BI dans les nuages

Olivier Bendavid, UM2

Prof. A. April, ÉTS

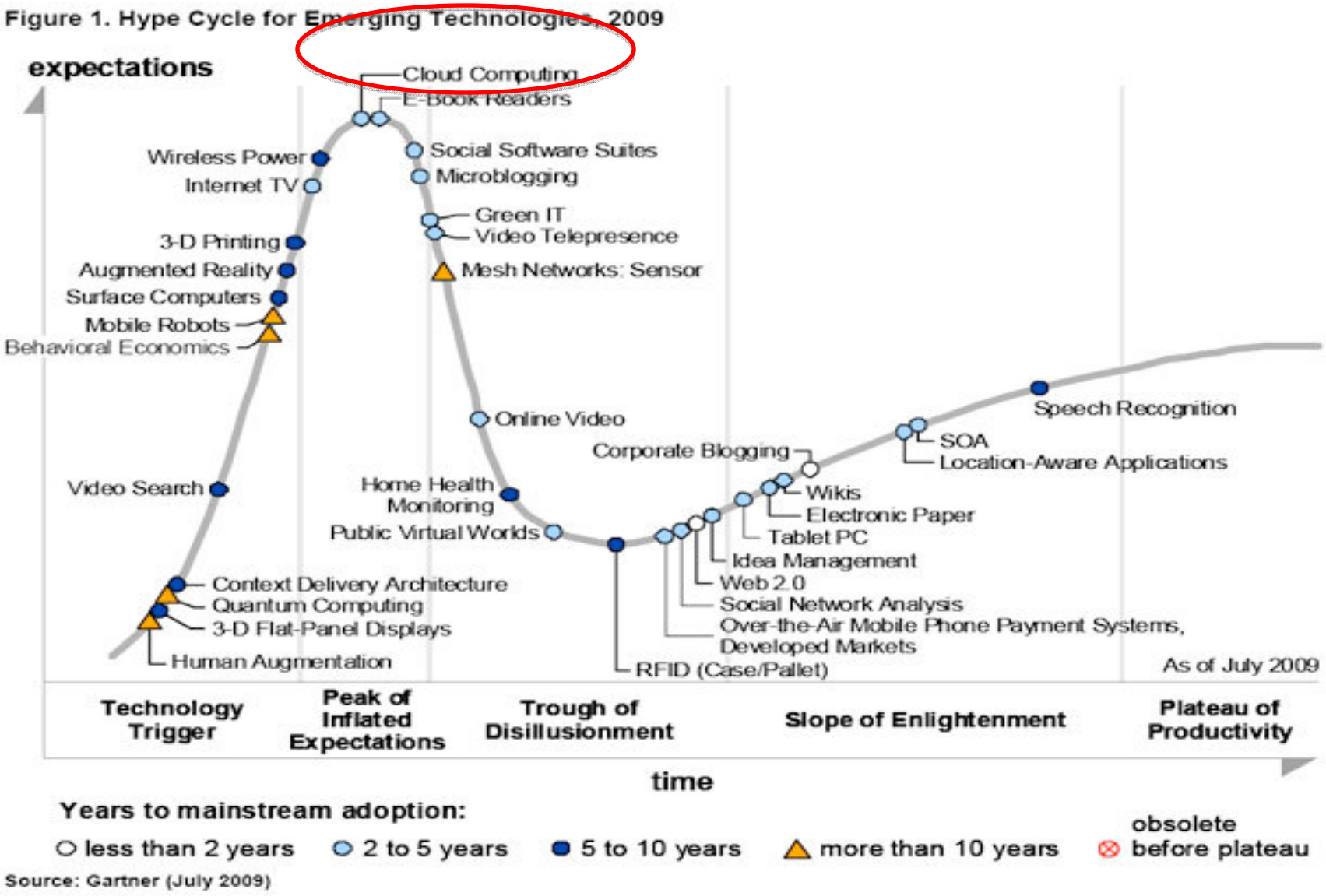
Table des matières

Introduction

- Quelles sont les défis actuels du BI ?
- Qu'est-ce que l'informatique en nuage (Cloud computing) ?
- Comment ces technologies émergentes peuvent répondre a ces défis?

Cycle de la surenchère (hype)

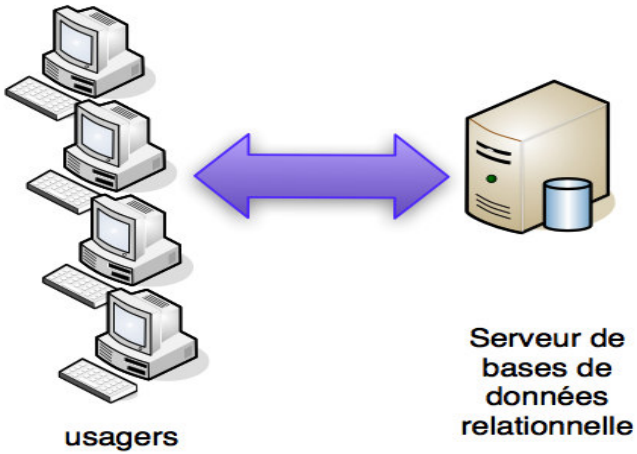
http://marketingtypo.com/2010/01/17/gartner-hype-cycle-2009-%E2%80%93-is-cloud-computing-over-hyped/



Quels sont les utilisateurs de ces technologies ?

Pourquoi migrer vers cette technologie ?

Débutent tous leurs architectures avec BD Relationnelles



Éprouvent progressivement des difficultés en analyses et gestion de BD:

- Volumes croissants
- Requêtes de plus en plus complexes
- Limité à quelques centaines de serveurs de BD
- La gestion devient progressivement impossible

Ex: Technologie BigTable de Google

Google a inventé une technologie pour traiter beaucoup de données (plusieurs pétabits). Une technologie qui se compose de :

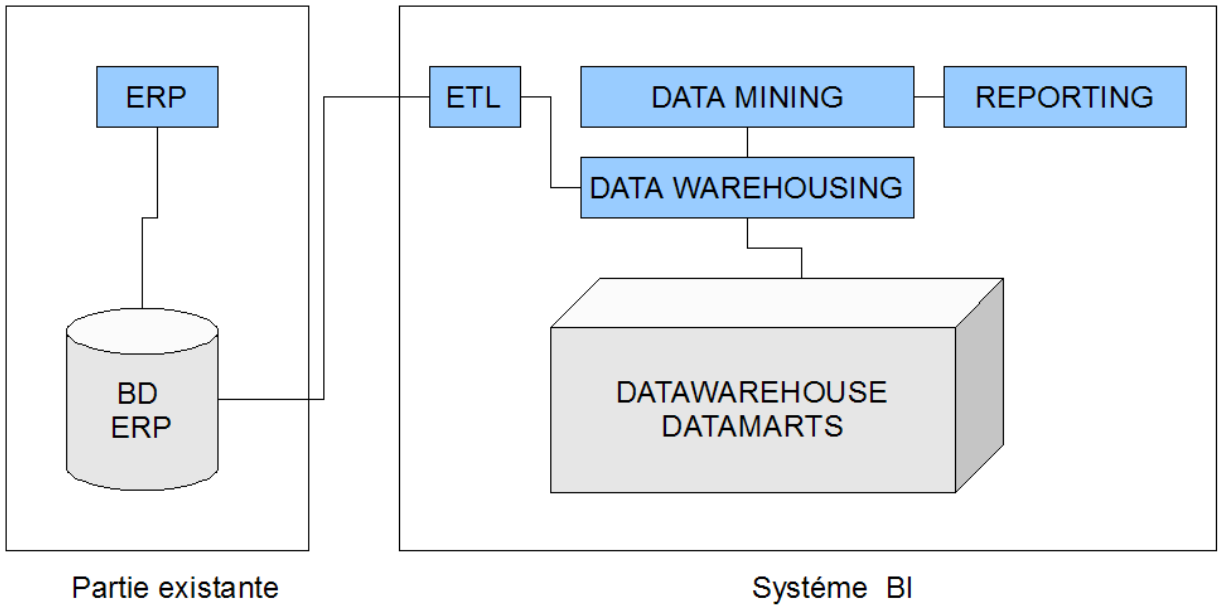
- ❑ Un système de fichier distribué **GFS**
- ❑ Un modèle de données **MAPREDUCE** pour traiter des grandes quantités de données en parallèle
- ❑ Une base de données distribuée (**BIGTABLE**) sur plusieurs milliers d'ordinateurs de commodités.
- ❑ Ordinateurs de **commodités** peu couteux (≈40 serveurs par rack, ≈150 racks par data center, 36+ data centers) 1 cluster = ≈2000 serveurs

Exemple de grappe de Google

2007 Seattle Scalability Conference:

Le projet
(CRM - analyse des ventes)

Étude de l'utilisation de cette technologie pour réaliser un système BI pour un CRM.



Le projet : Objectifs

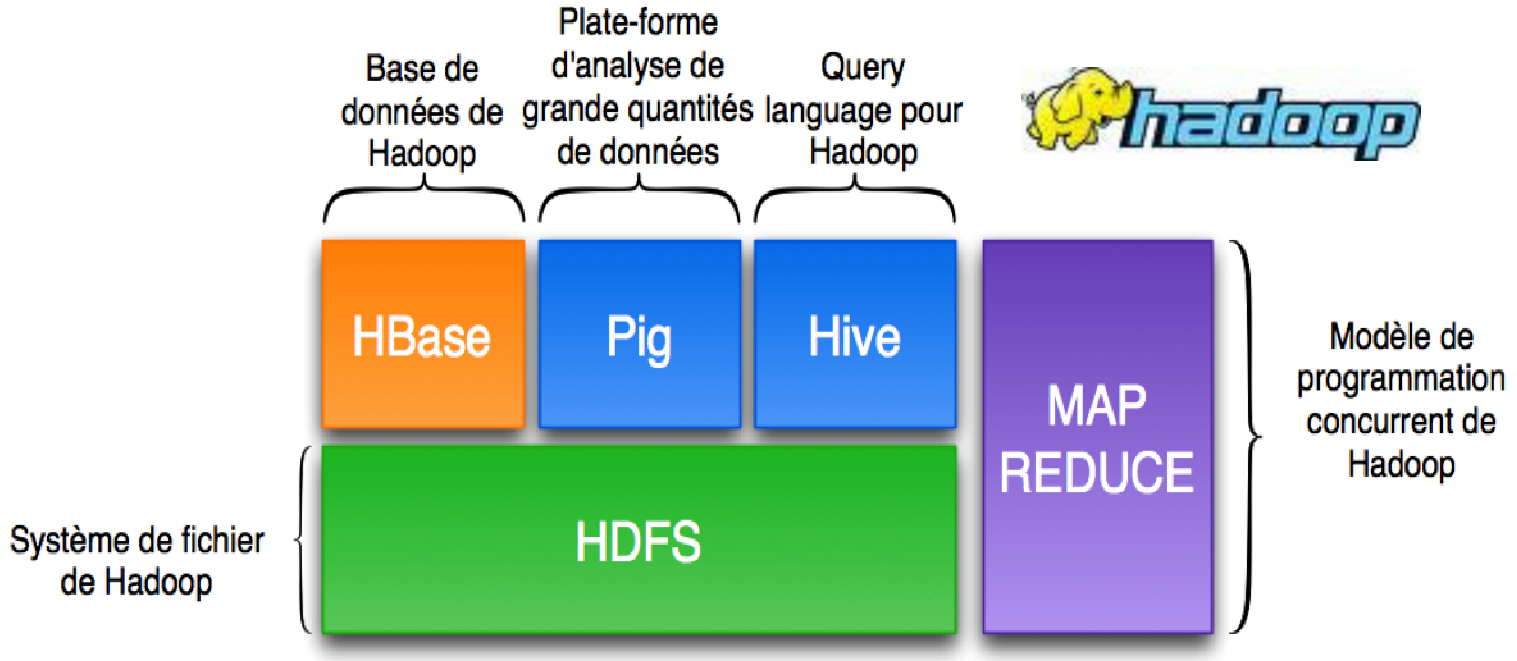
Objectifs a court terme :

- maîtriser les technologies, éliminer les problèmes de croissance et de gestion, diminuer les coûts d'acquisition d'équipements
- diminuer les temps de chargement et d'analyses des données pour obtenir une performance 'temps réel'

Objectifs à moyen terme :

- CRM (*incluant BI*) disponible '*on the Cloud*' pour compétitionner Sales*force*

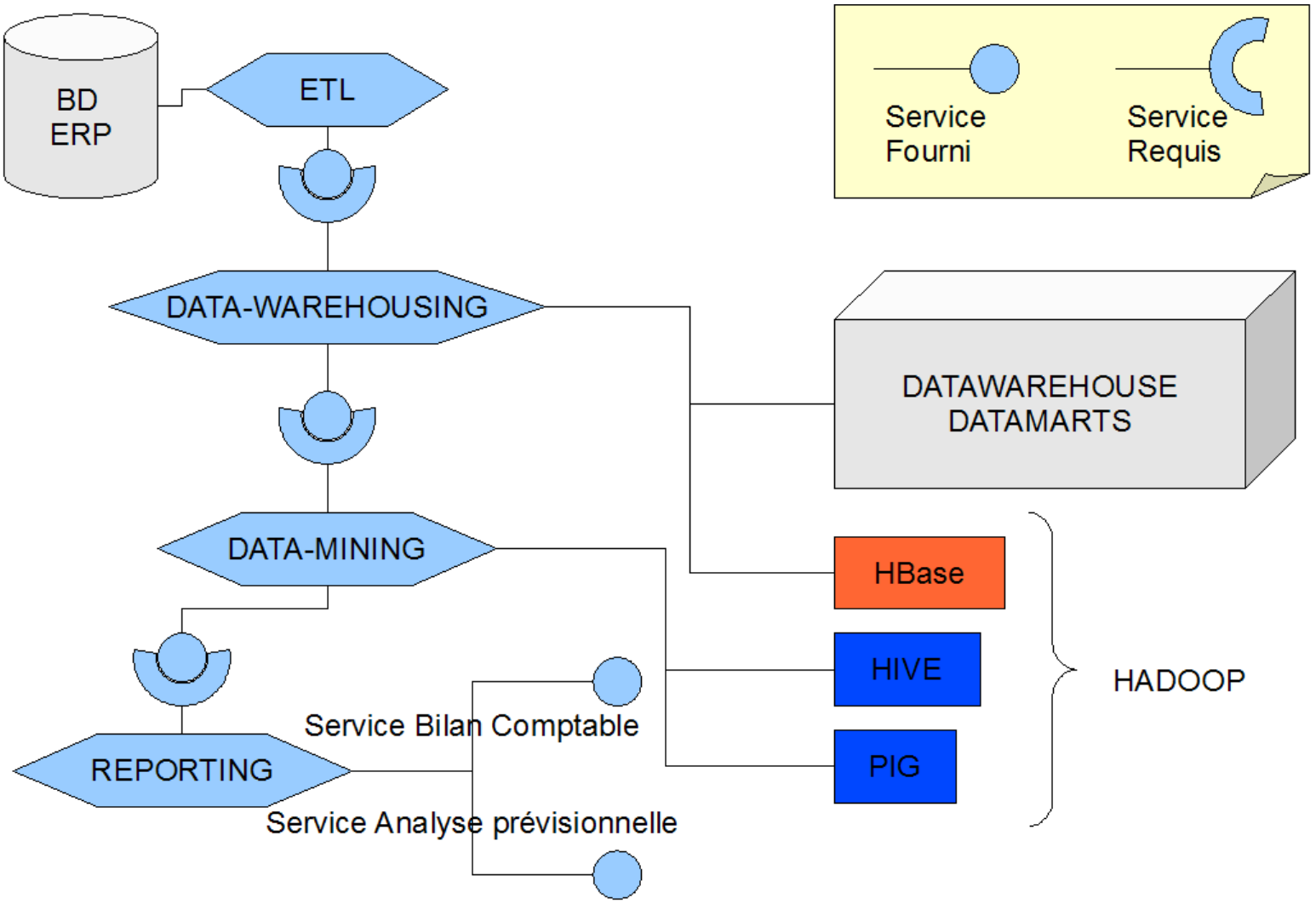
Logiciel libre (hadoop)



Vous pouvez retrouver ce projet Apache a l'adresse suivante : http://hadoop.apache.org/

Le Projet

L'architecture du projet et les liens avec hadoop



Étape 1 - ETL

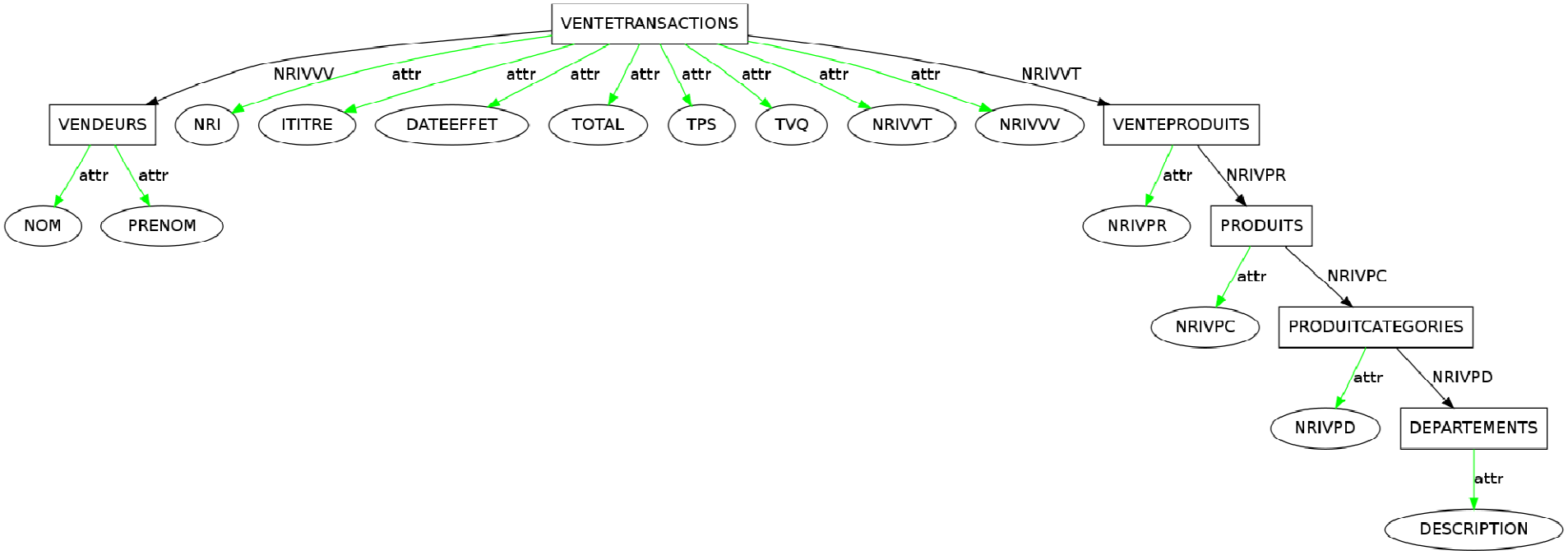
- ***Extraction :*** connexion à la BD source.
- ***Transformation :*** génération des requêtes pour préparer le chargement.
- ***Chargement :*** préparer l'écriture dans l'entrepôt de données.

2 activités principales :

- ✓ Le peuplement initial
- ✓ La synchronisation à la bd source

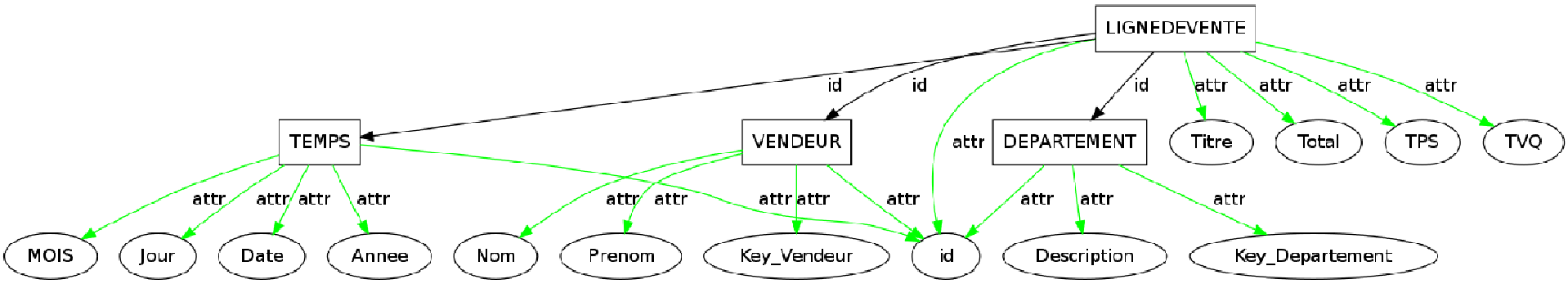
Procédure ETL
Utilisation des modèles

Représentation partielle du modèle relationnel de la bd source



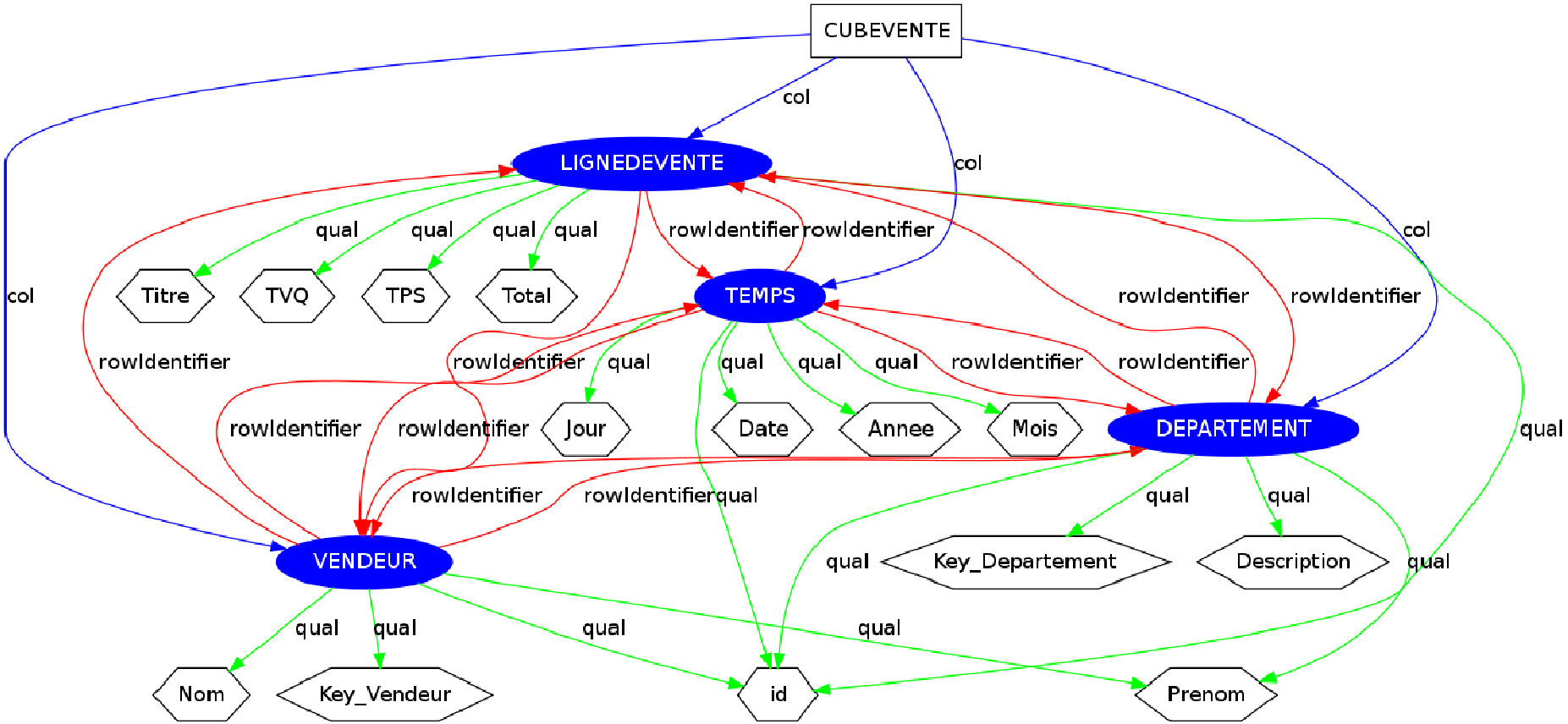
Procédure ETL
Utilisation des modèles

Représentation du modèle relationnel en étoile correspondant à la nouvelle organisation désirée



Procédure ETL et utilisation des modèles

Représentation sous forme de modèle d'un cube



Étape 1 - ETL

Expérimentation phase peuplement initial :

❖ Actuellement le chargement de 100 000 lignes de transactions en 50 minutes en utilisant un seul ordinateur de commodité et nous allons diminuer ce temps à quelques secondes dès lors que la grappe (cluster) sera mise en place.

❖ Test de la procédure de chargement basé sur les transformations de modèles pour le chargement de la table de faits du modèle étoile correspondant.

Étape 2 - Data Warehousing

- Administration de l'entrepôt de données

- Utilisation des modèles pour proposer des schémas de navigation

Expérimentation :

- ✓Création/Suppression de plusieurs entrepôts
- ✓Lecture/Écriture sur entrepôt

Étape 3 - Data Mining

- Analyse des données
- Utilisation des modèles

Première expérimentation réalisée sur 10 000 lignes :

Choix de générer plusieurs entrepôts (par transactions, par mois et un dernier par année) (création des 3 entrepôts en 16 min)

- ✓ Total par département
- ✓ Total par département par Mois
- ✓ Total par département par Année

De la même manière que pour l'ETL nous espérons obtenir un gain de temps (quelques secondes) avec la grappe en place. Dans notre test, ayant fait le choix de générer plusieurs entrepôts et de les maintenir. Cela nous permet d'obtenir les résultats de requêtes liées à la dimension temps en « temps réel ».

Étape 3 - Reporting

- Architecture REST
- Web services (basé wsdl)

Mise en place d'un serveur Tomcat6 de Apache et utilisation de Axis2 pour la mise en place des services web

Prochaines étapes

- Finaliser la partie REPORTING
- Déployer la grappe 

: ajouter des rack un à un pour obtenir la performance visée
- Augmenter les volumes de données (des centaines de téraoctets) et grossir la grappe: Objectif d'analyse de données en temps réel
- Intégration de l'architecture dans le *framework* et rendre disponible aux clients

Conclusions

- Le *Cloud Computing* a été conçu pour traiter des grandes quantités de données donc a un potentiel de régler des problèmes de BI – pas de hype !
- Une vision R&D permet :
  - ✓ Exploration de nouvelles technologies
  - ✓ Devancer la compétition
  - ✓ Financer les projet avec les crédit d'impôts RS&DE

Références

- Leavitt, N. **2010**. « **Will NoSQL Databases Live Up to Their Promise?** ». *Computer magazine, the flagship publication of the IEEE Computer Society, n° February 2010, p. 12–14.*
- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. WallachMike Burrows, Tushar Chandra, Andrew Fikes, Robert E.Gruber

  Bigtable: **A Distributed Storage System for Structured Data** [2006]

Des questions ?