Conception de systèmes d'information et d'entrepôts de données

Vers des structures informatiques malléables

André Miralles & Guilhem Molla

AgroParisTech ENGREF
cirad
irstea

Plan de l'exposé

- **Entrepôts de données**
  - Origine et architecture
  - Notion de Cube multidimensionnel
- **Les Outils ETL**
  - Exemple Talend
- **Projet SIE Pesticides**
  - Exemple de mise en œuvre des entrepôts de données environnementales
- **Conclusion**

Entrepôt de données

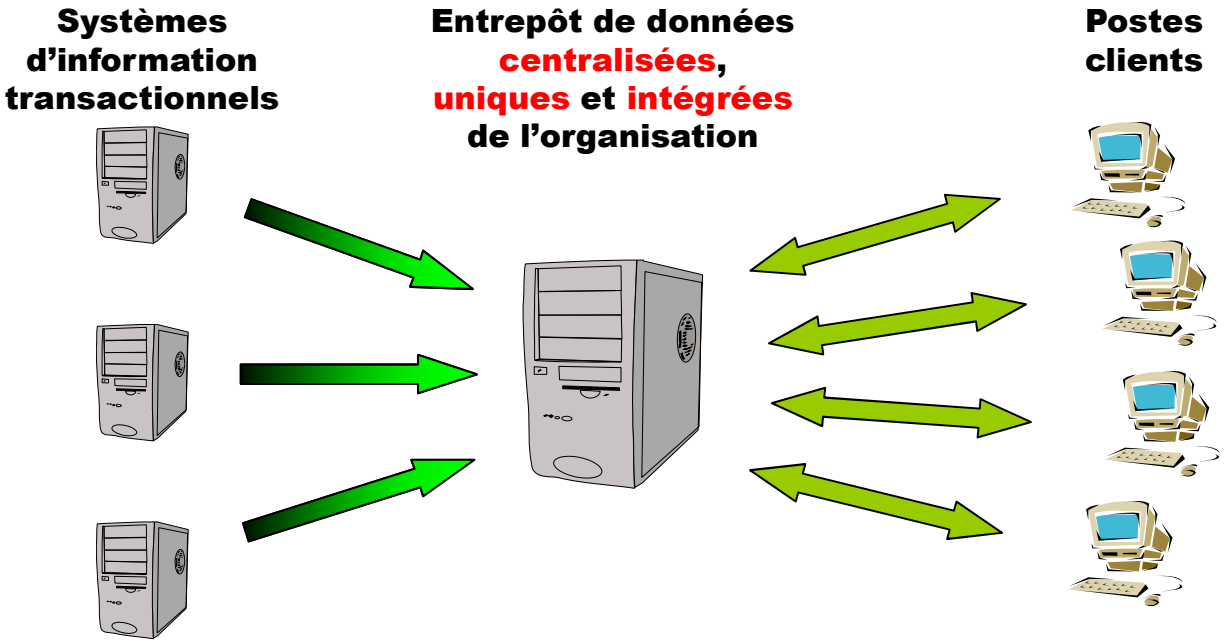
Origine et architecture

Rappel historique

- **Entrepôts de données : Famille d'outils informatiques**
  - **Conçus pour répondre au besoin de prise de décision rapide de la part de la Grande distribution**
    - **Gestion de flux financiers**
      - **Suivi du Chiffre d'affaire (CA)**
    - **Dédiés à l'aide à la décision**
      - **Temps de réponse rapide (de l'ordre de quelques secondes) et constants quelque soit la complexité des requêtes**

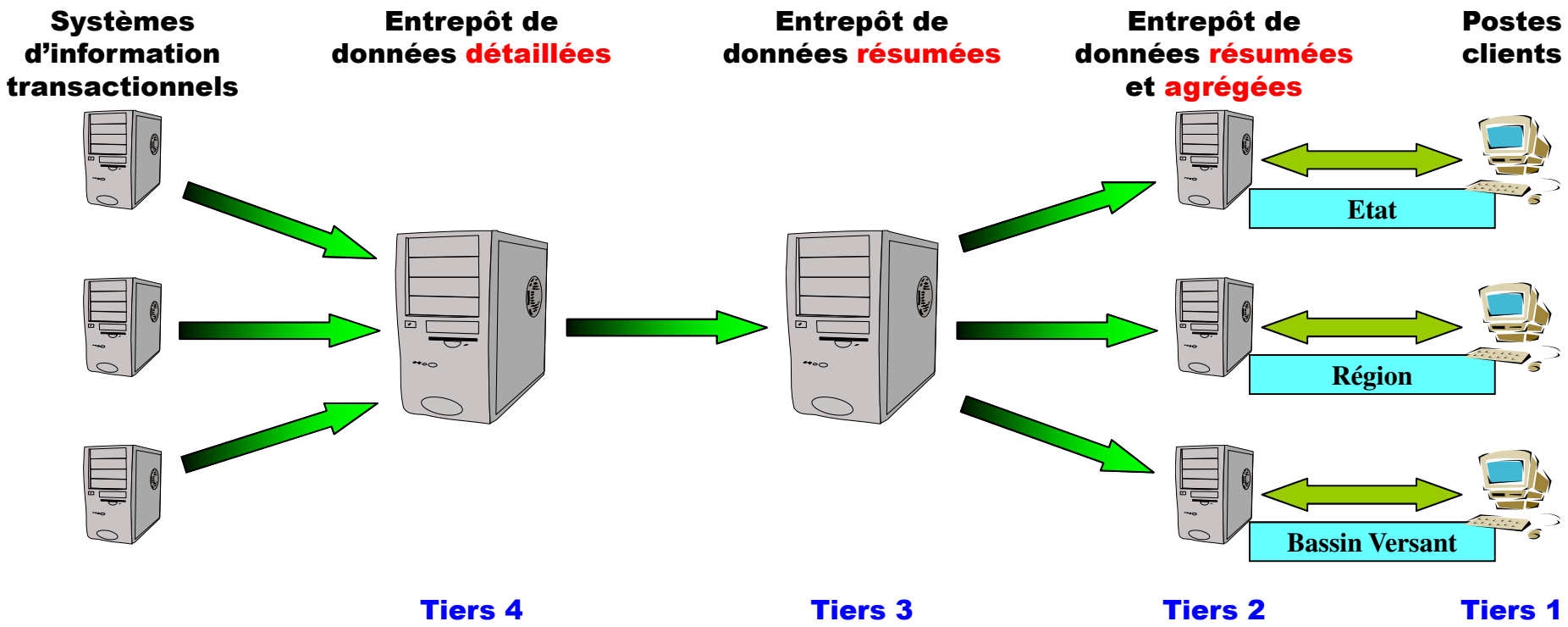
Architecture centralisée

- **Les données sont centralisées au sein d'une même plateforme**



Architectures n-tiers

- **Architecture Multi-tiers où les données sont résumées avant d'être agrégées**

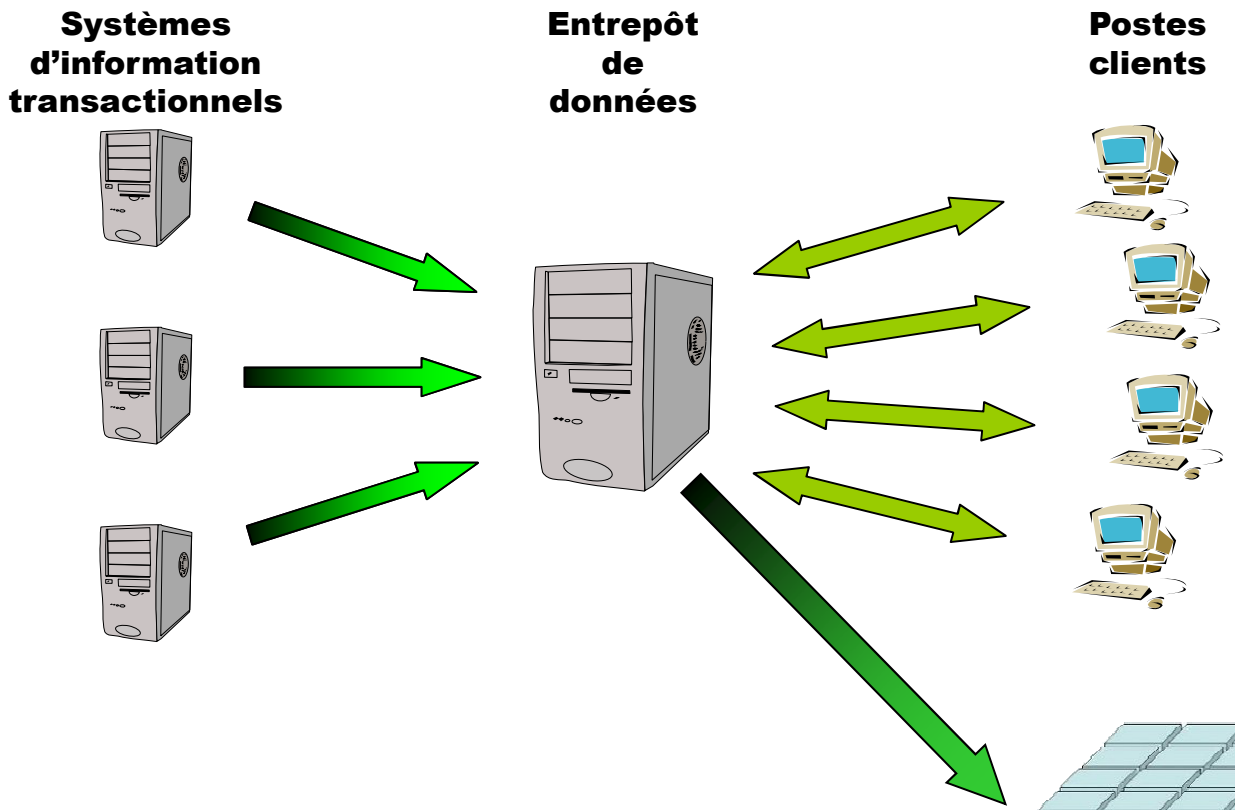


Entrepôt de données

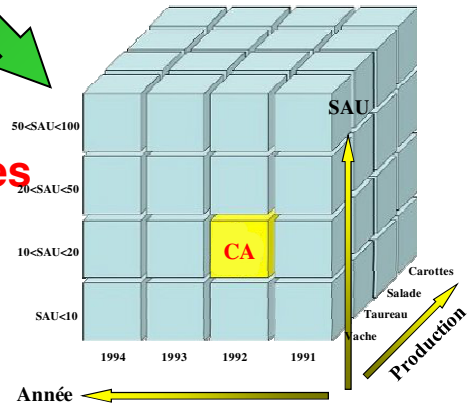
Notion de Cube multidimensionnel

AgroParisTech ENGREF
cirad
irstea

Notion de Cube multidimensionnel

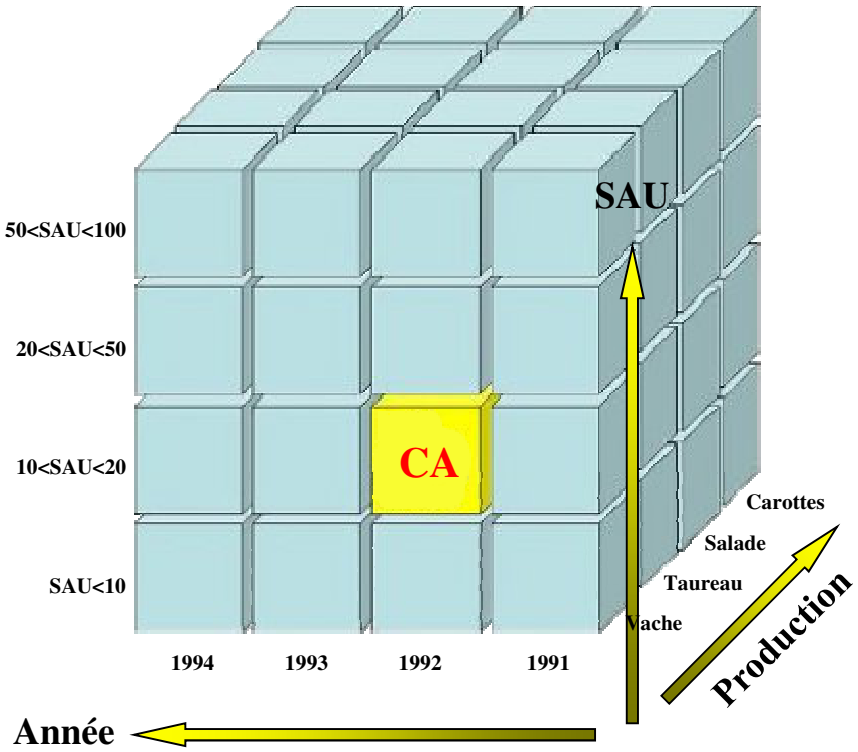


Chiffre d'affaire des Entreprises Agricoles

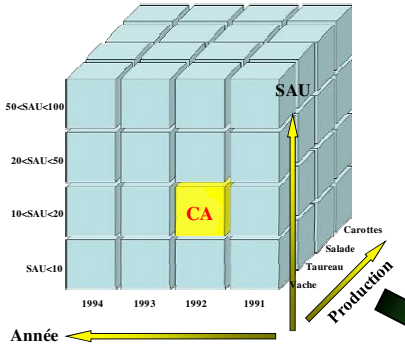


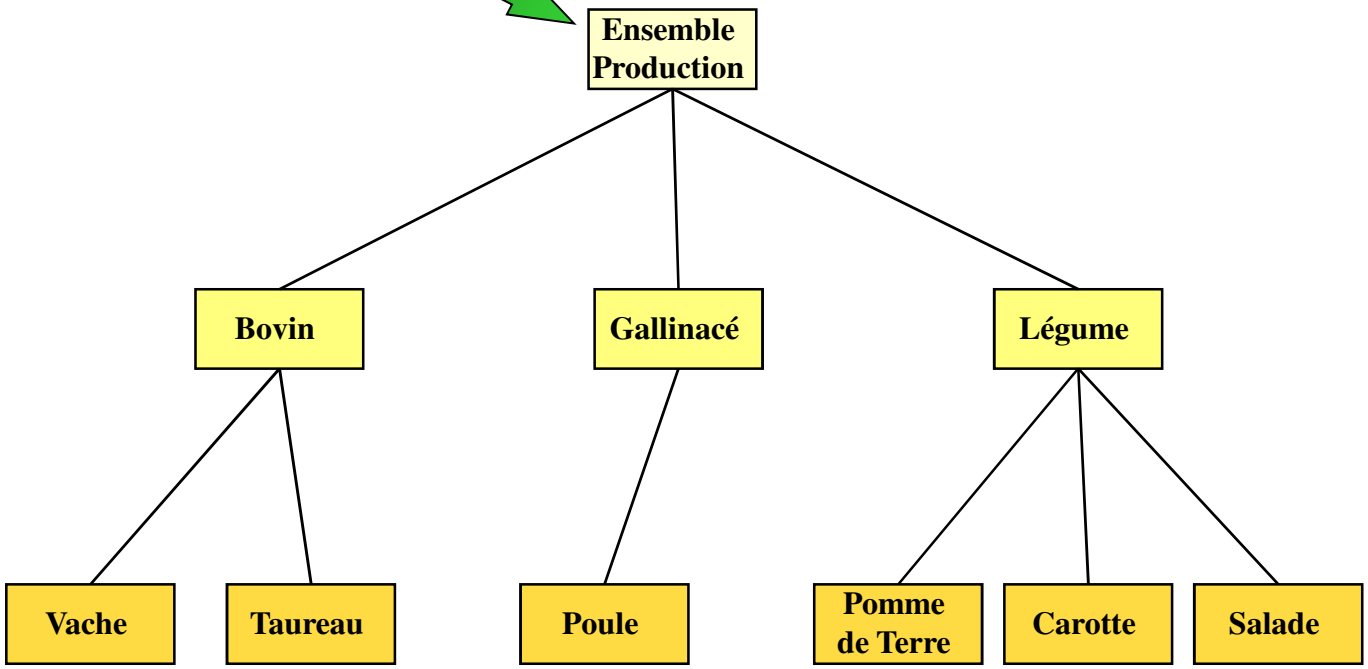
Notion Cube multidimensionnel

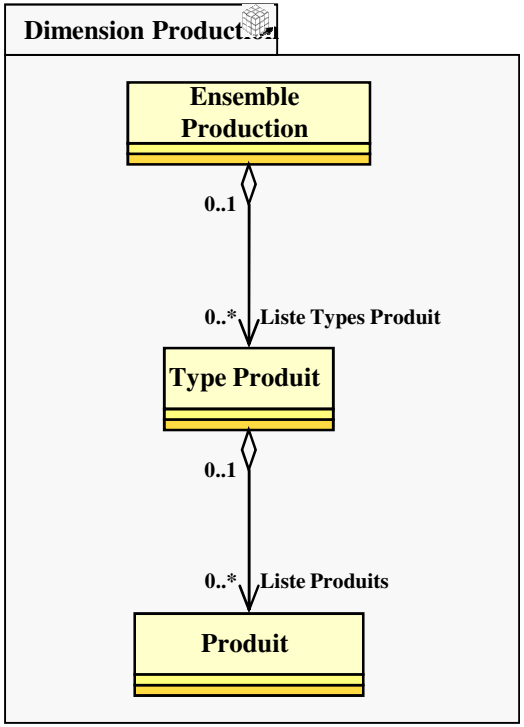
- **Exemple**
  - **Chiffre d'affaire (CA) d'une Entreprise Agricole**



Organisation Agrégative
Dimension Production Agricole

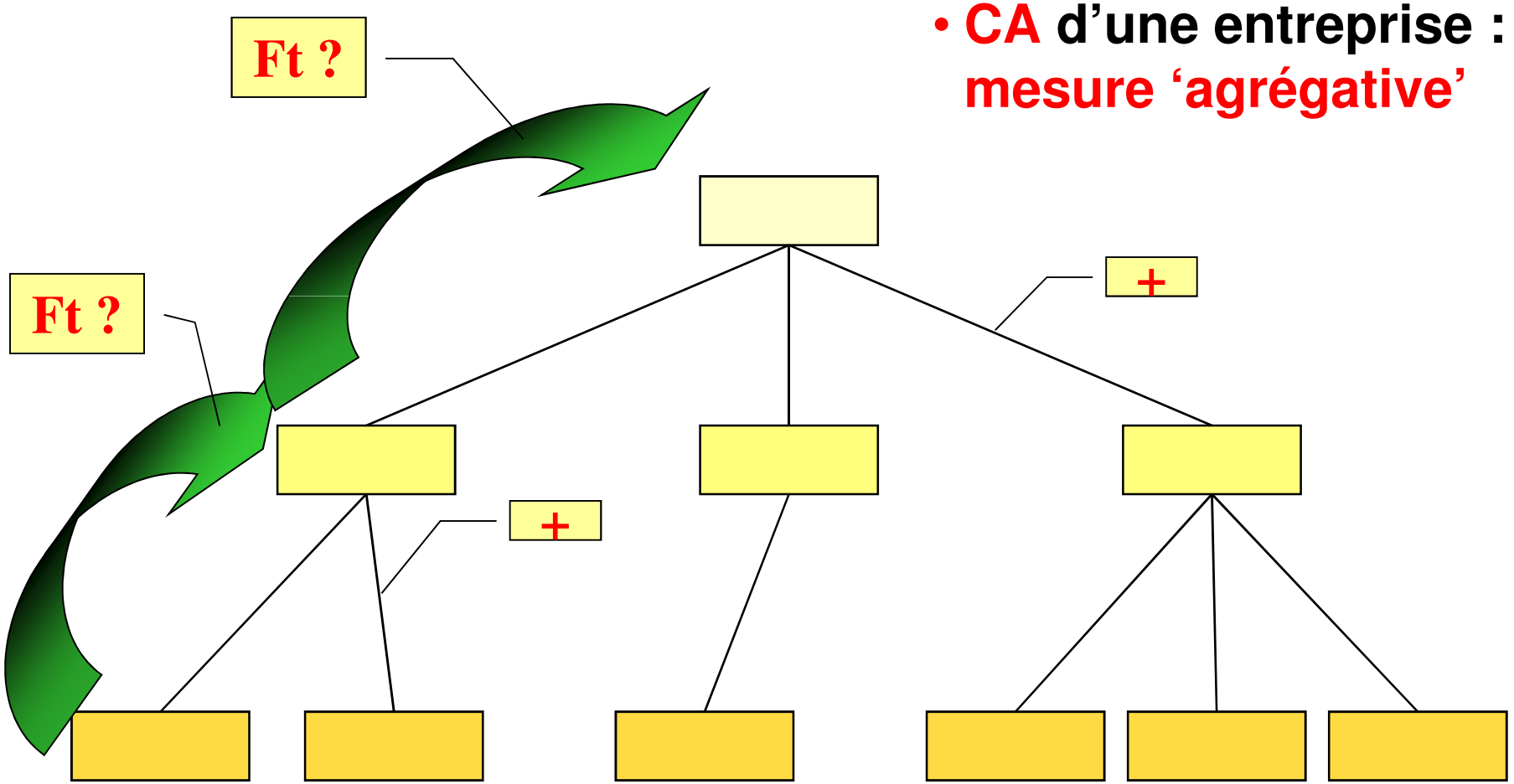






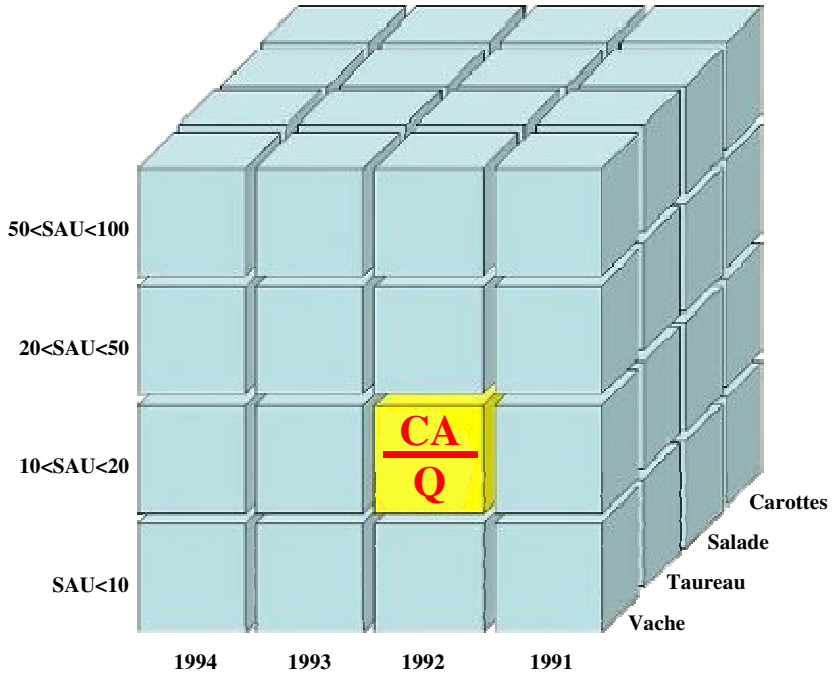
Agrégation : opérateur historique

- **CA d'une entreprise : mesure 'agrégative'**



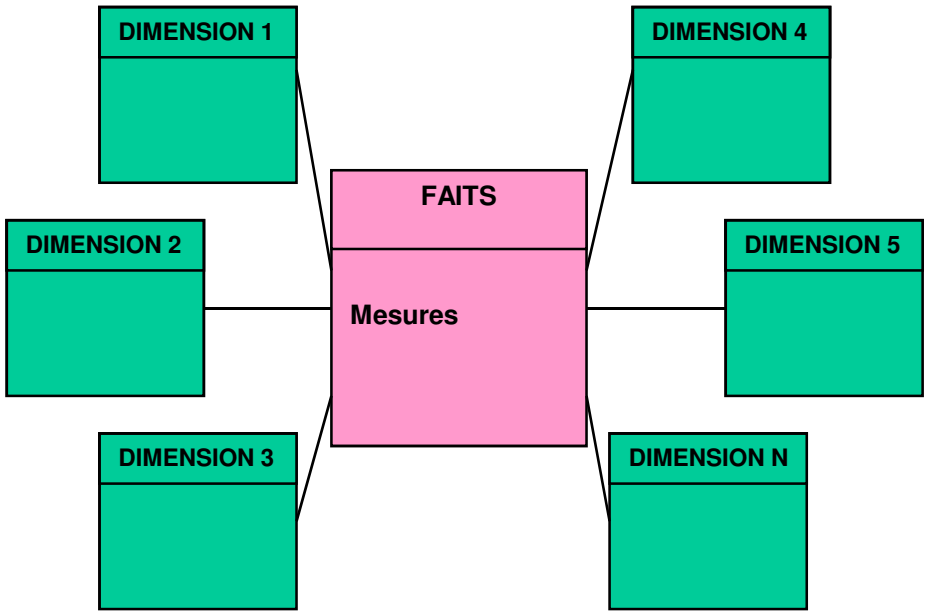
Notion Cube multidimensionnel

- **Exemple**
  - **Mesure : Chiffre d'affaire (CA) d'une Entreprise Agricole**
  - **Mesure : Quantité de produit (Q)**
    - **Mesure dérivée : Prix unitaire**



Les structures multidimensionnelles dans un SGBDR

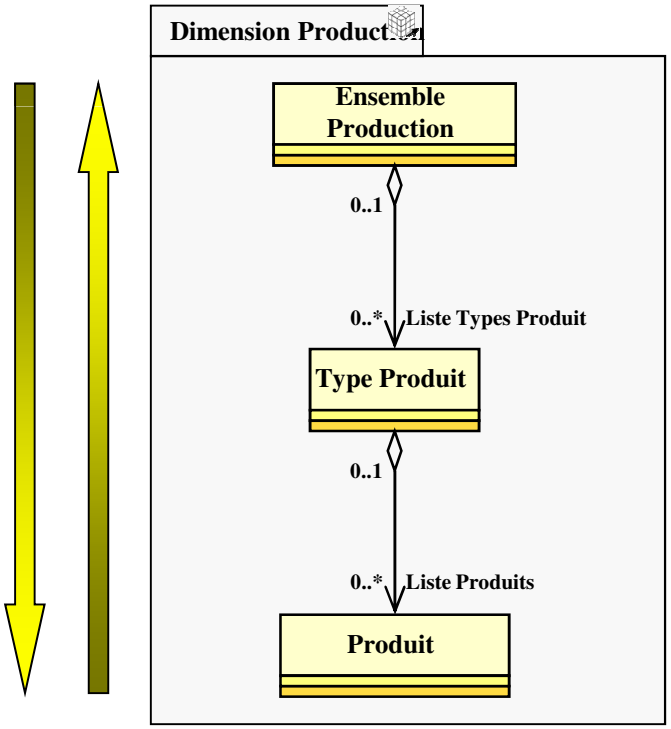
– **Modèle en étoile** (Star Schema)



– **Modèle en flocon (Snowflake Schema)**
– **Modèle mixte (Mixed Schema)**
– **Modèle en constellation (Fact Constellation Schema)**

Outils OLAP (On-Line Analytical Processing) / Spatial OLAP

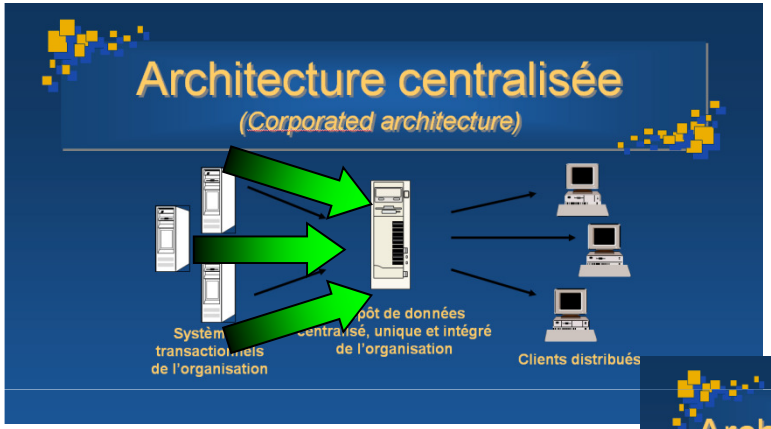
- **Opérateur de forage avant (Drill Down) / forage arrière (Drill Up)**
  - **Navigation à travers plusieurs niveaux d'une dimension**
    - **Niveau global vers niveau détaillé ou l'inverse**

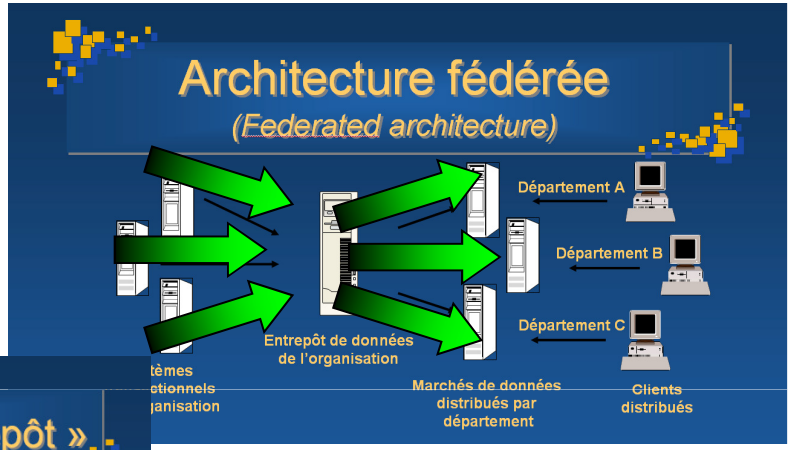


Les outils ETL

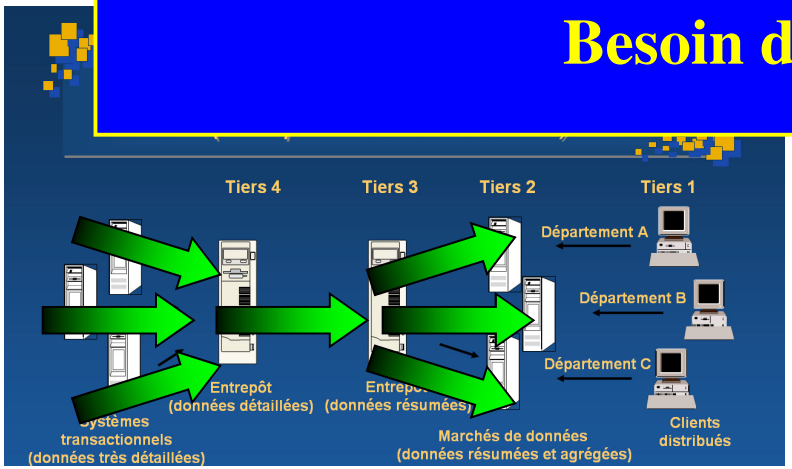
Exemple de Talend

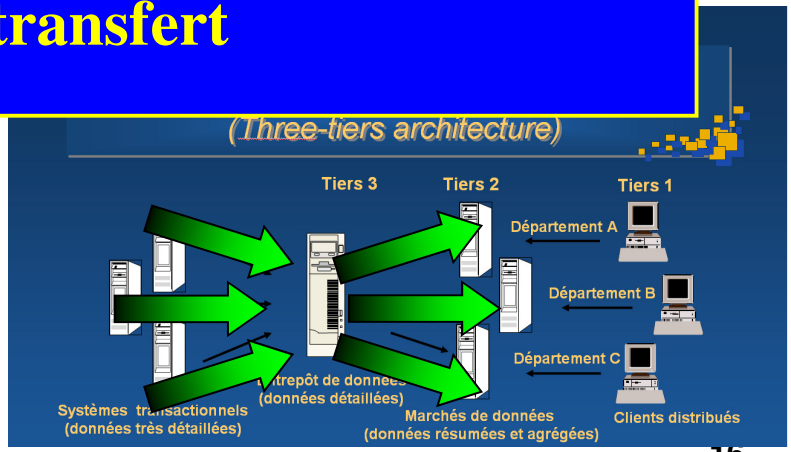
Transferts de données





Besoin d'un outil de transfert





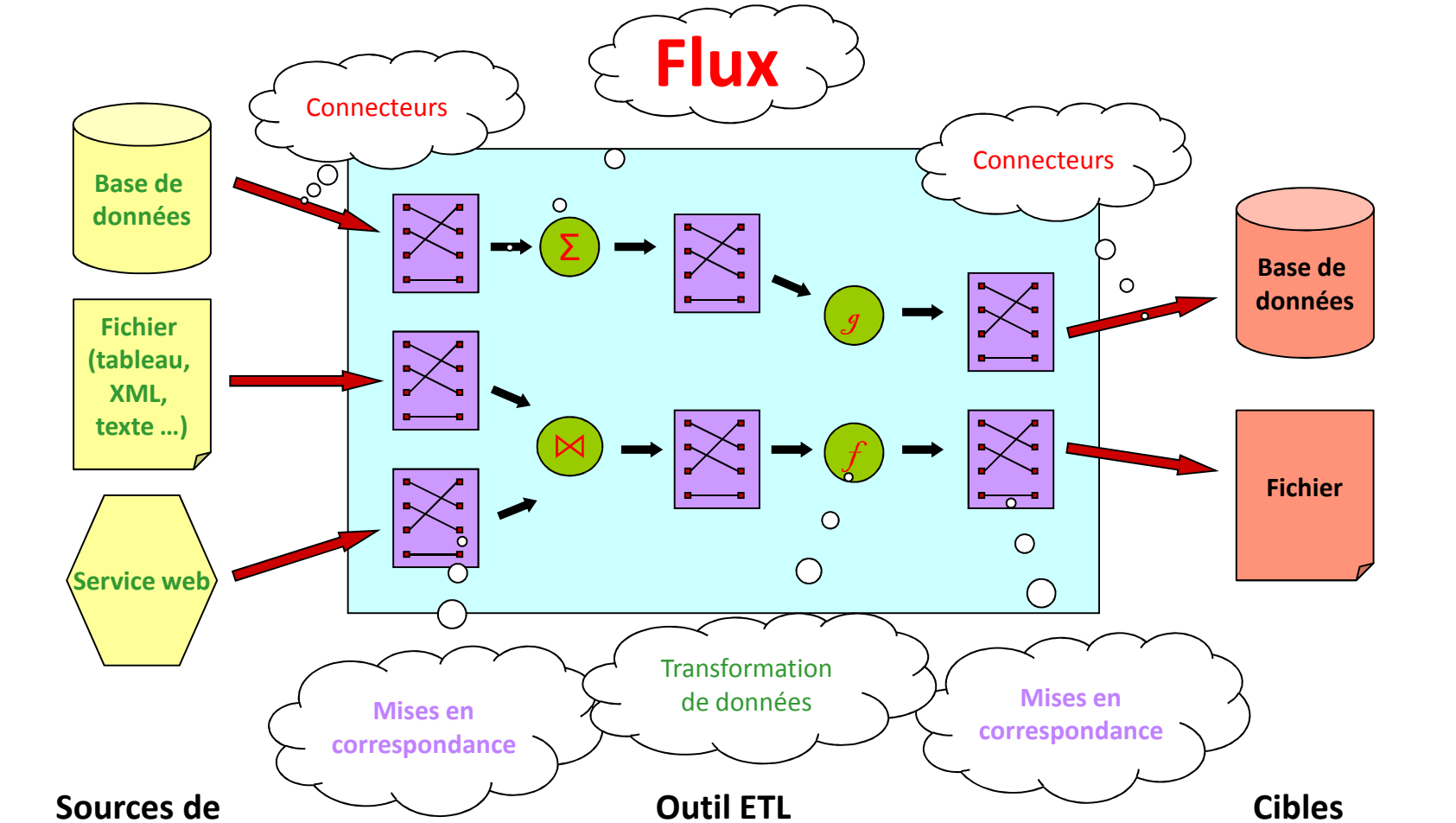
Définition d'un ETL

- **Technologie informatique intergicielle (middleware)**
  - Transformer des données sources en des données « cibles »
  - Synchroniser des sources de données entre elles

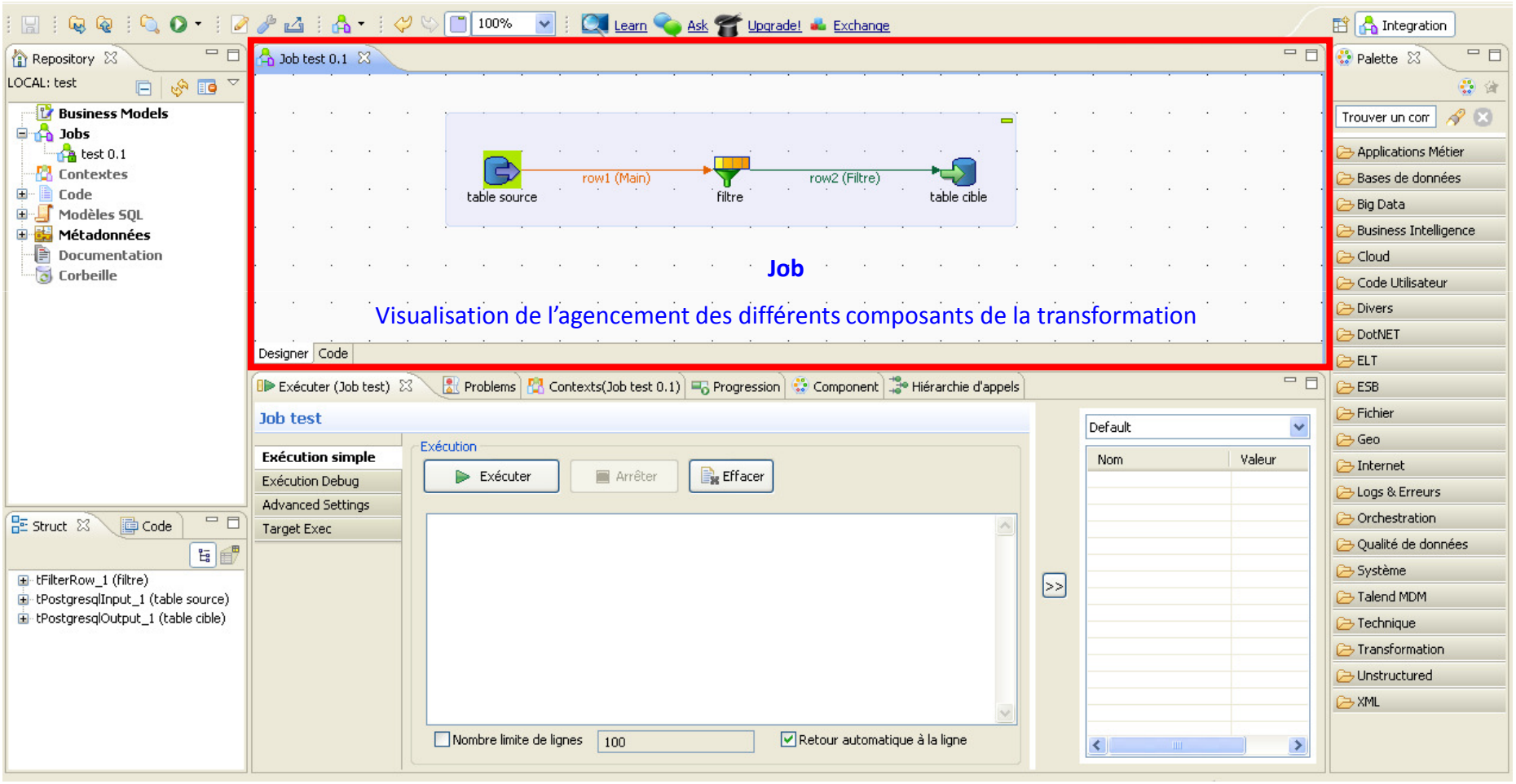
Définition d'un ETL

- **E : Extract**
  - Extraction depuis des sources de données
- **T : Transform**
  - Transformation des données : pour correspondre à un nouveau schéma/structure, pour les corriger, pour effectuer des calculs …
- **L : Load**
  - Chargement des données vers une source de données
  - Mise à jour éventuellement de données existantes

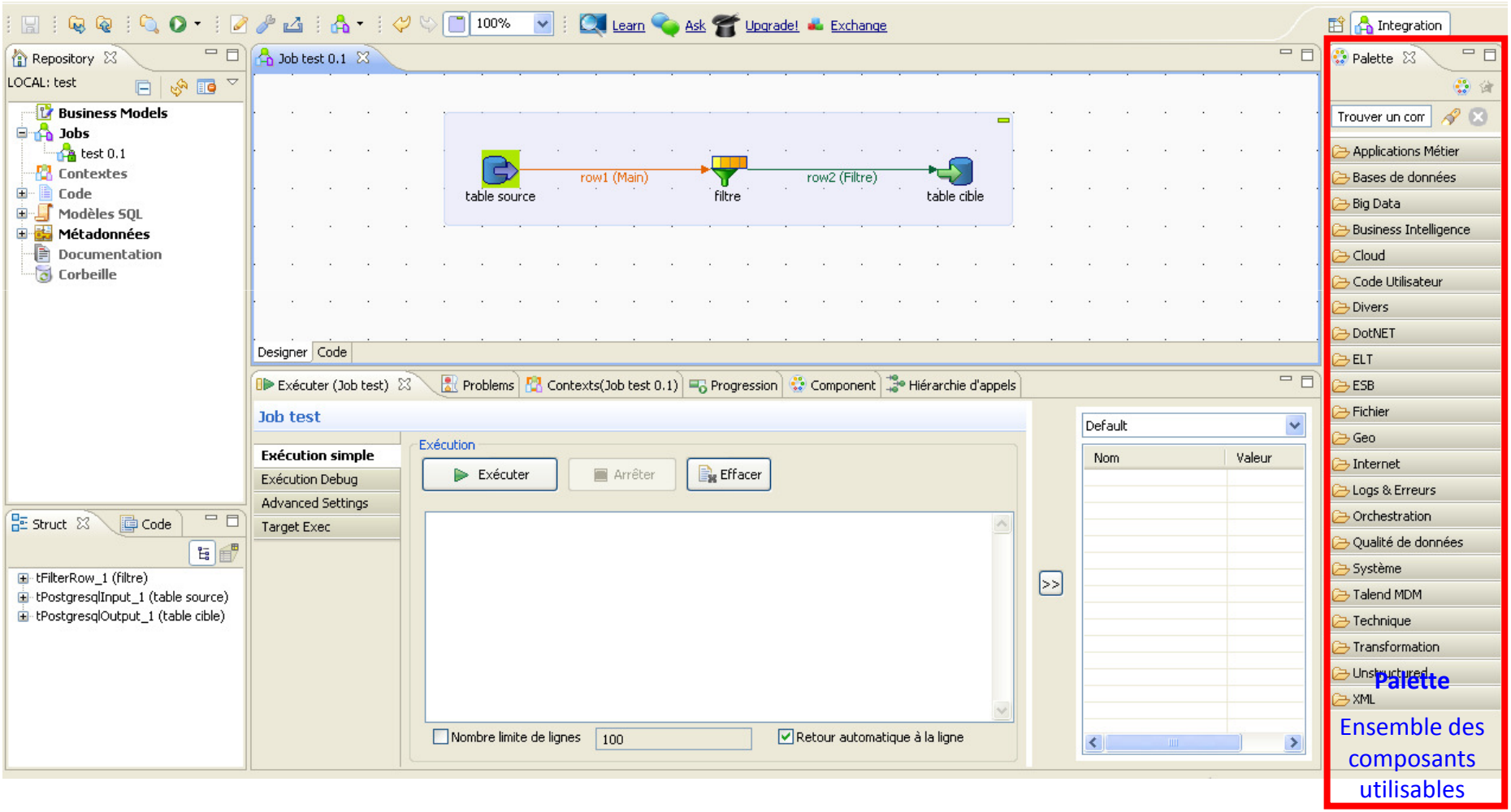
Processus de transformation



Environnement de travail de Talend

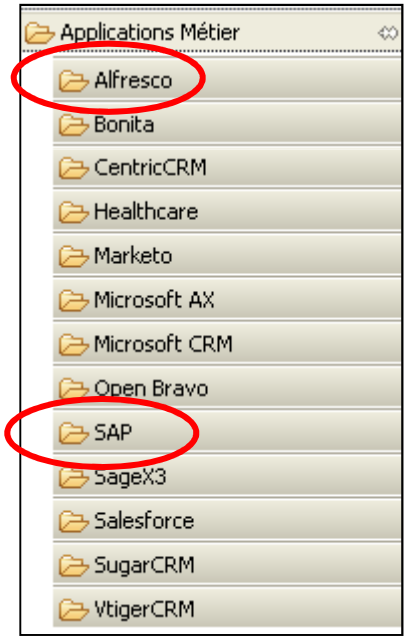


Environnement de travail de Talend



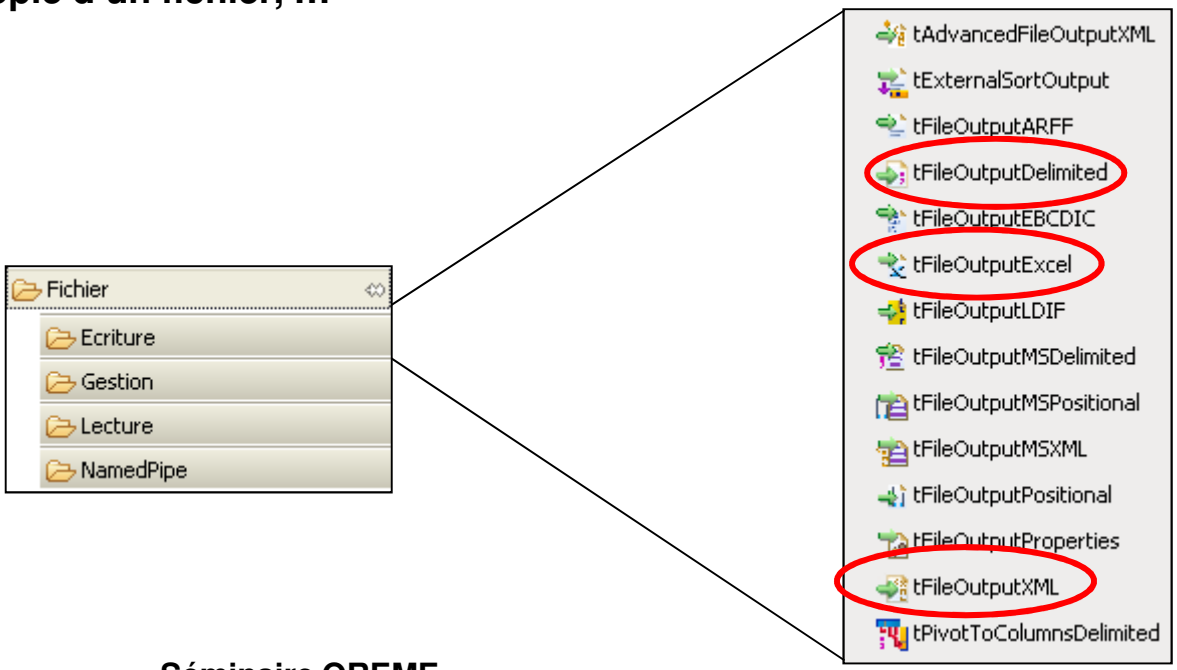
Composants disponibles dans Talend

- **Famille Applications métiers**
  - **Permettre de communiquer avec des applications extérieures comme des ERP par exemple (ici, SAP et Alfresco sont disponibles)**



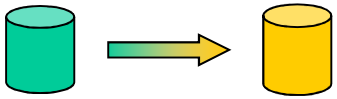
Composants disponibles dans Talend

- **Famille Fichier**
  - **Permettre le traitement de fichiers**
    - **Ecriture : Fonctionnalités d'écriture vers divers types de fichiers (Excel, XML, texte délimité, …)**
    - **Lecture : de la même façon, permet de consulter ces fichiers**
    - **Gestion : Permet le parcours d'un répertoire, la création de fichiers temporaires, la copie d'un fichier, …**



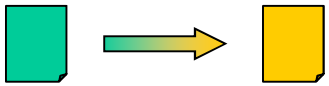
Quelques types d'utilisation

- **Intégration de données** dans une base

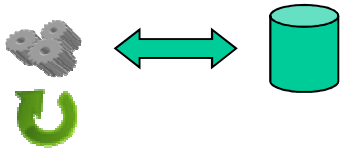


- Vérification de la **qualité** des données

- **Conversion de types** de fichiers



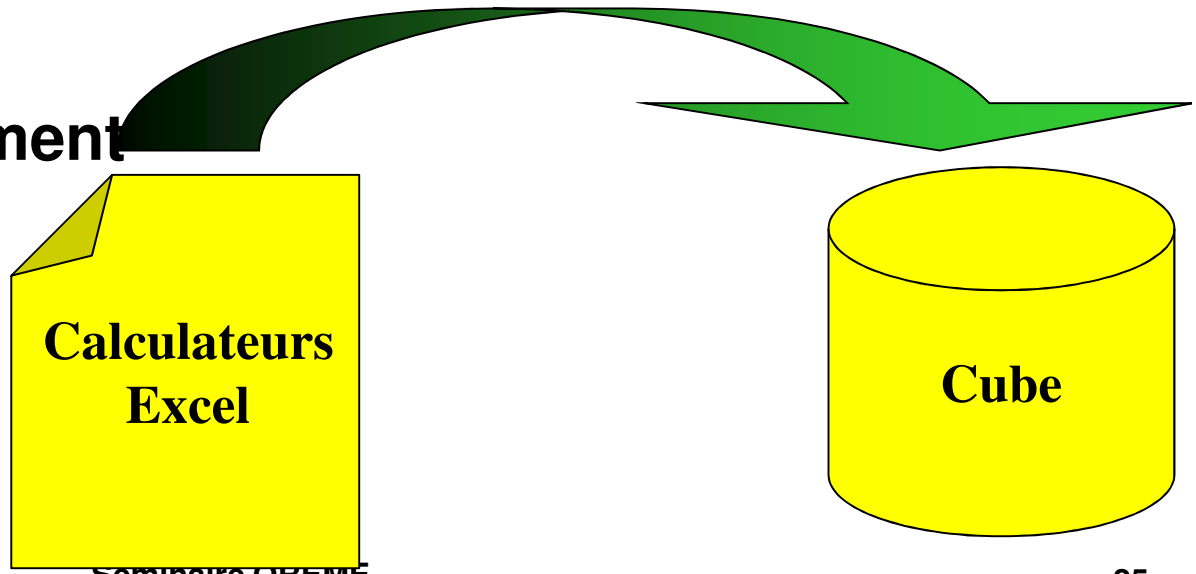
- **Automatisation de traitements** répétitifs



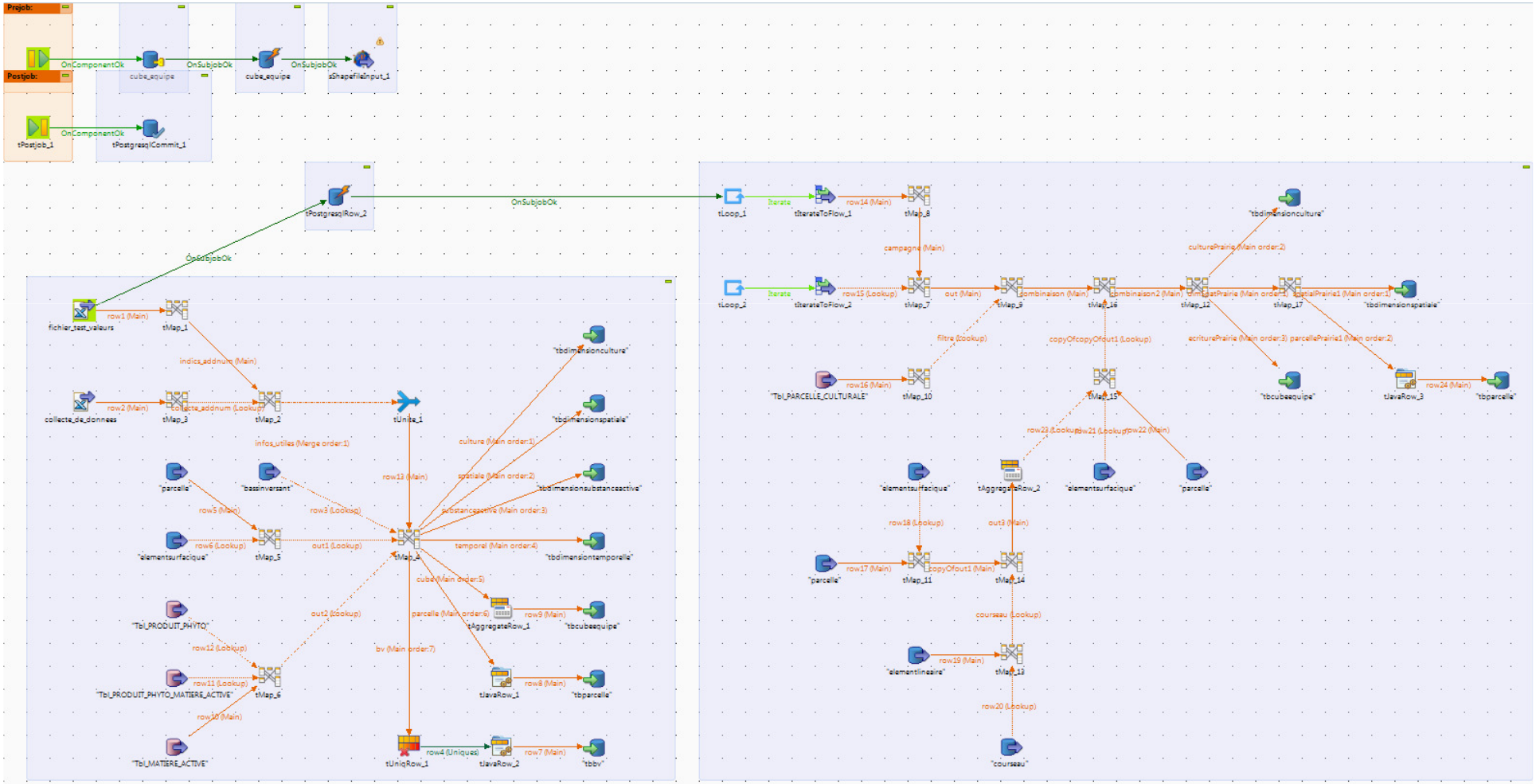
- …

Cube Multidimensionnel

- **Projet Equipe**
  - **Problématique**
    - **Analyser des indicateurs sur les dimensions (spatiale, temporelle, culture, matière active)**
- **Volume de données**
  - **8 indicateurs**
  - **~1400 tuples**
- **Durée de chargement**
  - **< 30 s**

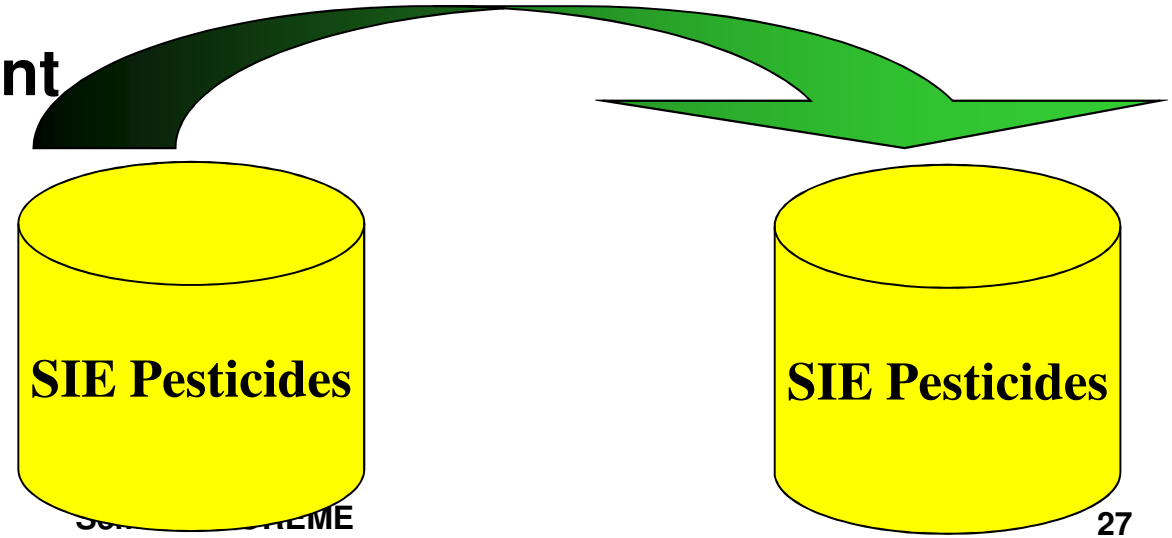


Cube Multidimensionnel

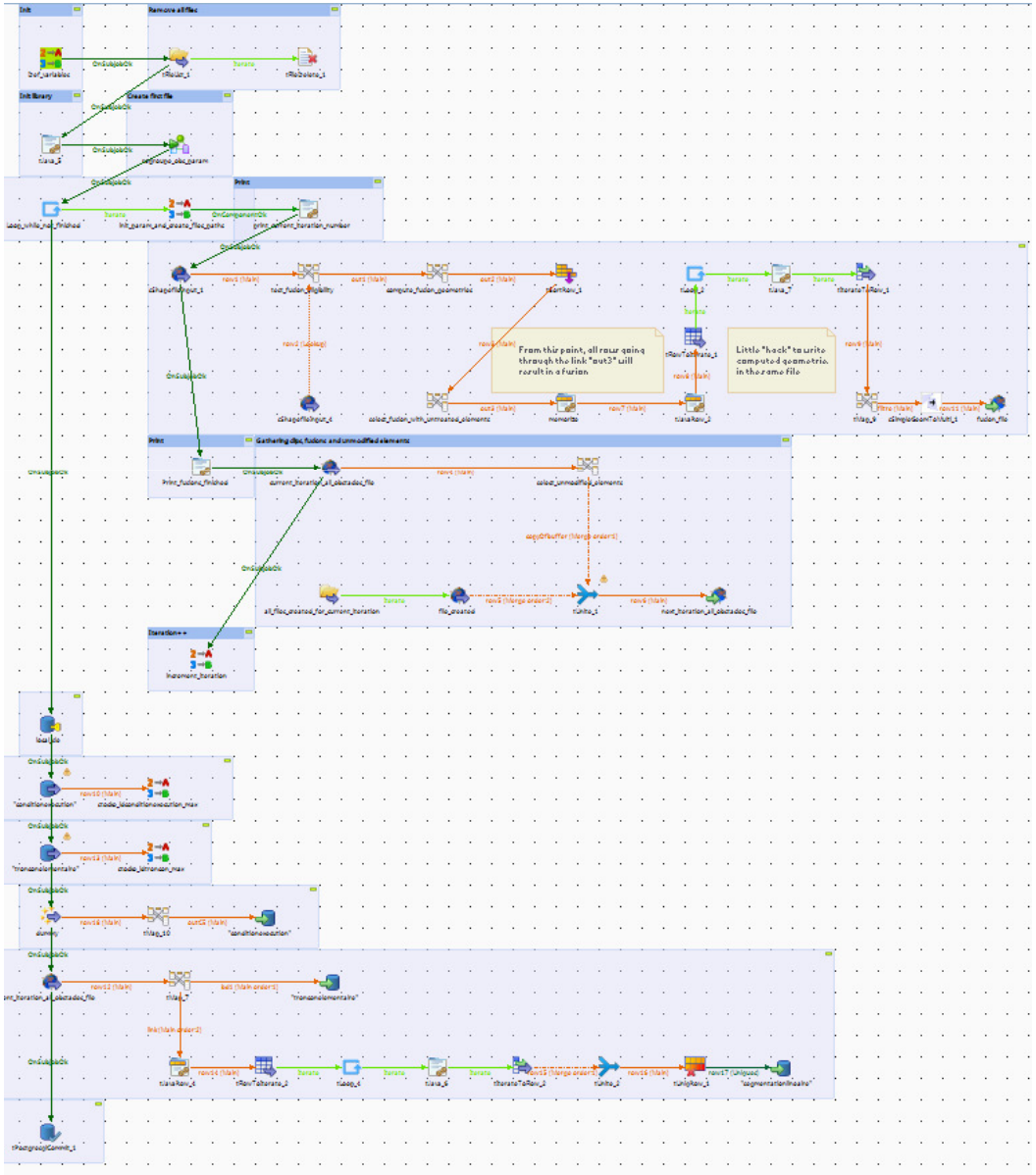


Fusion de linéaires d'un Bassin versant

- **Projet Miriphyque**
  - Problématique
    - Influence des linéaires (cours d'eau, routes, talus, fossés) sur le transfert de pesticides depuis les parcelles jusqu'au cours d'eau
    - Transfert vertical entre la surface et les subsurfaces proche et profonde
- **~350 éléments linéaires**
  - Fusion des éléments parallèles
- **Durée du traitement**
  - Environ 4 min 30 s

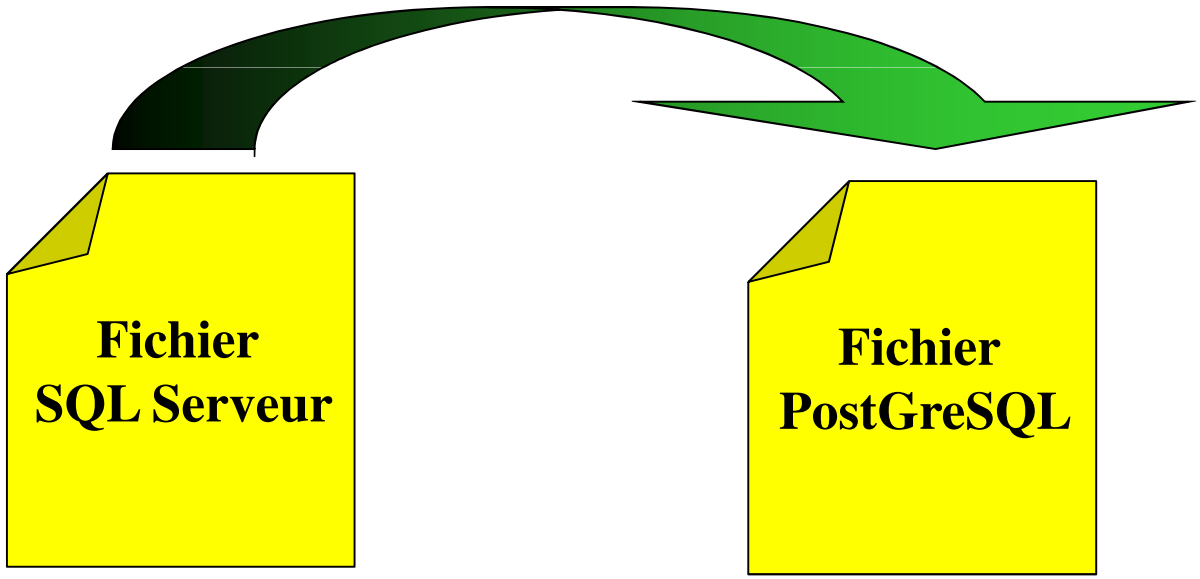


Fusion de linéaires d'un Bassin versant

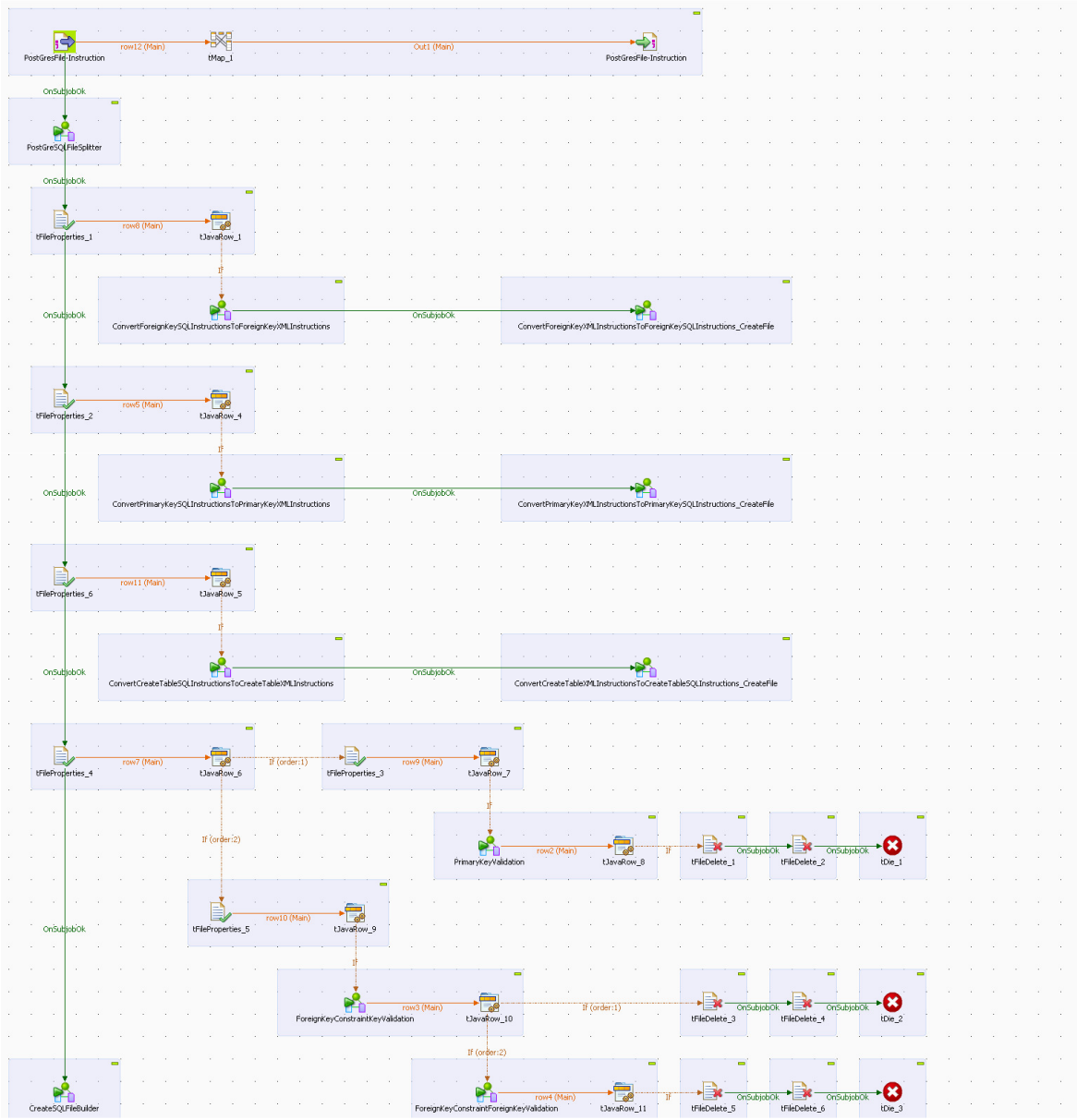


Transformation de fichiers

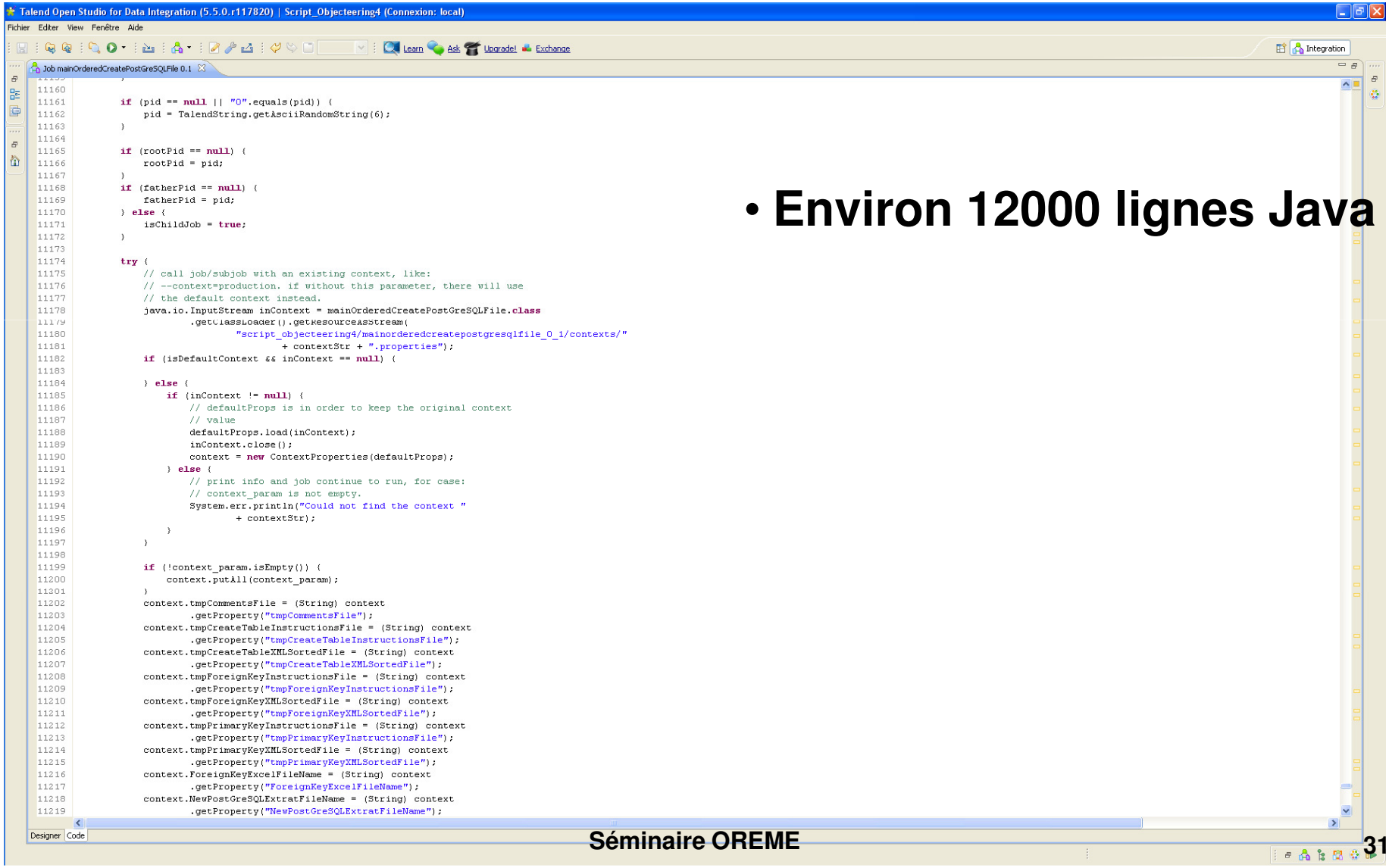
- **Fichiers SQL généré par AGL**
  - SQL Serveur -> PostGreSQL
  - Contrôle de la cohérence



Transformation de fichiers



Transformation de fichiers



- **Environ 12000 lignes Java**

Projet SIE Pesticides

Système d'Information Environnemental pour les Pesticides

AgroParisTech ENGREF

cirad

irstea

Contexte sociétal

- **Utilisation massive de pesticides**
  - 76 100 tonnes : 3° consommation mondial et 1° en Europe
  - Impact sur l'environnement
    - Positif : Selon certains experts, 50% de la surface de forêt actuelle a été préservée (Source ORP)
    - Négatif : Impacts sur les différents compartiments de notre environnement (Source ORP)

Plan EcoPhyto2018

Objectif : Réduite de 50% l'utilisation des pesticides en 2018

Objectif sociétal du projet SIE Pesticides

- **Réduire l'impact des produits phytosanitaires dans les différents compartiments de l'environnement en France**

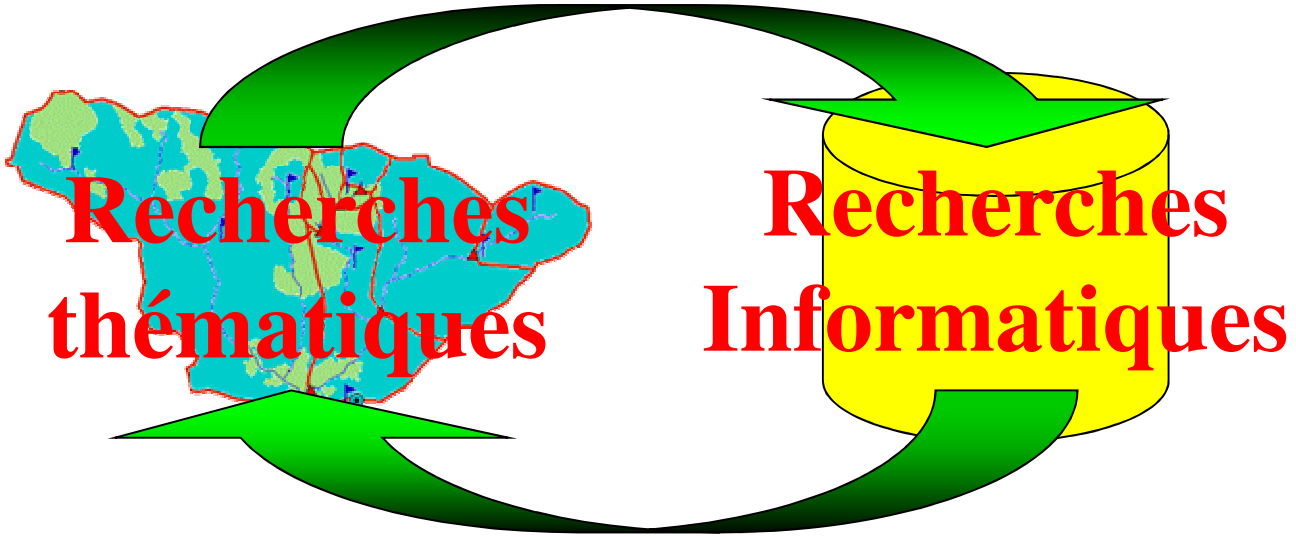
2 Objectifs opérationnels

- **Objectif de production de connaissance**
  - Faire évoluer et concevoir les méthodes, les indicateurs et les outils permettant d'identifier et d'évaluer l'impact des pesticides sur l'environnement

- **Objectif de structuration de la connaissance**
  - Concevoir un système d'information améliorant l'accès à la connaissance relative aux pesticides en vue de faciliter sa remobilisation dans un but d'information et de transparence vis-à-vis des acteurs

Contexte de recherche particulier

- Recherches informatiques menées en synergie avec des chercheurs « thématiciens »



Problématique : Concepts manipulés sont des objets de recherche

Défi informatique

Organiser l'évolution du SIE Pesticides

Défi informatique

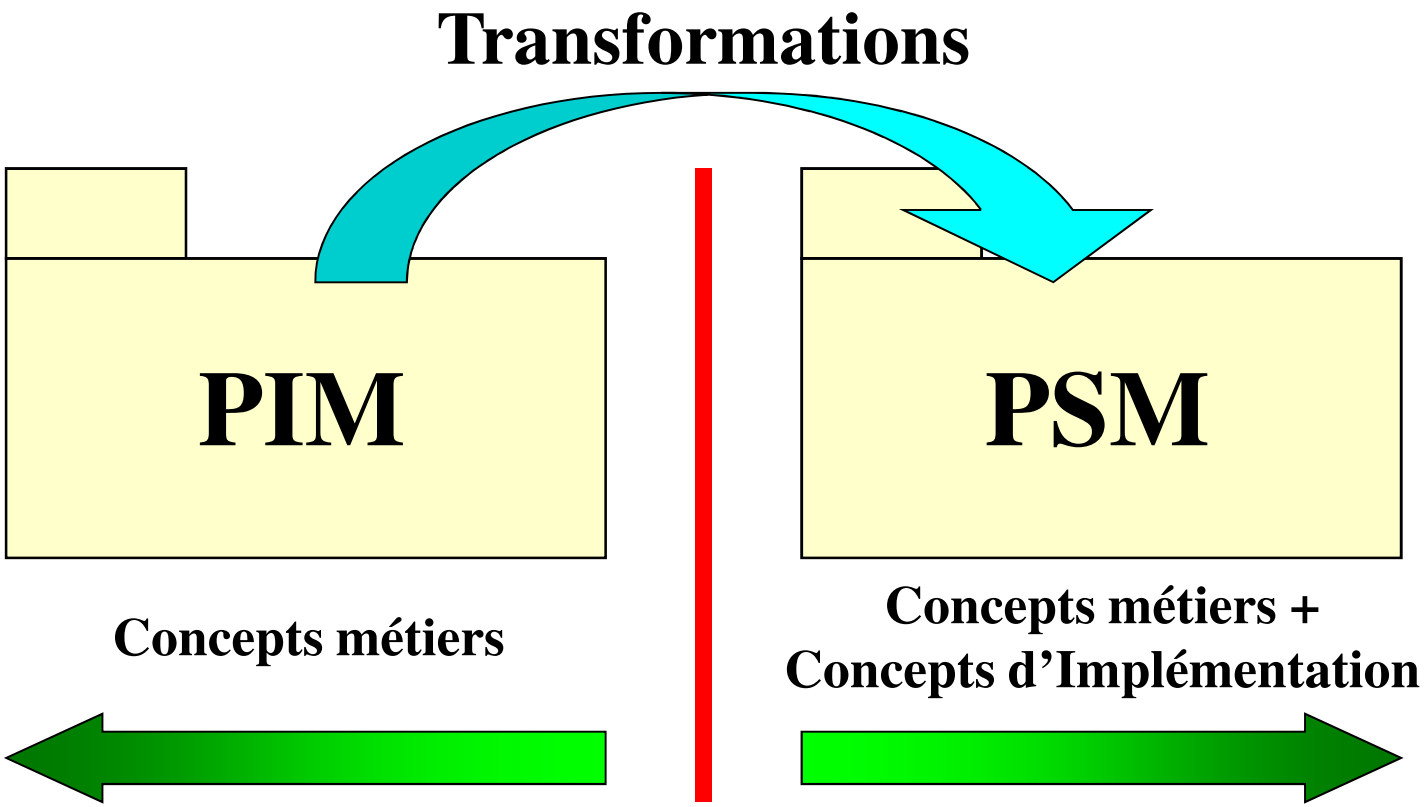
- **Concevoir et mettre en œuvre des méthodes et des outils permettant l'évolution**
  - **Du développement de l'application (code)**

Méthode itérative + Formalisme SIG

  - **Des structures de données**

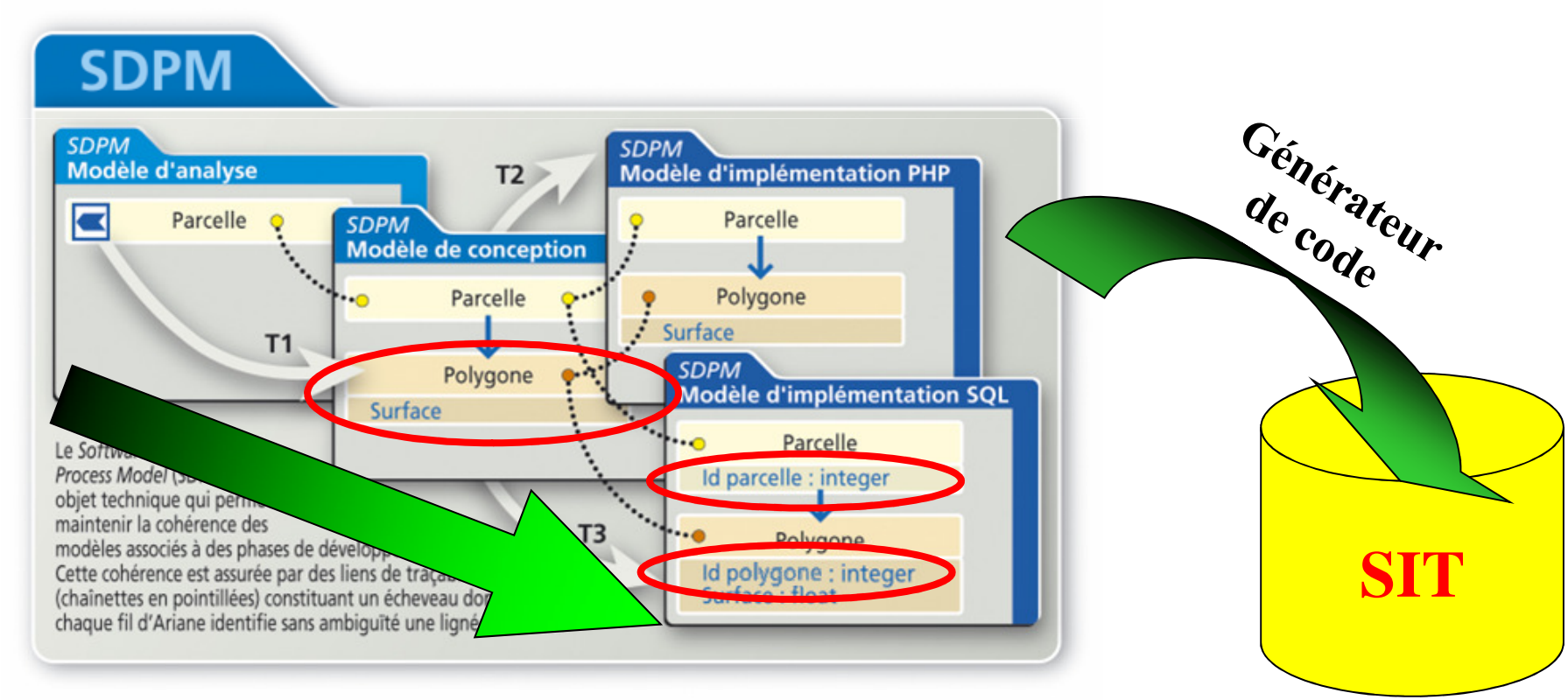
Entrepôt de données + SOLAP

Principe de l'Ingénierie Dirigée par les modèles



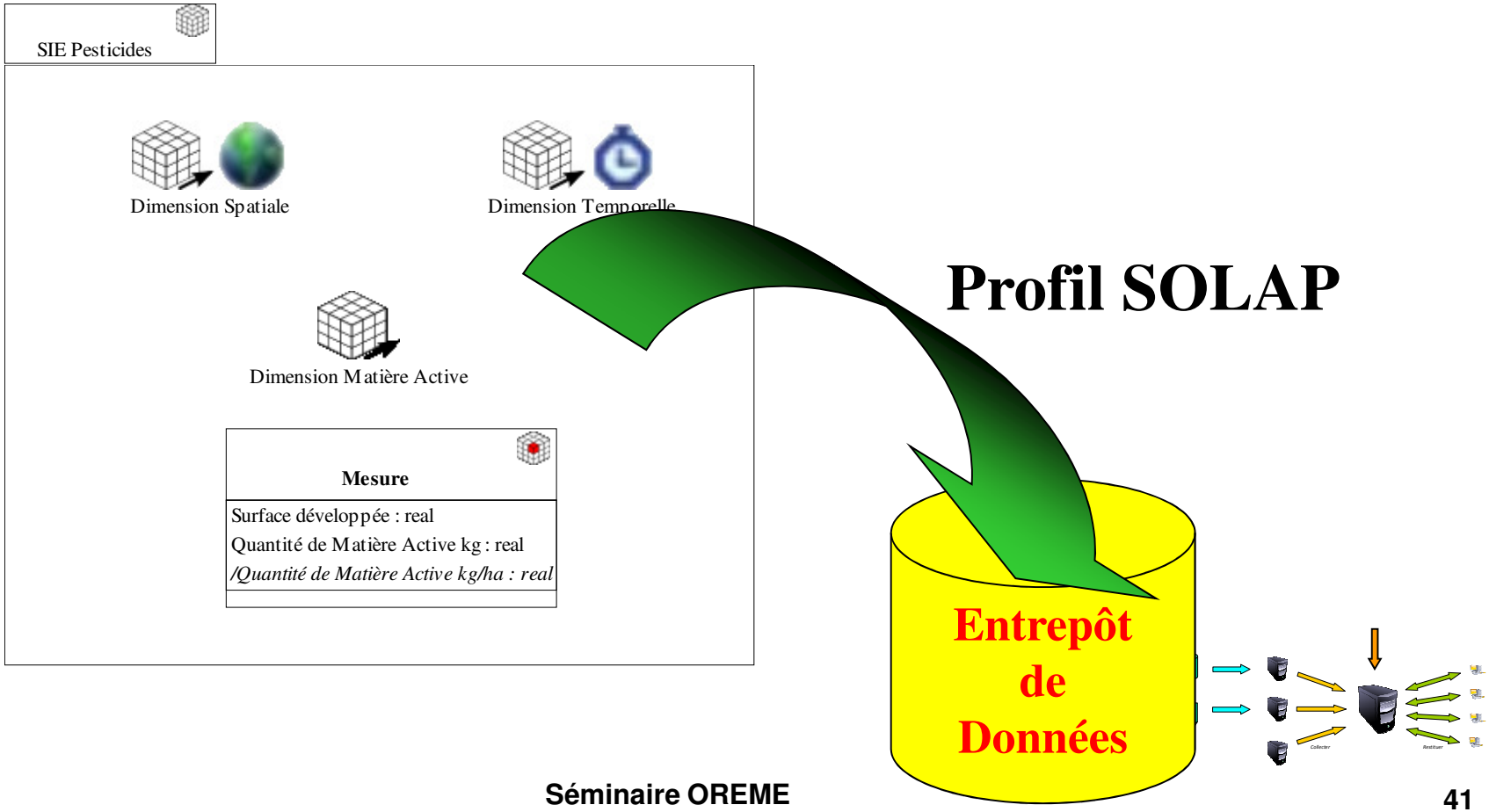
Profil SIG

- **Extension de l'AGL Objecteering**
  - **Formalisme pictogrammique**
    - **Propriétés spatiales et temporelles des concepts métiers**
  - **Ligne de production automatique du code SQL**

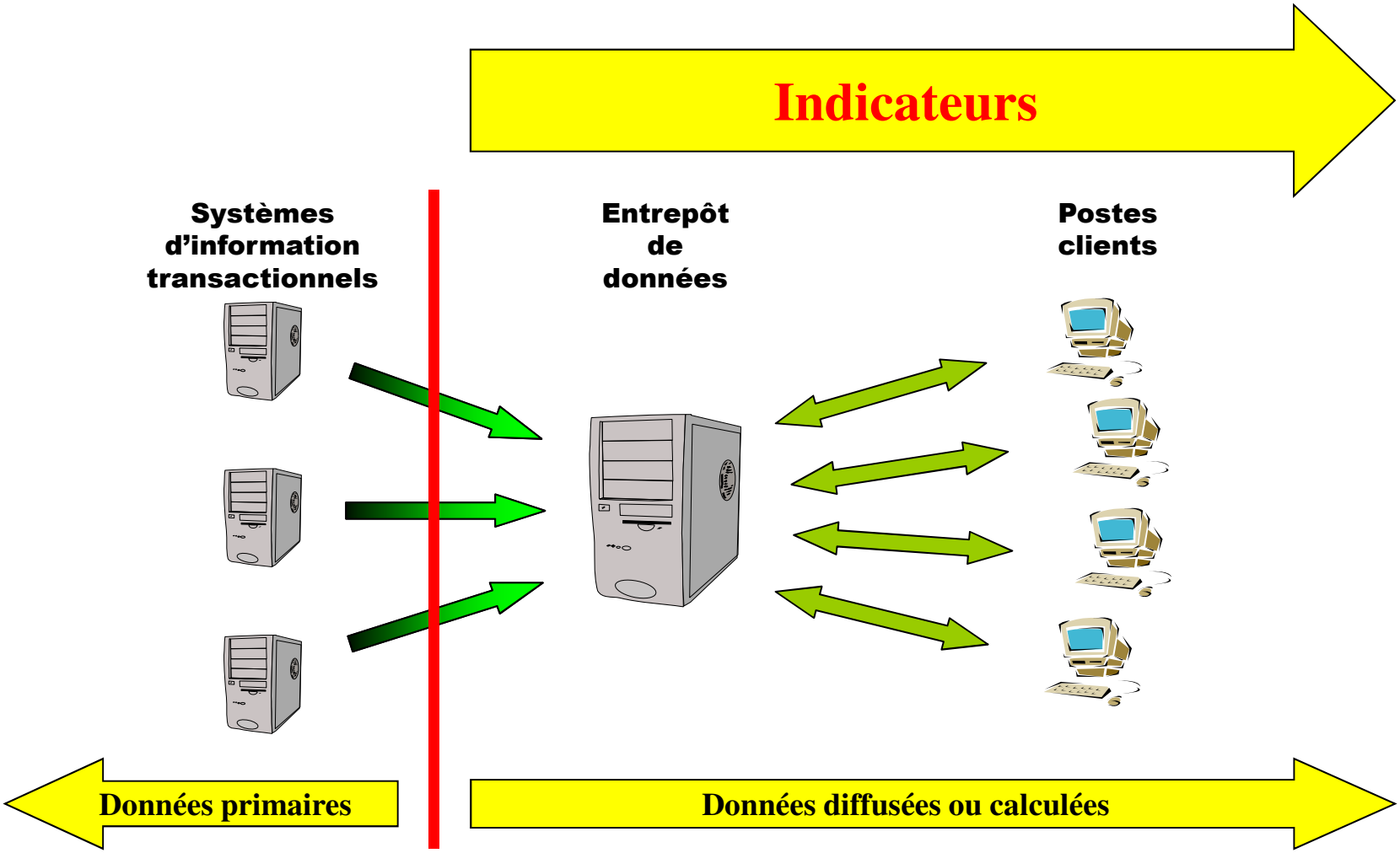


Génération automatique de l'Entrepôts de données

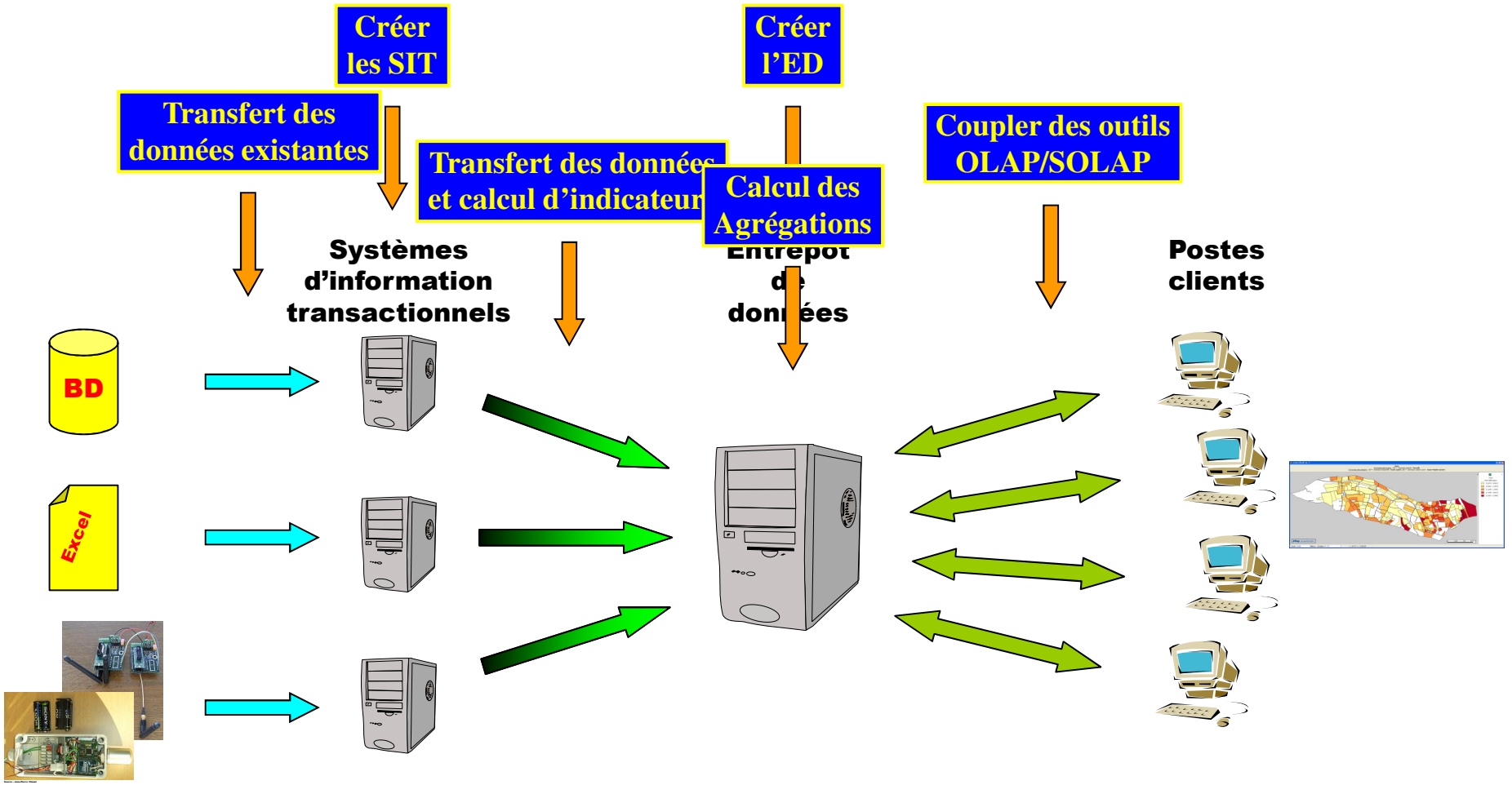
- **Ingénierie Dirigée par les modèles**



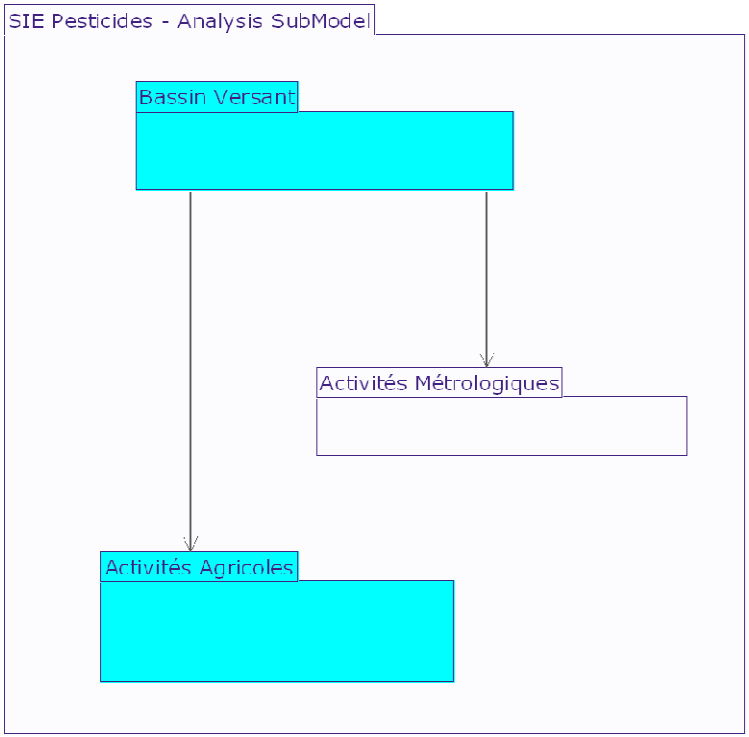
Dichotomie des données primaires / calculées

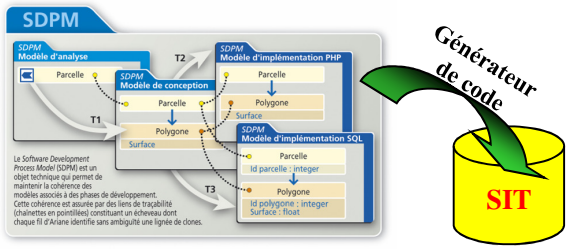


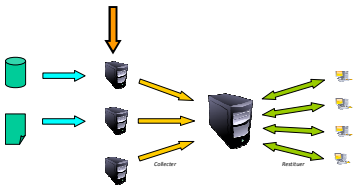
Chaîne de conception



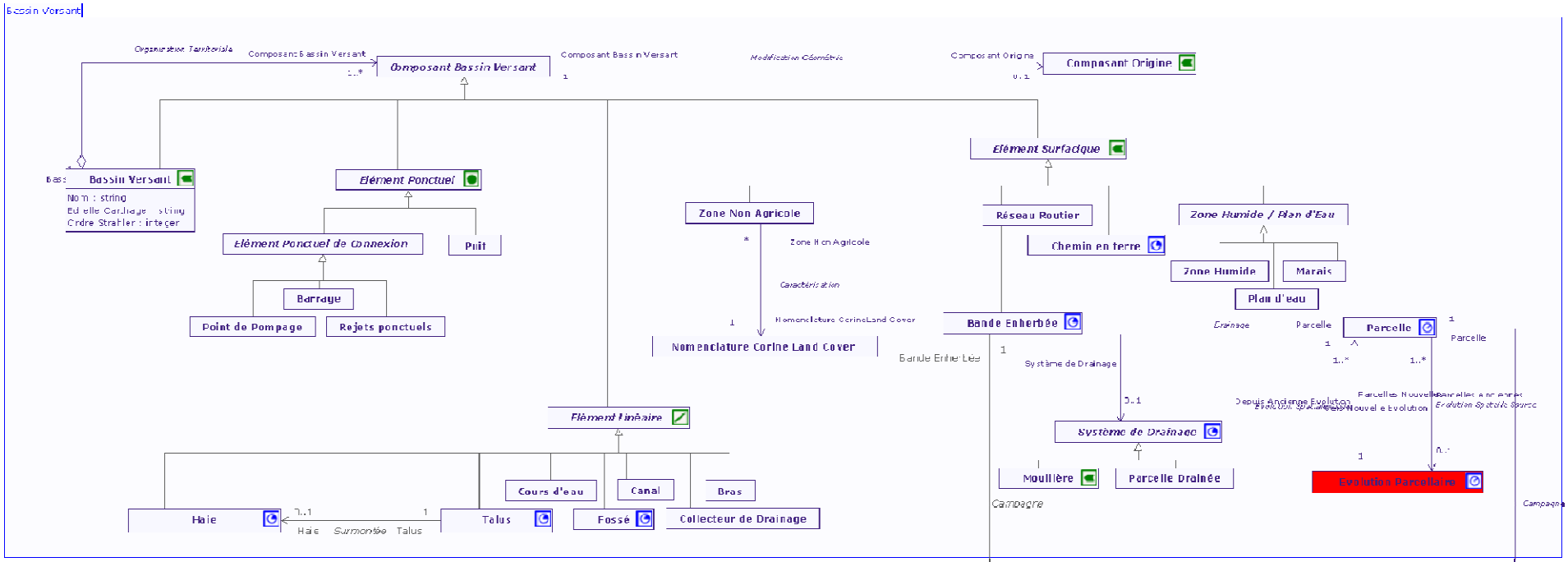
Modèle du SIE Pesticides

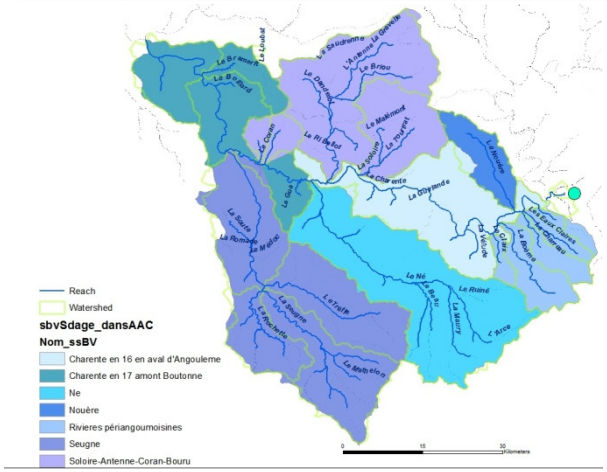




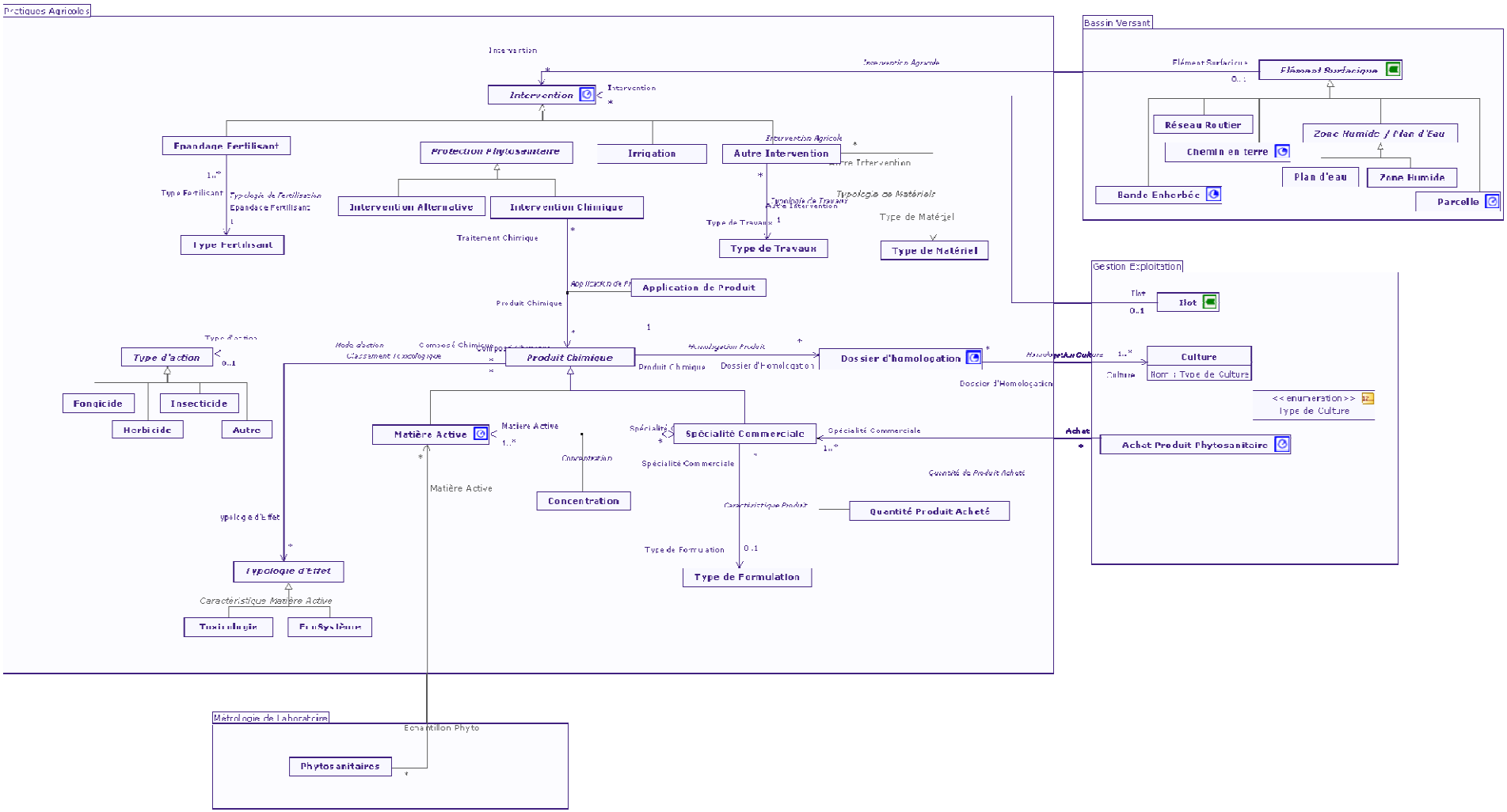


Modèle du Bassin versant

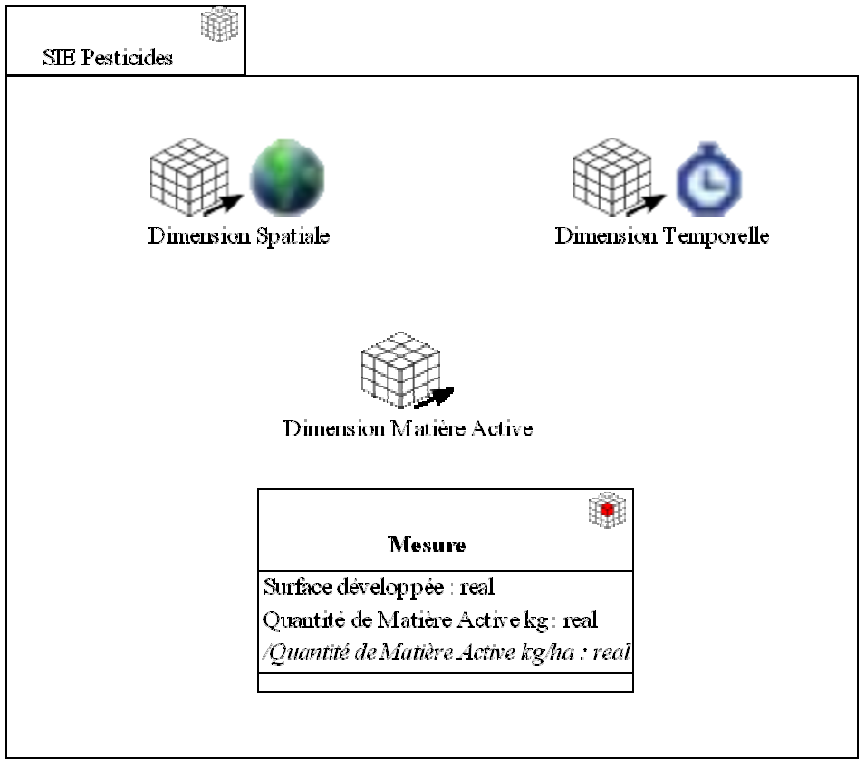


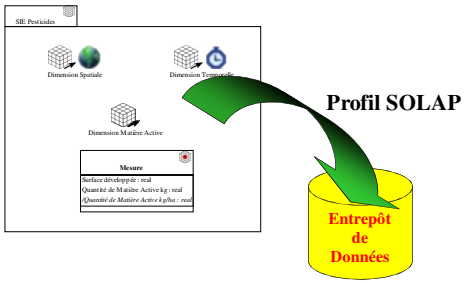


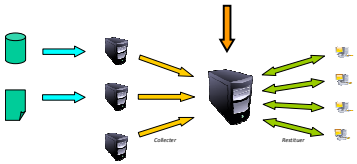
Modèle des Pratiques Agricoles



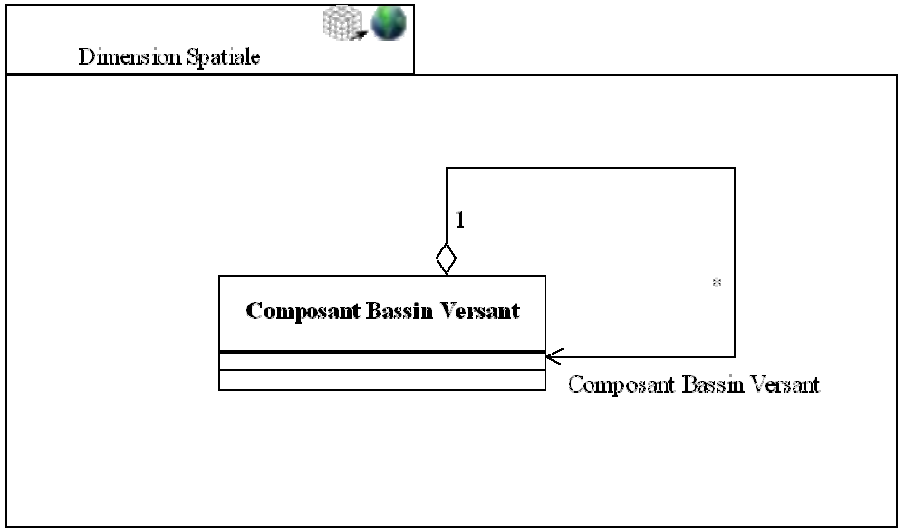
Modèle du Cube multidimensionnel

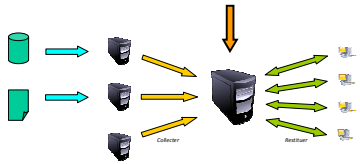




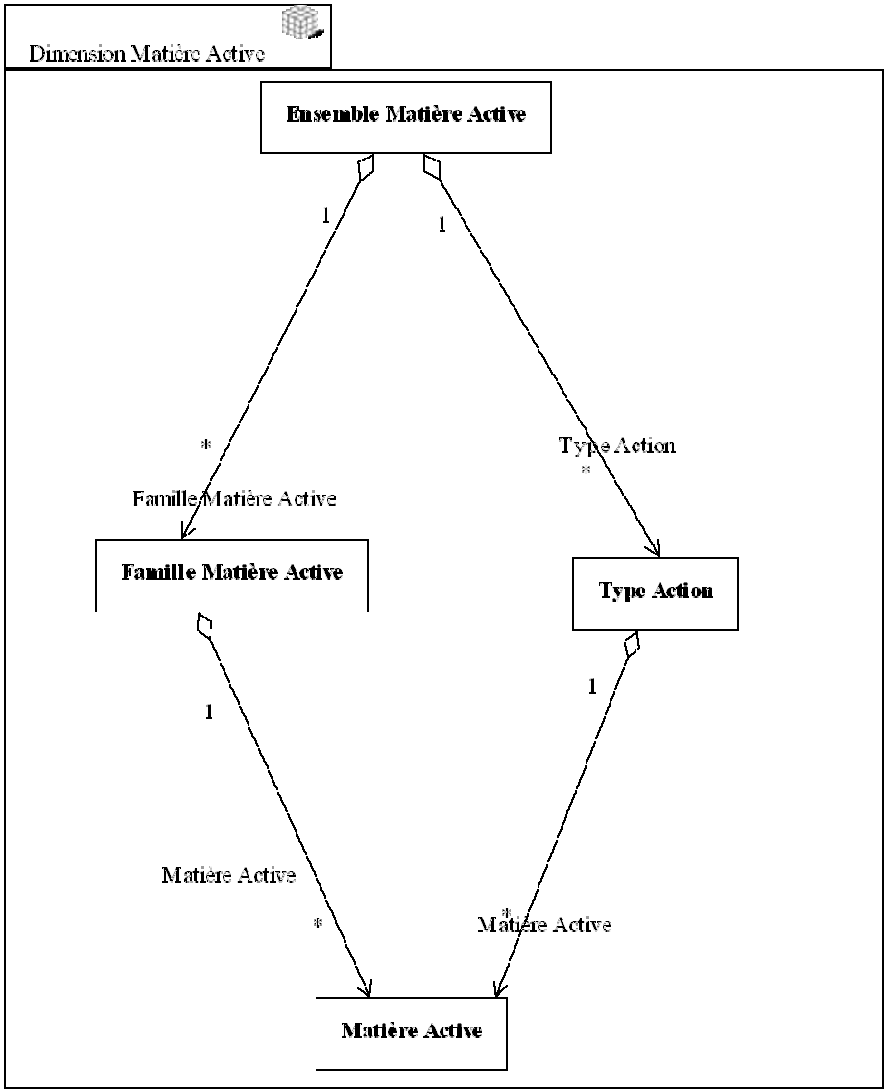


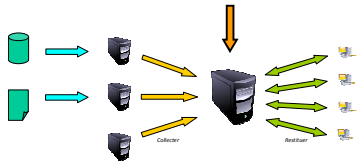
Modèle du Cube - Dimension Spatiale



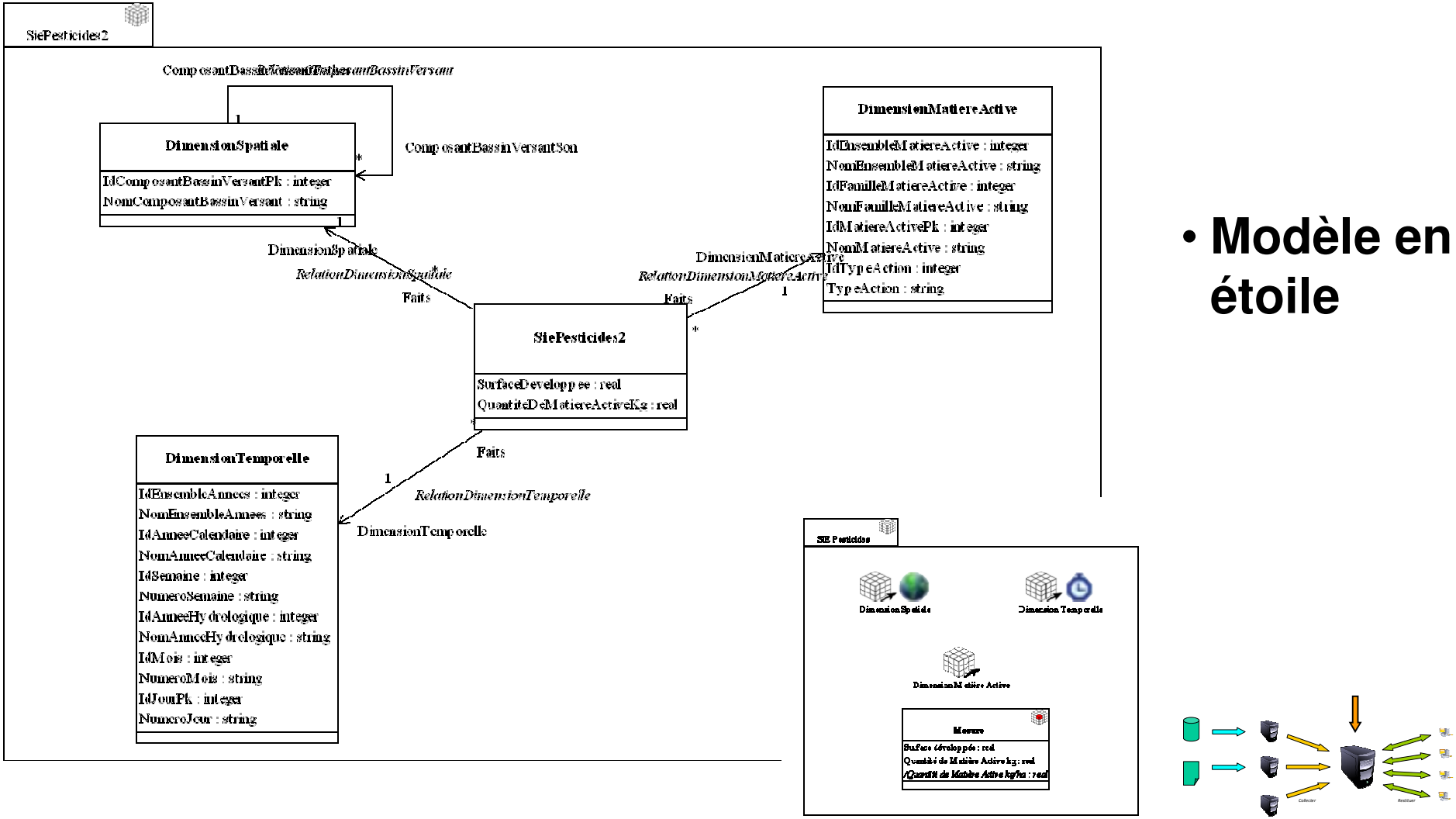


Modèle du Cube - Dimension Matière Active



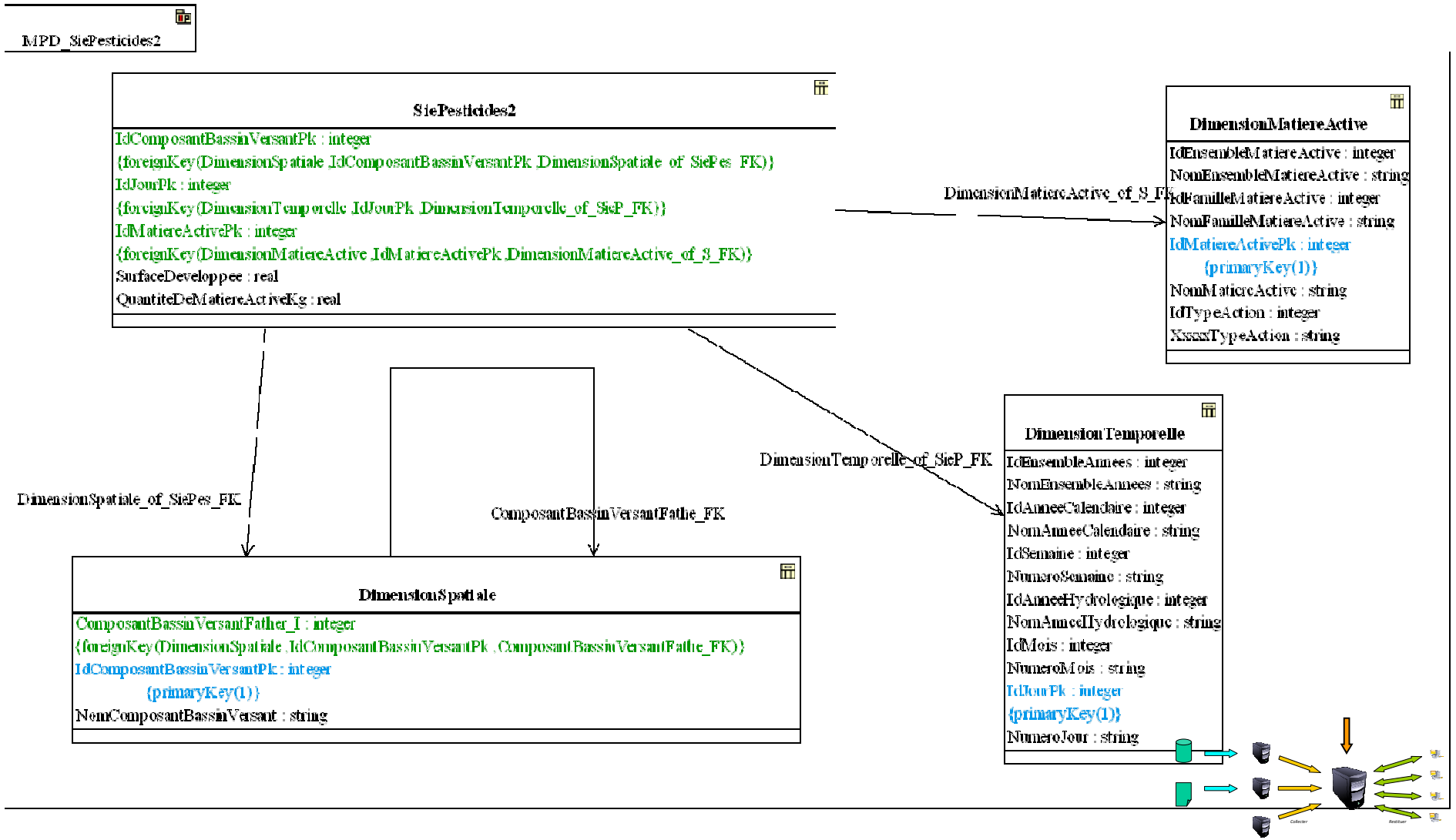


Modèle d'Implémentation de l'Entrepôt de données

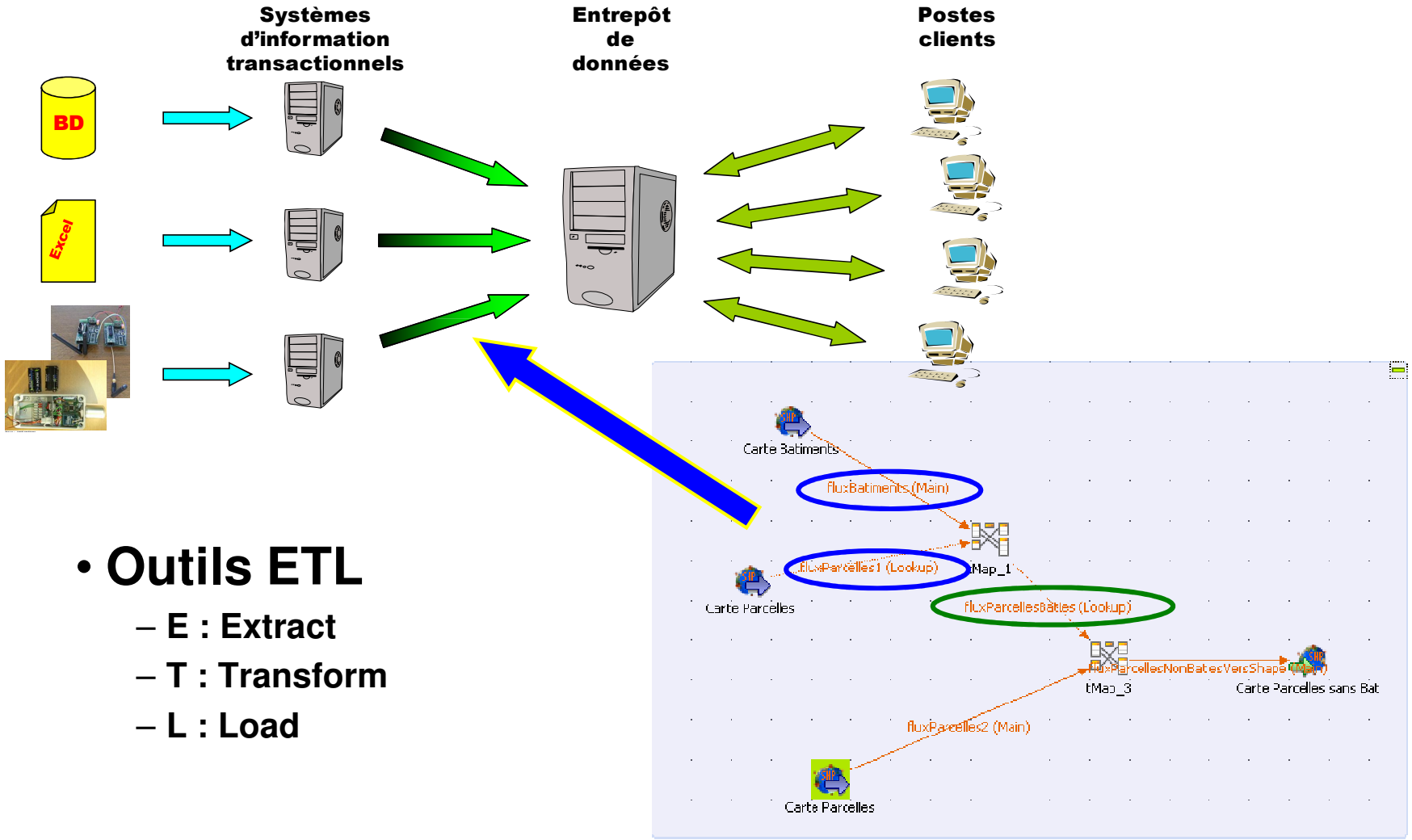


- **Modèle en étoile**

Modèle Physique des Données du Cube



Processus d'intégration d'un entrepôts de données



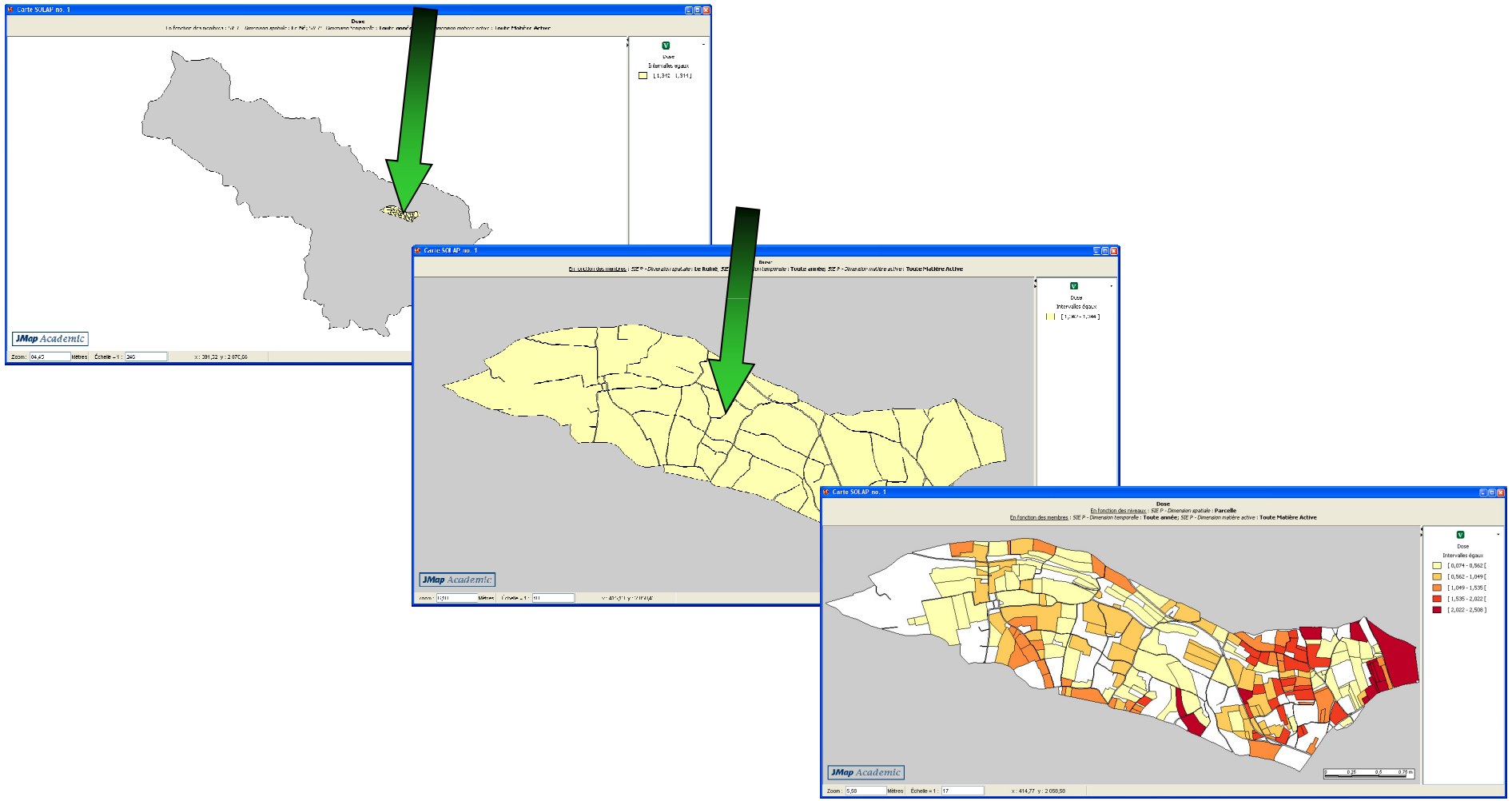
- **Outils ETL**
  - E : Extract
  - T : Transform
  - L : Load

Agrégation des données

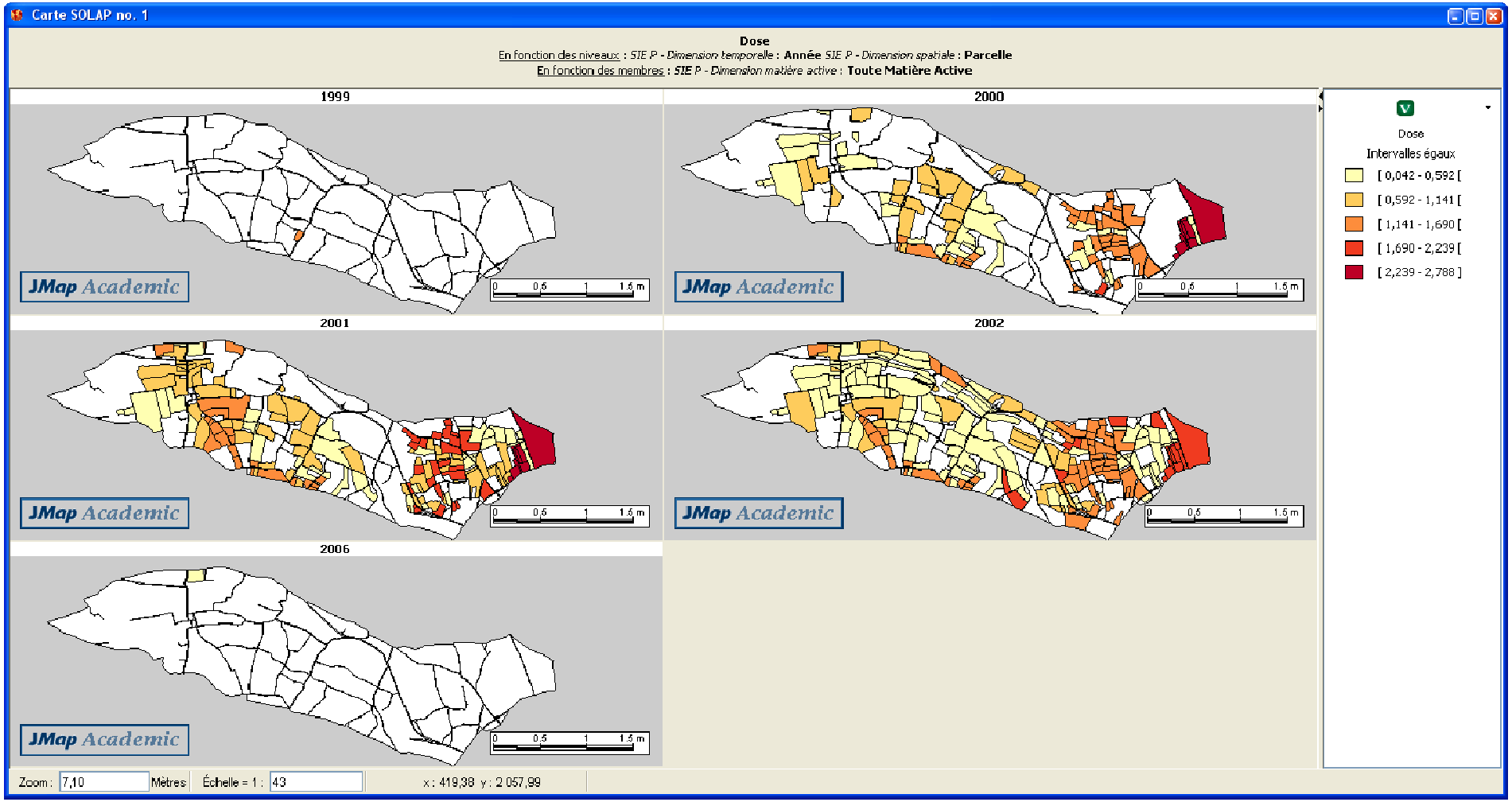
- **Effectuer par les outils OLAP**
  - **Configuration**

```
<?xml version="1.0"?>
<Schema name="SiePesticides">
   <Cube name="SiePesticides"  defaultMeasure="new">
      <Table name="siepesticides"/>

        <Dimension name="MatiereActive" foreignKey="idmatiereactivepk">
            <Hierarchy hasAll="true" primaryKey="idmatiereactivepk">
               <Table name="dimensionmatiereactive" />
                     <Level name="TypeAction" column="nomtypeaction" />
                     <Level name="NomMatiereActive" column="nommatiereactive" />
            </Hierarchy>
            <Hierarchy hasAll="true" primaryKey="idmatiereactivepk">
               <Table name="dimensionmatiereactive" />
                     <Level name="NomFamilleMatiereActive" column="nomfamillematiereactive" />
                     <Level name="NomMatiereActive" column="nommatiereactive" />
            </Hierarchy>
        </Dimension>

        <Dimension name="Time" foreignKey="idjourpk">
            <Hierarchy hasAll="true" primaryKey="idjourpk">
               <Table name="dimensiontemporelle" />
                     <Level name="NomMois" column="nommois" />
                     <Level name="NomJour" column="nomjour" />
            </Hierarchy>
        </Dimension>

        <Dimension name="BV" foreignKey="idcomposantbassinversantpk">
            <Hierarchy hasAll="false" primaryKey="idcomposantbassinversantpk">
               <Table name="dimensionspatiale" />
                     <Level name="BV" column="idcomposantbassinversantpk" nameColumn="nomcomposantbassinversant" parentColumn
            </Hierarchy>
        </Dimension>

         <Measure name="SurfaceDeveloppe" column="surfacedeveloppee" aggregator="sum" formatString="#.#"/>
         <Measure name="QuantiteDeMatiereActiveKg" column="quantitedematiereactivekg" aggregator="sum" formatString="#.#"/>

   </Cube>

</Schema>
```

Exemple d'emboîtements récursifs des BV



Evolution temporelle de la Matière Active appliquée

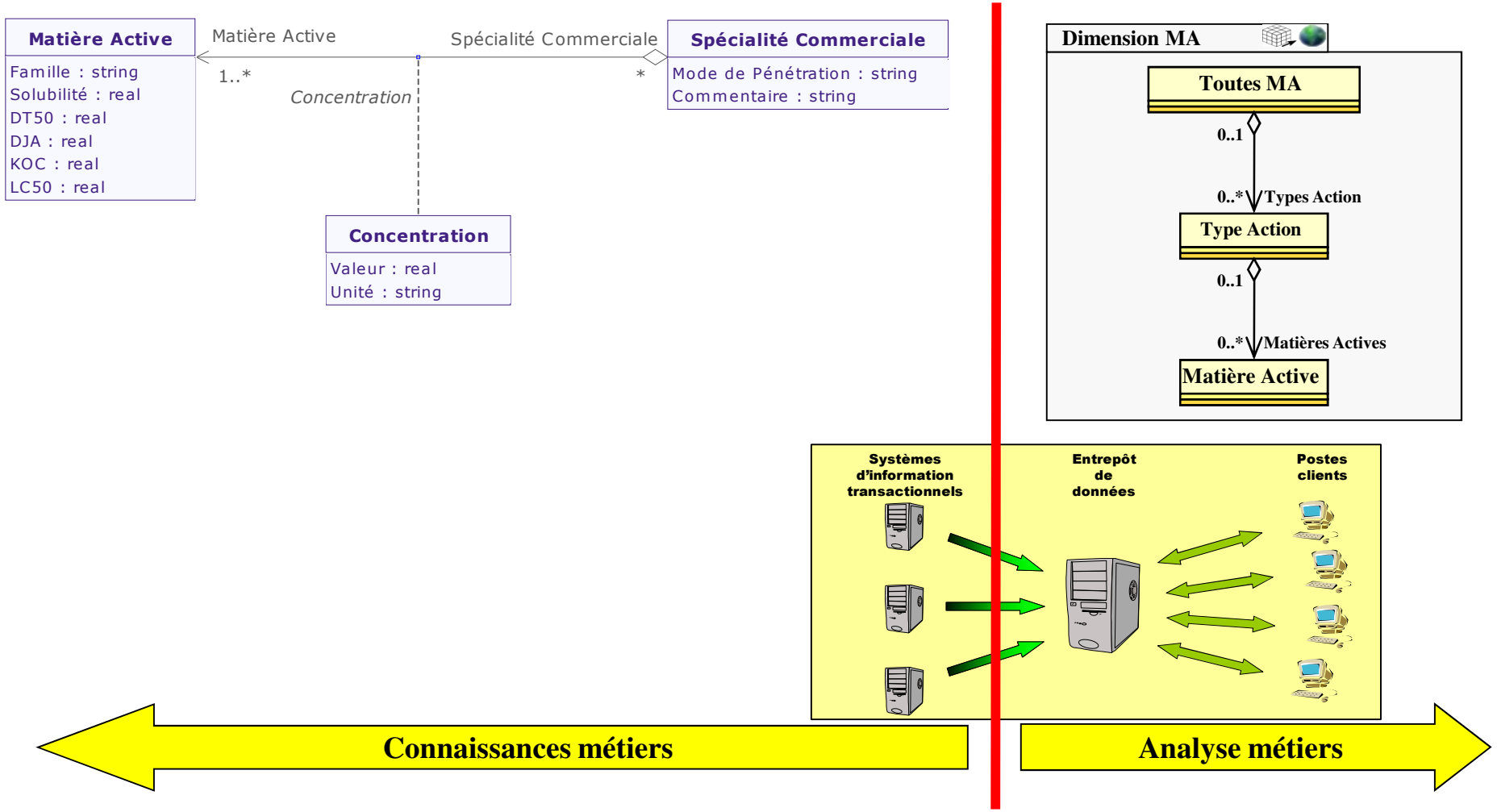


Conclusion

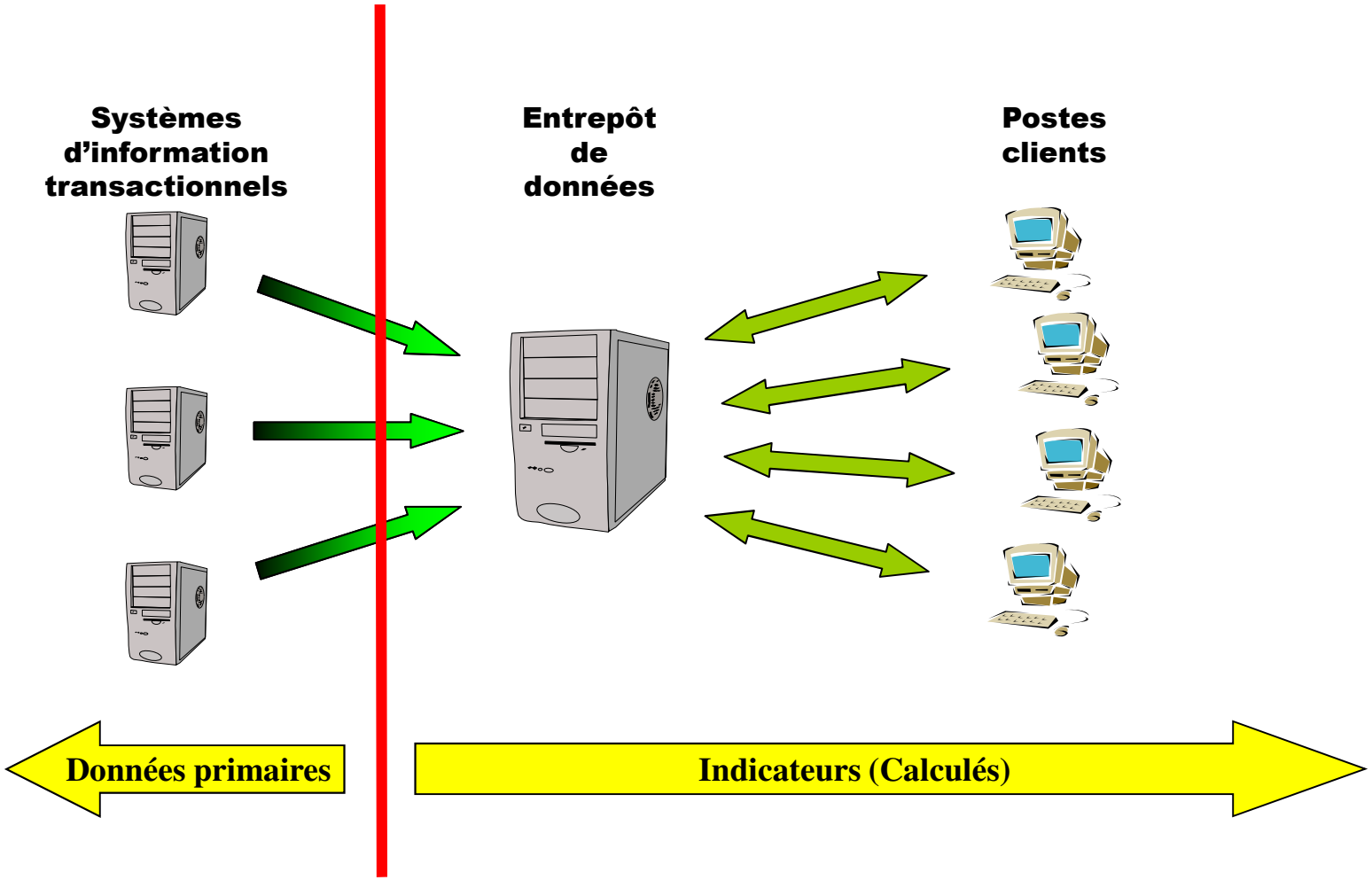
Apports de la méthode mise en œuvre

- **Capitalisation** des connaissances et des données
- **Traçabilité** des modèles et des transformations
- **Gain de productivité** et de **qualité**

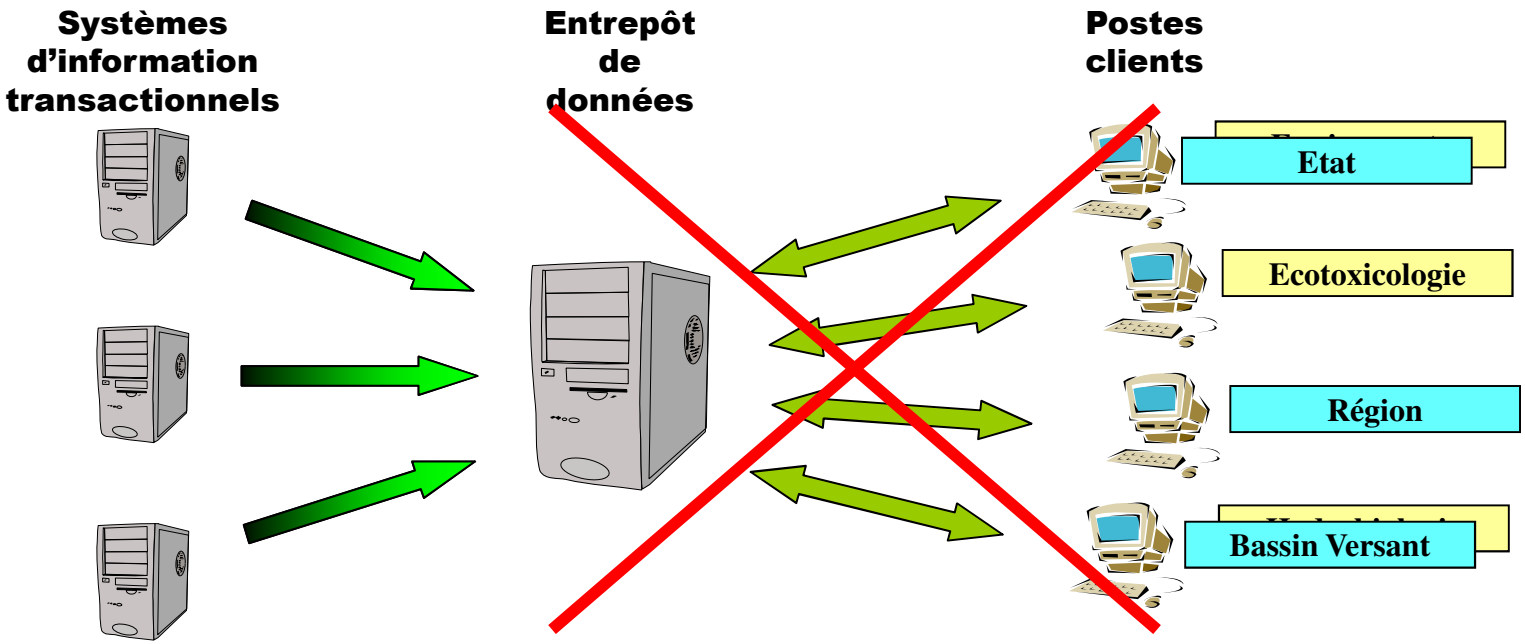
Capitalisation des connaissances et des modèles



Capitalisation des données et des indicateurs

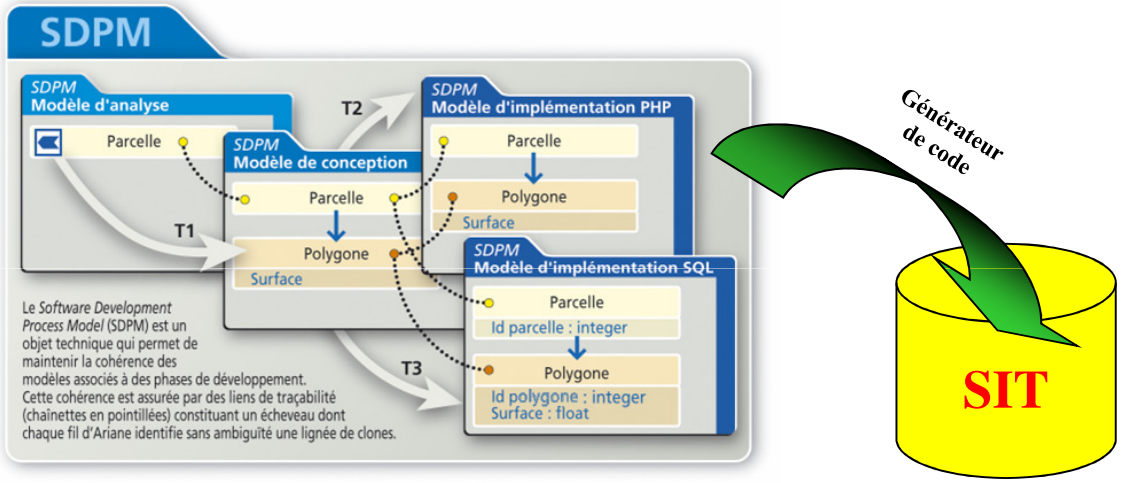


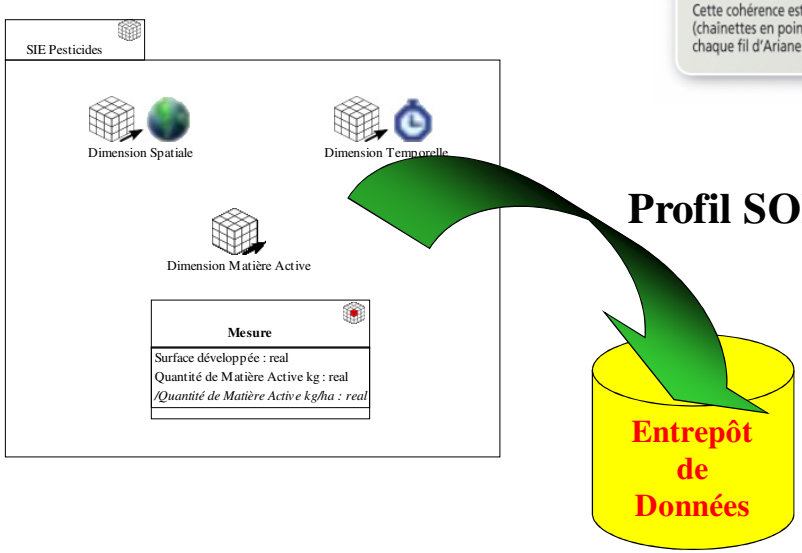
Malléabilité de l'Architecture informatique



Traçabilité des modèles

- **Atelier de Génie logiciel + Ingénierie Dirigée par les modèles**





Profil SOLAP

Entrepôt de Données

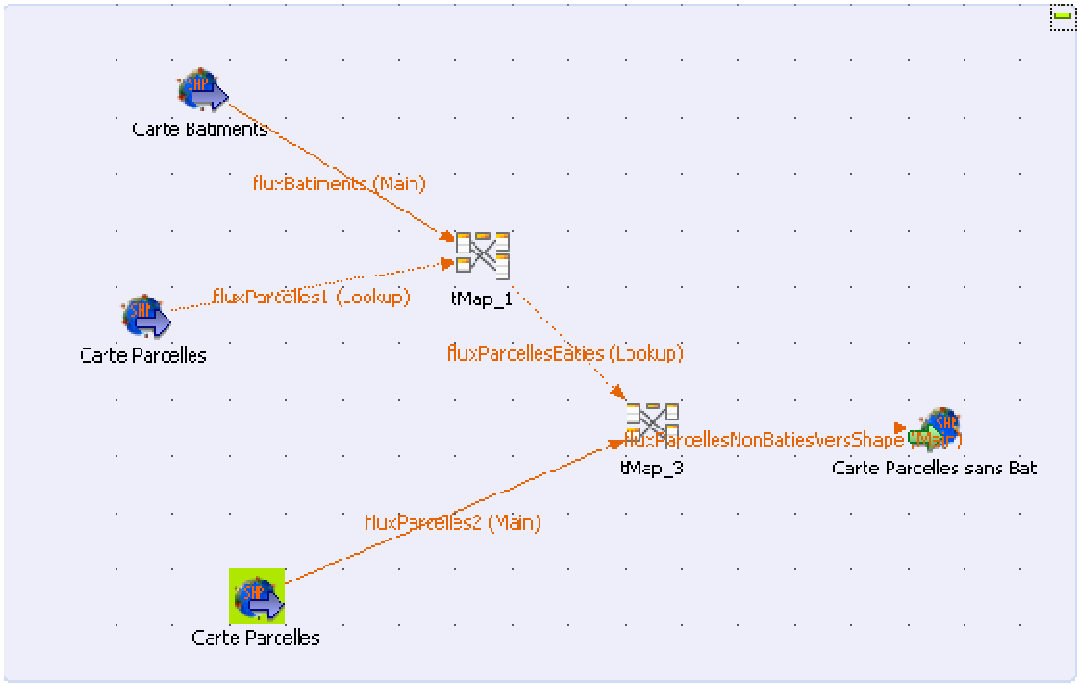
Gain de productivité et de Qualité

- **Cabinet conseil : The Middleware Company**
  - **Deux équipes de niveaux équivalents**
    - **Même application : Ventes d'animaux sur Internet**
    - **Même cahier des charges**

Équipe	Planifiée		Réalisée	
Equ. Tradition.	499 h		507 h	
Equ. MDA	442 h	-11%	330 h	-35%

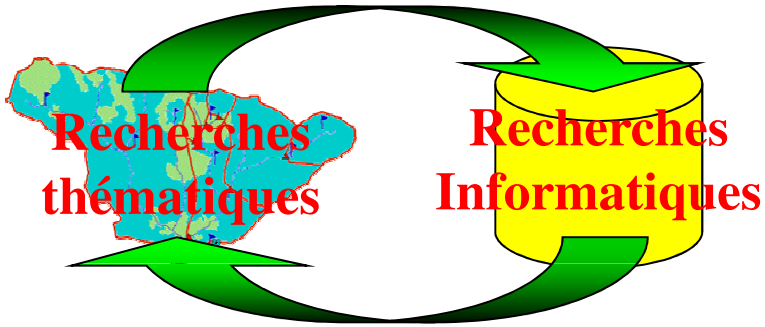
- **Résultat important**
  - **Equipe Traditionnelle : Corrections bogues**
  - **Equipe MDA : Pas de bogues**
  - ➔ **Meilleure qualité**

Traçabilité des transformations



Evolution des SI : une nécessité

- Concepts manipulés sont des objets de recherche



- Evolution des besoins
  - Nouvelles fonctionnalités, nouveaux indicateurs, etc.
- Evolution des textes réglementaires
  - Lois, normes, etc.
- Etc.

Merci de votre attention

- **Formations**
  - **Langage UML et Modélisation d'Applications Environnementales**
  - **Outils ETL pour système d'information décisionnel spatialisé**