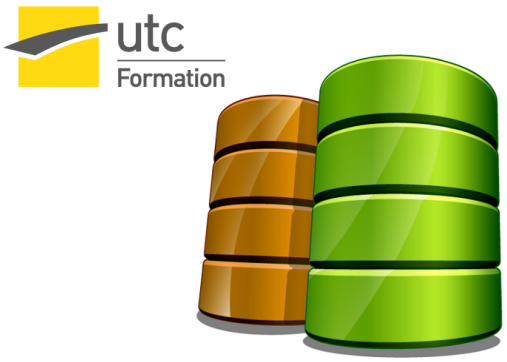
Introduction au domaine du décisionnel et aux data warehouses

Data warehouse
http://dwh.crzt.fr



Stéphane Crozat

1 septembre 2016

Table des matières

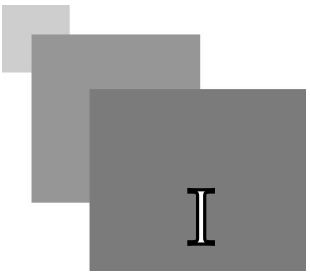
Objectifs

- Connaître les principaux domaines d'application des data warehouses
- Connaître le paradigme du décisionnel (et son articulation avec le paradigme transactionnel)
- Connaître les principes, les étapes et les méthodes de la modélisation dimensionnelle

Introduction

- Volume de cours : 2h

Le décisionnel



Objectifs

Connaître le paradigme du décisionnel (et son articulation avec le paradigme transactionnel)

Connaître les principaux domaines d'application des data warehouses

1. Décisionnel

Définition

«

> Le système d'information décisionnel est un ensemble de données organisées de façon spécifiques, facilement accessibles et appropriées à la prise de décision [...].
>
> La finalité d'un système décisionnel est le pilotage d'entreprise.
>
> Les systèmes de gestion sont dédiés aux métiers de l'entreprise [...].
>
> Les systèmes décisionnels sont dédiés au management de l'entreprise [...].
>
> (Goglin, 2001, pp21-22) *

 »

Synonymes : informatique décisionnelle, *business intelligence*, BI *

2. Enjeux du décisionnel

La prise de décisions stratégiques dans une organisation nécessite le recours et le croisement de multiples informations qui concernent tous les départements : production, RH, DAF, achats, ventes, marketing, service après-vente, maintenance, R&D...

Or ces données sont généralement :

- éparpillées au sein des départements et non connectées entre elles
- hétérogènes dans leurs formats techniques et leurs organisations structurelles, voire leurs sémantiques
- implémentées pour l'action (par construction) et non pour l'analyse
- volatiles, au sens où leur mise à jour peut conduire à oublier des informations obsolètes

Exemple

Un catalogue de produits sera conçu pour permettre de trouver facilement un produit en fonction de caractéristiques précises, de faire des mises à jour rapides et fiables, de gérer des stocks...

Mais un système décisionnel souhaitera :

- connaître l'organisation des produits selon certaines caractéristiques et regroupements qui ne sont pas forcément premiers dans la gestion quotidienne ;
- croiser le catalogue avec les ventes...

Fondamental

L'enjeu des systèmes décisionnels est de donner accès aux données existantes dans l'organisation, sous une forme intégrée, afin de faciliter leur interrogation croisée et massive.

Complément : Voir aussi

Différence entre un DW et un système transactionnel (cf. p.14)

3. Exploitation des données

Les données agrégées dans un système décisionnel servent à trois grandes catégories d'usage :

- La production de rapport récurrents (*reporting*)
- L'exploration manuelle
- L'analyse de données (descriptive ou prédictive)

Définition : Reporting

Le principe du *reporting* est d'agréger et de synthétiser des données nombreuses et complexes sous forme d'indicateurs, de tableaux, de graphiques permettant d'en avoir une appréhension globale et simplifiée.

Le *reporting* s'appuie principalement sur les agrégats (GROUP BY en SQL par exemple) afin de faire apparaître des comptages, sommes ou moyennes en fonction de critères d'analyses.

Le *reporting* est généralement récurrent, le même rapport sera produit à intervalles réguliers pour contrôler les variations des indicateurs.

Définition : Exploration manuelle

Une autre exploitation de données en contexte décisionnel consiste à pouvoir explorer les données de façon peu dirigée (heuristique) afin de trouver des réponses à des questions que l'on ne s'est pas posées (sérendipité). L'idée générale est plutôt que les réponses aux premières questions que l'on se posent conduiront à se poser de nouvelles questions.

L'exploration de données s'appuie sur des outils permettant de manipulation (IHM) et de visualiser (infovis) les données selon des requêtes dynamiquement produites par des utilisateurs experts du domaine.

Définition : Analyse de données

L'analyse de données est une branche de la statistique qui permet de mettre en évidence des tendances des données ou corrélations entre les données non évidentes a priori.

- Dans le cas de l'analyse descriptive, ll s'agit de rechercher une information statistique "cachée" que l'on ne connaît pas a priori.
- L'approche prédictive consiste à réaliser un modèle statistique des corrélations entre les données à partir d'échantillons d'apprentissage, puis à appliquer le modèle à des données nouvelles pour prédire leur comportement, avec des raisonnements du type "si ... alors" ; ou pour classifier des données (tel objet caractérisé par telles données appartient-il à telle classe ?). Les résultats sont généralement qualifiés par une probabilité d'occurrence.

4. Éthique et limites des systèmes décisionnels

Rationalisation excessive et processus complexes

Les systèmes décisionnels produisent des indicateurs ou s'appuient sur des modèles dont l'objectif est de simplifier la réalité pour aider à la prise de décision.

Mais la décision doit bien réintégrer des évaluations humaines qui la replacent dans sa réalité, qui est restée complexe.

- Le modèle ou l'indicateur n'est pas la réalité, s'en est une représentation.
- La décision ne s'applique pas à une représentation, mais à la réalité.

Sélectivité des données et organisations humaines

Les systèmes décisionnels s'appuient sur les données que l'on est en mesure de produire, mais ces données ne peuvent pas intégrer toutes les dimensions d'une organisation et de son environnement, en particulier les dimensions humaines.

Or ces dimensions cachées au système décisionnel déterminent de nombreux fonctionnements de l'organisation, et doivent continuer d'être prises en compte.

L'interprétation est humaine

Un système informatique produit des indicateurs qui nécessitent des interprétations humaines, expertes dans le cas du décisionnel. Un système informatique ne produit pas des directives qu'une organisation humaine doit suivre !

L'erreur est informatique

Les résultats produits par les systèmes décisionnels sont le résultat de conceptions informatiques et mathématiques complexes, qui peuvent receler des erreurs ou des raccourcis, par ailleurs les résultats sont souvent statistiques, donc non déterministes.

La possibilité d'une erreur ou d'une approximation inadaptée devra toujours être prise en compte dans les décisions.

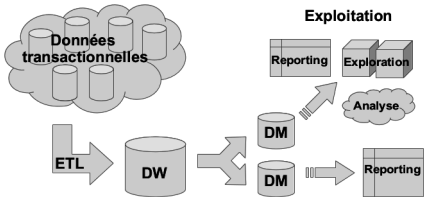
CNIL

Le fait de constituer des fichiers informatisés relatifs des personnes doit généralement faire l'objet d'une déclaration à la CNIL et nécessite le respect de certaines règles comme le droit de rectification et de radiation.

5. Architecture d'un système décisionnel

Tout système décisionnel est architecturé globalement de la même façon :

- En amont un accès au système transactionnel en lecture seule
- Un DW * fusionnant les données requises
- Un ETL * permettant d'alimenter le DW à partir des données existantes
- Des applications d'exploitation de *reporting*, exploration et/ou prédiction
- D'éventuels DM * permettant de simplifier le DW en vue de certaines applications



Architecture générale d'un systèmes décisionnel

Fondamental : Principe de fonctionnement

Le but du système est globalement d'être capable de présenter des tableaux de données (fichiers plats) en intrants des applications d'exploitation.

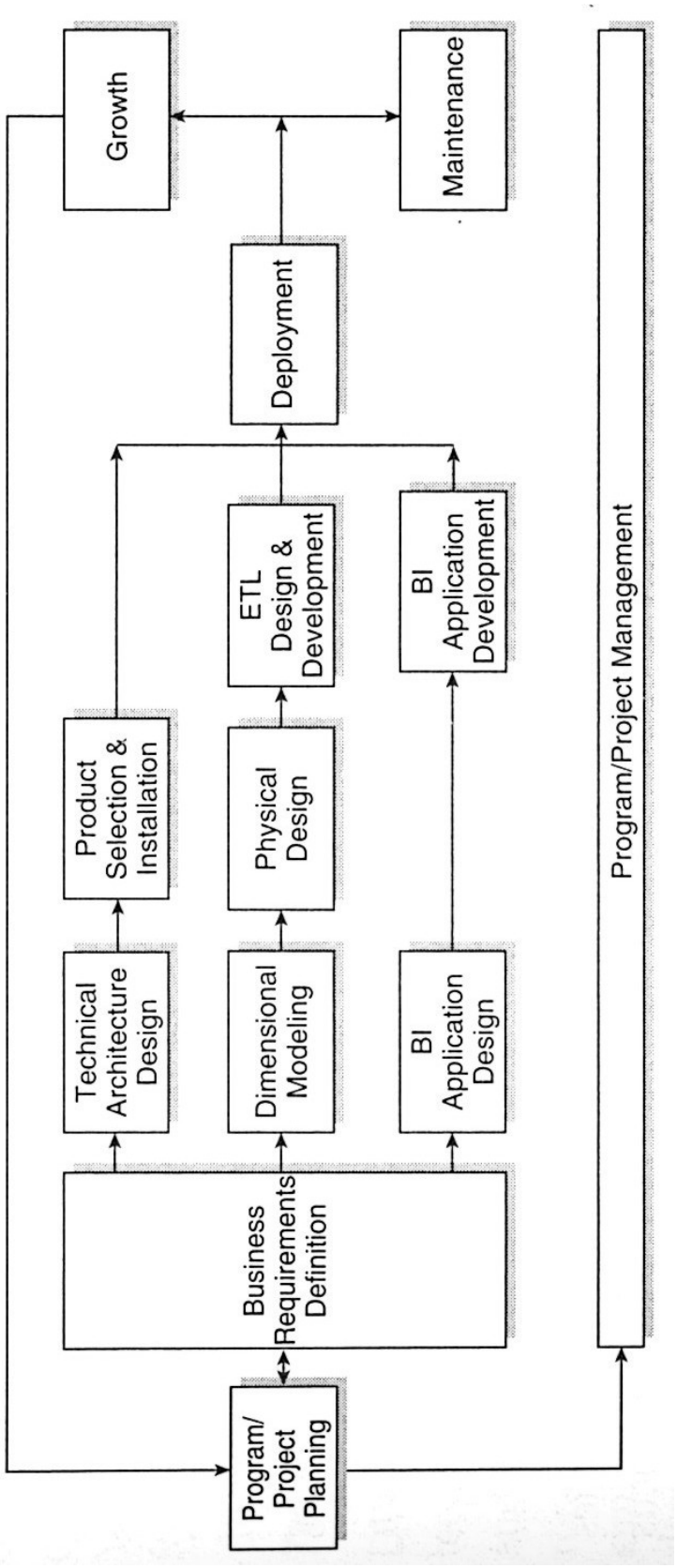
6. Conception d'un système décisionnel

Un projet de système décisionnel se structure selon quatre grands axes :

1. Étude des besoins et de l'existant
    - Étude des besoins utilisateurs
    - Étude des données existantes
2. Modélisation et conception
    - Modélisation dimensionnelle
    - Architecture technique
    - Spécification des outils d'exploitation
3. Implémentation du data warehouse
    - Implémentation du DW * et des DM *
    - Mise en place de l'ETL *
4. Implémentation des outils d'exploitation

- Implémentation des outils de *reporting*
- Implémentation des outils d'exploration
- Implémentation des outils de prédiction

Complément



Lifecycle approach to DW/BI (Kimball, 2008, p3)

Complément

(Kimball, 2008) *

7. Quelques exemples d'application

- Analyse du comportement de consommateurs ou de citoyens, en fonction de leurs caractéristiques (sexe, age...), de critères socio-économiques (profession...), géographiques...
- Analyse de ventes en fonction de l'implantation géographique de magasins (densité, caractéristiques des régions...), de l'organisation de magasins (rayonnage, marketing, RH...)
- Analyse des structures de paniers (quel produit est vendu en même temps que quel autre à quelles conditions ?)
- Prédiction de ventes en fonctions de données conjoncturelles, gestion des stocks, des approvisionnements
- Contrôle qualité et analyse de défaut des chaînes de production en fonction des centres de production, des organisations, des fournisseurs...
- ...

Le data warehouse

III

Objectifs

Comprendre ce qu'est et à quoi sert un data warehouse.

Comprendre les différences entre un data warehouse et une base de données transactionnelle.

1. Data warehousing

Définition : Définition historique de Inmon

« A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data.

(Inmon, 2002, p31) * »

Définition

Un data warehouse (DW *) est une base de données construite par copie et réorganisation de multiples sources (dont principalement le système transactionnel de l'entreprise), afin de servir de source de données à des applications décisionnelles :

- il agrège de nombreuses données de l'entreprise (*intégration*) ;
- il mémorise les données dans le temps (*historisation*) ;
- il les organise pour faciliter les requêtes de prise de décision (*optimisation*).

(Goglin, 2001, p27) *

Synonymes : entrepôt de données, base de données décisionnelle

Fondamental

L'objectif du data warehouse est de permettre des requêtes sur de grands ensembles des données, la plupart du temps sous forme d'agrégats (`GROUP BY`) afin d'en obtenir une vision synthétique (propre à la prise de décision).

Remarque

Le data warehouse dédié au décisionnel est séparé du système transactionnel dédié à la gestion quotidienne.

Complément : Voir aussi

Data warehouse et data mart (cf. p.15)

2. Différence entre un DW et un système transactionnel

BD transactionnelle

Une base données classique est destinée à assumer des *transactions* en temps réel :

- Ajout, mise à jour suppression de données
- Questions sur des données identifiées ou questions statistiques

Datawarehouse

Un DW * est uniquement destiné à l'exécution de *questions statistiques* sur des données statiques (ou faiblement dynamiques).

PRIMITIVE DATA/OPERATIONAL DATA	DERIVED DATA/DSS DATA
• application oriented	• subject oriented
• detailed	• summarized, otherwise refined
• accurate, as of the moment of access	• represents values over time, snapshots
• serves the clerical community	• serves the managerial community
• can be updated	• is not updated
• run repetitively	• run heuristically
• requirements for processing understood a priori	• requirements for processing not understood *a priori*
• compatible with the SDLC	• completely different life cycle
• performance sensitive	• performance relaxed
• accessed a unit at a time	• accessed a set at a time
• transaction driven	• analysis driven
• control of update a major concern in terms of ownership	• control of update no issue
• high availability	• relaxed availability
• managed in its entirety	• managed by subsets
• nonredundancy	• redundancy is a fact of life
• static structure; variable contents	• flexible structure
• small amount of data used in a process	• large amount of data used in a process
• supports day-to-day operations	• supports managerial needs
• high probability of access	• low, modest probability of access

Un changement d'approche, extrait de (Inmon, 2002, p15)

3. Implémentation du DW avec un SGBDR

Fondamental

Les deux problématiques fondamentales des DW * sont l'*optimisation* et la *simplification* : comment rester *performant* et *lisible* avec de très gros volumes de données et des requêtes portant sur de nombreuses tables (impliquant beaucoup de jointures) ?

On utilise massivement :

- Les *vues concrètes* : Un data warehouse procède par copie depuis le ou les systèmes transactionnels
- La *dénormalisation* : Un data warehouse est hautement redondant

Fondamental

Le caractère *statique* du data warehouse efface les inconvénients de ces techniques lorsqu'elles sont mobilisées dans des systèmes transactionnels.

Rappel

Dénormalisation (cf. p.)

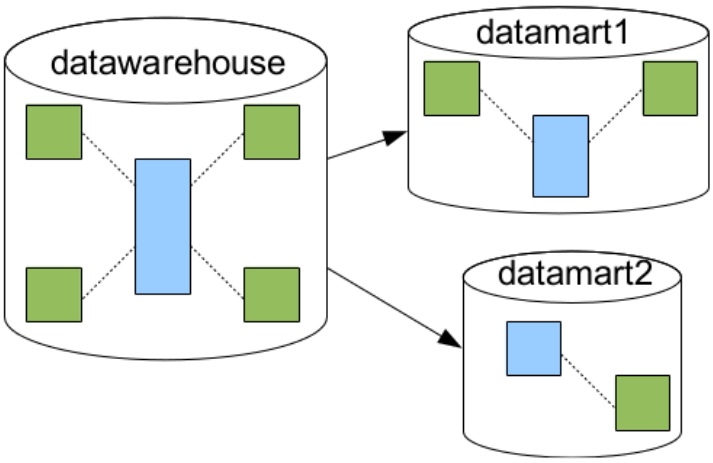
Vues concrètes (cf. p.)

4. Data warehouse et data mart

Un data warehouse et un *data mart* se distinguent par le spectre qu'il recouvre :

- Le *data warehouse* recouvre l'ensemble des données et problématiques d'analyse visées par l'entreprise.
- Le *data mart* recouvre une partie des données et problématiques liées à un métier ou un sujet d'analyse en particulier

Un *data mart* est fréquemment un sous-ensemble du *data warehouse* de l'entreprise, obtenu par extraction et agrégation des données de celui-ci.



Datawarehouse et datamarts

Pourquoi des data marts ?

Les *data marts* sont destinés à pré-agréger des données disponibles de façon plus détaillée dans les *data warehouse*, afin à traiter plus facilement certaines questions spécifiques, critiques, etc.

Exemple : Ticket de caisse

Si un *data warehouse* enregistre un ensemble de ventes d'articles avec un grain très fin, un *data mart* peut faciliter une analyse dite de *ticket de caisse* (co-occurrence de ventes de produits par exemple) en adoptant un grain plus grossier (le ticket plutôt que l'article).

Complément : Ticket de caisse

La lecture de Entrepôts de données : guide pratique de modélisation dimensionnelle * est recommandée pour la modélisation dimensionnelle des tickets de caisse (en particulier pages 31-60 sur la grande distribution).

Le modèle en étoile

III

Objectifs

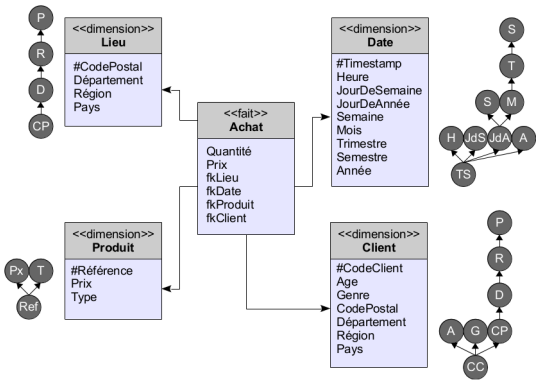
Connaître les principes de la modélisation dimensionnelle

1. Modélisation logique de données en étoile

Définition : Le modèle en étoile

Le *modèle en étoile* est une représentation fortement *dénormalisée* qui assure un haut niveau de performance des requêtes même sur de gros volumes de données.

Exemple



Exemple de modèle dimensionnel en étoile

Complément : Modèle en flocon

Le modèle en flocon est aussi un modèle dénormalisé, mais un peu moins que le modèle en étoile : il conserve un certain niveau de décomposition pour chaque dimension prise isolément.

Complément : Voir aussi

Modélisation en flocon (cf. p.)

2. Objectifs du modèle dimensionnel

La modélisation par schéma en étoile, par opposition aux schémas normalisés en 3NF, permet de répondre à deux besoins caractéristiques des systèmes décisionnels : la *performance* et la *simplicité* des requêtes.

Performance

En effet en tant que structures *redondantes* les schémas en étoiles permettent d'agréger la table des faits avec n'importe qu'elle dimension en *une seule opération de jointure* (deux ou trois pour les schémas en flocons).

Ce gain de performance est souvent critique puisque les volumes de données sont généralement d'un ordre de grandeur très supérieur à celui des systèmes transactionnels.

Cette redondance ne pose pas les mêmes problèmes que dans les systèmes transactionnels, en effet :

- les données étant statiques (importées), il n'y a pas de risque de divergence d'information lors de mises à jour
- l'usage du *datawarehouse* étant essentiellement statistique (regroupement), la conséquence d'une éventuelle erreur n'est pas du même ordre que dans un système transactionnel.

Simplicité

La présentation en étoile des données, avec les faits au centre et les dimensions autour, est particulièrement adaptée à *l'écriture rapide de requêtes simples* pour agréger des données de la table des faits selon des regroupements sur les tables de dimensions.

L'enjeu est de pouvoir répondre simplement et rapidement à une question simple, tandis qu'un modèle transactionnel, qui répond à d'autres contraintes, nécessitera souvent un code SQL complexe et des opérations multiples pour répondre à la même question. Cela permet notamment aux utilisateurs finaux de construire facilement de nouvelles requêtes au fil de leur exploration des données.

Fondamental : Caractéristiques d'un bon modèle décisionnel

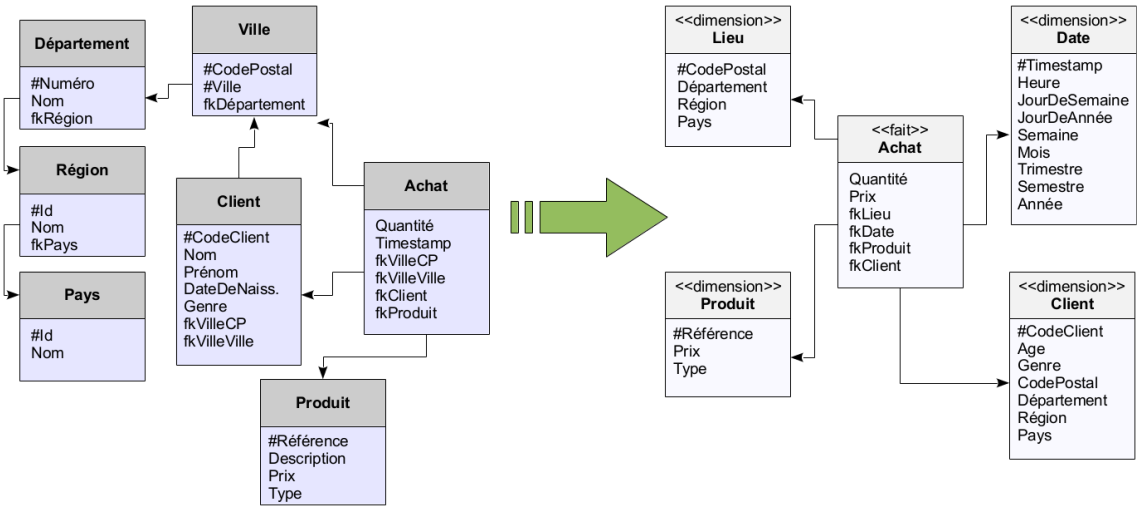
- Être performant pour le calcul d'agrégats sur de gros volumes de données (exploration de données, *reporting*)
- Être appréhendable par un utilisateur final, en particulier pour formuler facilement des requêtes (exploration de données)
- Être suffisamment performant au chargement pour répondre aux sollicitations de mise à jour (ETL *)
- Être évolutif en fonction des évolutions amont (sources transactionnels) et aval (besoins d'exploitation) (ETL, métadonnées)

3. Extraction Transformation Loading

Définition : ETL

L'ETL (*Extraction Transformation Loading*) est le processus de copie des données depuis les tables des systèmes transactionnels vers les tables du modèle en étoile du data warehouse.

Exemple



Exemple de modèle dimensionnel en étoile

Remarque

Les tables du modèle dimensionnel peuvent être vues comme des *vues concrètes* sur le systèmes transactionnel, à la nuance que des transformations (correction d'erreur, extrapolation...) peuvent avoir été apportées dans le processus ETL.

Les outils du décisionnel

IV

Objectifs

Connaître les grandes classes d'outils du domaine du décisionnel

Connaître quelques outils du marché

1. ETL, reporting, exploration, analyse

Fondamental : Principaux types d'outils d'une architecture décisionnel

- ETL *
- *Reporting*
- Exploration
- Analyse

(Smile, 2012) *

Exemple : ETL

Ils permettent de concevoir et d'organiser les processus de migration du système transactionnel vers le système décisionnel.

Exemple : Outils de reporting

Ils permettent :

- la création graphique de rapport
- l'accès aux sources de données via des API dédiées
- ...

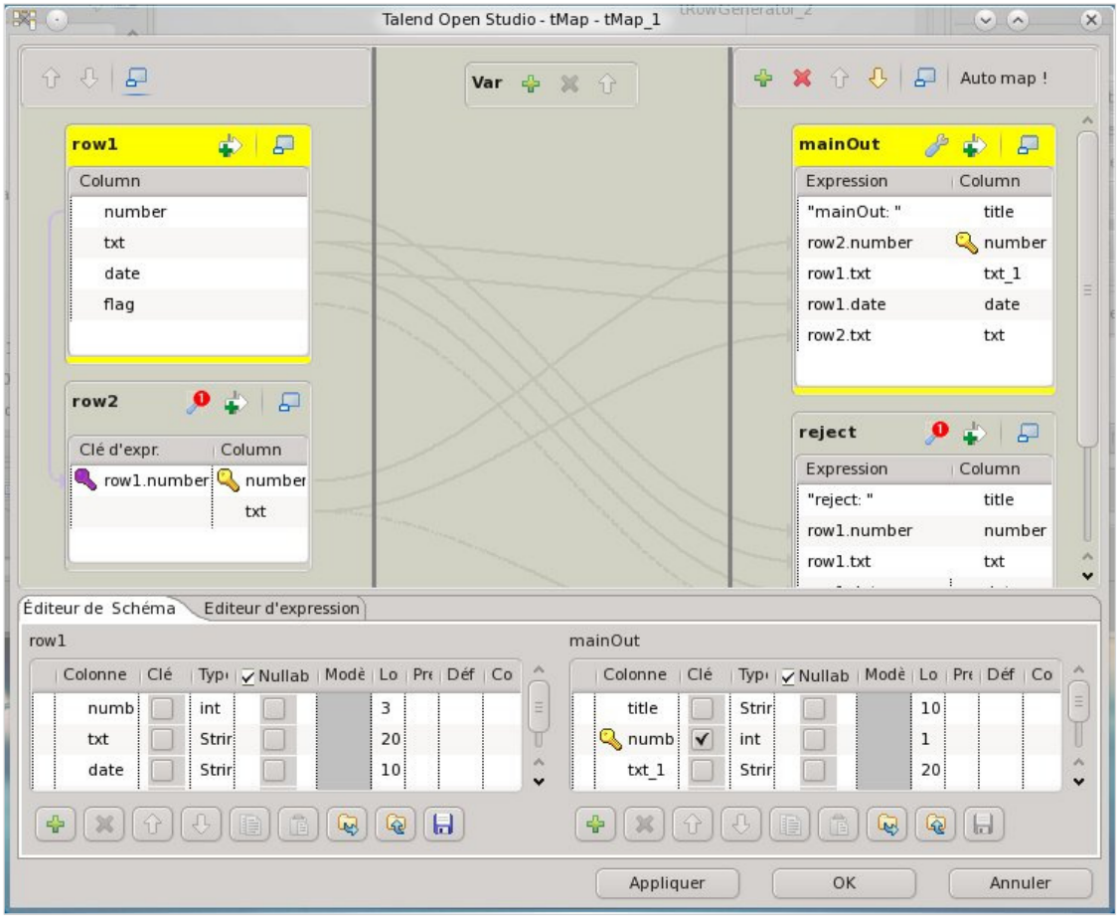
Exemple : Outils d'exploration

Ils permettent de manipuler interactivement des cubes multidimensionnels (choix des dimensions à croiser et des types d'agrégations à effectuer)

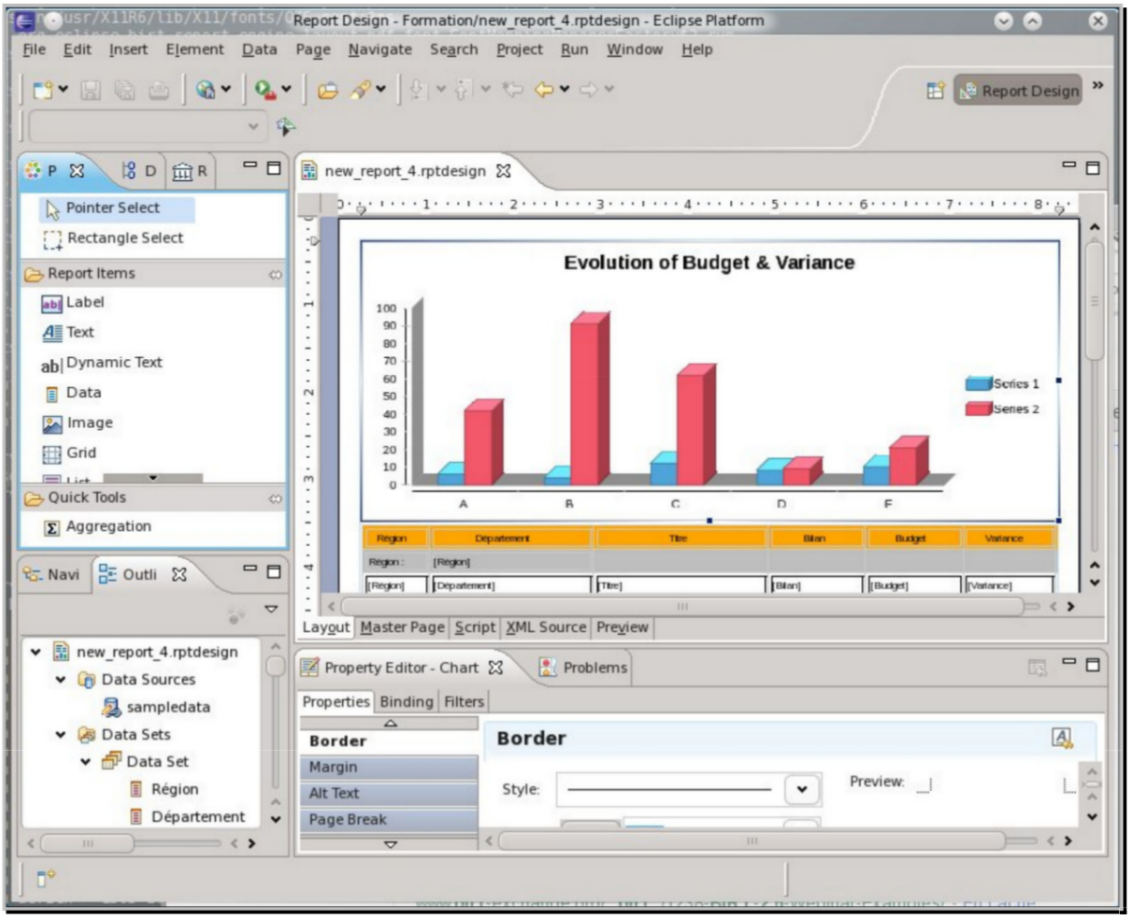
Exemple : Outils d'analyse

Ils permettent l'analyse statistique de données.

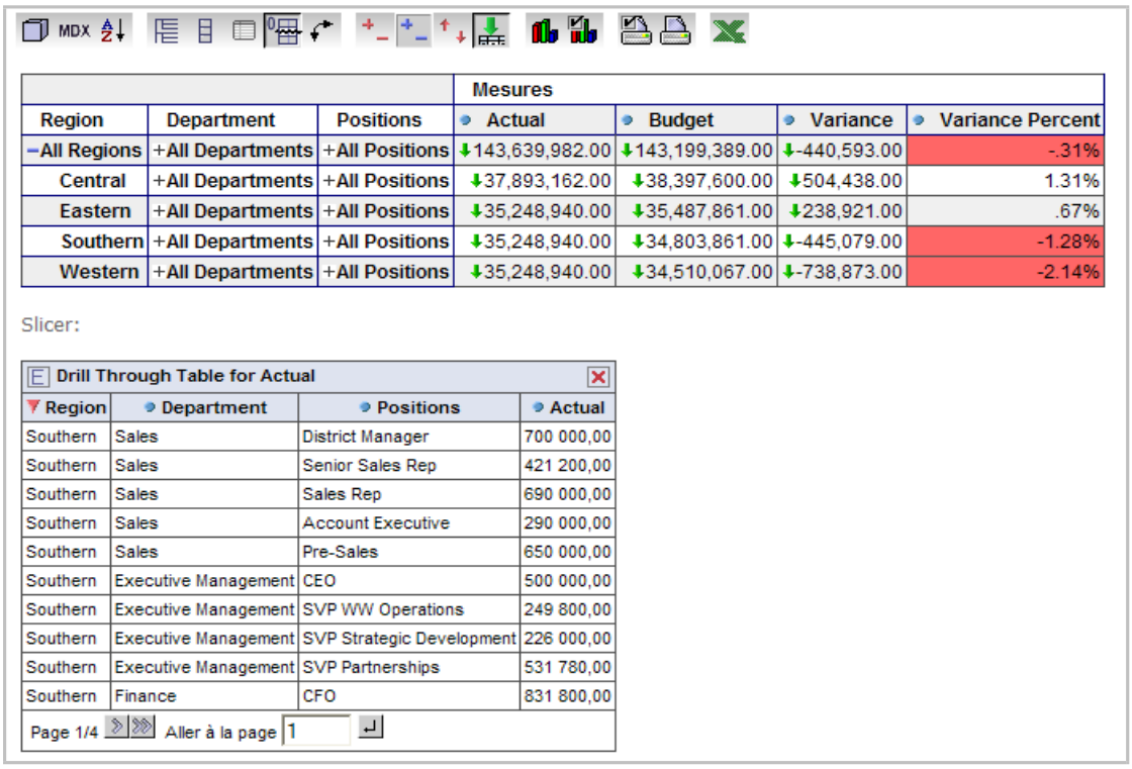
Exemples d'outils Open Source



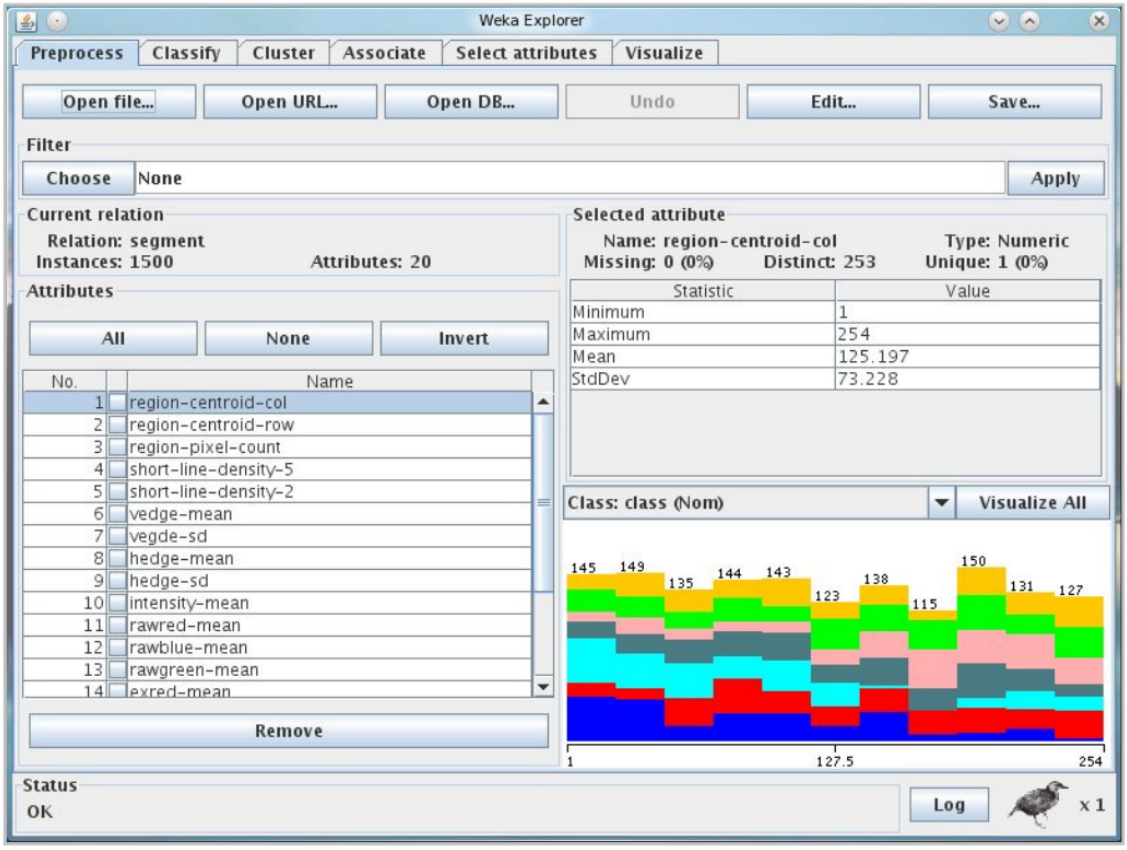
Outil d'ETL Talend



Outil de reporting Birt



Outil d'exploration de données JPivot



Outil d'analyse statistique Weka

2. SGBD orientés décisionnel

Il est possible d'utiliser une base relationnelle classique pour implémenter un entrepôt de données modélisé en étoile (c'est même aujourd'hui encore la forme la plus largement mobilisée).

Il existe également des technologies dédiées (qui peuvent s'appuyer sur des bases relationnelles ou sur des structures de données dédiées).

Le mouvement NoSQL réintègre progressivement des problématiques décisionnelles, reconfigurant petit à petit les approches technologiques liées à ce domaine.

Exemple : Teradata

Teradata est une technologie dédiée aux BD massivement parallèles, c'est à dire capable de faire exécuter une requête par plusieurs machines en parallèle, afin d'en accélérer le traitement. C'est à la fois un SGBD, un OS dédié (Unix) et des machines dédiées.

Complément

Voir Entrepôts de données : guide pratique de modélisation dimensionnelle [*], p.14.

Abréviations

BI : Business Intelligence

DM : Data Mart

DW : Data Warehouse

ETL : Extraction, Transformation, Loading

Bibliographie

Goglin J.-F. (2001, 1998). *La construction du datawarehouse : du datamart au dataweb*. Hermes, 2ème édition.

Inmon W.-H. (2002, 1990). *Building the data warehouse*. Wiley Publishing, 3rd edition.

Kimball R., Ross M. (2003). *Entrepôts de données : guide pratique de modélisation dimensionnelle* . Vuibert.

Kimball R., Ross M., Thornthwaite W., Mundy J., Becker B. (2008, 1998). *The Data Warehouse Lifecycle Toolkit*. Wiley Publishing, second edition.

Webographie

Smile (2012, 2006). *Décisionnel : le meilleur des solutions open source*. http://www.smile.fr/Livres-blancs/ERP-et-decisionnel/Le-decisionnel-open-source.