Extract – Transform – Load

Issam El Hachimi
Ludovic Schmieder

27 janvier 2011

Sommaire

- Le Business Intelligence
- Les ETL
  - PDI
  - Talend
- Démo : 2 exemples
- Constat et comparatif
- Conclusion

Entreprise: Stratégies, Marketing

« *Il faut connaitre le client pour connaitre ses besoins* »

Avant : Sondage : Coûteux, faible échantillon

Aujourd'hui : Profils déterminés en fonction des données récoltées

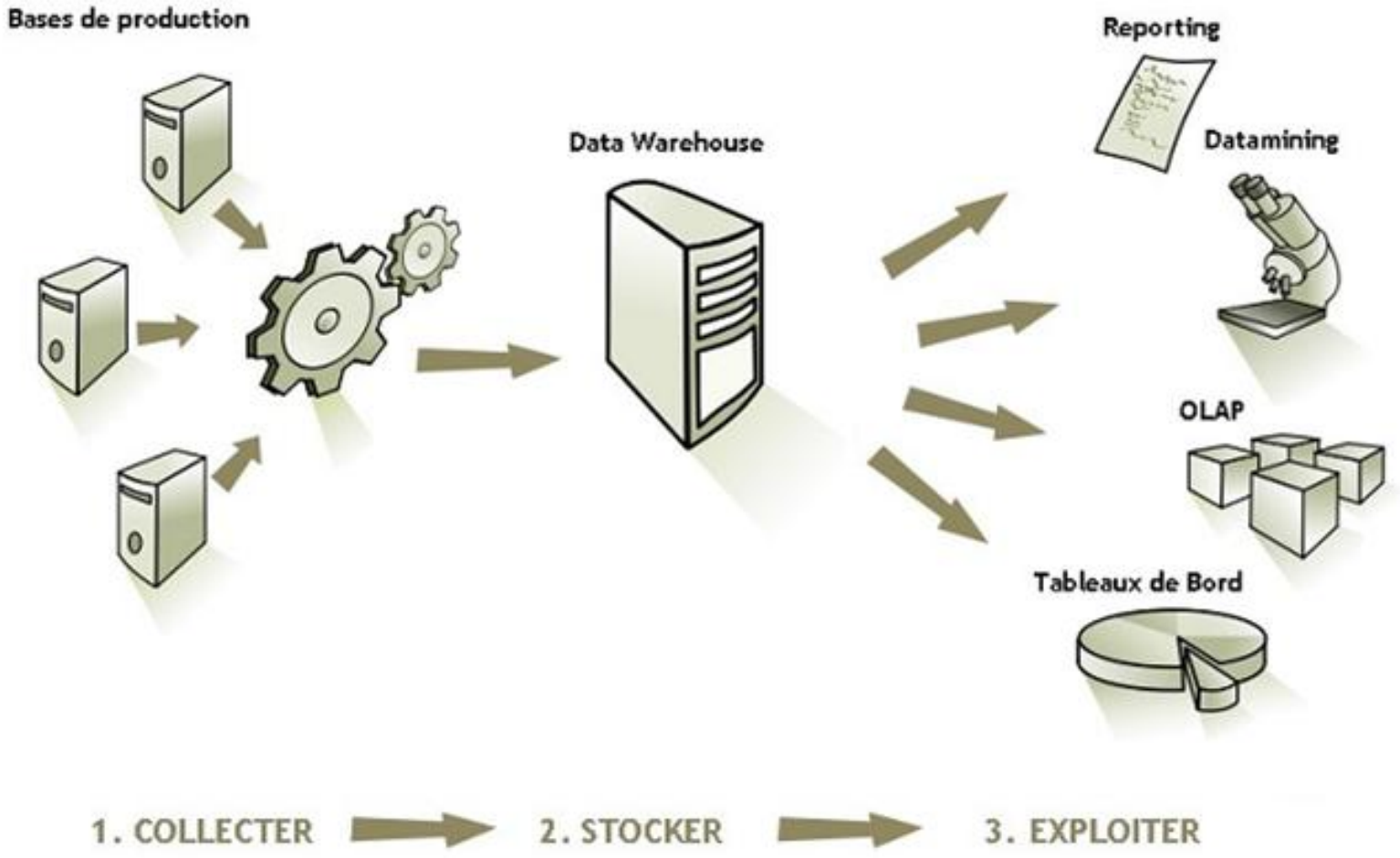
Le marketing suit une tendance reposant sur l'analyse des données.

Analyse des données

Pour comprendre les tendances, analysons :

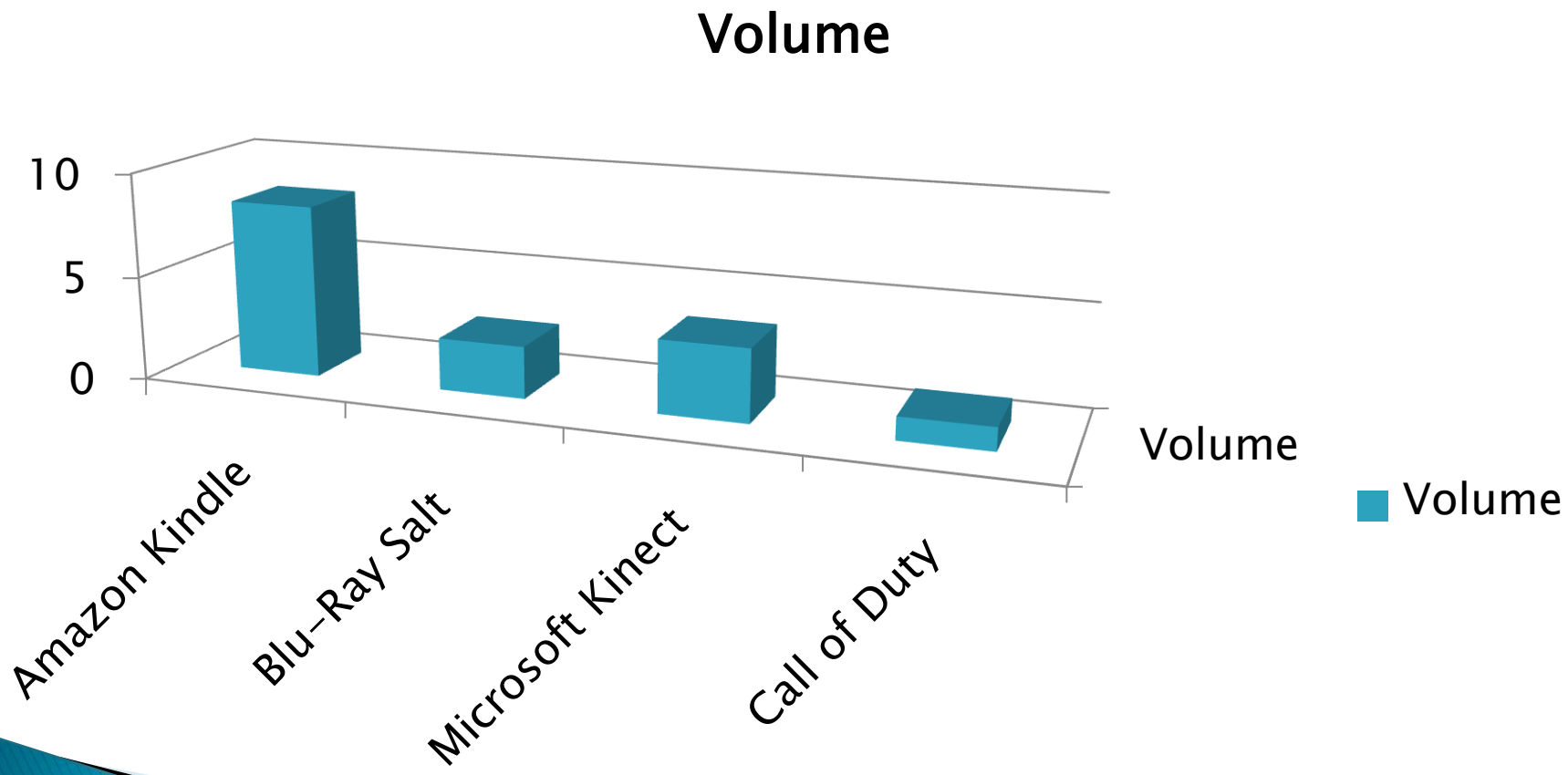
- Forfait téléphonique & Age du client
  - Identification de profils clients
- Produit & Localisation d'un supermarché
  - Identification de zones de vente
- Période & Transport ferroviaire
  - Identification des fréquentations

La Business Intelligence ?

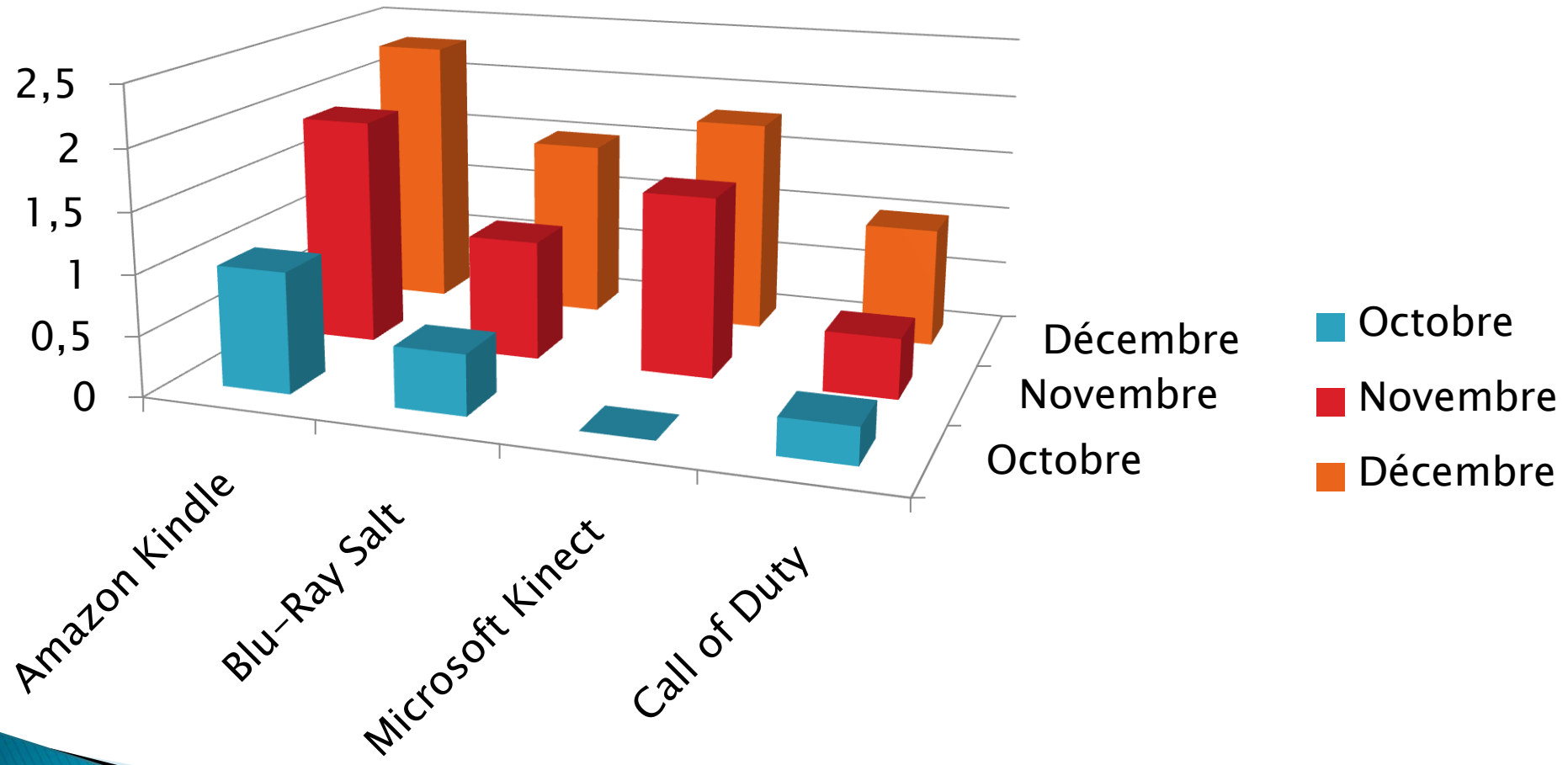


« C'est l'exploitation des données de l'entreprise dans le but de faciliter la prise de décision »

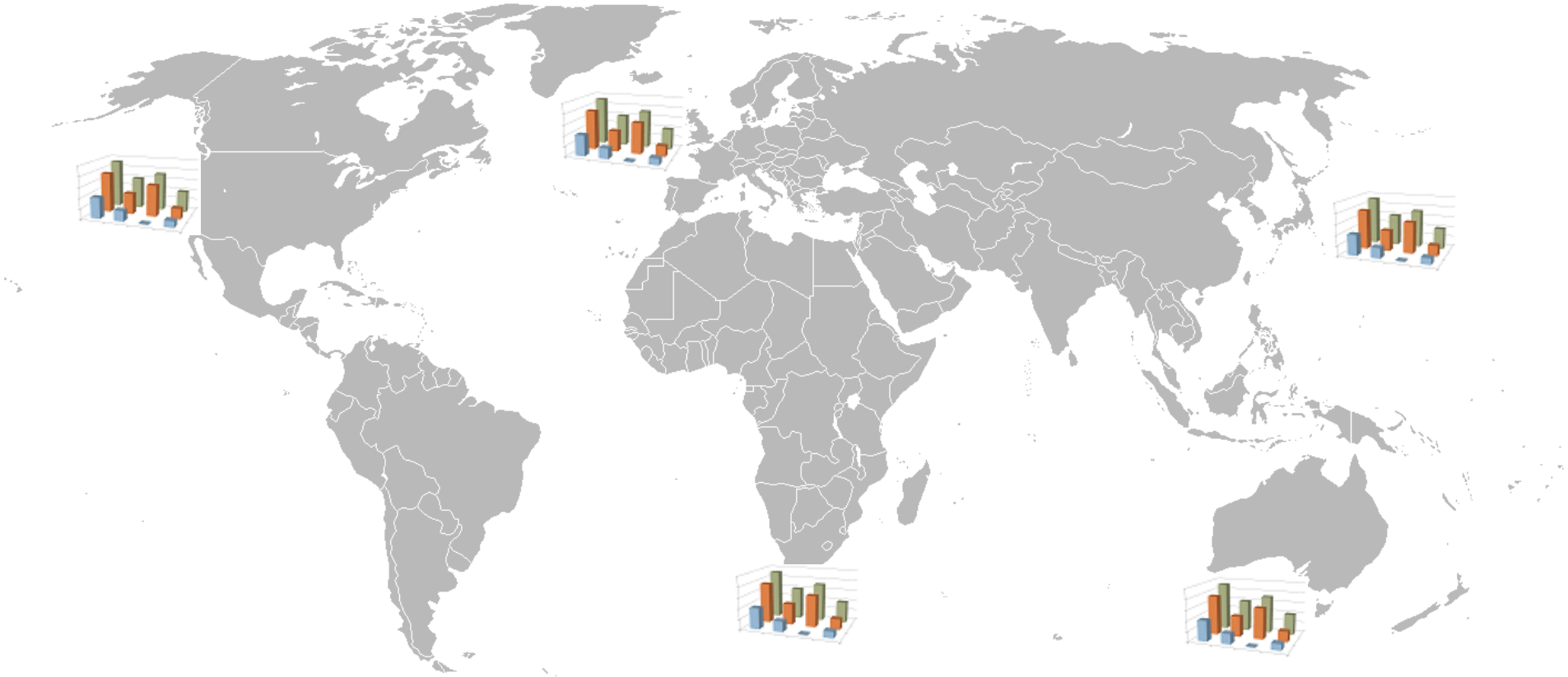
Exemple: Amazon.com



Nouvel axe : Le temps



Nouvel axe : La position géographique



Finalité

- Compréhension des données traitées
- Analyse de ces données
- Aide à la décision

- Exemple :
  - Dans un magasin, croisement des produits , des ventes et de leurs emplacements : constat de zones froides et chaudes (peu et beaucoup de ventes)

    $\Rightarrow$ Développement des ventes

Data Warehouse

- « Entrepôt de données opérationnelles et de gestion d'une entreprise »

- Base de données destinée à contenir des données :
  - historisées,
  - persistantes,
  - transverses au niveau métier,
  - **adaptées aux besoins d'analyse.**

Un besoin

- Besoin ressenti entre les sources de données et le Data Warehouse = transfert de données

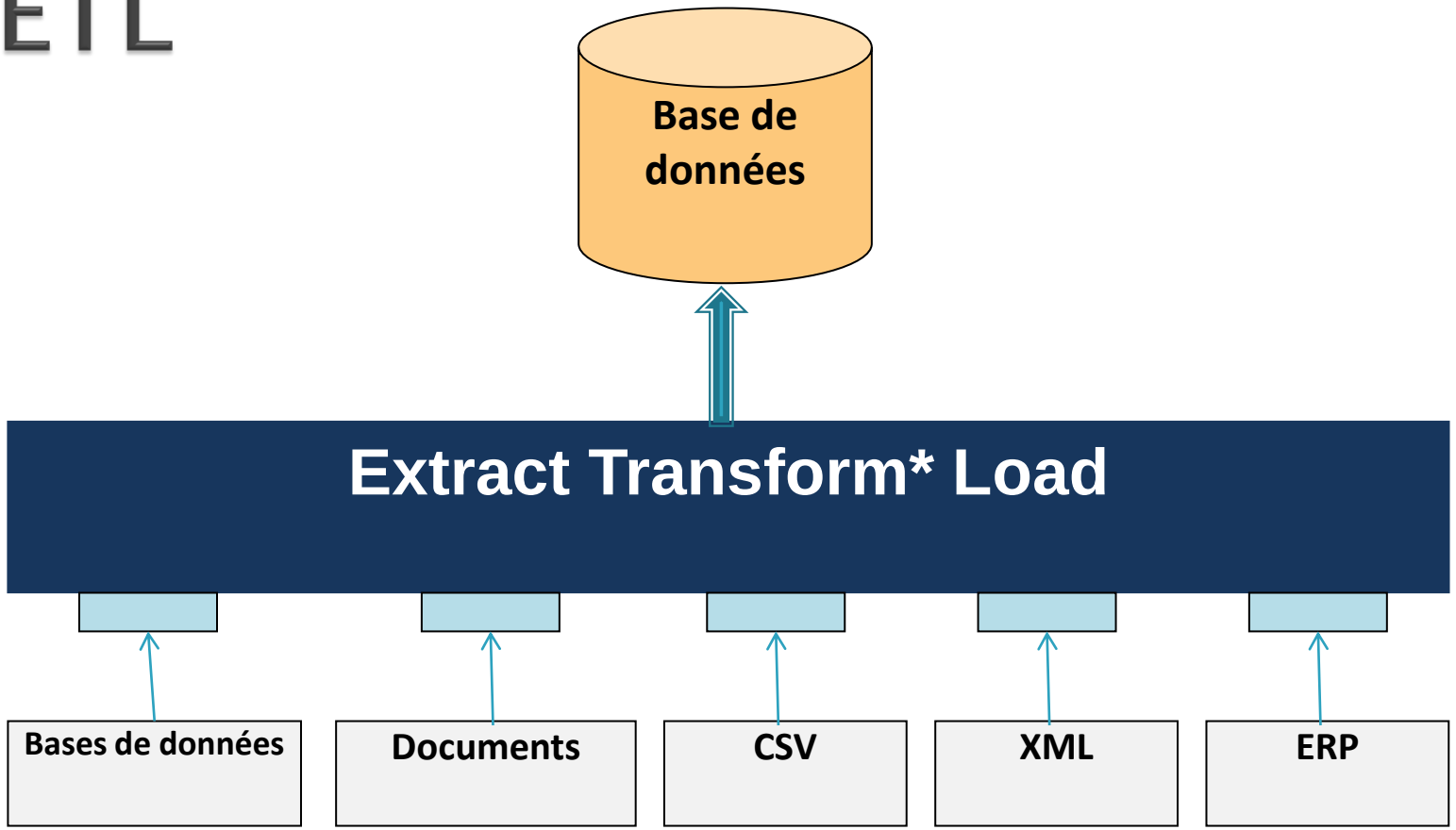
- Supports multiples:
  - Progiciels
  - Bases de données
  - Fichier XML, CSV
  - Fichier Excel
  - E-mail
  - ...

S'adapter aux systèmes opérationnels mis en place

Une réponse

- ETL : **E**xtract **T**ransform **L**oad
- Objectif : Réduire le coût de développement et de maintenance
- Avant d'être enregistrées dans le Data Warehouse, les données peuvent être :
  - Filtrées,
  - Agrégées,
  - Formatées,
  - Croisées

ETL



*Conversion de type, correction, filtre, agrégation, formatage

Ajouts majeurs des ETL

- Initiation facile à des non-informaticiens
- Maintenance moins coûteuse
- Conception rapide
- Lecture/Ecriture des données sur un large choix de supports
- Pas besoin de Data Warehouse pour les utiliser

Historique des ETL

5 générations d'ETL

- 1970 : Ordinateur Central (mainframe). Génération de COBOL.
- 1980 : Interface avec SQL. Usage limité avec SAP
- 1990 : Première IG. Connecteurs spécifiques
- 2000 : Gestion des flux temps-réel. Utilisation du XML
- 2005 : Informatique opérationnel.

ETL vs ELT

- Extract Load Transform : Même finalité mais fonctionnement différent

- ETL :Transformations effectuées ligne à ligne, par un moteur intermédiaire.
- ELT : Transformations en bloc, effectués par la source ou par la destination.

- Utilisation des outils fournis.
  - Bases de données

Choix en fonction du besoin ?

- ETL peu pertinent dans un environnement homogène
- On peut utiliser les deux !

ETL avec composants ELT pur :

- Oracle Data Integrator (Sunopsis)
- Stambia

Récapitulatif

ETL	ELT
Environnement hétérogène	Environnement homogène
Abstraction du support	Type de support restreint
Traitements effectués par le moteur intermédiaire	Traitement sur la source ou destination

Quand (ne pas) utiliser un ETL ?

ETL Tool versus Hand Coding ?
The answer is, “It depends".

- Les sources sont-elles susceptibles d'être modifiées ?
- Les supports sont-ils hétérogènes ?
- Les données doivent-elles être transformées ?

Les solutions...

Les open source...

Les retenus


pentaho™
open source business intelligence™


talend*
*open data solutions

PDI : Pentaho Data Integration

Anciennement Kettle, existe depuis 2000.
Open source, licence LGPL.

- Community Edition:
  - Gratuite.
- Entreprise Edition:
  - Version d'évaluation de 30 jours,
  - Support technique
  - Maintenance
  - …

Pentaho: Clients

PDI: Utilisation & Fonctions de bases

- Processus ETL créé avec *spoon*.
- Création de *transformations* (vue micro) :
  - traitement ligne à ligne de données.
- Création de *jobs* (vue macro) :
  - envois de mails, transfert de fichiers
  - appels de t*ransformations*, de *jobs*
- Planification avec *kitchen* et *pan*.

PDI: Fonctions avancées

- Utilisation de JavaScript pour le traitement spécifique de types primitifs.
- Envoi de mails
- Traitement FTP
- Communication avec serveur LDAP

PDI: Repositories

Centralisation : Les *jobs* et *transformations* peuvent être enregistrées dans un *repository*.

Les connexions aux bases sont aussi enregistrées.

PDI: Réaction à un dysfonctionnement

Si la source de données n'existe plus ?

Si on a modifié la structure de la source de données ?

Si une source (ou destination) est indisponible, l'outil est-il apte à en sélectionner une autre ?

Les administrateurs sont-ils informés ?

Talend Open Studio

talend*
*open data solutions

- Gratuit, Open source, Licence GPL
- Sortie officielle 5 oct. 2006

- Webinars, White Paper, Tutoriels, Documentation, Séminaires, Forum, Wiki

- International : USA, France, Japon, Autres pays européens

TOS : Suite

- Certification "Talend Certified"

- Payant :
  - **Talend Integration Suite** (Team, Pro, Entreprise...)
    - Travail collaboratif
    - Load balancing
    - Outils de déploiement
    - Support technique
    - Outils BI (Dashboard, Job analyses, Audit, SOA...)
    - Parallélisation des traitements

Talend : Clients

Talend: Utilisation & Fonctions de bases

- Basé sur Eclipse
- Génération de code Java ou Perl à partir des Jobs créées par l'utilisateur.
- Récupération des schémas des tables
- Génération de doc
- Traitements spécifiques par ajout de code en Java
- Composants principaux : SGBD, transformation, association, agrégation, filtrage, LDAP, XML, accès fichiers…

Talend: Fonctions avancées

- Mode debug/traçage
- Modifier le code
- Composants spécifiques : FTP, POP/SMTP, Shell…
- Composants personnels :
  - Création de ses propres composants (Java/Perl)
  - Utilisation de ceux de la communauté

Talend: Réaction à un dysfonctionnement

- Attrape les exceptions.
  - Sortie : Console ou Fichier
- Jouer un scénario différent en cas d'échec d'un composant.

Cas concret : Démo

Ce qu'on a constaté

- Durant l'installation ?
- Durant l'utilisation ?

- Deux philosophies différentes

- PDI:
  - NullPointerException
  - Problèmes d'actualisation lors de modification d'une base

- Talend:
  - Eclipse et TOS doivent être lancés sur deux workspace différents.

Ce qu'on a gagné

- Exemple 1 (PDI) :
  - Création et alimentation des données : 1 heure
  - Conception Job : 1 heure
  - Estimation en Java : 1,5 jours

- Exemple 2 (Talend) :
  - Création et alimentation bases de données : 1 heure
  - Conception Job : 2 heures
  - Estimation en Java : 2 jours

Ce qui ne va pas

- L'utilisation de code pour certains traitements => la partie la moins accessible à un non-informaticien.

- Les ETLs s'utilisent graphiquement : les développeurs peuvent avoir du mal.

PDI vs Talend

- Performances :
  - PDI plus rapide dans le cas de transferts sans traitement*.
  - Talend **beaucoup** plus rapide lors de croisements des données (facteur 20 dans certains cas)*.

*Benchmark indépendant.

PDI vs Talend

- Avantages pour PDI :
  - Communauté active.
  - Sauvegarde des tâches dans une base de données relationnelle.
  - Interface utilisateur plus appréciée.

- Avantages pour Talend :
  - Performances supérieures, selon de multiples benchmarks.
  - Prix
  - La solution évolue plus rapidement

Comparatif commercial

- Les services proposés par PDI :
  - service de maintenance
  - service d'hébergement
  - A partir de 45K € pour 3 ans.
- Les services proposés Talend :
  - Infrastructure indépendante au prix (intéressant pour les grandes infrastructures)
  - Tarifs axés sur le nombre de développeur. Concurrence : nombre de cœur CPU.
  - 5000€*/développeur/an. Dégressif à partir de 3 licences.
- Même tarif pour 3 développeurs.

*Valeur supposée

Pour conclure, les ETLs…

- Sont des outils en plein essor,
- Fournissent une grande assistance lors du **développement** et de la **maintenance**,
- Ainsi que **l'abstraction** des supports.

- Mais : les traitements complexes ne seront jamais **entièrement** accessibles à des non-informaticiens.

Bibliographie

- Livres :
  - Le data webhouse – Analyser les comportements client sur le Web, Ralph Kimball et Richard Merz, Edition Eyrolles.
  - Le Data Wharehouse – Le Data Mining, Jean-Michel Franco et EDS-Institut Prométhéus, Edition Eyrolles.
  - Data mining et scoring – Bases de données et gestion de la relation client – Stéphane Tufféry, Edition Dunod.
- Mémoire :
  - Le positionnement des systèmes ETL – Berthier Rodolphe
- Web :
  - http://etl-tools.info
  - http://en.wikipedia.org/wiki/Extract,_transform,_load
  - http://www.stylusstudio.com/etl/
  - http://blog.atolcd.com

Merci, des questions ?