

Les data Warehouse

Master 2 IDO

Le contexte

- **Besoin:**
- prise de décisions stratégiques et tactiques
- **Pourquoi:**
- besoin de réactivité
- **Qui:**
- les décideurs (non informaticiens)
- **Comment:**
- réponse qualitative

Qui sont mes
meilleurs
clients?

analyse des données, dégager

Pourquoi et
comment le
chiffre
d'affaire a
baissé?

Qui parmi nos
clients
consomment
beaucoup plus
de poisson?

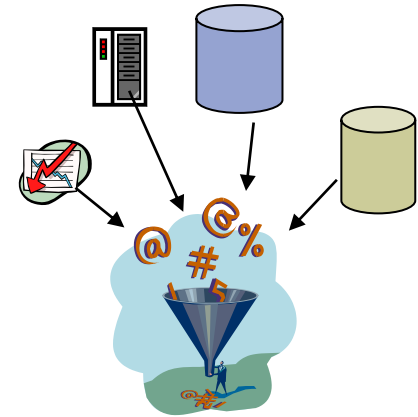


A combien
s'élèvent mes
ventes
journalières?

Les données utilisables par les décideurs

■ Données opérationnelles (de production)

- Bases de données (Oracle, SQL Server)
- Fichiers, ...
- Paye, gestion des RH, gestion des commandes...



■ Caractéristiques de ces données:

- Distribuées: systèmes éparpillés
- Hétérogènes: systèmes et structures de données différents
- Détaillées: organisation des données selon les processus fonctionnels
- Peu/pas adaptées à l'analyse : les requêtes lourdes peuvent bloquer le système transactionnel
- Volatiles: pas d'historisation systématique

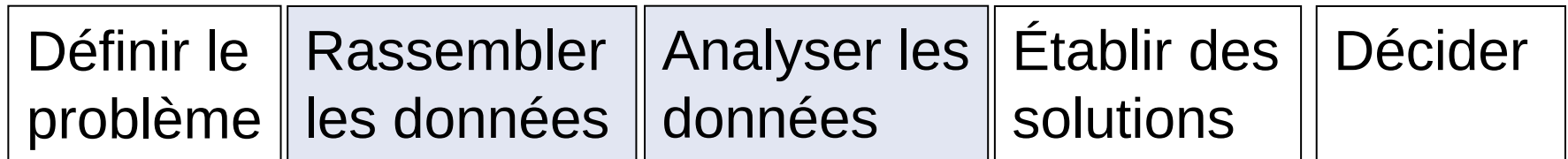
Problématique

- Comment répondre aux demandes des décideurs?
 - En donnant un accès rapide et simple à l'information stratégique
 - En donnant du sens aux données

➡ Mettre en place un système d'information dédié aux applications décisionnelles:
un data warehouse

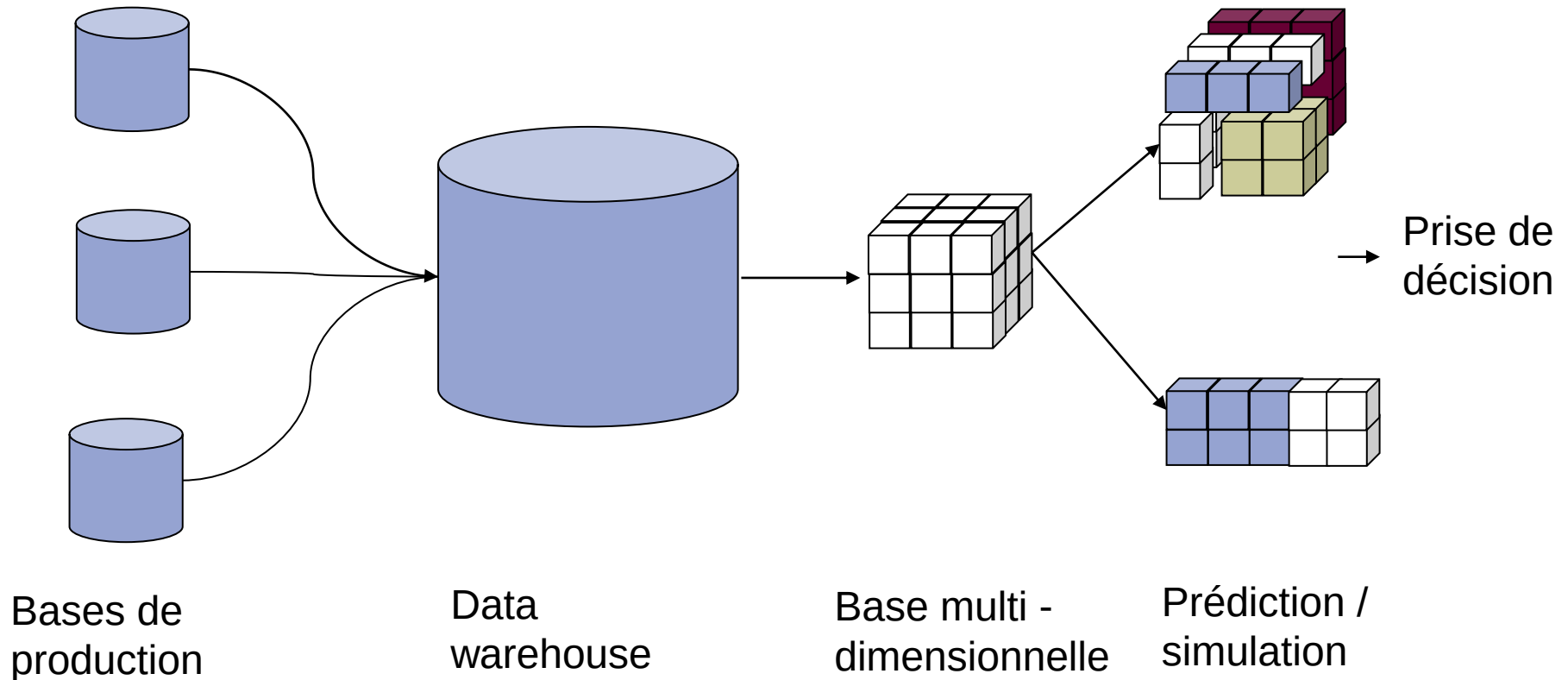
Le processus de prise de décision

Champs d'application des
systèmes décisionnels



Temps de prise d'une décision

Le processus de prise de décision



Domaines d'utilisation des DW

- Banque
 - Risques d'un prêt, prime plus précise
- Santé
 - Épidémiologie
 - Risque alimentaire
- Commerce
 - Ciblage de clientèle
 - Déterminer des promotions
- Logistique
 - Adéquation demande/production
- Assurance
 - Risque lié à un contrat d'assurance (voiture)
- ...

Quelques métiers du décisionnel

- **Strategic Performance Management**

- Déterminer et contrôler les indicateurs clé de la performance de l'entreprise

- **Finance Intelligence**

- Planifier, analyser et diffuser l'information financière. Mesurer et gérer les risques

- **Human Capital Management** (gestion de la relation avec les employés)

- Aligner les stratégies RH, les processus et les technologies.

- **Customer Relationship Management** (gestion de la relation client)

- Améliorer la connaissance client, identifier et prévoir la rentabilité client, accroître l'efficacité du marketing client

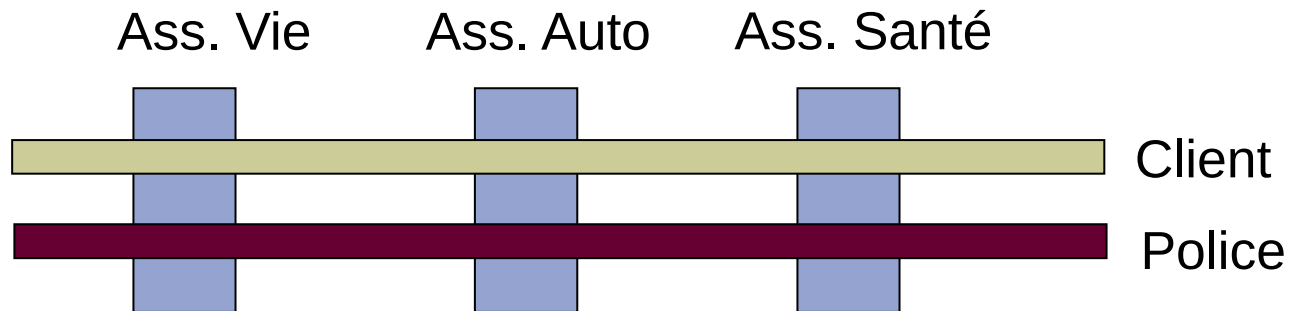
- **Supplier Relationship Management** (gestion de la relation fournisseur)

- Classifier et évaluer l'ensemble des fournisseurs. Planifier et piloter la stratégie Achat.

Les 4 caractéristiques des data warehouse

1. Données orientées sujet:

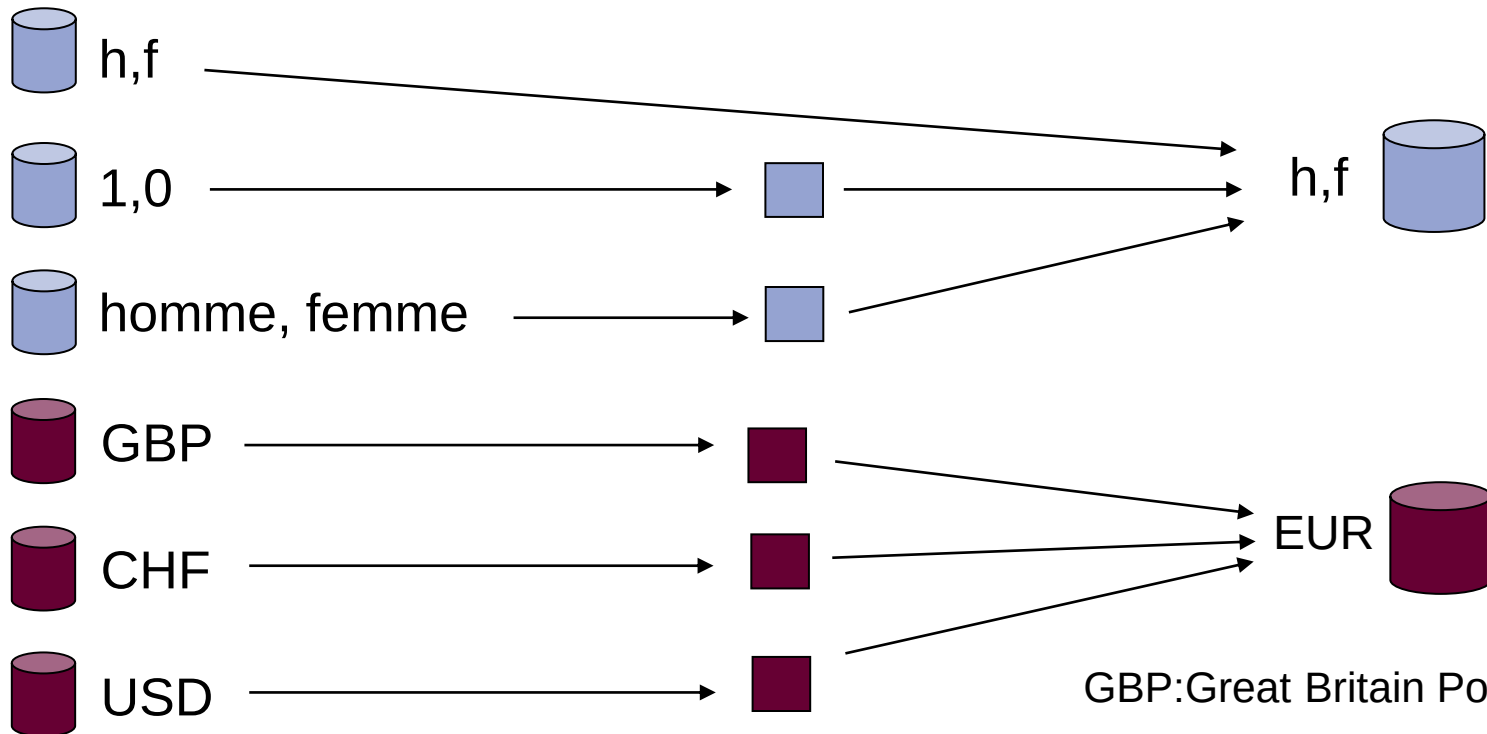
- Regroupe les informations des différents métiers
- Ne tiens pas compte de l'organisation fonctionnelle des données



Les 4 caractéristiques des data warehouse

2. Données intégrées:

- Normalisation des données
- Définition d'un référentiel unique

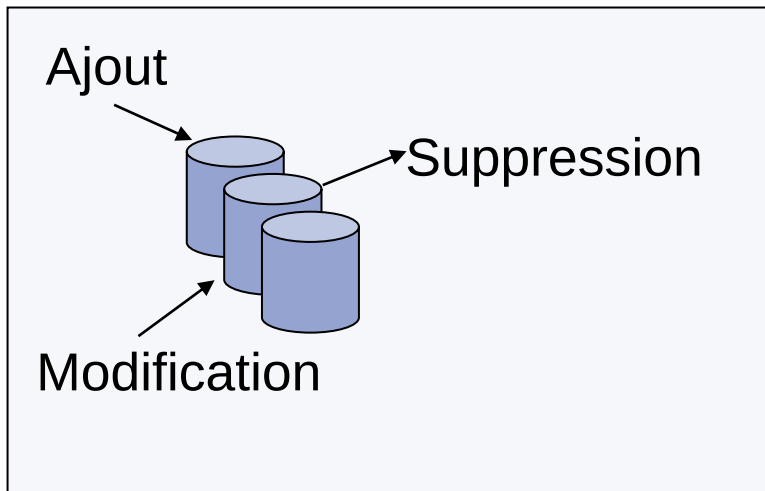


Les 4 caractéristiques des data warehouse

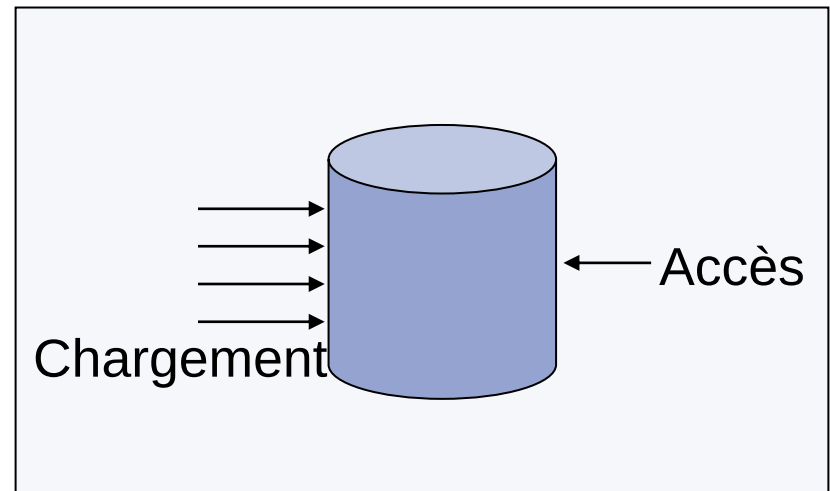
3. Données non volatiles

- Traçabilité des informations et des décisions prises
- Copie des données de production

Bases de production



Entrepôts de données



Les 4 caractéristiques des data warehouse

4. Données datées

- Les données persistent dans le temps
- Mise en place d'un référentiel temps

Base de
production

Image de la base en Mai 2005

Répertoire

Nom	Ville
Dupont	Paris
Durand	Lyon

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Dupont	Marseille
Durand	Lyon

Entrepôt
de
données

Calendrier

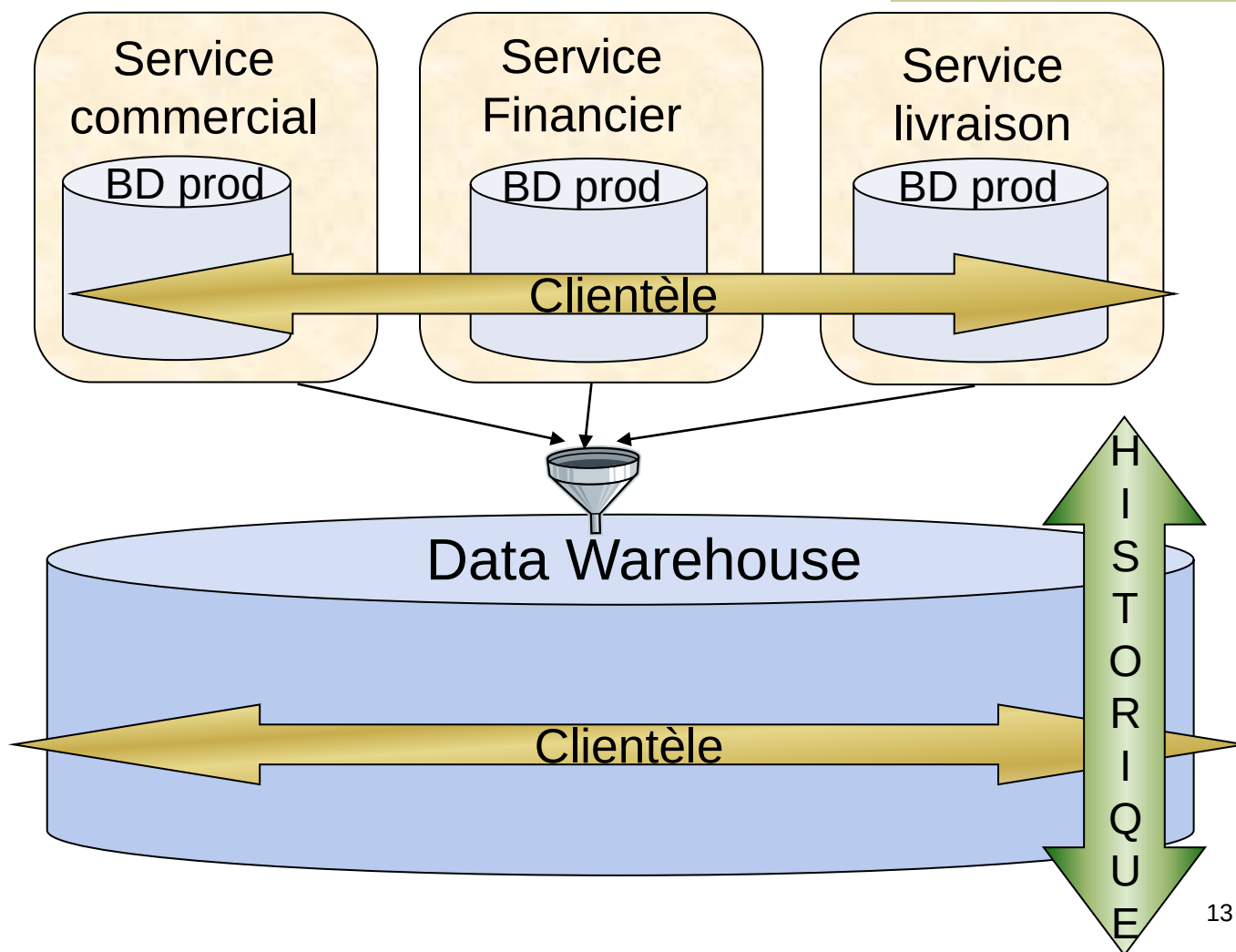
Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

Code	Nom	Mois
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

SGBD et DW

OLTP: On-Line
Transactional
Processing



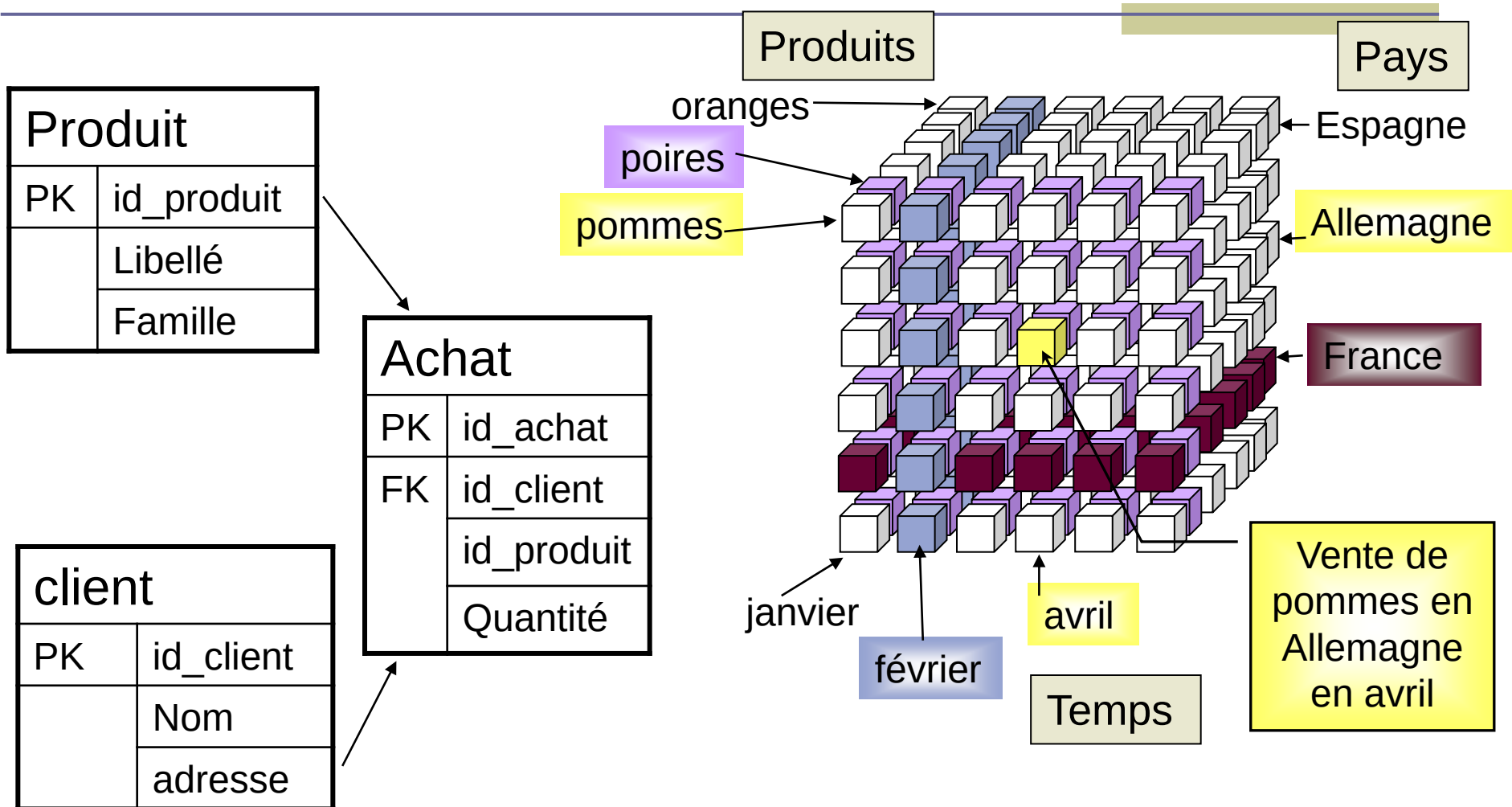
OLTP VS DW

OLTP	DW
Orienté transaction	Orienté analyse
Orienté application	Orienté sujet
Données courantes	Données historisées
Données détaillées	Données agrégées
Données évolutives	Données statiques
Utilisateurs nombreux, administrateurs/opérationnels	Utilisateurs peu nombreux, manager
Temps d'exécution: court	Temps d'exécution: long

TABLE 3.5 A Comparison Between OLTP and OLAP

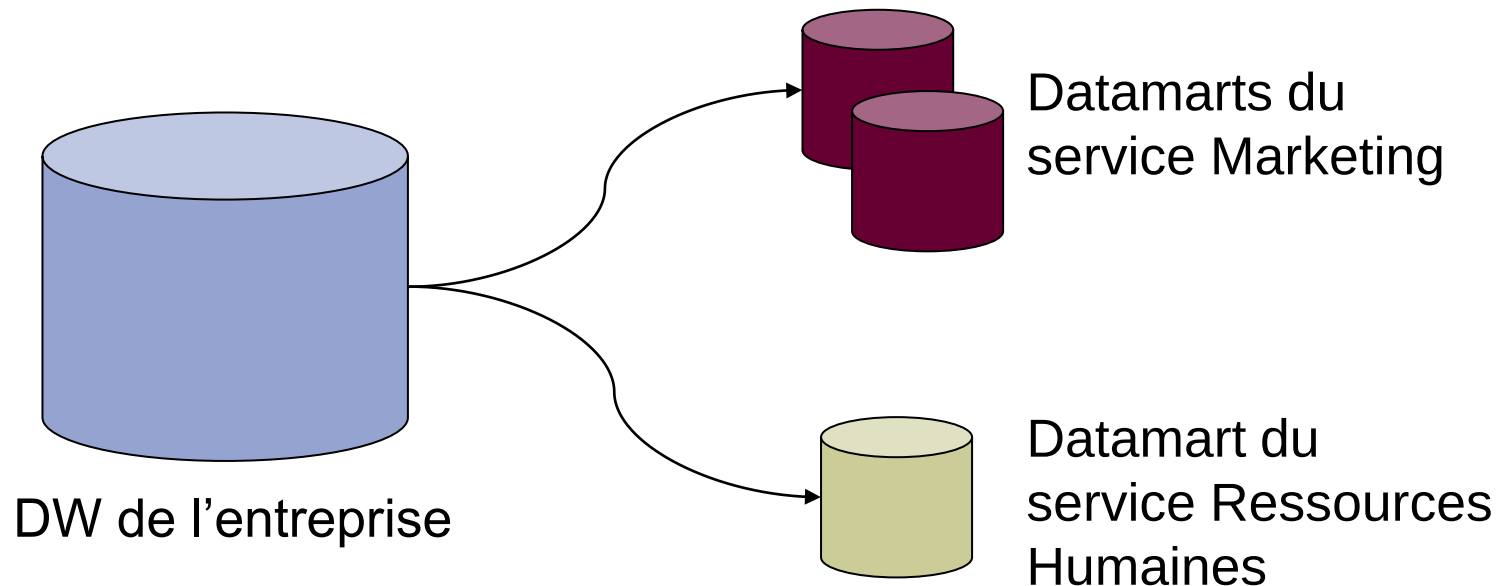
Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or data mart (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business transactions and routine reports)	Slow (resource intensive, complex, large-scale queries)

OLTP VS OLAP



Datamart

- Sous-ensemble d'un entrepôt de données
- Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
- Point de vue spécifique selon des critères métiers



Intérêt des datamart

- Nouvel environnement structuré et formaté en fonction des besoins d'un métier ou d'un usage particulier
- Moins de données que DW
 - Plus facile à comprendre, à manipuler
 - Amélioration des temps de réponse
- Utilisateurs plus ciblés: DM plus facile à définir

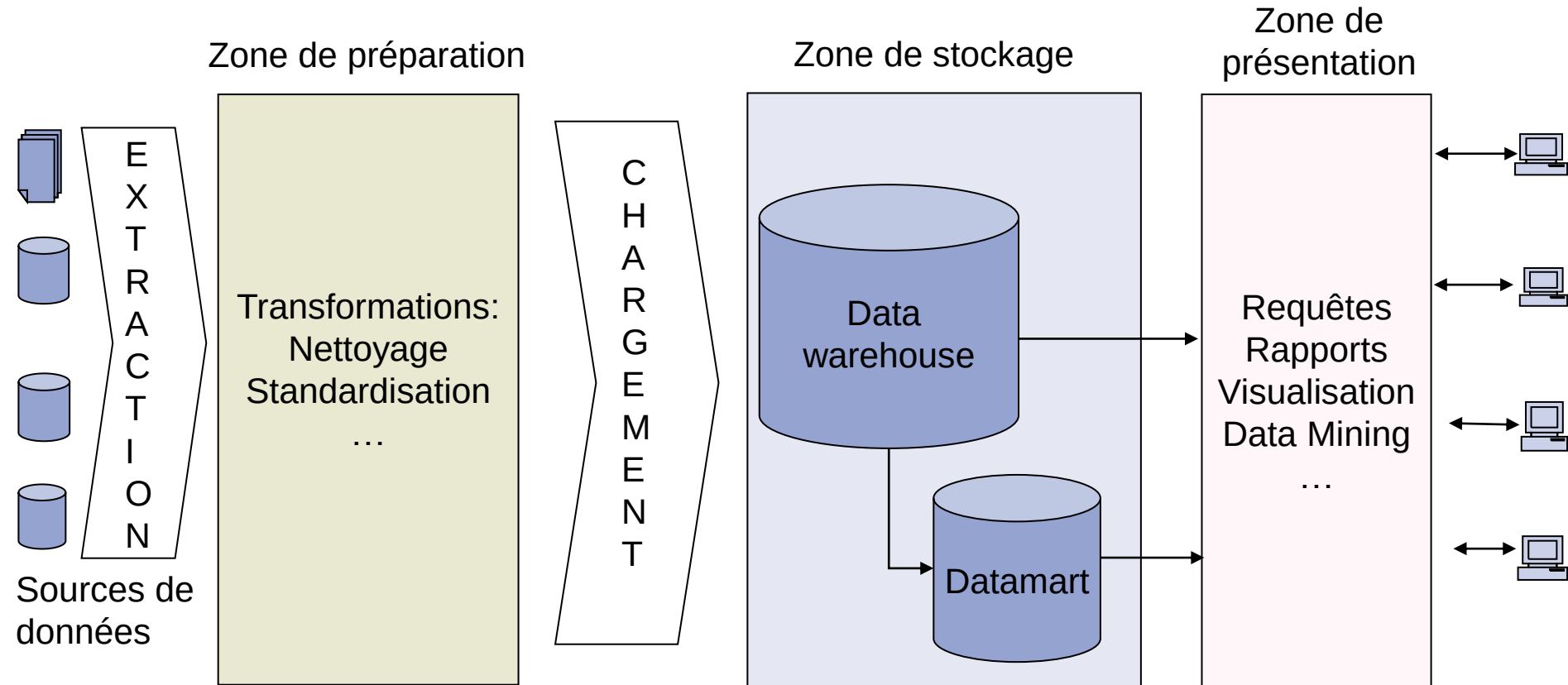
Les ETLs

Master 2 BI

II. Qu'est-ce qu'un « ETL » ?

- **ETL** est l'acronyme de « **Extract – Transform - Load** ».
- Un ETL permet ainsi l' **Extraction**, la **Transformation** et le **Chargement** de données depuis des sources diverses (bases de données, fichiers) vers des cibles préalablement définies.

Architecture générale



Les flux de données

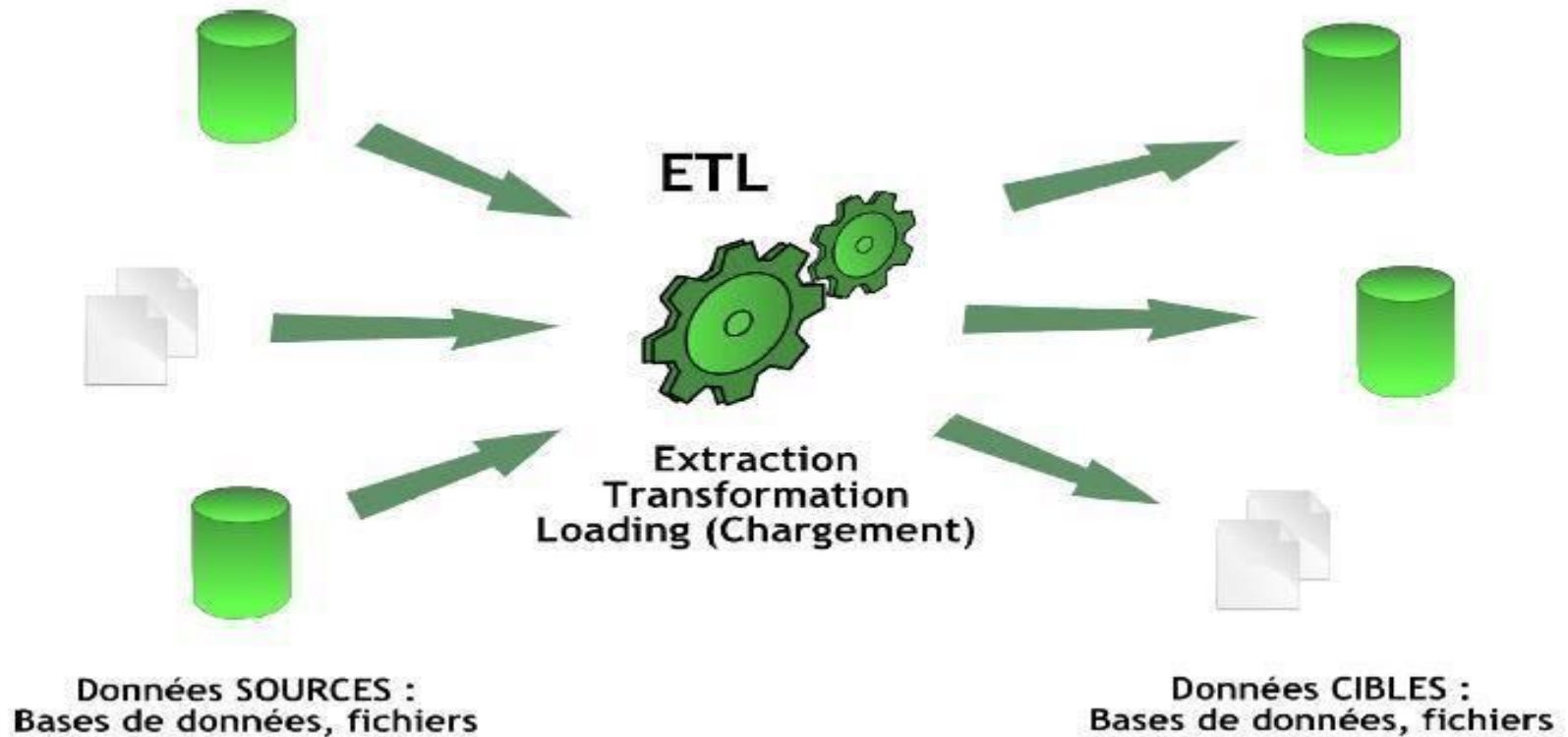
■ Flux entrant

- Extraction: multi-source, hétérogène
- Transformation: filtrer, trier, homogénéiser, nettoyer
- Chargement: insertion des données dans l'entrepôt

■ Flux sortant:

- Mise à disposition des données pour les utilisateurs finaux

Les flux de données



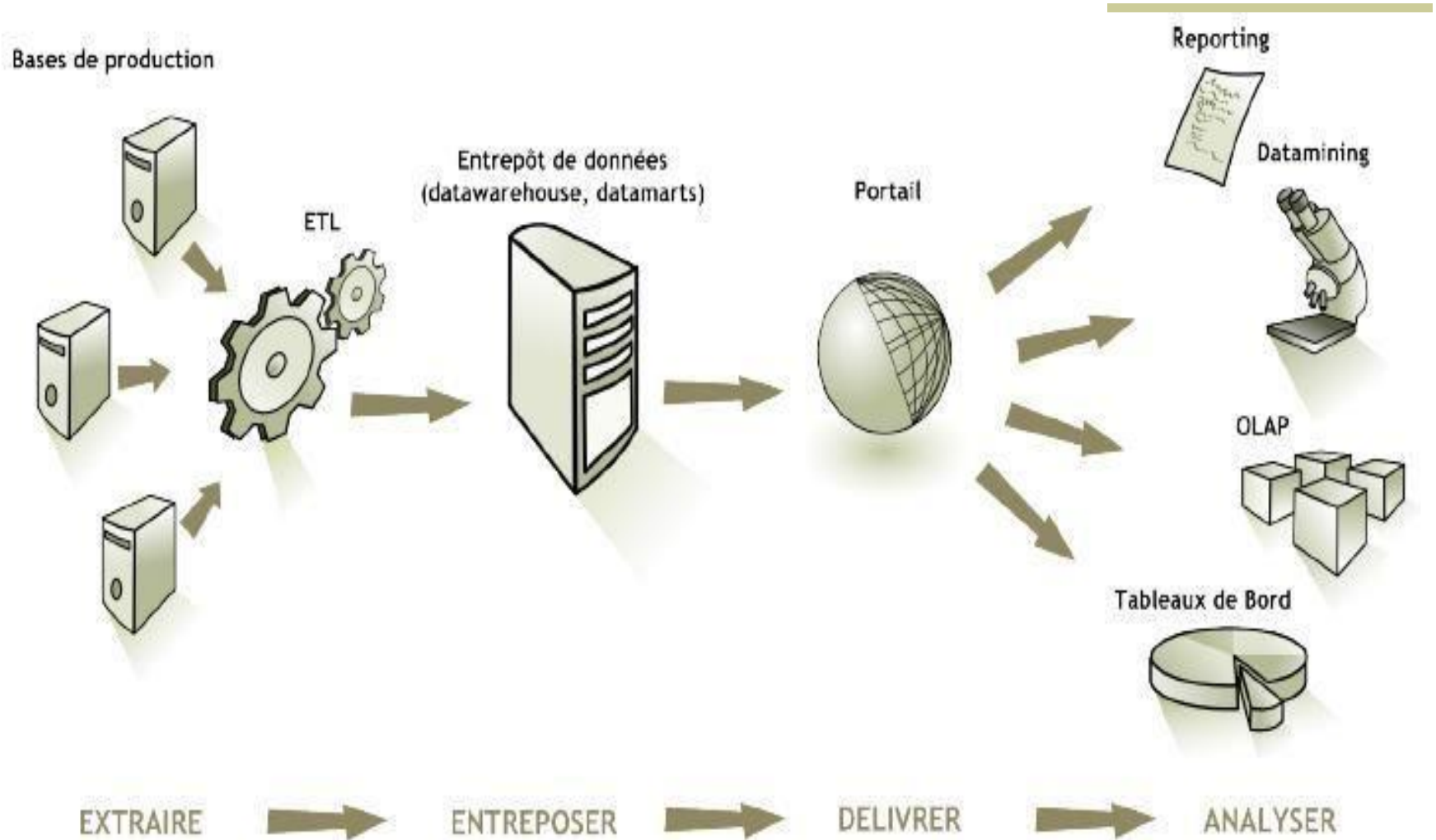
Les sources de données

- De nombreux systèmes de gestion de bases de données sont supportés nativement en lecture/écriture (Oracle, MS Sql Server, DB2, Postgresql, MySql,...).
- De nombreux types de fichiers peuvent également être lus ou écrits: Csv, Excel, Txt, Xml,
- Notons que la plupart des ETL disposent d'une interface graphique permettant l'élaboration des différents scénarios d'intégration.

Les flux de données sortant

- Les ETL sont, en générale, utilisés dans l'informatique décisionnelle afin de permettre l'alimentation des datawarehouses (entrepôts de données).
- Ces derniers servent de supports pour l'analyse des données sous plusieurs formes :
 - Rapports et états,
 - Tableaux de bords (dashboards, balanced scorecard),
 - Indicateurs de performance (« KPIs »),
 - Analyse multi-dimensionnelle (OLAP) ,
 - Analyse exploratoire (Data-Mining).

Alimentation des datawarehouses



Pourquoi ?

- Le travail des développeurs en est ainsi grandement facilité, tant au niveau de la conception que de la maintenance des traitements de données.
- Les ETL sont communément utilisés dans l'informatique décisionnelle afin de permettre l'alimentation des datawarehouses (entrepôts de données) mais aussi les CRM, ERP, ...
- Les volumes de données traités sont toujours très importants, ainsi les critères de performance sont primordiaux dans le choix d'un ETL.

ETL : Créer ou acheter ?

- Le choix le plus difficile dans tout projet décisionnel ou d'intégration/migration de données consiste à déterminer quelle méthode doit être mise en oeuvre :
- Faut-il créer du code spécifique (procédures SQL, code Java ou autre) ?
- Faut-il acheter un ETL propriétaire (Informatica, Oracle Warehouse Builder, BO Data Integrator ou autre) ?

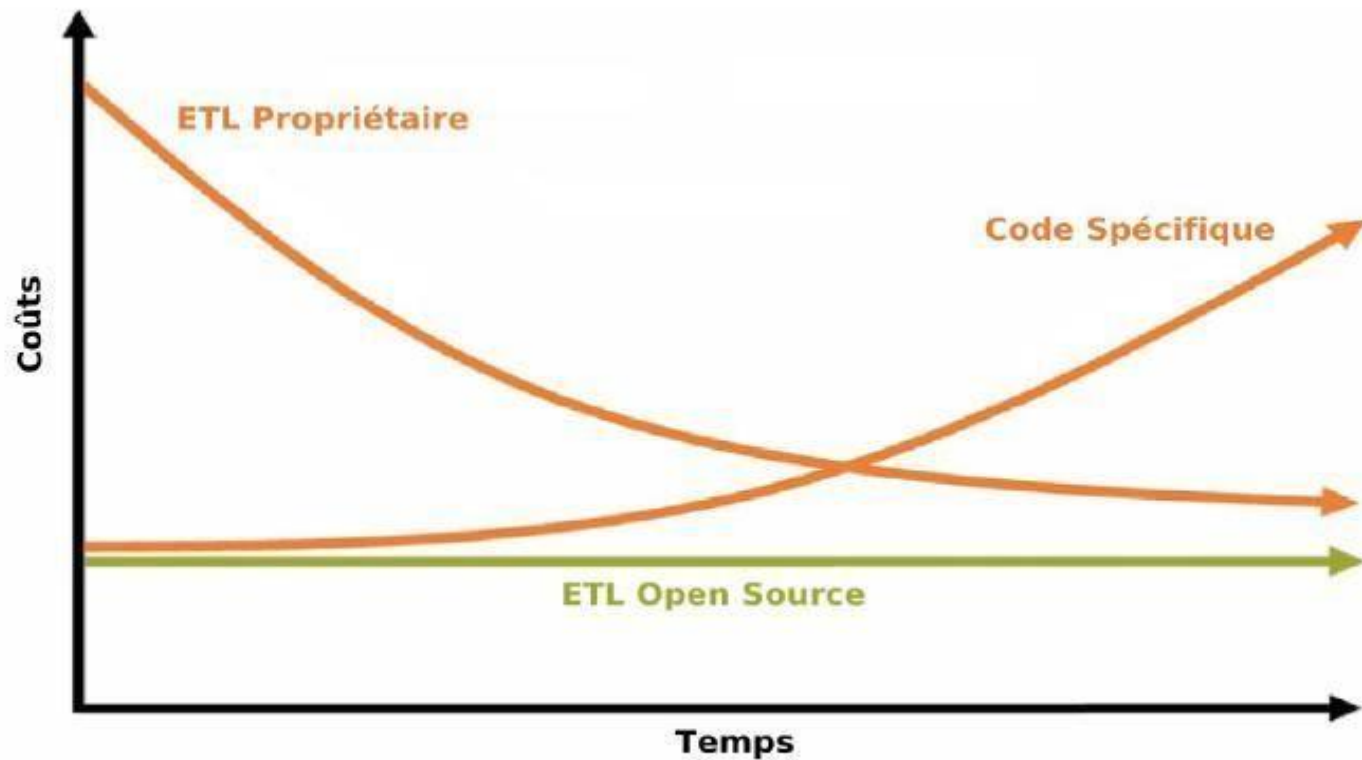
Créer un ETL

- Rester au plus près des spécificités métiers des données à traiter, tout en s'affranchissant des contraintes liées à l'achat et l'utilisation d'un ETL propriétaire coûteuse à long terme, tout simplement car l'évolutivité constante des données métier entraîne une nécessaire adaptation des traitements d'intégration.

Acheter un ETL

- La deuxième solution va permettre de mettre en œuvre très rapidement les traitements d'intégration, avec cependant des coûts élevés (achat des licences, formations,...) et ceci dès la phase de démarrage du projet.

ETL Open Source



(Graphique extrait de la doc. technique Pentaho)

Principaux acteurs

- ETL Open Source
 - **Talend Open Studio**
 - **Pentaho Data Integration**
- ETL propriétaires
 - Informatica Power Center,
 - Oracle Warehouse Builder,
 - Cognos Decision Stream

Types d'ETL

- ETL du type « générateur de code ». Pour chaque traitement d'intégration de données, un code spécifique est généré, ce dernier pouvant être en Java, Perl, ... Les données traitées et les traitements effectués sont donc intimement liés.
- ETL du type « moteur de transformation » les données traitées et les traitements à effectuer sont parfaitement séparés.

ETL : Outils et fonctionnalités

- Interface graphique qui permet la création des processus de manipulation de données.
- Facilite la construction des requêtes dans les bases de données en détectant le schéma et les relations entre tables.
- Un référentiel permet de stocker les méta-données afin de pouvoir les exploiter dans différents jobs.
 - Par exemple on peut sauvegarder le type et le format des données d'entrée d'un fichier CSV afin de pouvoir les exploiter ultérieurement.

ETL : Outils et fonctionnalités

- Se connecter aux principaux SGBD (Oracle, DB2, MS SQL Server, PostgreSQL, MySQL,...) ainsi que pour traiter tous les types de fichiers plats (CSV, Excel, XML), aussi bien en lecture qu'en écriture.
- composants et connecteurs, fournissant des fonctions basiques telles que des associations, transformations, agrégation et recherches; des fonctions spécialisées comme le filtrage de données, le multiplexage de données...

Exemple de composantes

Step	Etape	Description
Extraction depuis EXCEL		Extraction de données depuis un ou plusieurs onglets d'un fichier XLS
Extraction depuis ACCESS		Extraction de données depuis une table d'une base Access
Extraction depuis une base de données		Extraction de données depuis une table d'un SGBD (Oracle, MySql, Postgres, SQLServer, DB2, AS400, SAP, etc...) : 30 SGBD supportés via JDBC et/ou ODBC
Extraction depuis un fichier		Extraction de données depuis un fichier CSV ou de type « délimité »
Extraction depuis un fichier XML		Extraction de données depuis un fichier XML

Exemple de composantes

Extraction depuis un annuaire		Extraction de données depuis un annuaire de type LDAP
Alimentation base de données		Insertion ou mise à jour d'une table d'un SGBD (insert/update)
Recherche dans base de données		Recherche des enregistrements dans une base de données selon une liste de valeurs
Recherche dans un flux		Recherche des enregistrements dans un flux de données selon une liste de valeurs

Critères de comparaison

- Accès aux données relationnelles (SGBD)
 - Lecture de tables complète
 - Appel de procédure stockée
 - Ajout de clause Where/Order by
 - Exécution de requête, ...
- Fichiers plats : Csv, Excel, XML, ...
- Connecteurs : CRM, ERP, Cubes OLAP, services Web, ...

Critères de comparaison

■ Déclenchement des processus par message :

- CORBA
- XML RPC, ...

Traitement des données

- Présentation graphique des flux
- Visualisation des données en cours de traitement
- Outil de debugging
- Génération de documentations techniques
- Gestion des erreurs d'intégration

Exemple : Pentaho Data integration

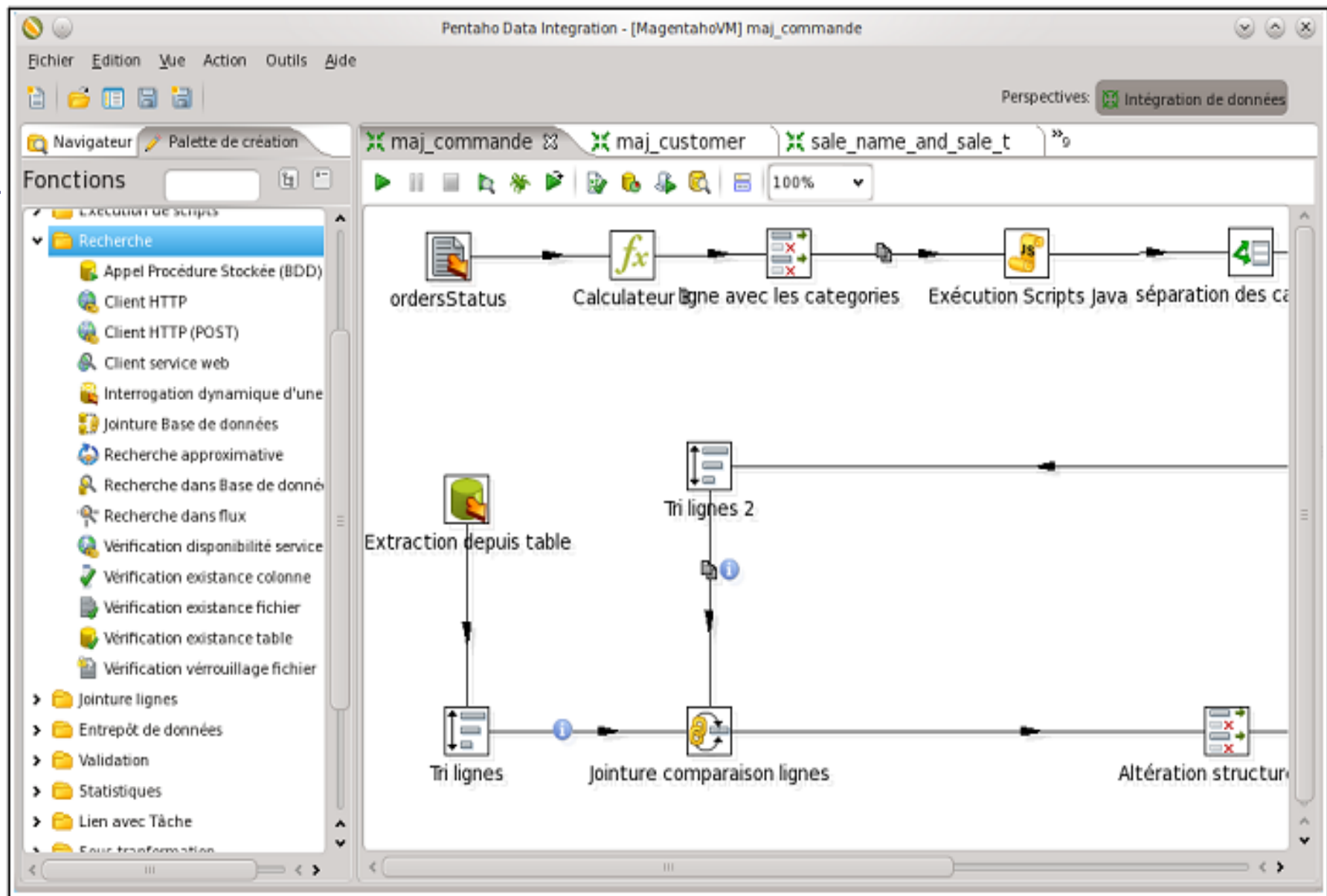
- ETL du type « moteur de transformation »
- il est possible de créer sans programmation des processus composés d'imports et d'exports de données, et de différentes opérations de transformation, telles que des conversions, des jointures, l'application de filtres, ou même l'exécution de fonctions javascript.

Pentaho Data integration

- PDI permet de créer deux types de processus :
 - **Les transformations** : traitements effectués au niveau d'une ou plusieurs bases de données comprenant des opérations de lecture, de manipulation et d'écriture.
 - **Les tâches** : traitements de plus haut niveau, combinant des actions telles que l'exécution d'une transformation PDI, l'envoi d'un mail, le téléchargement d'un fichier ou le lancement d'une application. Il est possible d'exécuter des actions différentes en fonction de la réussite ou de l'échec de chaque étape.

Pentaho Data integration

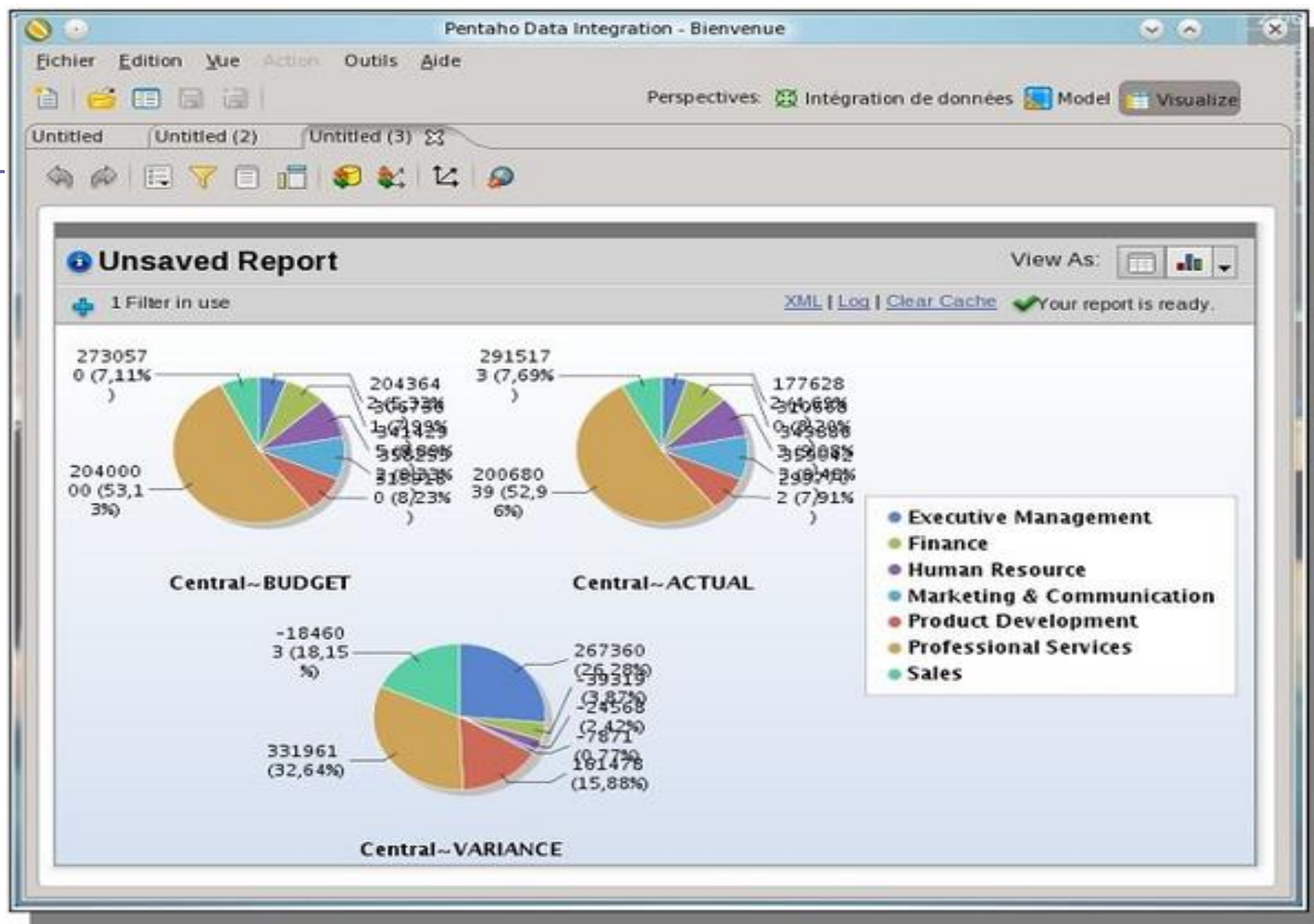
- PDI est également utilisable avec un référentiel partagé. Ainsi plusieurs développeurs peuvent utiliser des objets communs.
- PDI peut se connecter sur un grand nombre de bases de données, dont Oracle, Sybase, MySQL, PostgreSQL, Informix, SQL Server et bien d'autres, et peut également utiliser des données provenant de fichiers texte, XML et Excel.



Exemple de transformation

Exemple : PDI (suite)

- Les transformations et les tâches sont créées par glisser- déposer des différentes étapes du processus. Des assistants sont disponibles à chaque étape et permettent par exemple de générer automatiquement les requêtes de lecture et de création de tables.



Module AgileBI de Pentaho Data Integration