on veut dire une un enregistrement de la dimension complète. (Ceci implique que chaque dimension dispose d'un horodate dans le système source qui décrit la date et l'heure de la création de l'enregistrement dans ce dernier).

12 - Gestionnaire de dimension à hiérarchie fixe. [TRANSFORMATION]

- Créer et gérer les dimensions à hiérarchie fixe (Une hiérarchie fixe est une hiérarchie dont le nombre de niveau est fixe et ne change pas dans le temps d'exécution). (many-to-one).

13 - Gestionnaire de dimension à hiérarchie variable. [TRANSFORMATION]

- Créer est gérer les dimensions à hiérarchie variables (Une hiérarchie variable est une hiérarchie dont la profondeur ou le nombre de niveaux est variable comme par exemple l'organigramme d'un entreprise.

14 - Gestionnaire de dimensions multivaluées (Brige table). [TRANSFORMATION]

- Créer et gérer les tables associatives utilisées pour décrire les relations many-to-many entre les dimensions ou entre les faits et les dimensions. (la dimension médicament est multivaluée, car un médecin peut prescrire plusieurs médicaments lors d'une visite médicale).
  Inclure le facteur de pondération. (Optionnel).

15 - Gestionnaire des Junk dimensions. [TRANSFORMATION]

- Créer et gérer les junk dimension (voire différents types de dimensions).

16 - Système de chargement des tables de faits au niveau de détail le plus fin (grain) [CHARGEMENT]

- Insérer et mettre à jour les tables de faits au niveau du grain;

Les 38 sous systèmes ETL (Kimball 2004) 3/5

17 - Système de chargement périodique des tables de fait au niveau de détail le plus fin (grain). [CHARGEMENT]

- Insérer et mettre à jour d'une façon périodique les tables de fait dans le détail du niveau de grain.
- Manipuler les indexes et les partitions;
- Utiliser le gestionnaire des lookup.(voire sous-système 19).

18 - Système de chargement des tables de fait cumulatives au niveau de détail le plus fin (grain). [CHARGEMENT]

- Mettre à jour des tables de faits cumulatives;
- Manipuler les indexes et les partitions;
- Utiliser le gestionnaire des lookup. (voire sous-système 19).

19 - Gestionnaire des lookup. [TRANSFORMATION]

- Remplacer les clés d'affaires par les clés de substitution;
- Être performant lors de la substitution (Multithreaded process)

20 - Gestionnaire des faits arrivants en retard.[TRANSFORMATION]

- Insérer et mettre à jour des enregistrements de fait qui arrivent en retard

21 - Gestionnaire d'agrégation.[TRANSFORMATION]

- Créer et maintenir des structures d'agrégation qui sont utilisées conjointement avec le mécanisme du Query-Rewrite;
- Inclure les vues matérialisées

22 - Gestionnaire de cubes multidimensionnels. [TRANSFORMATION]

- Créer et gérer la fondation du schéma en étoile pour alimenter les cubes dimensionnels (Cubes OLAP);
- Préparer les hiérarchies pour alimenter les cubes selon la suite BI utilisée.

23 - Gestionnaire des partitions en temps réel.[TRANSFORMATION]

- Maintenir en mémoire seulement les partitions des données des faits qui arrivent depuis la dernière mise à jour.

24 - Système de gestion des dimensions.[TRANSFORMATION]

- Répliquer les dimensions conformes à partir d'un emplacement centralisé vers le fournisseur des tables de fait. (Voir le sous-système 25).

25 - Système de gestion des tables de faits.[TRANSFORMATION]

- Utiliser les dimensions conformes transmises par le système de gestion des dimensions (24).
- Substituer les clés étrangères;

Les 38 sous systèmes ETL (Kimball 2004) 4/5

26 - Ordonnanceur des processus ETL.[OPÉRATION]

- Ordonnancer et lancer les processus ETL;
- Être capable de coordonner les processus en tenant compte de différentes conditions de succès ou d'échec de processus;
- Produire des alertes et envoyer des messages.

27 - Système de surveillance du flux des processus ETL.[OPÉRATION]

- produire des tableaux de bord et des rapports d'audit pour tous les processus ETL en exécution incluant les horodates, les nombres d'enregistrements traités, les erreurs, les actions réalisées par le moteur ETL (rejet des enregistrements non concordant lors des lookups...).

28 - Système des recouvrement et reprise.[OPÉRATION]

- Reprendre l'exécution d'un processus au même endroit que celui-ci a planté;
- Offrir la possibilité d'arrêter (selon une condition) un processus ETL et le ré-exécuter.

29 - Gestionnaire de parallélisme et de pipelines.[OPÉRATION]

- Offrir les avantages d'utiliser des processeurs multiples ou l'informatique en grille (Grid computing);
- Offrir la possibilité de transmission continue de données (pipeline);
- Offrir le parallélisme automatique et conditionnel des processus ETL.

30 - Système de gestion des erreurs.[OPÉRATION]

- Gérer les erreurs;
- Aviser les personnes concernées;
- Journaliser les erreurs;
- Système de gestion des erreurs.

31 - Système de contrôle des versions. [DÉVELOPPEMENT]

- Gérer les versions du projet ETL;
- Réserver et replacer les composantes du projet ETL (Check-out, ckeck-in...);
- Comparaison des différentes versions d'un projet ETL.

32 - Système de déploiement.[OPÉRATION]

- Migration de l'environnement de développement vers celui de test et de production;
- S'intégrer ou intégrer le système de contrôle de version pour;
- Configurer les connexions pour la version;

Les 38 sous systèmes ETL (Kimball 2004) 5/5

33 - Système d'analyse de correspondance et de dépendance.[DÉVELOPPEMENT]

- Afficher les sources de données et les transformations subies par un élément de données spécifique (une colonne);
- Analyser l'impact de changer un élément de données;

34 - Gestionnaire de conformité aux règles.[OPÉRATION]

- Prouver que les données et les transformations n'ont pas changé et sont conformes aux règles établies;
- Surveiller les accès et les modifications aux données pour prouver que les données et les transformations n'ont pas changées.

35 - Système de sécurité.[OPÉRATION]

- Administrer la sécurité sur les données et les méta données des processus ETL;
- Offrir la possibilité de prouver que la version d'un processus ETL n'a pas changé;
- Afficher qui a effectué les changements.

36 - Système de sauvegarde.[OPÉRATION]

- Sauvegarder les données et les méta données pour le recouvrement, la sécurité et les besoins de conformité.

37 - Gestionnaire de référentiel du méta données.[DÉVELOPPEMENT]

- Collecter et maintenir les méta-données concernat le projet ETL, incluant les processus ETL, les transformations...

38 -Système de gestion de projet.[DÉVELOPPEMENT]

- Surveiller toutes les activités de développement, de test du projet ETL

Plan de projet ETL 1/3

Phase	Description de la phase	Tâche	Description de la tâche	Responsable
I	Mise en place de l'environnement de dévelopement	1	Configurer l'infrastructure matérielle	DBA
		2	Installation des logiciels et outils	DBA / A-ETL
		3	Mettre en place les documents sur les meilleures pratiques	G-ETL / A-ETL
II	Analyse des besoins métier	1	Revue de la doc existante avec le data modeler	A-ETL / A-Système
		2	Définition et documentation des règles métier	A-ETL / A-Système
		3	Analyse des systèmes sources	A-ETL / A-Système
		4	Définition de la portée des phases de projet	G-ETL
III	La conception des mises en correspondance des données	1	Revue du modèle de données de l'entrepôt de données	A-ETL
		2	Revue des règles métier	A-ETL
		3	Analyse des systèmes sources	A-ETL
		4	Création du document de mise en correspondance des données	A-ETL

Plan de projet ETL 2/3

Phase	Description de la phase	Tâche	Description de la tâche	Responsable
IV	Stratégie de qualité des données	1	Définition des règles de qualité des données	G-ETL / S-Q-D
		2	Documentation des défauts de donnés	G-ETL / S-Q-D
		3	Affectation de la responsabilité des défauts de données	G-ETL / S-Q-D
		4	Création du document de mise en correspondance des données	G-ETL / S-Q-D
		5	Sensibilisation des utilisateurs finaux des défauts des données	G-ETL / S-Q-D
		6	Intégration des règles de qualité dans le document de mise en correspondance	G-ETL / S-Q-D
V	Développement des processus ETL	1	Revue du document de mise en correspondance	D-ETL
		2	Développement des dimensions simples	D-ETL
		3	Développement des dimensions SCD-2 (historique)	D-ETL
		4	Développement des dimensions SCD-2 (incrémental)	D-ETL
		5	Développement des tables de faits (historique)	D-ETL
		6	Développement des tables de faits (incrémental)	D-ETL
		7	Automatisation des processus	D-ETL

Plan de projet ETL 3/3

Phase	Description de la phase	Tâche	Description de la tâche	Responsable
VI	Tests unitaires / Tests d'assurance qualité / Tests d'acceptance	1	Mise en place de l'environnement de test	DBA / A-ETL
		2	Création des plans de test et les scripts	A-Système
		3	Chargement des données	D-ETL
		4	Exécution des scripts de tests unitaires	A-Système
		5	Contrôle de la qualité des données	A-Système
		6	Validation des données	A-Système
		7	Validation des règles métier	A-Système
		8	Obtention de l'acceptance	G-ETL
VII	Déploiement	1	Création des documents de support	A-ETL
		2	Création des documents des mécanismes de	A-ETL
		3	Mise en place de l'environnement de production	A-ETL
		4	Chargement des données historiques	A-ETL
		5	Ordonnancement des processus incrémental	A-ETL
VIII	Maintenance	1	Développement des rapports d'audit pour les	A-ETL
		2	Vérification des journaux d'exécution	A-ETL
		3	Mise en place de l'environnement de production	A-ETL