



Cours 1 : Introduction au domaine du décisionnel et aux data warehouses

Riadh ZAAFRANI

Janvier 2022

2^{ème} année Licence Computer Science - GLSI

1

1

Objectifs du cours

- 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

2

2

Plan

- Le décisionnel
- Le data warehouse
- Le modèle en étoile
- Les outils du décisionnel
- Bibliographie

3

3

Objectifs

- Connaître le **paradigme du décisionnel** (et son articulation avec le paradigme transactionnel)
- Connaître les principaux **domaines d'application des data warehouses**

4

4

Présentation du décisionnel



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 ⁵

5

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

6

6

Enjeux du décisionnel

👉 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...

7

7

Enjeux du décisionnel

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.

8

8

Enjeux du décisionnel



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.

9

9

Enjeux du décisionnel



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 pose 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.

10

10

Enjeux du décisionnel



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, il 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.

11

11

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.

12

12

Éthique et limites des systèmes décisionnels

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.

13

13

Éthique et limites des systèmes décisionnels

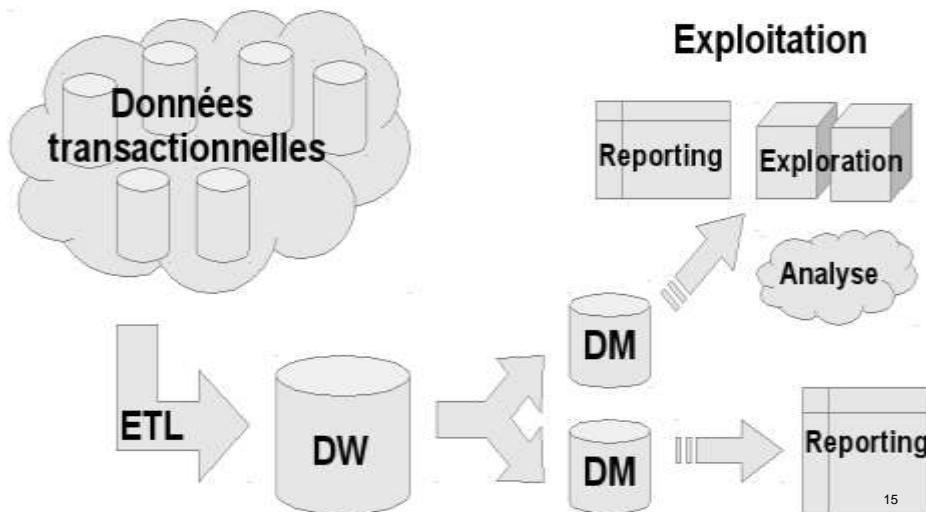
L'interprétation est humaine

- 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 !

14

14

Architecture d'un système décisionnel



15

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

Principe de fonctionnement

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

16

16

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

17

17

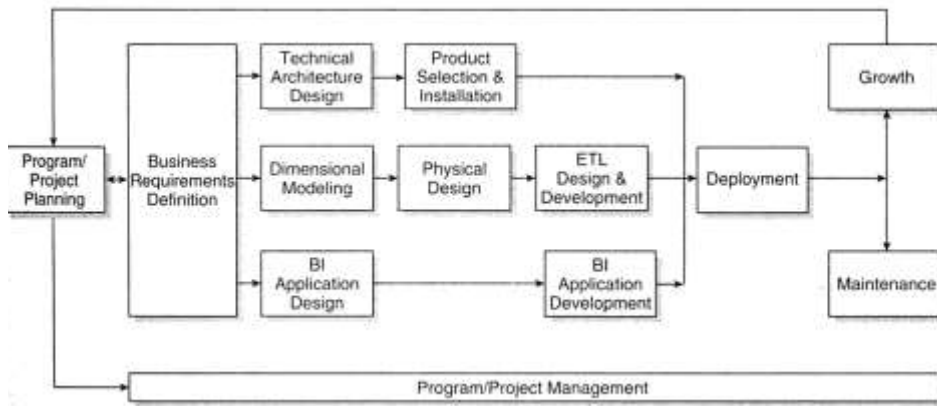
Conception d'un système décisionnel

- 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

18

18

Conception d'un système décisionnel



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

19

19

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 qu'un autre, à quelles conditions ?)

20

20

Quelques exemples d'application

- 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...
- ...

21

21

Plan

- Le décisionnel
- **Le data warehouse**
- Le modèle en étoile
- Les outils du décisionnel
- Bibliographie

22

22

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.

23

23

Présentation du data warehousing



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.

[\(Inmon, 2002, p31\)](#)

24

24

Présentation du data warehousing

 **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

25

Présentation du data warehousing

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).

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

26

26

Différence entre un DW et un système transactionnel

BD transactionnelle

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

- Ajout, modification et 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).

27

27

Différence entre un DW et un système transactionnel

PRIMITIVE DATA/OPERATIONAL DATA

- application oriented
- detailed
- accurate, as of the moment of access
- serves the clerical community
- can be updated
- run repetitively
- requirements for processing understood a priori
- compatible with the SDLC
- performance sensitive
- accessed a unit at a time
- transaction driven
- control of update a major concern in terms of ownership
- high availability
- managed in its entirety
- nonredundancy
- static structure; variable contents
- small amount of data used in a process
- supports day-to-day operations
- high probability of access

DERIVED DATA/DSS DATA

- subject oriented
- summarized, otherwise refined
- represents values over time, snapshots
- serves the managerial community
- is not updated
- run heuristically
- requirements for processing not understood a priori
- completely different life cycle
- performance relaxed
- accessed a set at a time
- analysis driven
- control of update no issue
- relaxed availability
- managed by subsets
- redundancy is a fact of life
- flexible structure
- large amount of data used in a process
- supports managerial needs
- low, modest probability of access

28

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

28

Implémentation du DW avec un SGBDR

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

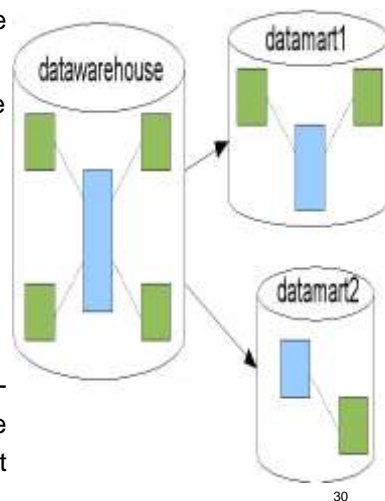
Le caractère **statique** du data warehouse efface les inconvénients de ces techniques lorsqu'elles sont mobilisées dans des systèmes transactionnels.

29

29

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.



30

Data warehouse et data mart

Pourquoi des data marts ?

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



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).

31

Plan

- Le décisionnel
- Le data warehouse
- **Le modèle en étoile**
- Les outils du décisionnel
- Bibliographie

32

32

Objectifs

- Connaître les principes de la **modélisation dimensionnelle**

33

33

Présentation de la modélisation en étoile



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.



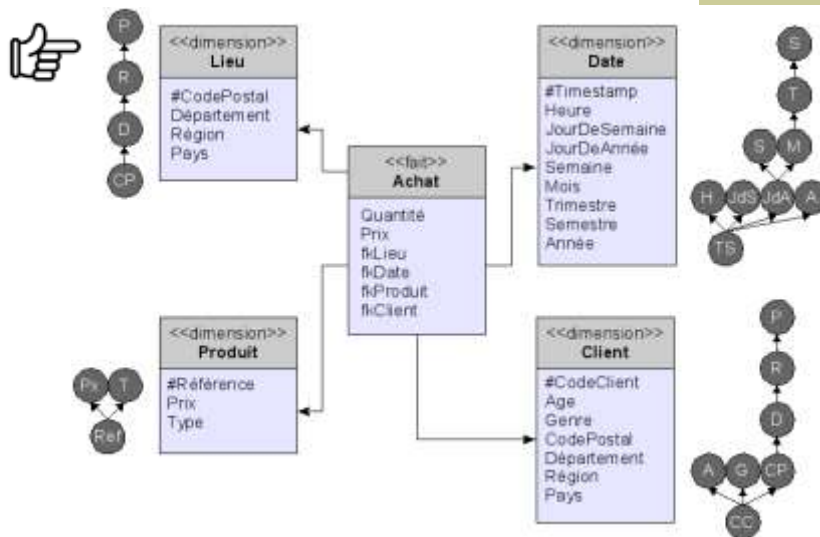
Le modèle en flocon

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.

34

34

Présentation de la modélisation en étoile



35

35

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.

36

36

Objectifs du modèle dimensionnel

Performance

En effet en tant que structures **redondantes** les schémas en étoiles permettent d'agréger la table des faits avec n'importe quelle 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.

37

37

Objectifs du modèle dimensionnel

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.

38

38

Objectifs du modèle dimensionnel

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)

39

39

Extraction Transformation Loading

ETL

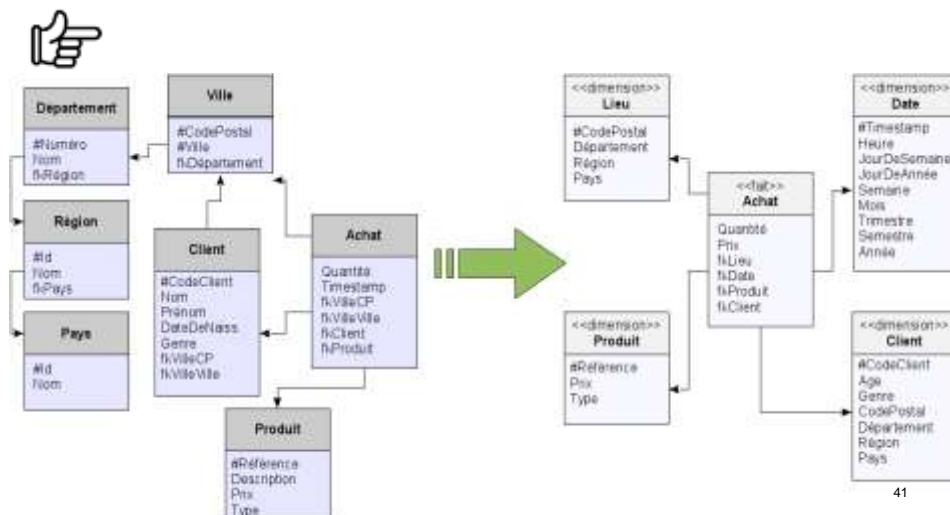
L'ETL (Extraction Transformation Loading)

est le processus de copie des données depuis les tables des systèmes transactionnels vers les tables du modèle en étoile du data warehouse.

40

40

Extraction Transformation Loading



41

Extraction Transformation Loading

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.

42

42

Plan

- Le décisionnel
- Le data warehouse
- Le modèle en étoile
- Les outils du décisionnel
- Bibliographie

43

43

Objectifs

- Connaître les grandes **classes d'outils** du domaine du décisionnel
- Connaître **quelques outils** du marché

44

44

ETL, reporting, exploration, analyse

Principaux types d'outils d'une architecture décisionnel :

- ETL
- *Reporting*
- Exploration
- Analyse



ETL

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

45

45

ETL, reporting, exploration, analyse



Outils de reporting

Ils permettent

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



Outils d'exploration

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



Outils d'analyse

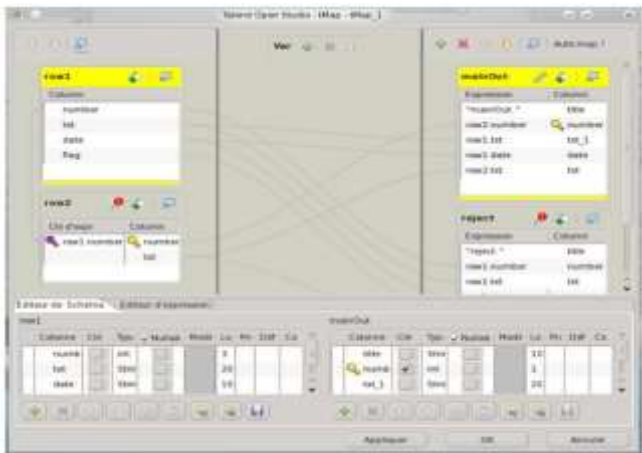
Ils permettent l'analyse statistique de données.

46

46

ETL, reporting, exploration, analyse

Exemples d'outils Open Source



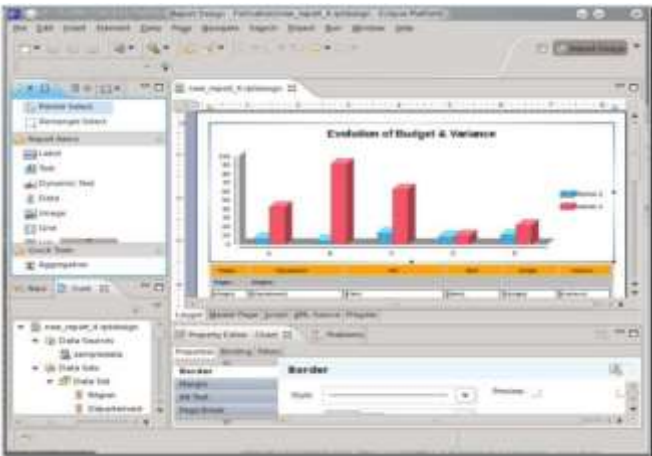
Outil d'ETL Talend

47

47

ETL, reporting, exploration, analyse

Exemples d'outils Open Source



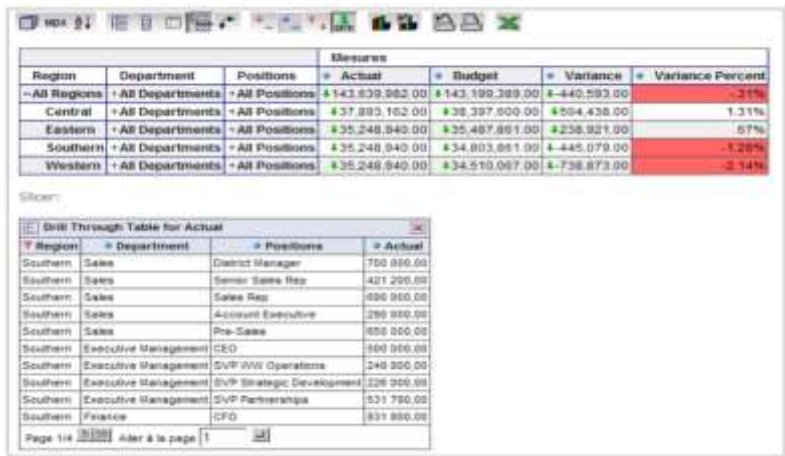
Outil de reporting Birt

48

48

ETL, reporting, exploration, analyse

Exemples d'outils Open Source

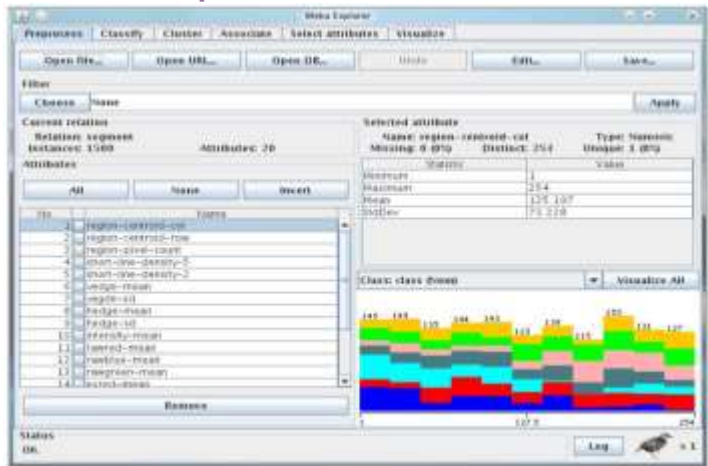


Outil d'exploration de données JPIVOT

49

ETL, reporting, exploration, analyse

Exemples d'outils Open Source



Outil d'analyse statistique Weka

50

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 s'appuient généralement *in fine* sur une base relationnelle).
- Les principes de modélisation dimensionnels sont indépendants de la technologie choisie pour l'implémentation.
- Le mouvement NoSQL réintègre progressivement des problématiques décisionnelles, reconfigurant petit à petit les approches technologiques liées à ce domaine.

51

51

SGBD orientés décisionnel



Bases de données massivement parallèles

SGBD capable de faire exécuter une requête par plusieurs machines en parallèle, afin d'en accélérer le traitement.



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.

- SGBD
- OS dédié (Unix)
- Machines dédiées

52

52

Plan

- Le décisionnel
- Le data warehouse
- Le modèle en étoile
- Les outils du décisionnel
- **Bibliographie**

53

53

Bibliographie & Webographie

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

54

54