Entrepôts de données agro-environnementales

Sandro Bimonte

Equipe COPAIN, UR TSCF, Dept. MathNum, INRAE Clermont-Ferrand

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

Plan

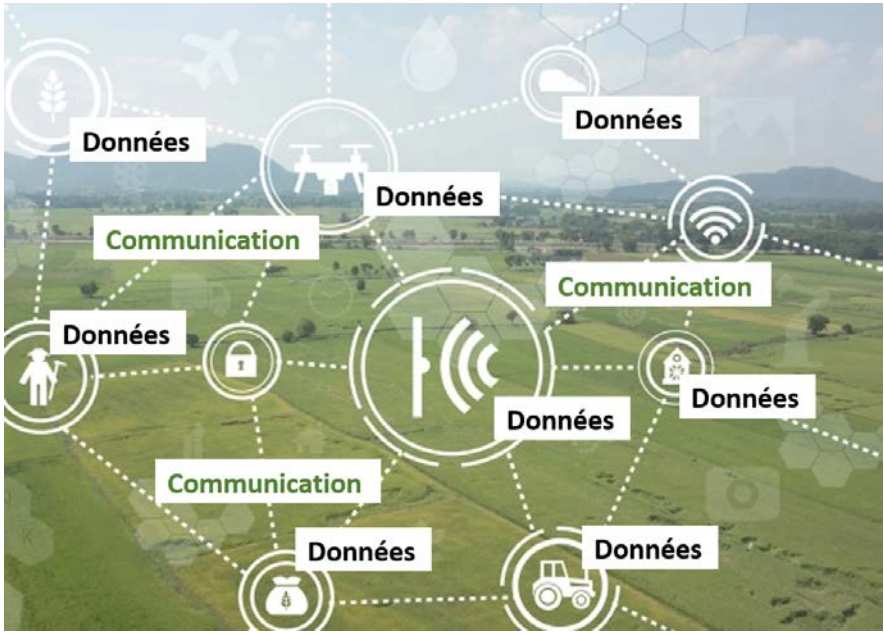
- Contexte
- Entrepôts de données et OLAP
- Un exemple de cas d'étude : agro-biodiversité (ANR VGI4bio)
- Des verrous de recherche

Contexte

Contexte : *Big Data*

- **Big Data** : Révolution numérique - **nouvelles possibilités d'analyse**
  - Santé, Transports, Economie, …

  , mais aussi
  - Agriculture numérique,
  - Etude du vivant,
  - Risques, **…**
- **Nouveaux défis scientifiques multi- et inter- disciplinaires**

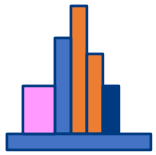


Contexte : … *et à INRAE?*

Applications agro-environnementales

- **Complexité des données**
- **Particularité des décideurs**

INRAE ?



Entrepôts de données

Analyse
OLAP
Data Mining
Stats,
…

Integration

Dictionnaires de données/ Ontologies

Données sources (Lac de données)

INRAE ?

Contexte : *Motivations*

- **Agroenvironnement nécessite une approche systémique :**
  - Approche multi-contexte,
  - Approche système,
  - Approche d'évaluation multicritère, …
- **Couplage des entrepôts de données et Big Data** pour l'**agroenvironnement** :
  - Le Big Data permet de mobiliser toutes les données nécessaire à l'approche systémique de l'agroenvironnement
  - Les entrepôts de données permettent :
    - intégrer et gérer de ces données
    - mettre à disposition les données et les indicateurs à tous les acteurs du terrain (ex : appuyer des choix d'actions)
      - Analyse statistique descriptive simple MAIS exploratoire, multi-échelle et multi-variable
        - Confirmer/infirmer des hypothèses et connaissances en agroécologie
          - Faire apparaitre des nouvelles pistes de recherche en agroécologie
            - support à d'autres méthodes d'analyse de données

INRAE

Contexte : *Exemples d'applications en agroenvironnement*

- Pratiques agricoles alternatives à faible dépendance aux pesticides
  - ex: « *Est-ce que la rotation induit un usage moins important de pesticides pour les différentes variétés de blé?* »
- Pratiques agricoles alternatives préservant l'environnement
  - ex : « *Est-ce que le lieu limitrophe impacte la biodiversité ?* »
- Organiser les systèmes agricoles à l'échelle d'un territoire pour améliorer l'environnement
  - ex: « *Quel est le lien entre pollution de l'eau et pratiques agricoles?* »
- Etude de l'impact environnemental des pratiques agricoles robotisées
  - ex : « *Comment alimenter les inventaires de cycle de vie en ACV* » ?
- Gestion des variabilités spatiales et temporelles par des actions ciblées et localisées
- …

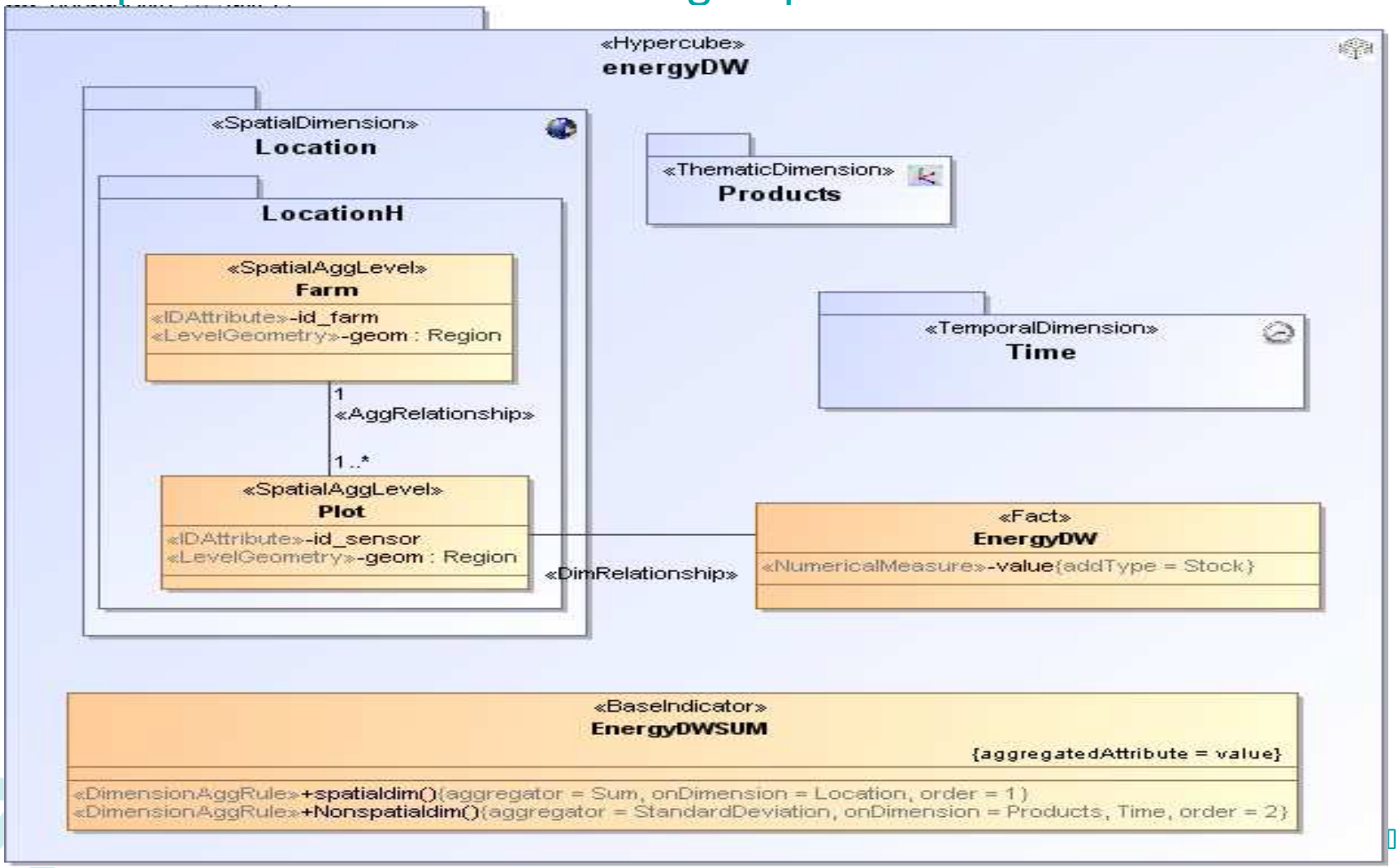
Entrepôts de données et OLAP

Entrepôts de données et OLAP

- Un Entrepôts de données est « *une collection de données, intégrées, non volatiles et historisées pour la prise de décision* »
- Le Modèle multidimensionnel
  - Faits et mesures, dimensions et hiérarchies
  - Agrégation (SUM, MIN, MAX, AVG, COUNT)
- L'analyse OLAP
  - Forage, coupe, pivot, etc.
  - Interactive

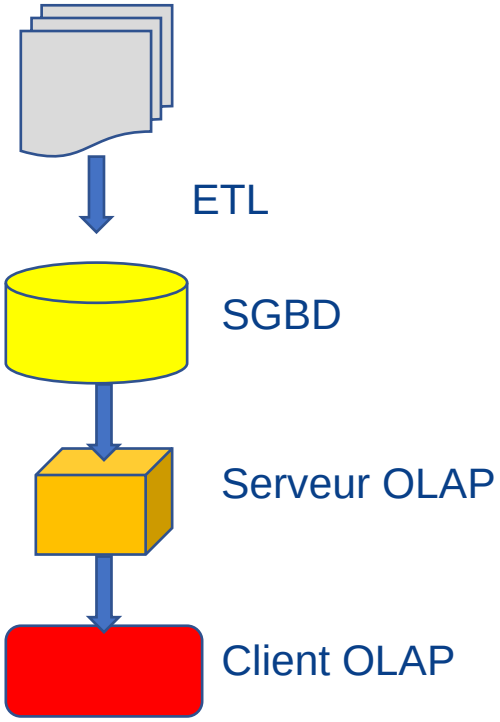
Entrepôts de données et OLAP

- Exemple de consommation énergétique



Architecture : Entrepôts de données et OLAP

- Multi-niveaux



Entrepôts de données

- Faits
- Mesures
- Dimensions
- Faits et dimensions
- Hiérarchies

Les Faits

- La définition
  - Un fait est la plus petite information analysable. C'est une information qui contient les données observables (les faits) que l'on possède sur un sujet et que l'on veut étudier, selon divers axes d'analyse (les dimensions)

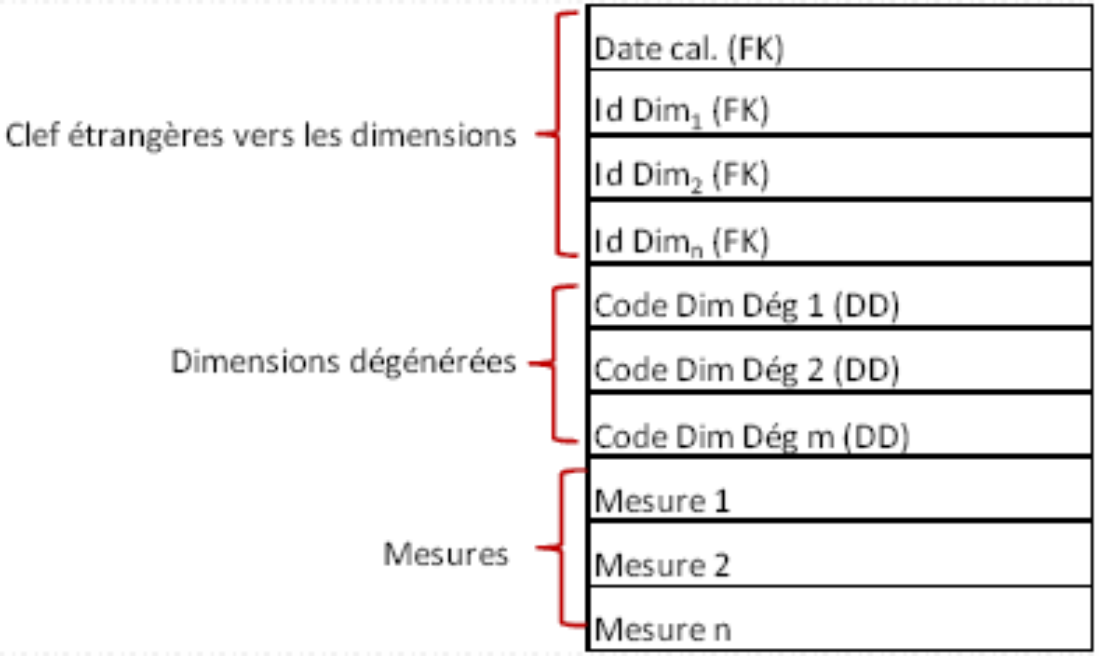
- Les « faits » dans un entrepôt de données, sont normalement numériques, puisque d'ordre quantitatif
  - ex : montant en argent des ventes, du nombre d'unités vendues d'un produit, etc.

Les Faits

- Les faits représentent des associations dont l'existence d'une occurrence dépend de l'existence des occurrences correspondantes parmi les descripteurs dimensionnels.
    - C'est-à-dire, la ''table'' de faits contient l'ensemble des mesures correspondant aux informations de l'activité à analyser.
- la table des faits est la matérialisation d'une association entre n entités.

Les Faits

- Structure de base d'une ''table'' des faits :



Les mesures

- Mesure additive
  - additionnable suivant toutes les dimensions
- Mesure semi additive
  - additionnable seulement suivant certaines dimensions
- Mesure non-additive
  - non additionable quelque soit la dimension

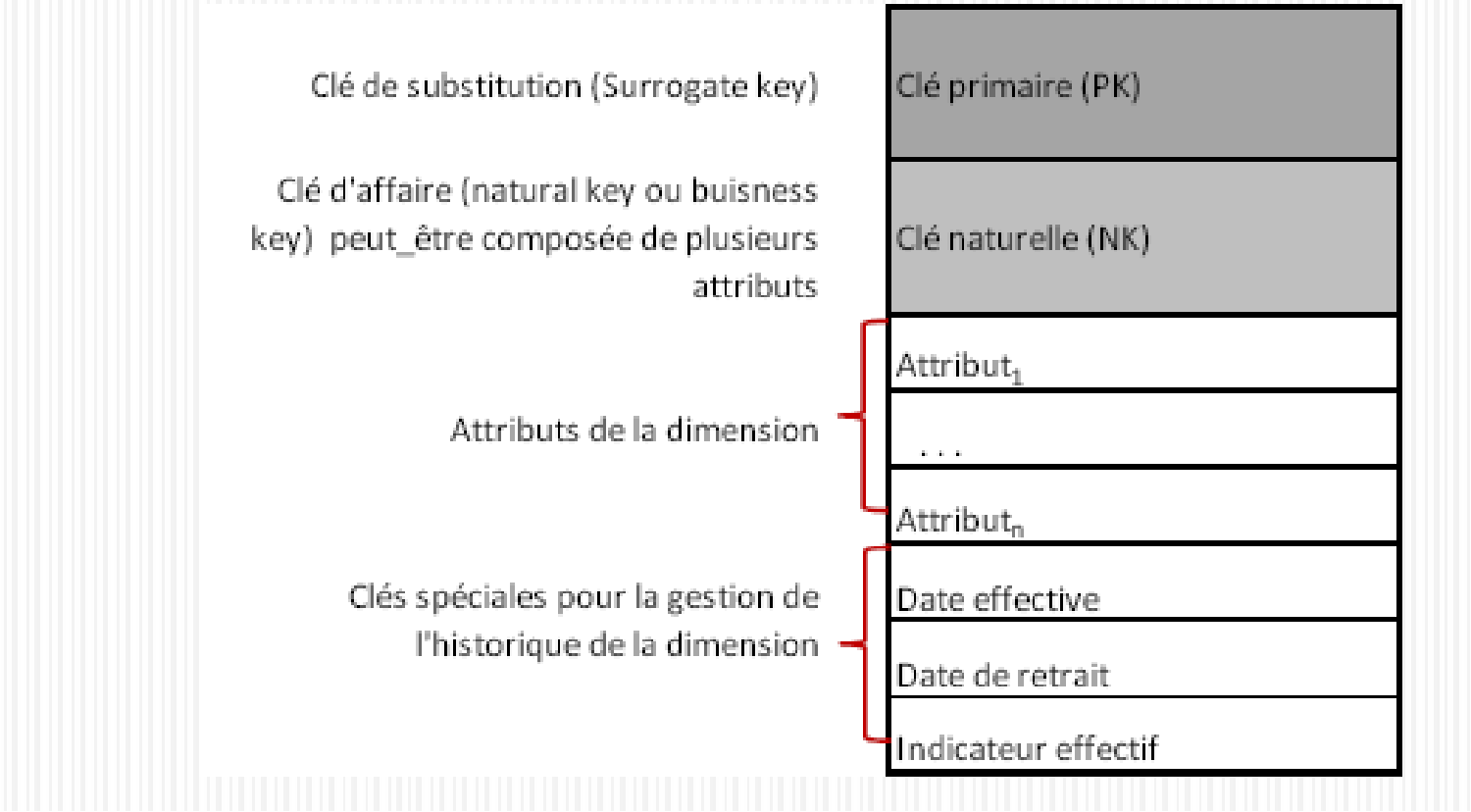
Dimensions

- La définition
  - Une dimension est une ‘’table’’ qui représente un axe d'analyse selon lequel on veut étudier des données observables (les faits) qui, soumises à une analyse multidimensionnelle, donnent aux utilisateurs des renseignements nécessaires à la prise de décision.

- On appelle donc « dimension » un axe d'analyse.
  - Il peut s'agir des Clients ou des Produits d'une entreprise, du Temps, etc.

Dimensions

- **Structure de base d'une dimension**



Dimensions

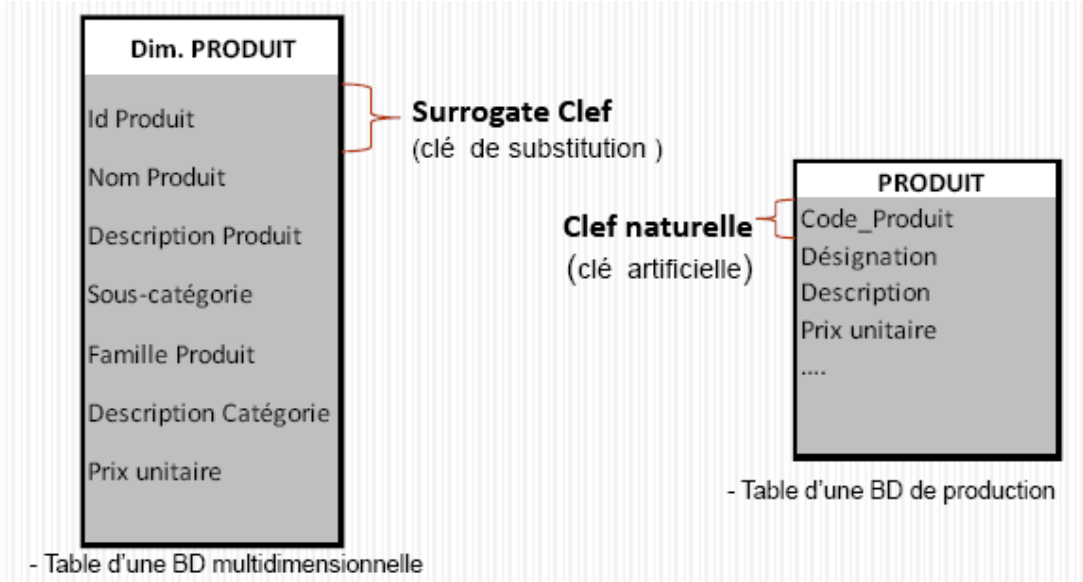
- Caractéristique d'une dimension
  - Une table de dimension contient les informations descriptives des valeurs numériques de la table des faits
  - Une table de dimension contient en général beaucoup moins d'enregistrements qu'une table des faits

Dimensions

- Composantes:
  - Composante 1 : Surrogate key ou clé de substitutions
  - Composantes 2 : Attributs
  - Composantes 3 : Clés spéciales

Dimensions

- Composante 1 : surrogate key ou clé de substitution



Dimensions

- La Définition
  - Une clé de substitution (Surrogate key) est une clé non intelligente utilisée afin de substituer la clé naturelle (Business Key) qui provient des systèmes opérationnels.
  - La clé naturelle est en général composée de plusieurs colonnes.

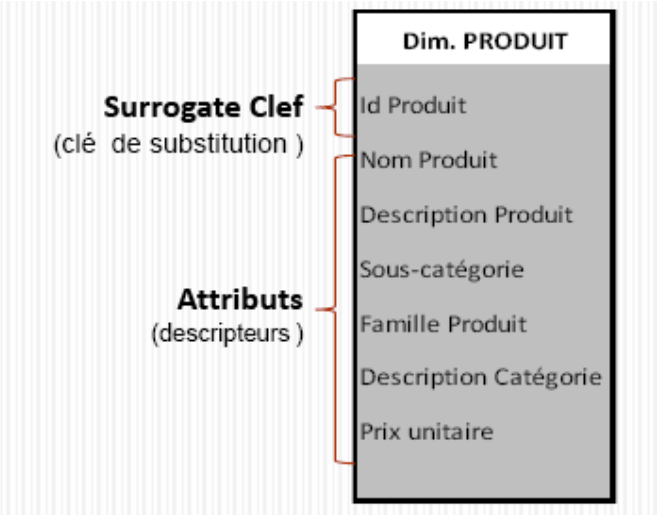
- Dans un système opérationnel, on utilise une clé artificielle afin d'identifier d'une façon unique un élément de l'entité : (client_id pour l'entité Client, emp_id pour l'entité Employé).

- La clé de substitution ne doit pas être confondue avec la clé artificielle attribuée par les systèmes opérationnels.

- La clé de substitution est alors utilisée dans un entrepôt de données pour remplacer la clé artificielle du système opérationnel afin de rendre un élément unique dans la dimension

Dimensions

- Composantes 2 : attributs
  - En plus de la clé de substitution ou de la clé naturelle, d'autres attributs sont ajoutés à la dimension.
  - Ces attributs sont descriptifs et représentent l'information utile sur la dimension et représentent les hiérarchies de la dimension



INRAE

Dimension conforme

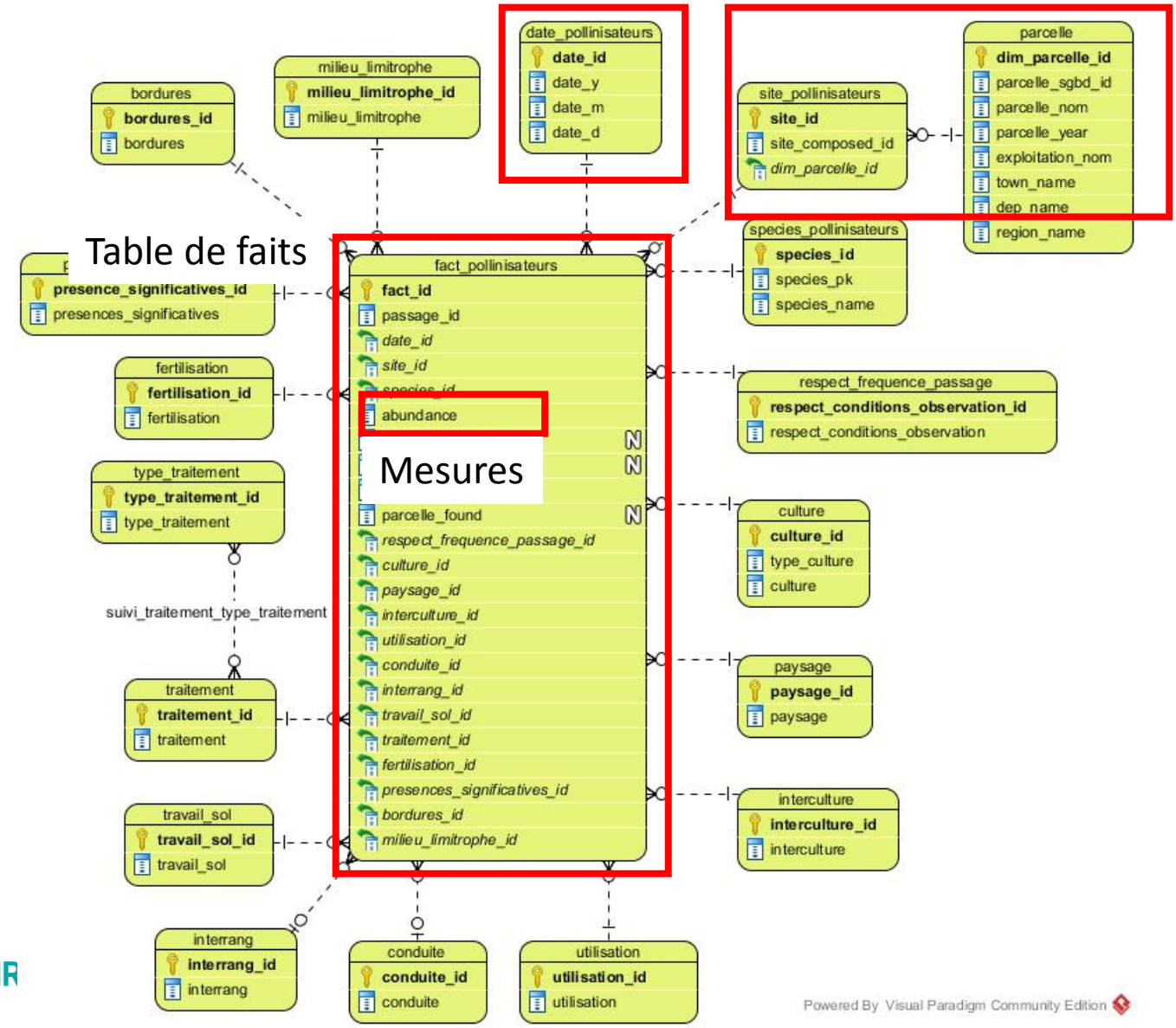
- Une dimension conforme ou partagée est une dimension utilisée par les faits de plusieurs entrepôts de données
  - Opération de drill-across
    - Pour expliquer un résultat, il est parfois nécessaire de le comparer avec d'autres faits

Faits et dimensions

- 3 formes de modèles multidimensionnels :
  - Le modèle en étoile (Star schema)
  - Le modèle en flocon de neige (Snowflake schema)
  - Le modèle en constellation (Factflake schema)

Modélisation logique

Tables de dimensions



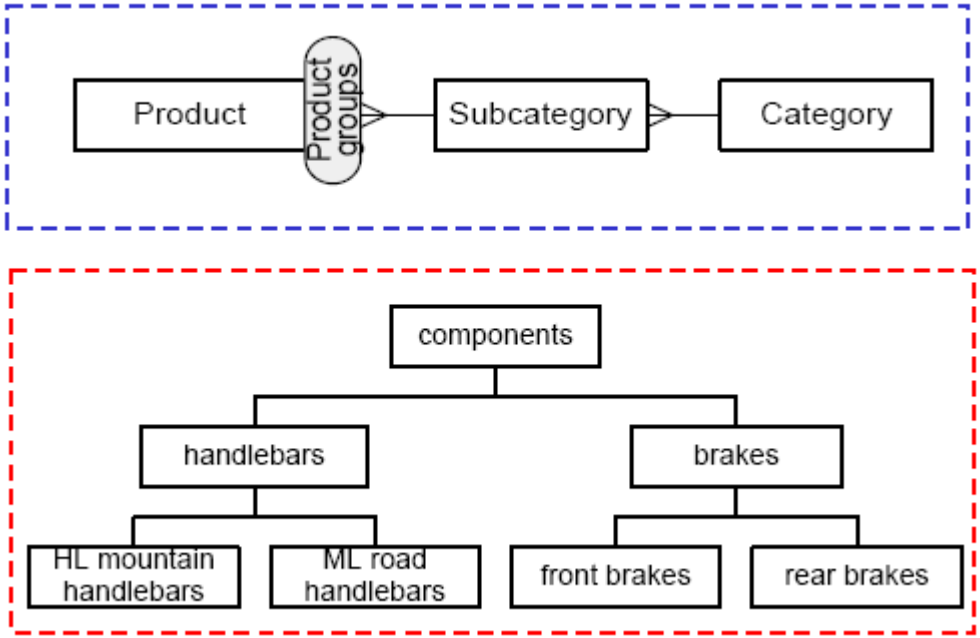
Hiérarchies

- Une dimension peut être composé par *n* hiérarchies
  - Balanced
  - Unbalanced
  - Recursive
  - Non-covering
  - Non-strict

Hiérarchies

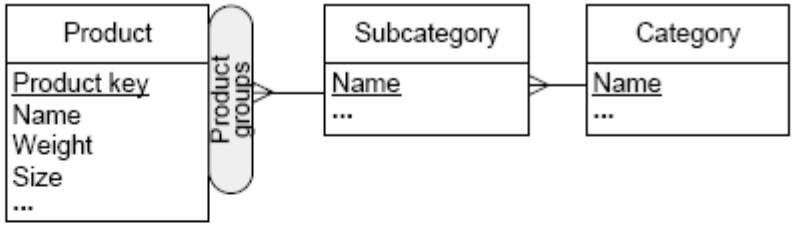
Balanced

- Schéma : un seul chemin
- Instance : les membres forment un « balanced tree »

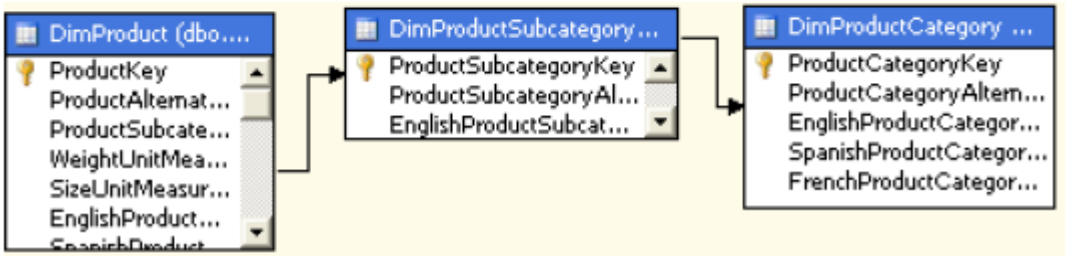


Hiérarchies

- Balanced
  - MultiDim



  - Relational model: snowflake schema



Serveur OLAP

- Mapping entre le client OLAP et le SGDB
- Implementation des operators OLAP
- Techniques d'optimisation

Serveur OLAP : Exemple avec Mondrian

<Measure name="Promotion Sales" aggregator="sum" column =“salesP”
formatString="#,###.00“/>

- <Dimension name="Gender" foreignKey=“factcustomer_id">
  <Hierarchy hasAll="true" primaryKey="customer_id">
  <Table name="customer"/>
  <Level name="Gender" column="gender"
  uniqueMembers="true"/>
  </Hierarchy>
  </Dimension>

Serveur OLAP

- Les requêtes OLAP définissent un « Pivot table »

		Unit Sales	Store Cost	Store Sales	Sales Count	Store Sales I
1997	Canada					
	USA	266.773,00	225.627,23	565.238,13	86837	339.610,90
1998	Canada	46.157,00	39.332,57	98.045,46	14632	58.712,89
	USA	259.916,00	220.645,11	550.808,42	84470	330.163,31

1. Définies en langage Serveur OLAP : MDX
2. Traduites en SQL par le Serveur OLAP

Exemple MDX

```
SELECT
{([Measures].[Unit Sales])} ON COLUMNS,
[Time].[Year].Members ON ROWS
FROM SALES
```

- Aucune vision du SGBD, mais que du modèle multidimensionnel

Méthodologies de conception

- Guidées par les données : exclusivement les sources de données disponibles
- Guidées par les utilisateurs : exclusivement les besoins des utilisateurs décisionnels
- ***Hybrides*** : guidées par les données et guidées par les utilisateurs

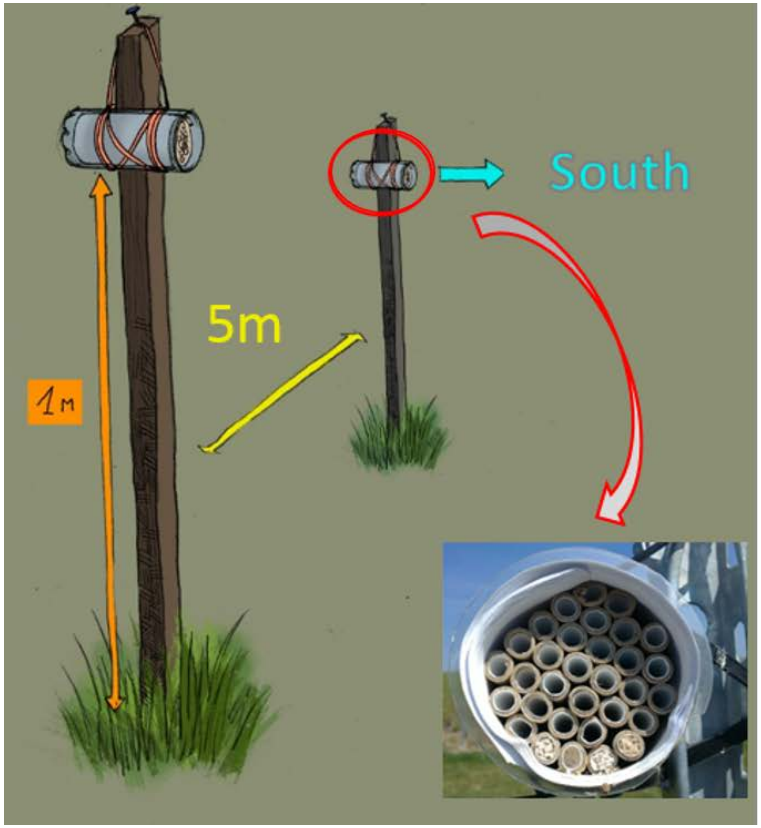
Modèle physique

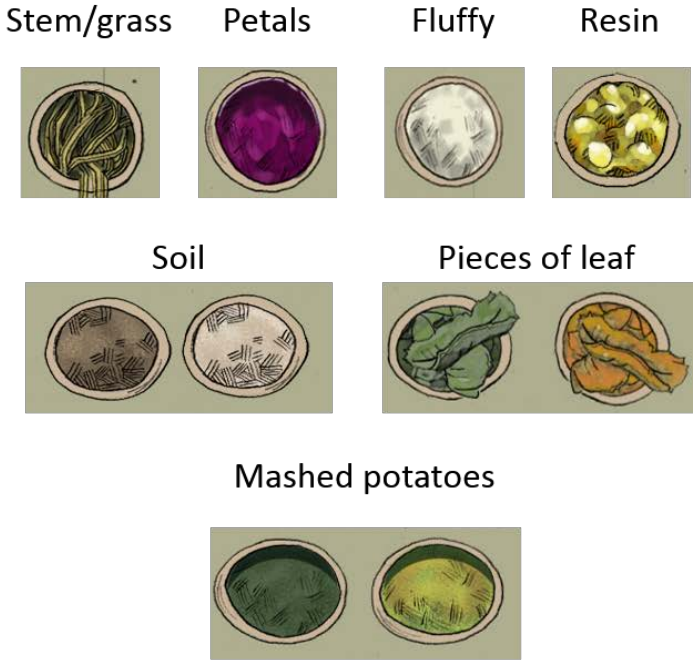
- Indexes (ex: Bitmap)
- Partitions
  - Jointures
  - Sélection
- Vues matérialisées
  - Agrégations

Cas d'étude

Observatoire Agricole de la Biodiversité

- Collecte de données à l'échelle de la France depuis 10 ans
- Pollinisateurs, vers de terre, invertébrés et papillons



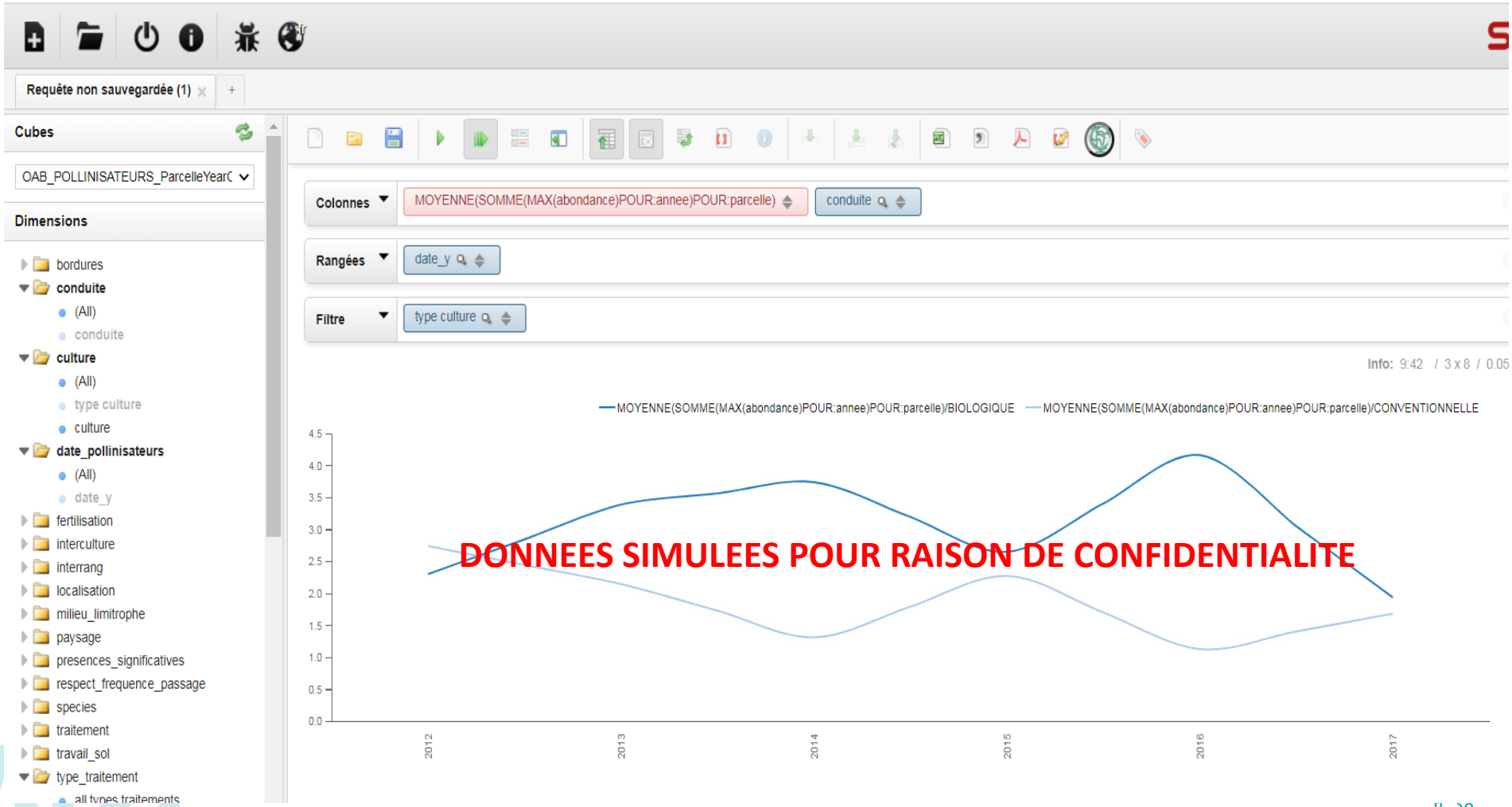


Exemple Pollinisateurs

INRAE

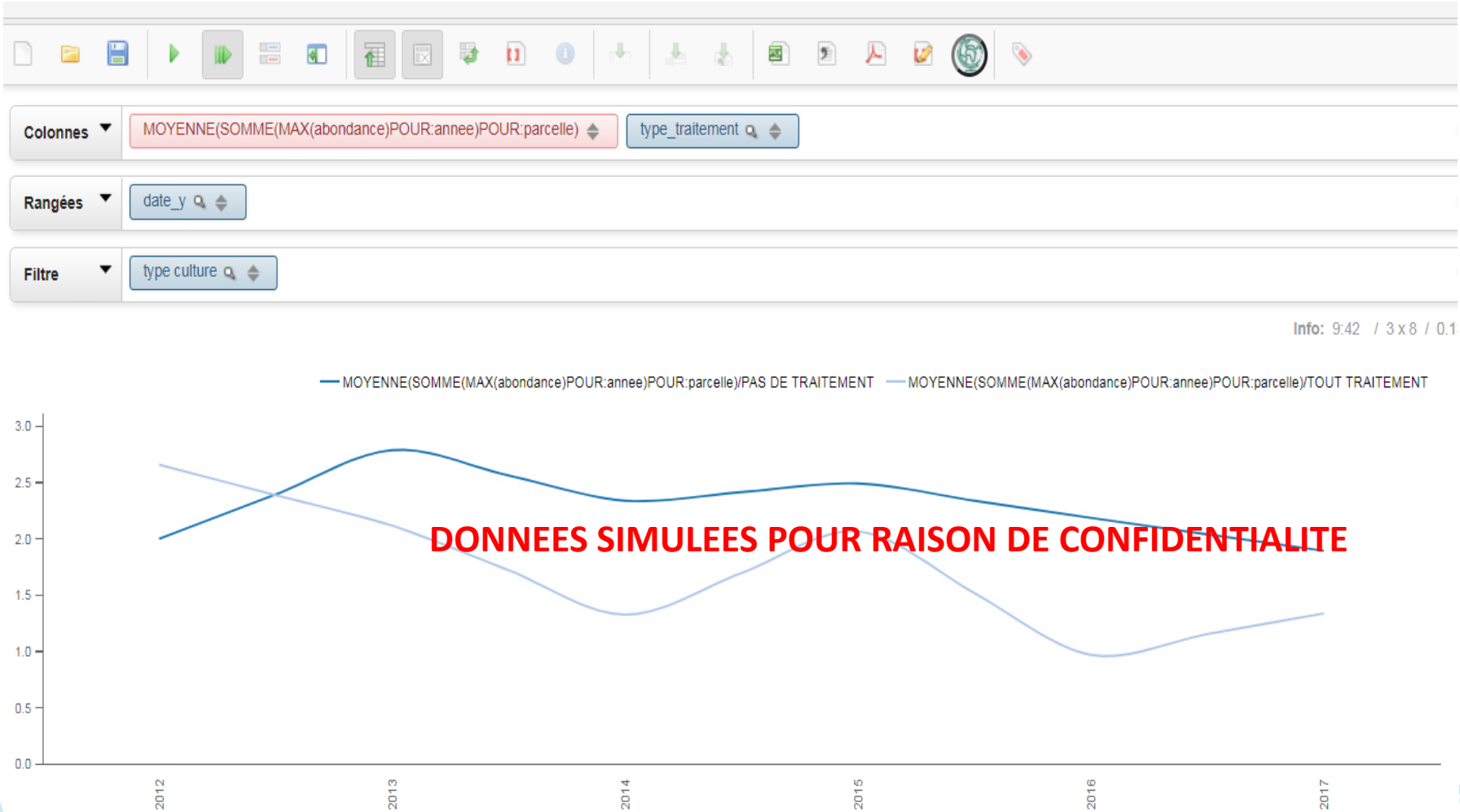
Analyse OLAP (ANR VGI4bio)

- Evolution temporelle de l'abondance des pollinisateurs par type de conduite pour les grandes cultures



Analyse OLAP (ANR VGI4bio)

- Evolution temporelle de l'abondance des pollinisateurs par type de traitement (TOUS, AUCUN) pour les grandes cultures



La recherche en entrepôts de données et OLAP

Pistes de recherche

- Nouveaux SGDB : NoSQL (document, graph, Key Value, etc.)
- Analyse OLAP « aidée »
- Méthodologies de conception collaboratives
- « Augmented BI » : intégration avec méthodes de Data science
- Entreposage de données IoT
- Entrepôts de données sémantiques
- Multi-modèle

Entrepôts de données multi-modèle

Entrepôt de données

?

Variété (XML, JSON, etc.)

Notre vision

Sandro Bimonte, Yassine Hifdi, Mohammed Maliari, Patrick Marcel, Stefano Rizzi:
To Each His Own: Accommodating Data Variety by a Multimodel Star Schema. DOLAP 2020: 66-73

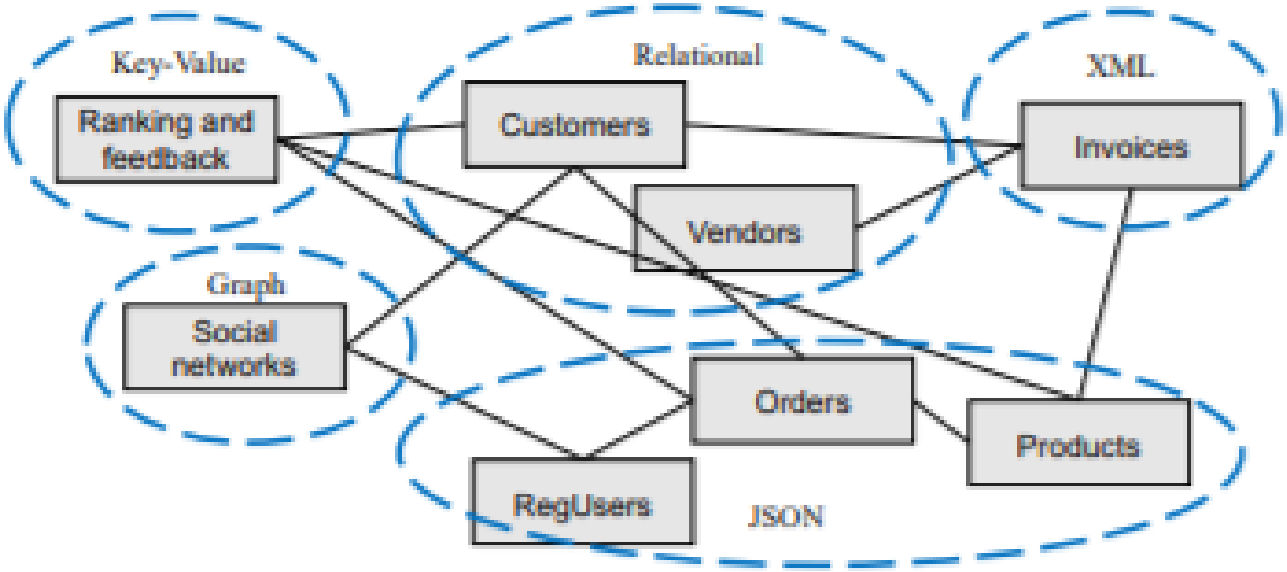
- L'entrepôt de données multi-modèle peut stocker des données selon le modèle multidimensionnel et, en même temps, laisser chacun de ses éléments être représenté nativement à travers le modèle le plus approprié

- AVANTAGES :
  - Manipulation de la variété
    - et volume et la vitesse
  - Combler le fossé architectural entre les lacs de données et les ED
  - Réduire le coût des transformations de données ETL
  - Grâce à l'utilisation de modèles sans schéma :
    - Extensibilité
    - Évolutivité

Cas d'étude : Unibench

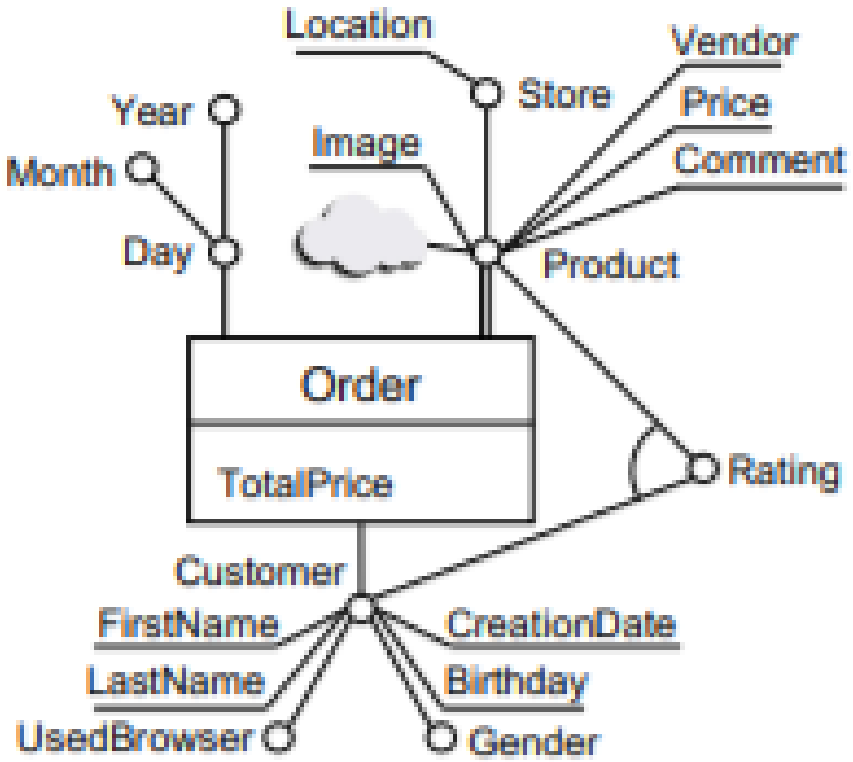
- UniBench est un benchmark pour les bases de données multi-modèles

- UniBench n'a pas été conçu pour les requêtes OLAP



Cas d'étude : Unibench

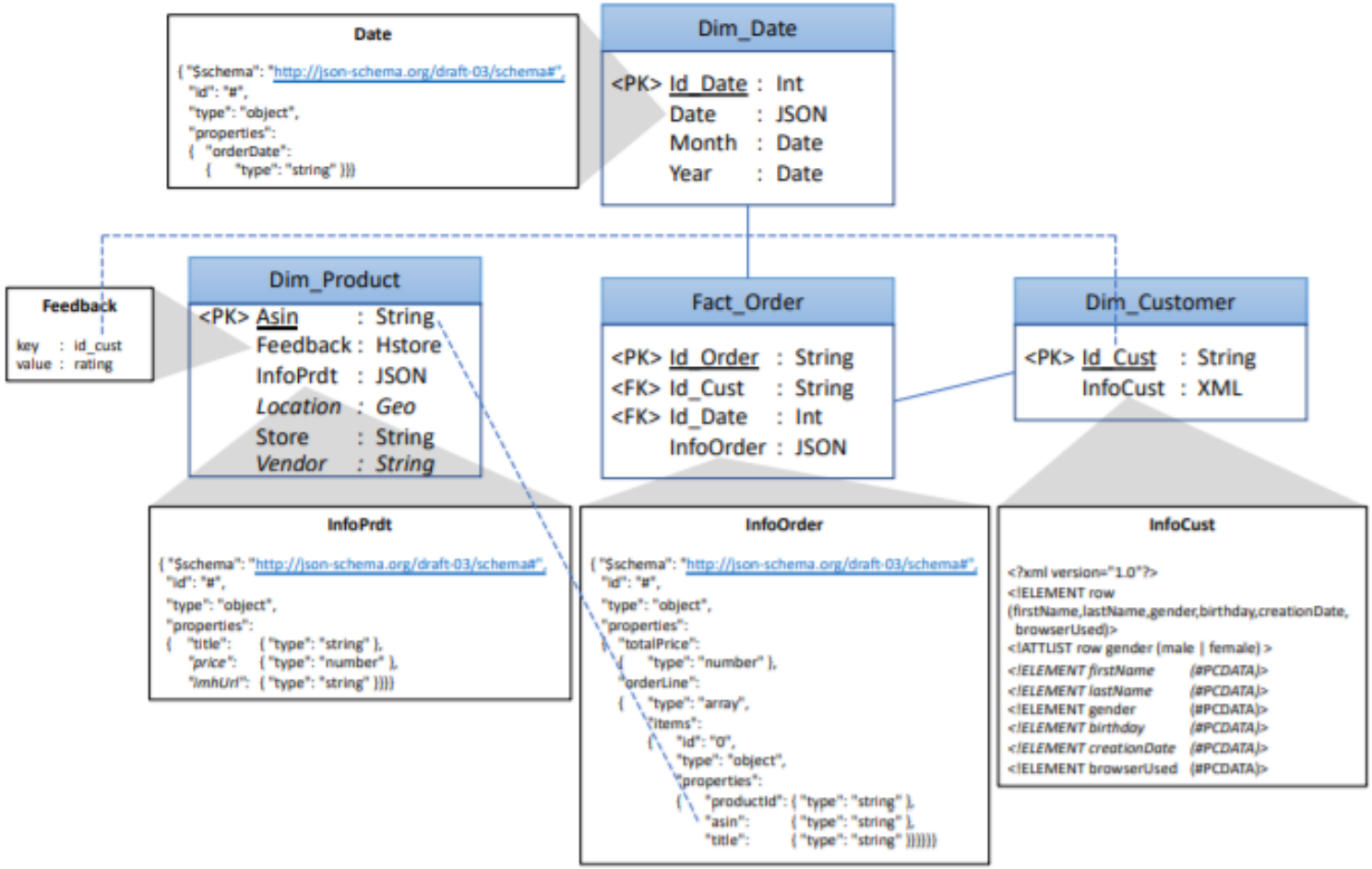
- Schéma multidimensionnel dérivé de Unibench



Multi-model star schema

- Nous utilisons un schéma en étoile classique avec des tables de faits et de dimensions, étendues avec :
  - Données semi-structurées sous forme JSON et XML
  - Données spatiales
  - Données de clé-valeur

Multi-model star schema



Merci pour votre attention

Sandro.bimonte@inrae.fr