UNIVERSITÉ HASSAN II DE CASABLANCA
FACULTÉ DES SCIENCES AÏN CHOCK

Informatique Décisionnelle (BI)

ABDELTIF EL BYED
Professeur à l'université Hassan II – Faculté des sciences Ain Chok (FSAC)
Aelbyed@gmail.com / Abdeltif.elbyed@univh2c.ac.ma
Site: https://sites.google.com/site/aelbyed

Année: 2019-2020

Plan général du Module

Partie1: Informatique Décisionnelle

1. Introduction à l'Informatique Décisionnelle (BI)
2. Entrepôt de données (DW)
3. OLAP: On-Line Analytical Processing
4. Les système OLAP: ROLAP, MOLAP & HOLAP

Partie 2: IBM Cognos Report Studio (En Anglais)

1. Introduction to the Reporting Application
2. Create List Reports
3. Focus Reports Using Filters
4. Create Crosstab Reports
5. Present Data Graphically
6. Focus Reports Using Prompts

Plan général du Module

Partie1: Informatique Décisionnelle

1. Introduction à l'Informatique Décisionnelle (BI)
2. Entrepôt de données (DW)
3. OLAP: On-Line Analytical Processing
4. Les système OLAP: ROLAP, MOLAP & HOLAP

Partie 2: IBM Cognos Report Studio (En Anglais)

1. Introduction to the Reporting Application
2. Create List Reports
3. Focus Reports Using Filters
4. Create Crosstab Reports
5. Present Data Graphically
6. Focus Reports Using Prompts

Partie 3: Data Mining

1. Objectifs de la fouille de données
2. Les algorithemes de data mining
3. Le DW et le data mining
4. IA et l'apprentissage

Plan général du Module

Partie1: Informatique Décisionnelle

1. Introduction à l'Informatique Décisionnelle (BI)
2. **Entrepôt de données (DW)**
3. OLAP: On-Line Analytical Processing
4. Les système OLAP: ROLAP, MOLAP & HOLAP

Partie 2: IBM Cognos Report Studio (En Anglais)

1. Introduction to the Reporting Application
2. Create List Reports
3. Focus Reports Using Filters
4. Create Crosstab Reports
5. Present Data Graphically
6. Focus Reports Using Prompts

UNIVERSITÉ HASSAN II DE CASABLANCA
FACULTÉ DES SCIENCES AÏN CHOCK

Introduction aux Entrepôts de Données (ED)

ABDELTIF EL BYED
Professeur à l'université Hassan II – Faculté des sciences Ain Chok (FSAC)
aelbyed@gmail.com / a.elbyed@fsac.ac.ma

1. Introduction et définition d'un entrepôt de données
2. Architecture fonctionnelle d'un entrepôt
3. Modélisation d'un entrepôt de données
4. Implantation d'un ED
5. Alimentation d'un entrepôt de données
6. Exploitation d'un entrepôt
7. Domaines d'application des entrepôts, « succès stories » …

Plan

1. Introduction et définition d'un entrepôt de données (ED)
2. Architecture fonctionnelle d'un ED
3. Modélisation d'un ED
4. Implantation d'un ED
5. Alimentation d'un ED
6. Exploitation d'un ED
7. Domaines d'application des entrepôts et « succès stories » …

1. Introduction et définition d'un entrepôt de données

1. Définition d'un entrepôt de données
2. Entrepôt de données versus bases de données opérationnelles
3. Entrepôt de données versus infocentre
4. Processus général de construction et exploitation d'un entrepôt

Définition d'un entrepôt de données (Data Warehouse)

- «L'entrepôt de données (ED) est une collection de données
  - ***thématiques,***
  - ***intégrées,***
  - ***non volatiles et***
  - ***historisées***

  organisées pour le support d'un processus d'aide à la décision»

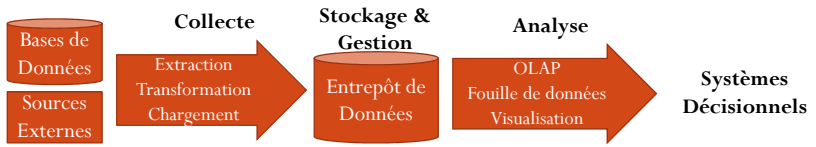
Définition de Inmon (1992)

Caractéristiques des données d'un ED

- ***Orientées sujet*** : un ED rassemble et organise des données associées aux différentes structures fonctionnelles de l'entreprise, pertinentes pour un sujet ou un thème nécessaire aux besoins d'analyse.
- ***Intégrées*** : les données résultent de l'intégration de données provenant de différentes sources pouvant être hétérogènes.
- ***Historisées*** : les données d'un ED représentent l'activité d'une entreprise durant une certaine période (plusieurs années) permettant d'analyser les variations d'une donnée dans le temps.
- ***Non-volatiles*** : les données de l'ED sont essentiellement utilisées en interrogation (consultation) et ne peuvent pas être modifiées (sauf certain cas de rafraîchissement).

De l'entrepôt à l'aide à la décision

- Entreposage des données : avant d'être chargées dans l'entrepôt, les données sélectionnées doivent être :
  - Extraites des sources (internes : BD opérationnelles, externes : BD et fichiers notamment issus du Web)
  - Soigneusement épurées afin d'éliminer des erreurs et réconcilier les différentes sémantiques associées aux sources)
- Exploitation des données de l'ED : systèmes décisionnels
  - A partir d'un ED diverses analyses peuvent être faites:
    - « On-Line Analitical processing » (OLAP); de fouille de données (Data Mining) et de visualisation.
  - Les informations et les connaissances obtenues par exploitation de l'ED ont un impact direct sur les bénéfices de l'entreprise
    - (augmentation des ventes par un marketing plus ciblé, amélioration de la rotation des stocks, …)

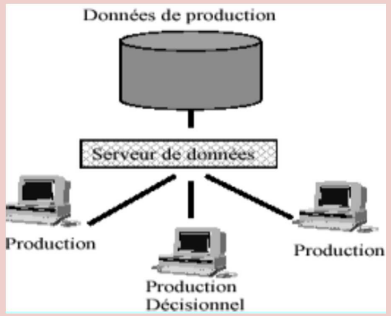


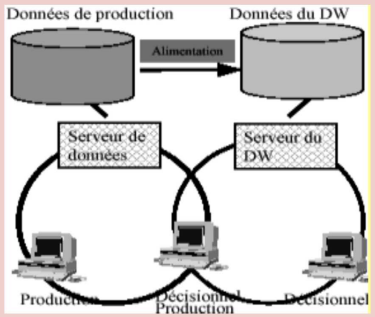
Entrepôt de données versus Bases de données opérationnelles

	BD opérationnelles	Entrepôt de données
Niveau de détail des informations	Très détaillé	Données agrégées, métadonnées
Homogénéité des informations	Information pas nécessairement homogènes	• Informations homogènes • Intégration de données souvent nécessaire
Fonctions de l'entreprise concernées par les données	Données organisées par processus fonctionnel	Données orientées sujet
Comparaison de données sur plusieurs années	Non : Archivage ou mise à jour des données	Oui : Données non volatiles, données historisées
Opérations réalisées sur les données	Consultation, mais surtout mise à jour et ajout de données	Consultation de données uniquement

Entrepôt de données vs Infocentre

Infocentre	Entrepôt
Collection de données orientées sujet, intégrées, volatiles, actuelles, organisées pour le support d'un processus de décision ponctuel	Collection de données orientées sujet, intégrées, ***non volatiles et historisées***, organisées pour le support d'un processus d'aide à la décision



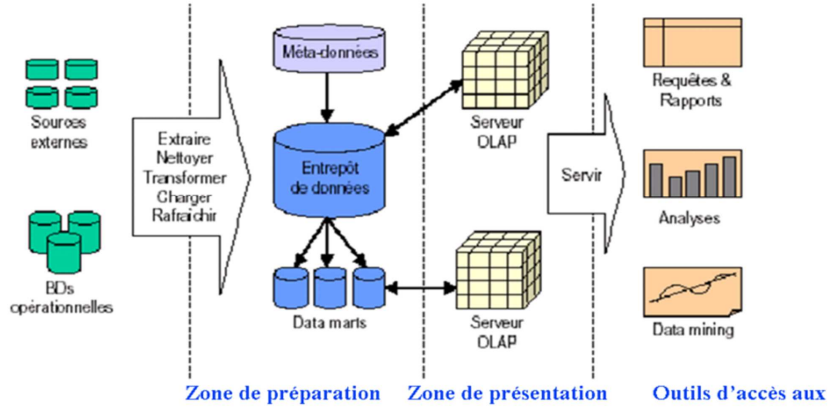


Processus général de construction et exploitation d'un ED

Processus en 3 phases :

1. Construction de la BD décisionnelle :
   - Modélisation conceptuelle des données multiformes et multi-sources
   - Conception de l'entrepôt de données
   - Alimentation de l'entrepôt (extraire, nettoyer, transformer, charger)
   - Stockage physique des données
2. Sélection des données à analyser :
   - Besoins d'analyse de l'utilisateur
   - Magasins de données (Data Marts)
   - Cubes multidimensionnels
   - Tableaux ou tables bidimensionnels
3. Analyse des données :
   - Statistiques et reporting, OLAP, Data Mining

Processus général de construction et exploitation d'un ED

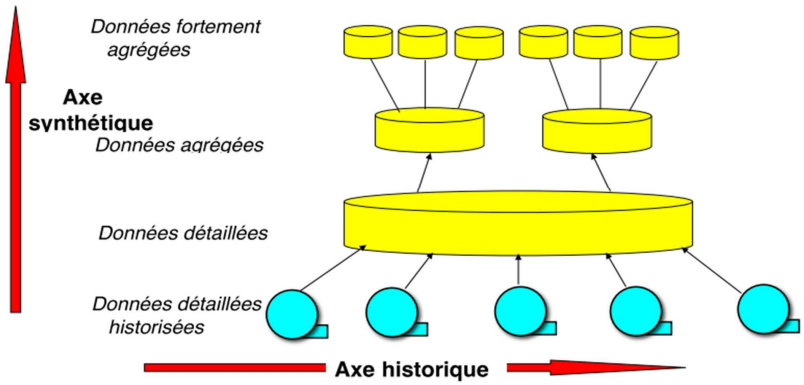


2. Architecture fonctionnelle d'un entrepôt de données

1. Axes historique et synthétique des données
2. Entrepôts de données (ED) et Magasins de données (MD)
3. Architecture fonctionnelle d'un ED
4. Composants logiciels d'un ED

Axes historique et synthétique des données d'un ED (1)

Les données d'un ED se structurent selon 2 axes : synthétique et historique :



Axes historique et synthétique des données d'un ED (1)

Axe synthétique

- Établit une hiérarchie d'agrégation comprenant:
  - Les **données détaillées** représentant les événements les plus récents au bas de la hiérarchie
  - Les **données agrégées** synthétisant les données détaillées
  - Les **données fortement agrégées** synthétisant à un niveau supérieur les données agrégées

Axe historique

- Comprenant les données détaillées historisées représentant les événements passés
  - Nécessaire de stocker des méta-données :
    - informations concernant les données de l'ED (provenance, structure, méthode utilisées pour l'agrégation, …)

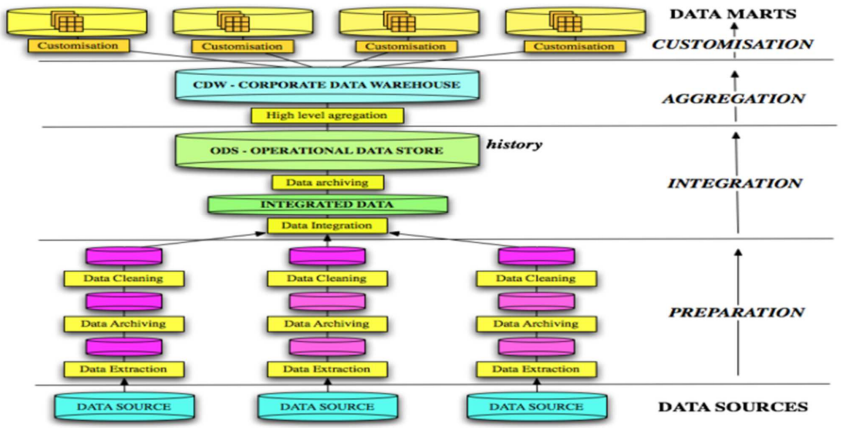
Entrepôt et Magasins de données (1)

- L'entrepôt de données - **ED** (Data Warehouse - **DW**) :
  - Collecte l'ensemble de l'information utile aux décideurs à partir des sources de données (BD opérationnelle, BD externes, Web …)
  - Centralise l'information décisionnelle en assurant l'intégration des données extraites, leur pérennité dans le temps
- Les magasins de données – **MD** (Data Marts - **DM**) :
  - Objectif : supporter efficacement des processus d'analyse de type OLAP
  - Extraire pour chacun une partie de l'information décisionnelle de l'entrepôt d'une partie des données utile :
    - pour une classe d'utilisateurs ou
    - pour un besoin d'analyse spécifique
  - Ils sont orientés sujet

Entrepôts et magasins de données (2)

- Les entrepôts de données (Data Warehous) :
  - Est le lieu de stockage centralisé d'un extrait des bases de production.
  - Nécessitent de puissantes machines pour gérer de très grandes bases de données contenant des données de détail historisées
  - L'organisation des données est faite selon un modèle facilitant la gestion efficace des données et leur historisation.
- Les magasins de données (Data Marts) :
  - Sont de petits entrepôts nécessitant une infrastructure plus légère et sont mis en œuvre plus rapidement (6 mois environs)
  - Conçus pour l'aide à la décision à partir de données extraites d'un ED plus conséquent ou de BD sources existantes
  - Les données extraites sont adaptées pour l'aide à la décision (pour classe de décideurs, usage particulier, recherche de corrélation, logiciel de statistiques,…)
  - L'organisation des données est faite selon un modèle facilitant les traitements décisionnels

Entrepôts et magasins de données (3)



- ODS Operational Data Store : regroupe les données intégrées récupérées des sources
- CDW Corporate Data Warehouse : regroupe les vues agrégées

Architecture fonctionnelle d'un entrepôt : 3 niveaux

Niveau extraction :

- Extraction de données des BD opérationnelles (SGBD traditionnel en OLTP) et de l'extérieur :
  - Approche « push » : détection instantanée des mises à jour sur les BD opérationnelles pour intégration dans l'ED
  - Approche « pull » : détection périodique des mises à jour des BD opérationnelles pour intégration dans l'ED

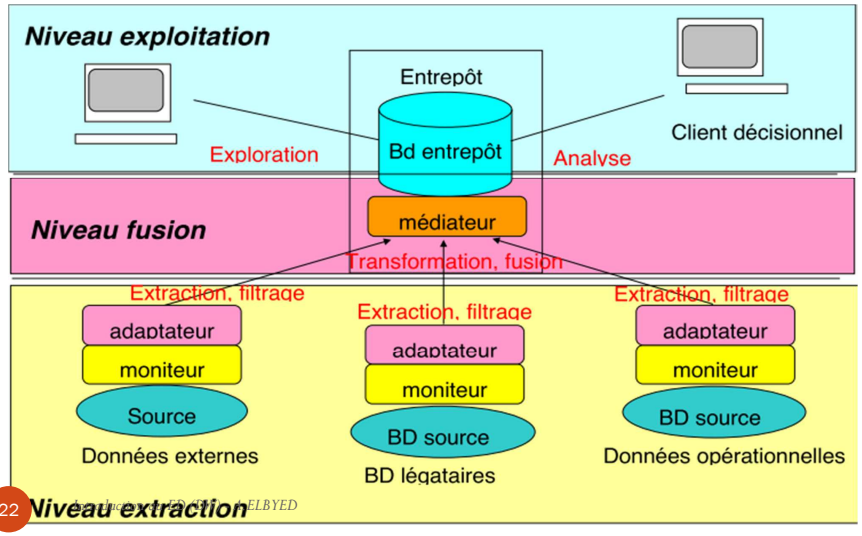
Niveau fusion :

- Intégration, chargement et stockage des données dans la BD entrepôt organisée par sujets
- Rafraîchissement au fur et à mesure des mises à jour

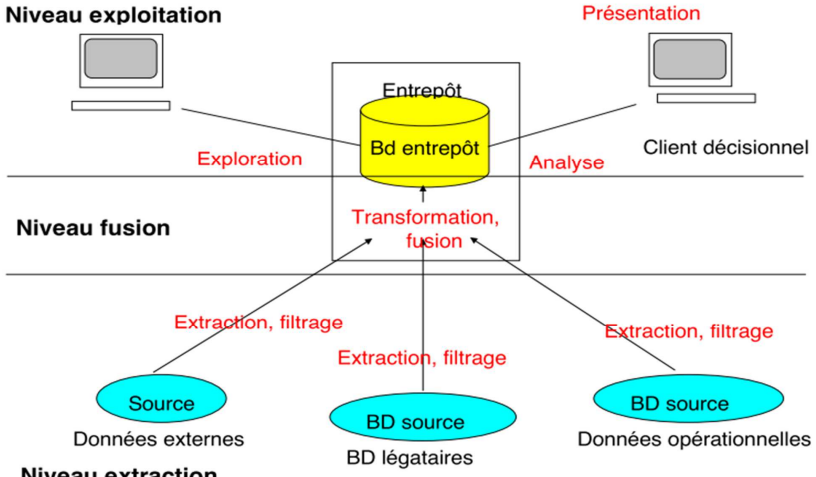
Niveau exploitation :

- Rapports, tableaux de bords, visualisation graphiques diverses, …
- Analyse et l'exploration des données entreposées (OLAP)
- Requêtes complexes pour analyse de tendance, extrapolation, découverte de connaissance, Fouille de données,…

Composants logiciels d'un ED



Architecture fonctionnelle d'un ED : 3 niveaux



Niveau extraction : sources d'informations hétérogènes

- Les données sources alimentant l'ED sont :
  - Généralement modifiées quotidiennement
  - Fortement hétérogènes :
    - issues de différentes sources : BD relationnelles, BD objets, BD réseaux, fichiers (flat files), documents HTML, bases de connaissances,
    - issues de différents environnements

Source d'information	Environnement
gestion commerciale	progiciel sybase/unix
gestion marketing	progiciel SQL server/NT
gestion financière paye	mainframe DB2/IBM
suivi de production	oracle/NT
contrôle qualité	oracle/NT
gestion du temps	progiciel oracle/unix
gestion des stocks	progiciel oracle/HP
fichier mailings	fichier ASCII
références nationales	document excel

➔**Nécessité de composants d'alimentation pour l'homogénéisation et l'intégration de données**

Niveaux extraction : Moniteur et Adaptateur de sources

- ***Le moniteur (source monitor)*** : composant logiciel détectant les mises à jour effectuées sur la source d'information et repérerant les données à envoyer à l'ED pour sa mise à jour ultérieure :
  - Utilisation de Triggers si les SGBD en disposent
  - Sinon interrogation périodique de chaque base locale ou son journal afin de récupérer les mises à jour effectuées durant la dernière période
- ***L'adaptateur de source*** (source wrapper) : composant logiciel traduisant les requêtes et les données depuis le modèle d'une source d'information locale vers le modèle de l'ED et vice-versa :
  - Les bases locales préexistent et sont souvent relationnelles, voire hiérarchiques ou réseaux ou parfois des fichiers

Niveau fusion : Médiateur

- ***Le médiateur (mediator)*** : composant logiciel capable de :
  - Donner une vision intégrée des différentes sources d'information
  - Extraire par des requêtes des parties de ces vues intégrées :
    - avant d'être déversées dans l'ED, les données doivent être nettoyées, transformées, réorganisées et souvent filtrées
    - les données, en provenance de sources multiples, doivent généralement être intégrées ou fusionnées
    - cette fusion en général assurée par union ou jointures de sources multiples, des sélections et agrégats
  - Le médiateur s'appuie principalement sur le SGBD de l'ED

Niveau exploitation : Moteur OLAP et Outils de fouille

- ***Moteur OLAP*** : composant logiciel permettant de:
  - Exécuter des requêtes interactives complexes
  - Analyser interactivement les données selon des axes d'analyse et des niveaux de détail particuliers :
    - changement de points de vue, de niveau de détail
  - Visualiser des résultats de ces analyses
  - Effectuer les opérations OLTP classiques
- ***Outils de fouille de données (Data Mining)*** : composants logiciels permettant de:
  - Extraire automatique les propriétés cachées
  - Extraction automatique de connaissances :
    - connaissances valides, nouvelles, compréhensibles, pertinentes, implicites, …

Dictionnaire et méta-données

- Le dictionnaire contient des informations (méta données) sur :
  - Toutes les données de l'ED.
  - Chaque étape lors de la construction de l'ED;
  - Le passage d'un niveau de données à un autre lors de l'exploitation de l'ED
- Le rôle de ces méta-données est:
  - La définition des données
  - La fabrication des données
  - Le stockage des données
  - L'accès aux données
  - La présentation des données

3 - Modélisation d'un entrepôt de données

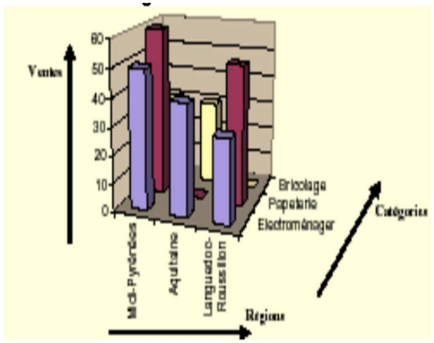
1. Problématique de la modélisation multidimensionnelle
2. Concept de fait
3. Concept de dimension
4. Paramètres d'hiérarchies de dimension
5. Opérations OLAP

Problématique de la modélisation multidimensionnelle

- Les analyses décisionnelles (OLAP) sont directement reliées à une modélisation de l'information conceptuelle :
  - Proche de la perception de l'analyste
  - Basée sur une vision multidimensionnelle des données
- La modélisation multidimensionnelle :
  - Considère un sujet analysé comme un point dans un espace à plusieurs dimensions
  - Les données y sont organisées de façon à mettre en évidence le sujet analysé et les différentes perspectives de l'analyse

Modélisation multidimensionnelle (1)

- Soit les données relatives aux ventes de 1999 d'une entreprise de distribution :



Catégories des produits	Régions	Montant des ventes
Electroménager	Midi-Pyrénées	50
Electroménager	Aquitaine	40
Electroménager	Languedoc-Roussillon	30
Papeterie	Midi-Pyrénées	60
Papeterie	Languedoc-Roussillon	50
Bricolage	Midi-Pyrénées	30
Bricolage	Aquitaine	30

On peut distinguer différentes perspectives pour observer ces données :

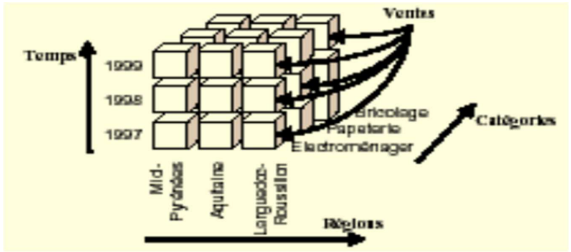
- Une dimension relative à la catégorie des produits
- Une dimension relative à la région

Modélisation multidimensionnelle (2)

Considérons plusieurs tables des ventes de chaque année entre 1997 et 1999, on peut alors observer les données dans un espace à 3 dimensions :

- La dimension catégories produit
- La dimension régions
- La dimension temps

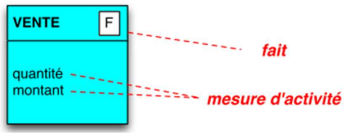
Chaque intersection de ces dimensions représente une cellule comportant le montant des ventes :



La modélisation multidimensionnelle a donné naissance aux concepts de fait et de dimension [Kimball 1996]

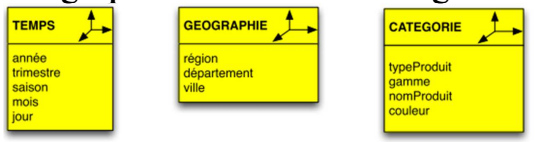
Concept de fait

- Un ***fait*** :
  - Modélise le ***sujet*** de l'analyse
  - Est formé de ***mesures*** correspondant aux informations de l'activité analysée.
  - Ces mesures sont ***numériques*** et généralement ***valorisées de façon continue***
    - on peut les ***additionner***, les ***dénombrer*** ou bien ***calculer*** le minimum, le maximum ou la moyenne.
- Exemple : le fait de « Vente » peut être constitué des mesures d'activités suivantes :
  - Quantité de produits vendus et
  - Montant total des ventes



Concept de dimension

- Le sujet analysé, le fait, est analysé suivant différentes perspectives ou axes caractérisant ses mesures de l'activité : on parle de dimensions.
- Une ***dimension*** :
  - Modélise un axe d'analyse
  - Se compose de paramètres correspondant aux informations faisant varier les mesures de l'activité.
- Ex: Le fait « Vente » peut être analysé suivant différentes perspectives correspondant à trois dimensions : la dimension **Temps**, la dimension **Geographie** et la dimension **Categorie** :

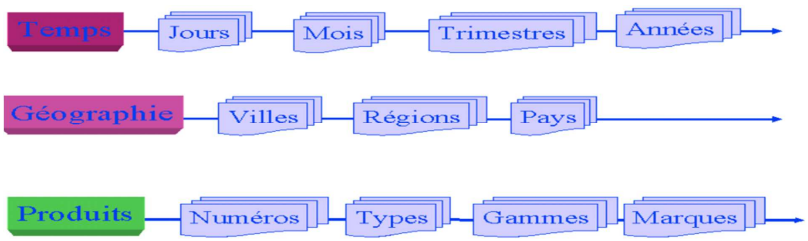


Agrégation des données

- Plusieurs niveaux d'agrégation
  - Les données peuvent être groupées à différents niveaux de granularité
  - Les regroupements sont pré-calculés,
    - Par exemple, le total des ventes d'un mois est calculé à partir de la somme de toutes les ventes du mois.
- Granularité = niveau de détail des données emmagasinées dans un Data Warehouse

Hiérarchie des paramètres d'une dimension

- En OLAP les mesures d'un fait sont généralement analysées selon les dimensions qui le caractérisent
- Nécessaire de définir pour chaque dimension ses différents niveaux de détail définissant ainsi une (ou plusieurs) hiérarchie(s) de paramètres
- La ***hiérarchie*** de paramètre d'une dimension :
  - Définis des niveaux de détail de l'analyse sur cette dimension



4 - Implantation d'un entrepôt de données

1. Stratégies d'implantation d'un ED
2. Schéma en étoile (star schema)
3. Schéma en flocon (snowflake schema)
4. Schéma en constellation (fact constellation schema)

Les trois stratégies d'implantation d'un ED

1. Usage d'un SGBD Relationnel (systèmes ***ROLAP***)
   - Les SGBDR représentant plus de 80% des SGBD : ils sont principalement envisagés pour le développement d'ED
   - Ils doivent cependant être adaptés car ils n'ont pas les caractéristiques adéquates pour répondre aux besoins des ED.
2. Usage d'un SGBD Multidimensionnel (systèmes ***MOLAP***)
   - Un SGBD Multidimensionnel (SGBDM) est un SGBD capable de stocker et traiter des données multidimensionnelles
   - A ce jour pas encore de cadre technologique commun pour le développement de tels systèmes : chaque produit est spécifique
3. Usage d'un SGBD Hybride (systèmes ***HOLAP***)
   - Tire profit des avantages des technologies ROLAP et MOLAP :
     - un ROLAP pour stocker, gérer les données détaillées ET
     - un MOLAP pour stocker, gérer les données agrégées

NIVEAU LOGIQUE: ROLAP

- OLAP Relationnel
- Données obtenues à partir de tables relationnelles et de jointures entre celles-ci.
- En fonction de la granularité, la requête générée est plus ou moins complexe.
- A chaque consultation, la requête est recalculée.
  - Les résultats ne sont pas stockées .
- Langage utilisé: SQL .

***Avantages*:**

- Faible cout (car il tire partie des ressources existantes).

***Inconvénients* :**

- Temps de réponse long
  - sollicitation de la base à chaque relance d'un rapport

NIVEAU LOGIQUE: MOLAP

- OLAP Multidimensionnel.
- Données stockées dans une base de données multidimensionnelle appelée CUBE .
- Tous les croisements possibles sont pré-calculés (Restitution des données instantanée).
- Langage utilisé: MDX

***Avantages*:**

- Temps de réponse très court
  - toutes les données et résultats sont stockés.

***Inconvénients* :**

- Coût élevé des licences pour les bases multidimensionnelles.
- Coût élevé de développement des cubes .
- Difficiles à mettre en place pour les gros volumes de données, à cause de tous les résultats précompilés.

NIVEAU LOGIQUE: HOLAP

- Association du ROLAP et du MOLAP
- Concept de Drill-through.
  - Accès aux données agrégées avec MOLAP (Cube).
  - Accès aux détails avec le ROLAP (tables relationnelles).

Avantages:

- Temps de réponse assez court
- Moins couteux que MOLAP car moins développé.

Inconvénients :

- Ne pourra pas être utilisé si les rapports sont trop complexes et font trop de croisement de données.

Schéma d'un entrepôt de données

Niveau logique « ROLAP » :

- 3 grands types de schémas :
  - schéma en ***étoile*** (star schema)
  - schéma en ***flocon*** (snowflake schema)
  - schéma en ***constellation*** (fact constellation)

- Le schéma en étoile est souvent utilisé pour l'implantation physique

Schéma en étoile (1)

Caractéristiques :

- Structure simple
- Une table centrale : la table des faits :
  - objets de l'analyse
  - taille très importante
  - nombreux champs
- Des table périphériques : les tables de dimensions :
  - dimensions de l'analyse
  - taille peu importante
  - peu de champs

Schéma en étoile (2)

- Ex 1 : Vente de médicaments dans des pharmacies

Schéma en étoile modélisant les analyses des **quantités** et des **montants** des médicaments dans les pharmacies selon 3 dimensions : le **temps**, la **catégorie** et la situation **géographique**

- Table de faits : Vente
- Tables de dimension : Temps, Catégorie, Géographie

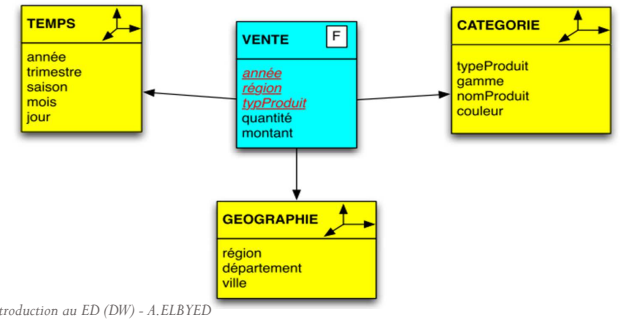


Schéma en étoile (3)

- Ex 2 : Ventes d'articles dans un supermarché

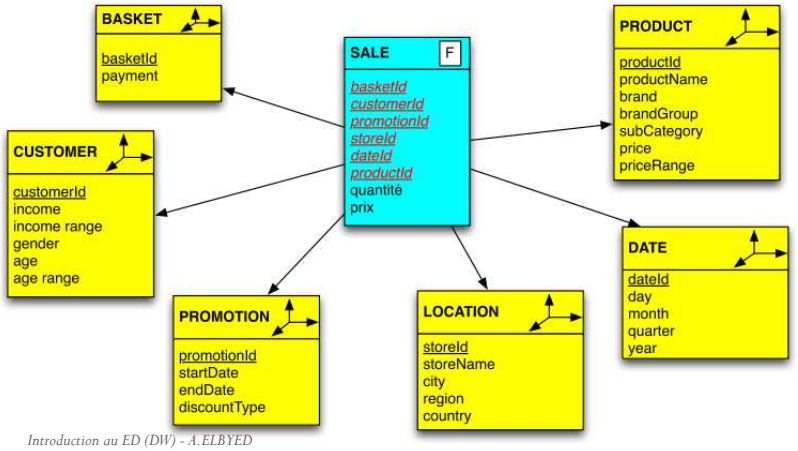


Schéma en étoile (4)

- Associé à Ex 2 :
- Un fait :
  - il a été acheté 3 exemplaires à 1 euro (SALE)
  - du produit pid3
  - par le client cid1
  - à la date did3
  - dans le magasin mid2 (store)
  - dans le chariot cid8 (basket)
  - correspondant à la promotion prid1
- Un élément de la dimension location :
  - store id mid2
  - store name rondpoint
  - city blois
  - région centre
  - country France

Schéma en étoile (5)

- Normalisation de la table de faits :
  - normalisation en Boyce-Codd Normal Form (BCNF)
  - Rappel : une relation R est en BCNF si :

$\forall x \rightarrow y$ DF définie sur R , x contient une clé de R soit : chaque attribut non clé dépend fonctionnellement de la seule clé de la relation

- Normalisation des tables de dimensions :
  - elles représentent une ou plusieurs hiérarchies
  - elles contiennent des données redondantes
  - faut-il les normaliser ?
    - la table des faits constitue l'essentiel du stockage
    - pas/peu de mises à jour des dimensions
    - la perte d'espace n'est donc pas significative

➔ tables de dimensions : NON normalisées

Schéma en flocon (1)

Un modèle en flocon : une évolution du schéma en étoile avec :

- Une décomposition des dimensions du modèle en étoile en sous hiérarchies.
- le fait est conservé et les dimensions sont éclatées conformément à sa hiérarchie des paramètres
- Cela conduit à une normalisation des tables de dimensions :
  - structure hiérarchique des dimensions
  - un niveau inférieur identifie un niveau supérieur

Avantage

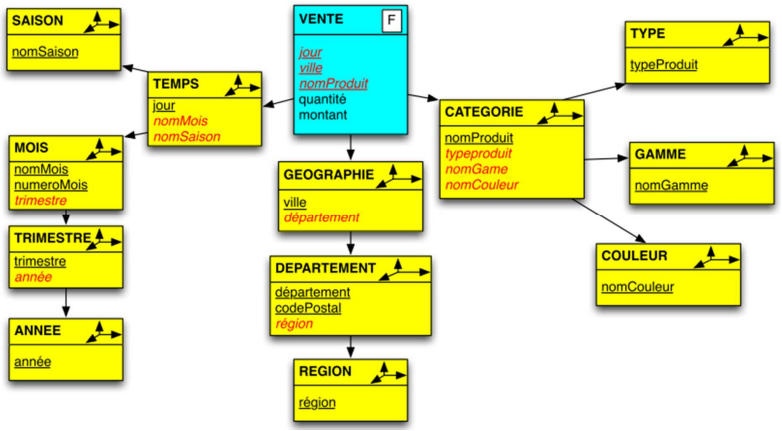
- Formaliser une hiérarchie au sein d'une dimension.
- Maintenance des tables de dimensions simplifiée
- Réduction de la redondance

Inconvénient

- Induit une dé-normalisation des dimensions générant une plus grande complexité en termes de lisibilité et de gestion.
- Navigation coûteuse
  - Beaucoup de jointure!

Schéma en flocon (2)

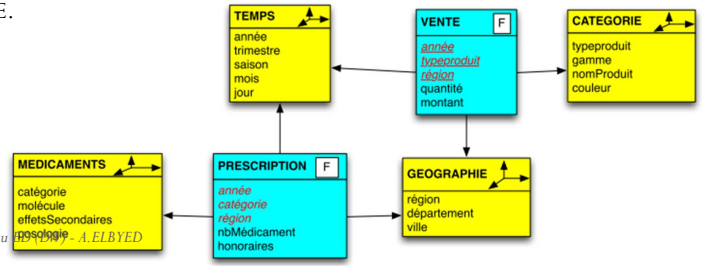
- Ex 3: Vente de médicaments dans des pharmacies



Chaque dimension du schéma en étoile précédent est dénormalisée

Schéma en constellation

- Un modèle en constellation :
  - Fusionne plusieurs modèles en étoile qui utilisent des dimensions communes.
  - Comprend en conséquence plusieurs faits et des dimensions communes ou non
- Ex : Vente de médicaments dans des pharmacies
  - Une constellation est constituée de 2 schémas en étoile :
    - l'un correspond aux ***VENTEs*** effectuées dans les pharmacies et
    - l'autre analyse les ***PRESCRIPTIONs*** des médecins
  - les dimensions ***Temps*** et ***Géographie*** sont partagées par les faits PRESCRIPTION et VENTE.



Pré-agrégation

- Agrégation des faits selon une ou plusieurs dimensions
- Deux moyens de les représenter :
  1. Une table des faits séparés/dédiés avec les tables pour les dimensions correspondantes
  2. Dans la même table des faits, en codant les niveaux hiérarchiques dans les tables de dimensions

5 - Alimentation d'un Entrepôt de données

1. Processus général d'alimentation d'un ED
2. Préparation des données
3. Intégration des données
4. Agrégation des données
5. Personnalisation des données (customisation)

Processus d'alimentation d'un ED

- Le processus d'alimentation d'un ED (ou entreposage des données) consiste à :
  - Rassembler de multiples données sources souvent hétérogènes
  - Les homogénéiser
- Homogénéisation faite selon des règles précises . Ces règles :
  - Sont mémorisées sous forme de méta-données (information sur les données) stockées dans le dictionnaire de données
  - Permettent d'assurer des tâches d'administration et de gestion des donnés entreposées.

Processus d'alimentation d'un ED

- Après avoir conçu le modèle des données, comment alimenter l'ED ?

➔Problématique de l'ETL (Extracting Transforming and Loading)

- ETL se fait en quatre étapes :
  1. Sélection des données sources
  2. Extraction des données
  3. Nettoyage et Transformation
  4. Chargement

ETL tools (Extract Transform Load)

- Support et/ou automatisation des tâches suivantes :

TACHES	SUPPORT
Extraction	accès aux différentes sources
Nettoyage	recherche et résolution des inconsistances dans les sources
Transformation	entre différents formats, langages, etc.
Chargement	des données dans l'entrepôt
Réplication	des sources dans l'entrepôt
Analyse	Ex : détection de valeurs non valides ou inattendues
Transfert de données haut débit	pour les très grands entrepôts
Test de qualité	Ex : pour correction et complétude
Analyse des méta données	aide à la conception

1. Tâche de sélection des données sources

- Quelles données de production faut-il sélectionner pour alimenter l'ED?
  - Toutes les données sources ne sont pas forcément utiles
    - Ex : Doit-on prendre l'adresse complète ou séparer le code postal ?
  - Les données sélectionnées seront réorganisées pour devenir des informations.
- La synthèse de ces données sources a pour but de les enrichir.
- La dénormalisation des données crée des liens entre les données et permet des accès différents.

2. Tâche d'Extraction des données

- Un extracteur (wrapper) est associé à chaque source de données :
  - Il sélectionne et extrait les données
  - Il les formate dans un format cible commun
  - Utilisation d'interfaces comme ODB, OCI, JDBC.
  - Le format cible est en général le modèle Relationnel

3. Tâche de Nettoyage et Transformation des données

- Objectifs du nettoyage :
  - Résoudre le problème de consistance des données au sein de chaque source
  - Une centaine de type d'inconsistances ont été répertoriées
  - 5 à 30 % des données des BD commerciales sont erronées
- Types d'inconsistances :
  - Présence de données fausses dés leur saisie :
    - fautes de frappe
    - différents formats dans une même colonne
    - texte masquant de l'information (e.g., "N/A")
    - valeur nulle
    - incompatibilité entre la valeur et la description de la colonne
    - duplication d'information, …
  - Persistance de données obsolètes
  - Confrontation de données sémantiquement équivalentes mais syntaxiquement différentes.
    - Ex, synonyme, hyperonyme, hyponyme

3. Tâche de Nettoyage et Transformation des données

- **Les problèmes sémantique des Données**

	Sémantique	Syntaxe	Exemple
Synonymes	Même	différent	
Antonymes	Contraire		Bon - mauvais
Homonymes homophones	différent	Différent, mais se prononcent de la même façon	ver, vair, vert, ver, vers
homonymes homographes	différent	Même	(Un) enseigne (Officier de la marine), (une) enseigne (Marque ou société)
Hyperonymes	inclut	différent	**siège** inclut le sens des mots: chaise, fauteuil, banc, banquette, pouf, canapé
Hyponymes	L'inverse de l'**Hyperonymes**		

3. Tâche de Nettoyage des données (2)

- Fonctions de ***normalisation***
- Fonctions de ***conversion***
- Usage de ***dictionnaires*** de synonymes ou d'abréviations
- Définition de table de règles

valeur source	Valeur cible
Mr	M
Masculin	M
monsieur	M
Msieur	M

3. Tâche de Transformation des données

- Objectifs : Suppression des incohérences sémantiques entre les sources pouvant survenir lors de l'intégration :
  - Des schémas :
    - problème de modélisation : différents modèles de données sont utilisés
    - problèmes de terminologie : un objet est désigné par 2 noms différents, un même nom désigne 2 objets différents
    - incompatibilités de contraintes : 2 concepts équivalents ont des contraintes incompatibles
    - conflit sémantique : choix de différents niveaux d'abstraction pour un même concept
    - conflits de structures : choix de différentes propriétés pour un même concept
    - conflits de représentation : 2 représentations différentes choisies pour les mêmes propriétés d'un même objet
  - Des données :
    - Equivalence de champs
    - Equivalence d'enregistrements : fusion d'enregistrements

4. Tâche de Chargement des données

- Objectif : charger les données nettoyées et préparées dans l'ED
- C'est une opération :
  - Qui risque d'être assez longue
  - Plutôt mécanique et la moins complexe.
- Il est nécessaire de définir et mettre en place :
  - Des stratégies pour assurer de bonnes conditions à sa réalisation
  - Une politique de rafraîchissement automatique ou semi-automatique.

6 – Exploitation d'un entrepôt de données

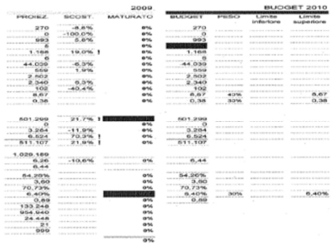
1. Stratégies d'implantation d'un ED
2. Exploitation d'un ED
3. Visualisation autour d'un ED

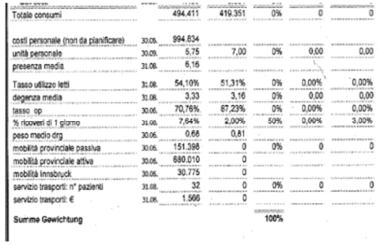
Principales applications autour d'un ED

- Réalisation de rapports divers (***Reporting***)
- Réalisation de tableaux de bords (***Dashboards***)
- Analyse en ligne diverses (***OLAP***)
- Fouille de données (***Data Mining***)
- Visualisations autour d'un ED (***Visualizations***)
- Etc.

Rapports (Reporting)

- Ils sont crées pour les utilisateurs qui ont besoin d'un accès régulier à des informations d'une manière presque statique
  - Ex: les hôpitaux doivent envoyer des rapports mensuels à des agences nationales
- Un rapport est défini par une requête (plusieurs requêtes) et une mise en page (diagrammes, histogrammes, etc)
- Les rapports peuvent être exécutés automatiquement ou manuellement





Tableaux de bords (Dashboards)

- Affichent une quantité limitée d'informations dans un format graphique facile à lire
- Fréquemment utilisé par les cadres supérieurs qui ont besoin d'un rapide aperçu des changements les plus importants
  - Ex : un aperçu en temps réel d'évolutions
- Pas vraiment utile pour une analyse complexe et détaillée

Analyse OLAP (On-Line Analytical processing)

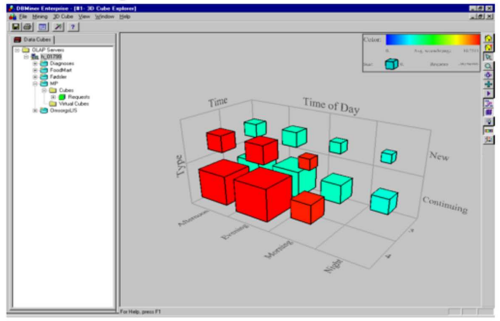
- Techniques OLAP apparues en recherche dans les années 70 mais ont été développées dans les années 90 dans l'industrie
- Permettent de réaliser des synthèses, des analyses et de la consolidation dynamique de données multidimensionnelles
- Constitue la façon la plus naturelle d'exploiter un ED du fait de son organisation multidimensionnelle

Fouille de données (Data Mining)

- Recherche de connaissance, sous forme de modèle de comportement, cachés dans les données
- Domaine jeune à l'intersection de l'Intelligence Artificielle, les Statistiques et les BD
- Nombreuses techniques de fouille :
  - Régression linéaire, induction d'arbres de décision, algorithmes génériques, réseaux de neurones, …
- Les techniques de fouille sont en pleine évolution et sont de plus en plus intégrées dans les ED

Visualisation autour d'un ED

- Facilitent l'analyse et l'interprétation de données
- Convertissent des données complexes en images, graphiques en 2 et 3 dimensions, voire en animations
- Sont de plus en plus intégrées dans les ED



7. Domaines d'application des entrepôts et « succès stories »

1. Les domaines privilégiés :
   - Domaine bancaire
   - Domaine de la grande distribution
   - Domaine des télécommunications
   - Domaines de l'assurance et de la pharmacie
   - Domaine de la santé, …
2. « Succès stories »:
   - Casino, Walmart, Camaieu, …
   - FranceTélécom, …

Domaines privilégiés : Bancaire

- Domaine bancaire : un des premiers utilisateurs des ED
- Pour une banque, il est important de pouvoir regrouper les informations relatives à un client afin de répondre à ses demandes de crédit par exemple
- Des mailing ciblés doivent aussi être rapidement élaborés à partir de toutes les informations disponibles sur un client lors de la commercialisation d'un nouveau produit
- L'utilisation de cartes de crédit nécessite des contrôles à posteriori.
  - Par exemple pour la recherche de fraudes : la mémorisation des mouvements peut rendre de grands services
- Les échanges d'actions et de conseils de courtages sont facilités par une mémorisation de l'histoire et une exploitation par des outils décisionnels avancés
  - par exemple pour déterminer des tendances de marchés

Domaines privilégiés : Grande distribution

- Domaine de la grande distribution fortement demandeur d'ED
- Intéressant de regrouper les informations de ventes pour déterminer les produits à succès, mieux suivre les modes, détecter les habitudes d'achats, les préférences des clients par secteur géographique
- La fouille de données (Data Mining) a permis de développer des techniques sophistiquées d'exploitation de données qui aident à mettre en évidence les règles de consommation
- Explorer le panier de la ménagère est devenu un exercice d'école :
  - il s'agit de trouver à partir de l'enregistrement des transactions quelles sont les habitudes d'achats, plus précisément quels sont les produits achetés en même temps
- Apports constatés dans la grande distribution :
  - Augmentation des ventes grâce à un meilleur marketing
  - Amélioration des taux de rotation de stocks
  - Élimination des produits obsolètes
  - Réduction des rabais, remises, ristournes
  - Meilleure négociation des achats

Domaines privilégiés : Télécommunications

- Domaine très concurrentiel des télécommunications : utilise beaucoup les ED
- Grande masse de données concernant les abonnés et les appels est enregistrée
- Description détaillée des appels comprenant,
  - pour chaque appel appelant, appelé, heure et durée sont disponibles chez les opérateurs
- L'exploitation de ces données regroupées en ED par des techniques d'analyse et d'exploration permet :
  - D'analyser le trafic
  - De mieux cerner les besoins des clients,
  - De classer les clients par catégories,
  - De comprendre pourquoi certains changent d'opérateurs et mieux répondre à leur besoins

Domaines privilégiés : Assurance et Pharmacie

- Domaines de l'assurance et de la pharmacie : très friands de techniques décisionnelles
- L'exercice de base de l'assureur est de déterminer le facteur de risque d'un assuré
- Celui d'un producteur pharmaceutique est de détecter l'impact d'un médicament
- Plus généralement, le suivi des informations relatives à la liaison produit- client sur un ED est souvent synonyme de gains importants :
  - Ex. meilleure connaissance des produits, détection des défauts, meilleure connaissance des clients, détection de rejets, ciblage du marketing, etc
- Le couplage aux technologies du Web ouvre aussi des horizons nouveaux pour le suivi des produits, des clients, des concurrents :
  - notion émergente de « ***Data Webhouse*** »

Succès story : Grande distribution (1)

Exemple du groupe Casino :

- Projet :
  - Un des premiers entrepôts en France
  - Plusieurs millions de dollars économisés en s'apercevant que les stocks de coca-cola faisaient souvent défaut…
  - 1994 : 80 Go et 50 utilisateurs
  - 2002 : + de 10 To, 1500 utilisateurs, 25000 requêtes/jour
- Solution : Teradata

Exemple du groupe Walmart :

- Projet :
  - Le plus gros entrepôt de données du monde, en 2006 : 0.5 Po de données
  - Distributeurs, magasins, clients (> 108), produits (> 109)…
  - Un des plus secret également…
- Solution :
  - Teradata

« Wal-Mart, for example, discovered that people who buy Pampers often buy beer, so they moved Pampers and beer close together. The result was that sales of both increased (Computer Business Review) »

Succès story: Grande distribution (2)

Exemple du groupe Camaieu

- Projet :
  - plusieurs systèmes de production (magasin, logistique, comptable, etc.)
- Solution :
  - 1996 : agrégés dans un entrepôt de données, via l'ETL Sunopsis
  - Base Oracle découpée en référentiels métier (datamarts achat, marketing…)
  - Consultation des datamarts via le système de reporting de Business Objects
  - 2003 : ajout d'un cube OLAP intégré à la base relationnelle Oracle9i :
    - Meilleure ergonomie,
    - permet des requêtes complexes avec prise en compte de plusieurs niveaux au sein de la BD (types d'articles, collections, produits, zones géographiques, …)
    - base de composants Java (BI Beans) livrée par l'éditeur au sein de son environnement de développement (JDeveloper).

Succès story : télécommunications

Exemple de France Télécom

- Le projet :
  - 12 BD sources
  - Récupération des données : 1,5 année
  - Données régionales et nationales
  - Parfois chez des prestataires de services /Parfois au prix d'un intense lobbying
  - En 2003 : environ 5 années de travail
- Solution :
  - Entreposage : SQL server
  - DW de 3 bimestres, vidé périodiquement
  - 1,2 million d'individus
  - 1 fait = 1 client
  - 250 colonnes
  - Intégration faite à la main périodiquement
- Exploitation : progiciel de DM développé spécifiquement

APPLICATION: entreprise « Bon Pieds »

L'entreprise française Au Bon Pieds, spécialisée dans la vente de chaussures désire suivre l'évolution de ses ventes.

- ***Modélisation:***

-Tables de faits: Ventes.

-Dimensions: >Magasin.

>Modele.

>Pointure.

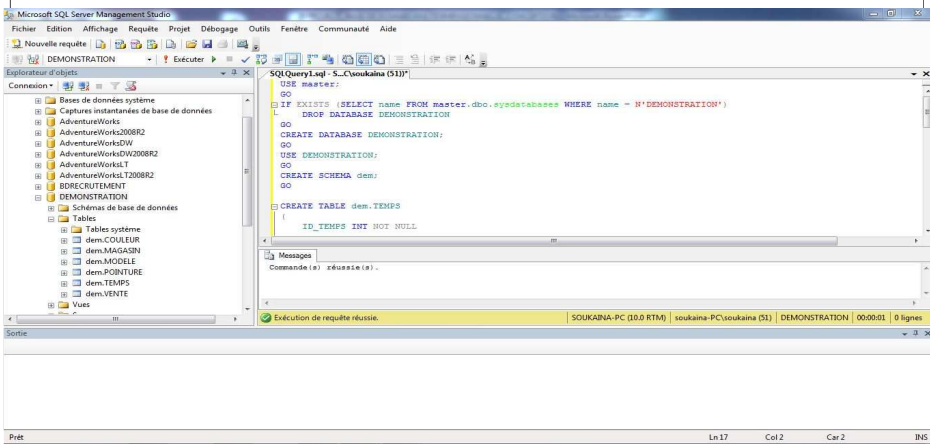
>Couleur.

>Temp.

APPLICATION:

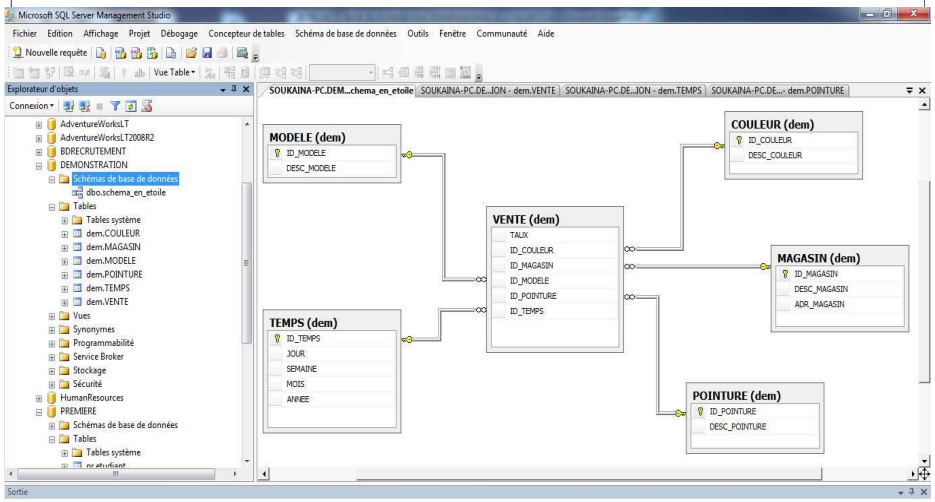
- ***Etape1:***

Création de la base DEMONSTRATION qui contient les différents tables.



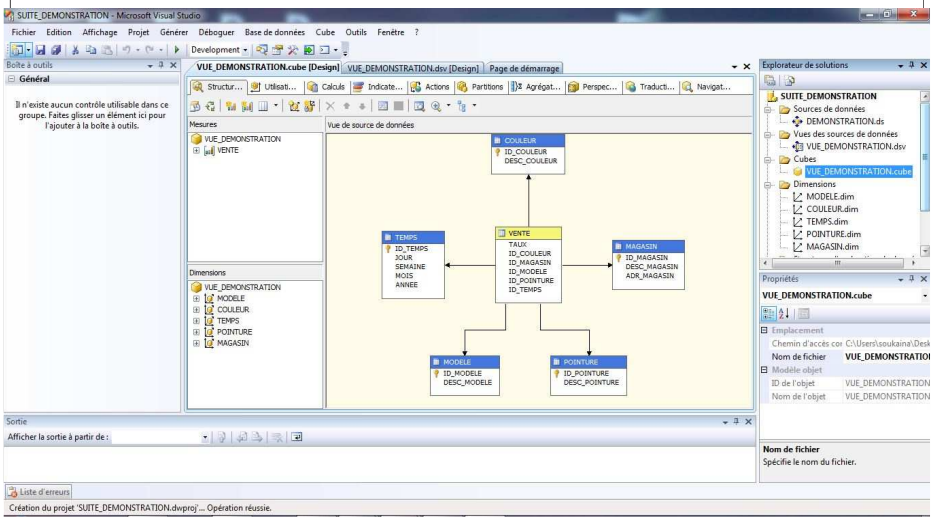
APPLICATION:

- ***Etape2:*** Génération d'un Schéma de base de données.



APPLICATION:

- ***Etape3: création du cube OLAP***



TD ED

1. EX2: opérateur Télécom

Exercice 2

Un opérateur Télécom propose à ses abonnés un boiter routeur ADSL qui groupe un accès à Internet, le téléphone et un bouquet de chaînes de télévision interactives. Nous ne nous intéresserons qu'à cette dernière fonction dans notre Datawarehouse (DW).

Les chaînes de télévision proposées par cet opérateur dans son bouquet sont des chaînes nationales et des chaînes à péage. Le téléspectateur peut réagir (interactivement) de plusieurs manières aux émissions et aux publicités qu'il reçoit au moyen de sa télécommande:

- A tout moment, il peut zapper vers une autre chaîne
- A tout moment, il peut éteindre son poste
- Pendant une publicité, il peut zapper à la pub suivante sans attendre la fin de la publicité en cours.
- Pendant une publicité, il peut demander plus d'information sur le produit présenté.

Le DW est centré sur les actions du téléspectateur. L'objectif de ce DW est de profiler au plus précis le foyer (ou plutôt le téléspectateur qui détient la télécommande) afin de maximiser la demande d'information sur les produits présentés par les publicités.

En effet, l'opérateur perçoit plus d'argent de la part de l'annonceur quand le spectateur demande plus d'information au moment au l'annonce est passée. Le schéma de ce DW est constitué des tables suivantes.

Date (CléDate, Année, Mois, JourDeMois, JourDeSemaine, TrancheHoraire, Heure, Minute)

Foyer (CléFoyer, NomAbonné, AnnéeNaissanceAbonné, Région, Département, District, Ville, Quartier, SituationFamille, RevenuFoyer, CS, NombreAdulte, NombreEnfant)

Emission (CléEmission, Chaîne, DateDébut, DuréeSeconde, TypeEmission, Catégorie, Annonceur)

Action (CléDate, CléFoyer, CléEmission, TypeAction, DuréeAction, DuréeRestante)

Quelle est la table de fait dans ce DW ? Justifiez votre réponse..

Exercice 3

Une université cherche à étudier les facteurs influant sur la réussite de ses étudiants aux examens.

Pour cela elle décide de construire un Datawarehouse.

Elle souhaite pouvoir répondre aux questions suivantes:

- Quel est le nombre de réussites aux examens par cours, pour l'année 2007?
- Quel est le nombre de réussites aux examens d'un cours obligatoire, pour l'année 2007?
- Quel est le nombre de réussites aux examens par sexe (féminin, masculin), pour l'année 2007?
- Combien d'étudiants ayant un âge de 22 ans ont réussi leurs examens de base de données relationnelle?
- Quel est le nombre de réussites aux examens pendant le semestre d'hiver 2006?

Pour cela elle dispose des données suivantes:

Pour chaque examen passé, on connaît l'âge et le sexe de l'étudiant, le nom du cours (les cours peuvent être regroupés en cours obligatoire et cours à option), la date de l'examen, la note obtenue et si l'examen est réussi ou non.

Proposez un modèle en étoile pour cette application. Recherchez tout d'abord les différentes dimensions et proposez une hiérarchie pour ces dimensions.