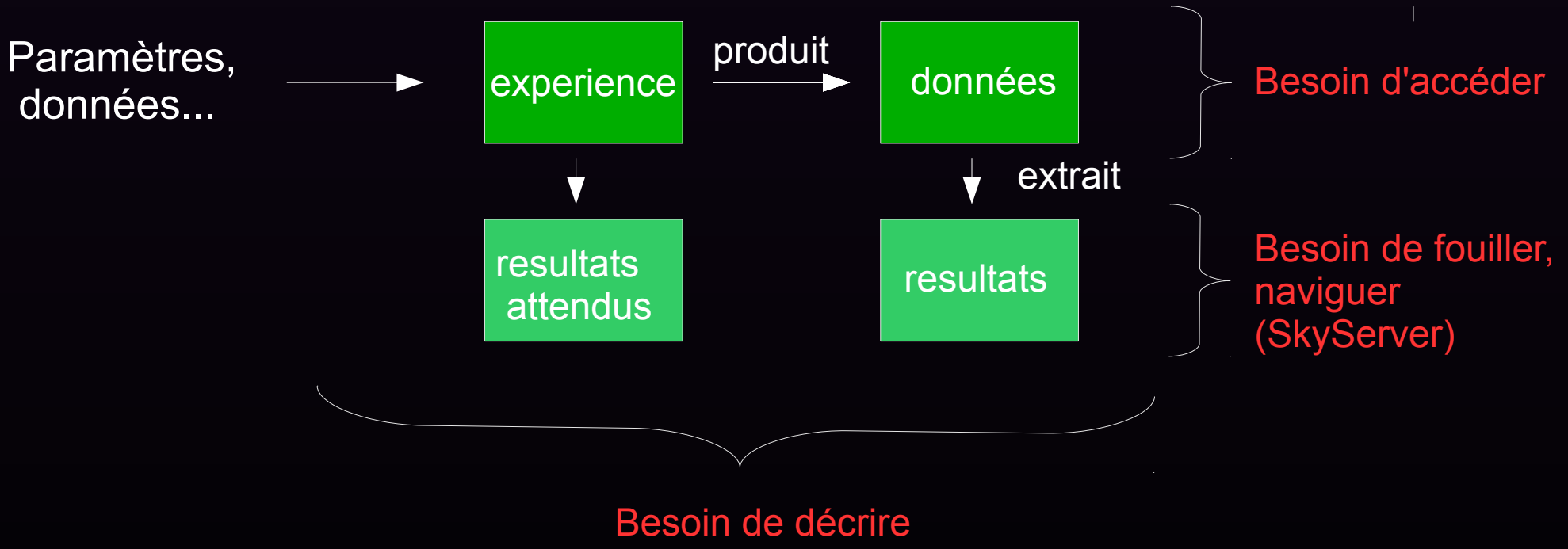
Les très grands ensembles de données des simulations numériques

David Languignon, Franck Le Petit

ASOV meeting - Paris - 17 Janvier 2012

Cas d'utilisation

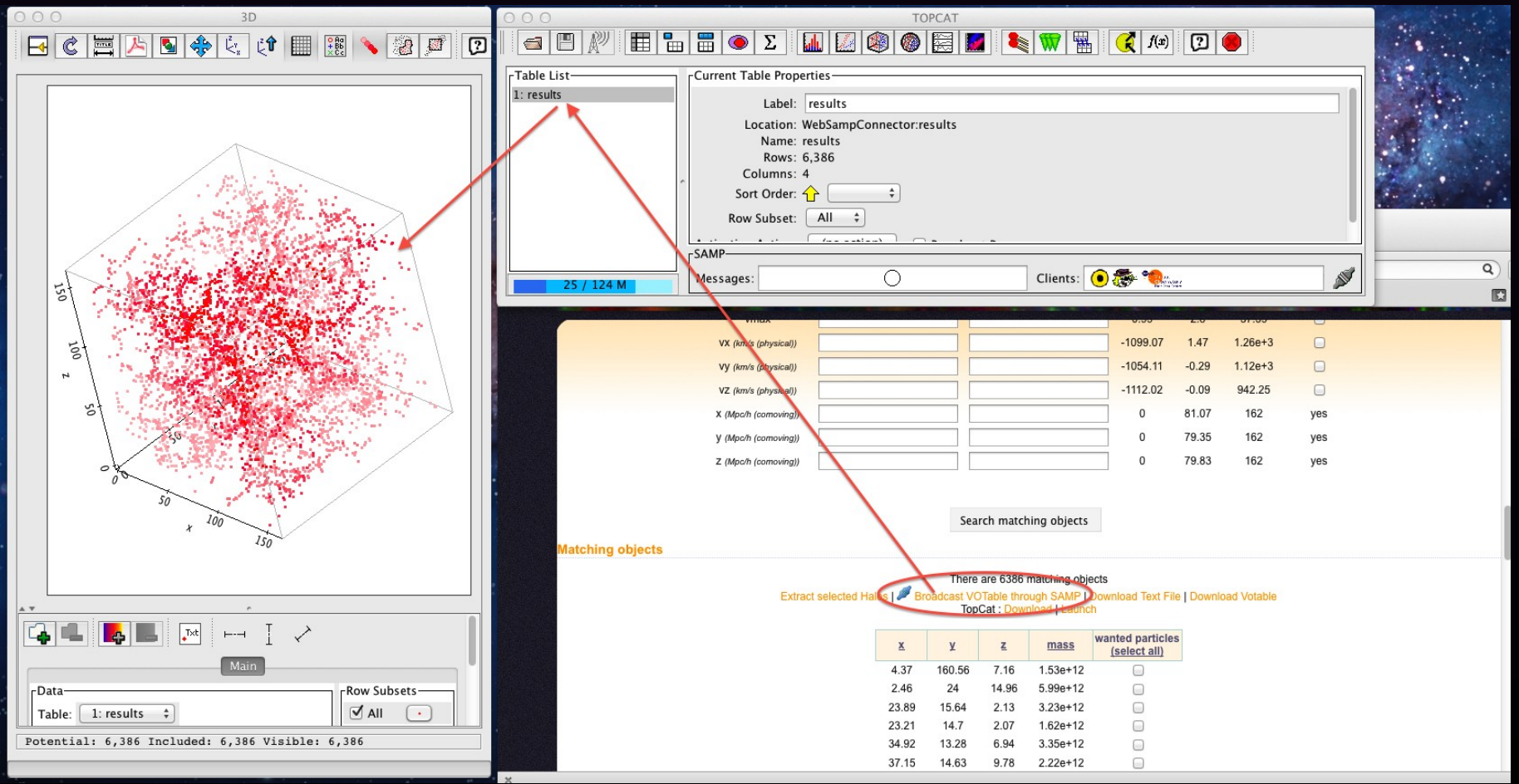


Actions de standardisations au sein de VO-Theory

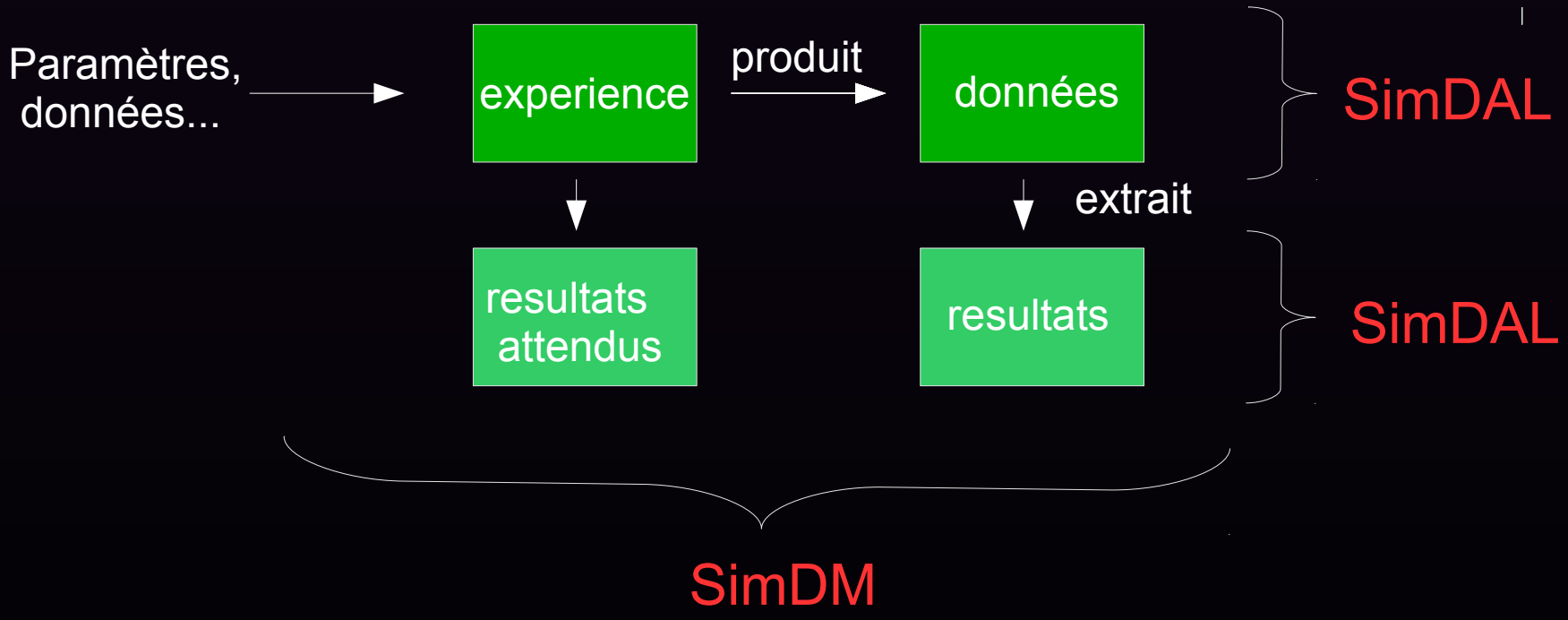
- Solutions pour décrire
  - Simulation Data Model (SimDM) Ok
- Solutions pour accéder/fouiller
  - Simulation Data Access Layer (SimDAL) En Cours
    - Adaptation des solutions DAL pour les simulations
    - Ex : TAP..
  - Couche Sémantique (skos concepts...) En cours

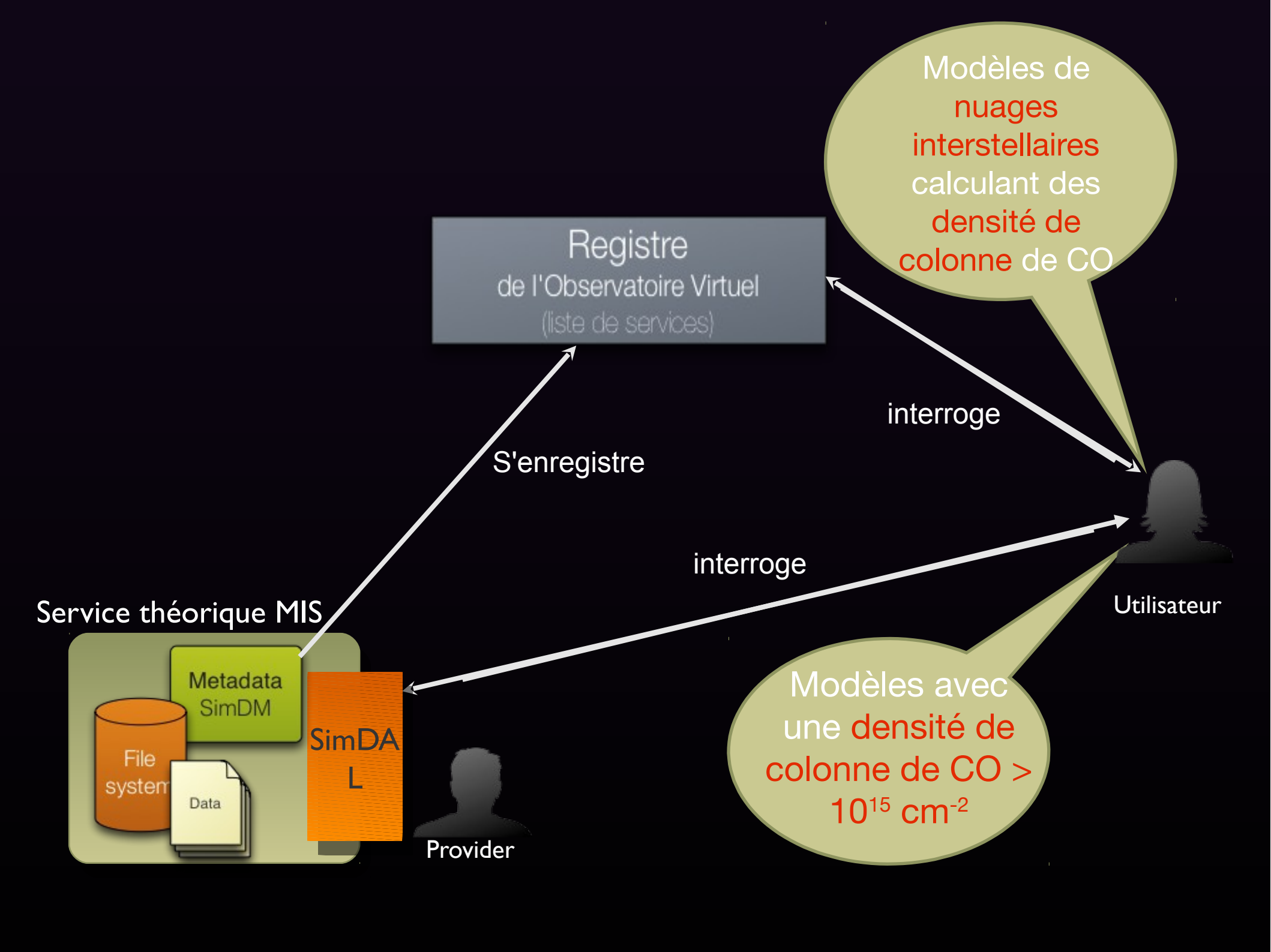
Ce qu'on peut faire aujourd'hui

- Décrire de manière standardisée (SimDM)
- Accéder de manière partiellement standardisée



Il manque un standard d'accès



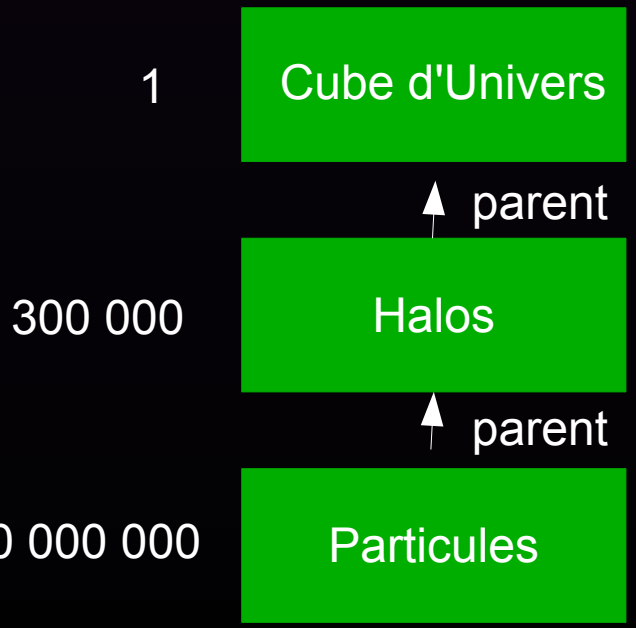

Modèles de nuages interstellaires calculant des densité de colonne de CO
Registre
de l'Observatoire Virtuel
(liste de services)
interroge
S'enregistre
interroge
Utilisateur
Service théorique MIS
Metadata
SimDM
File
system
Data
SimDA
L
Provider
Modèles avec une densité de colonne de CO > 10^{15} cm^{-2}

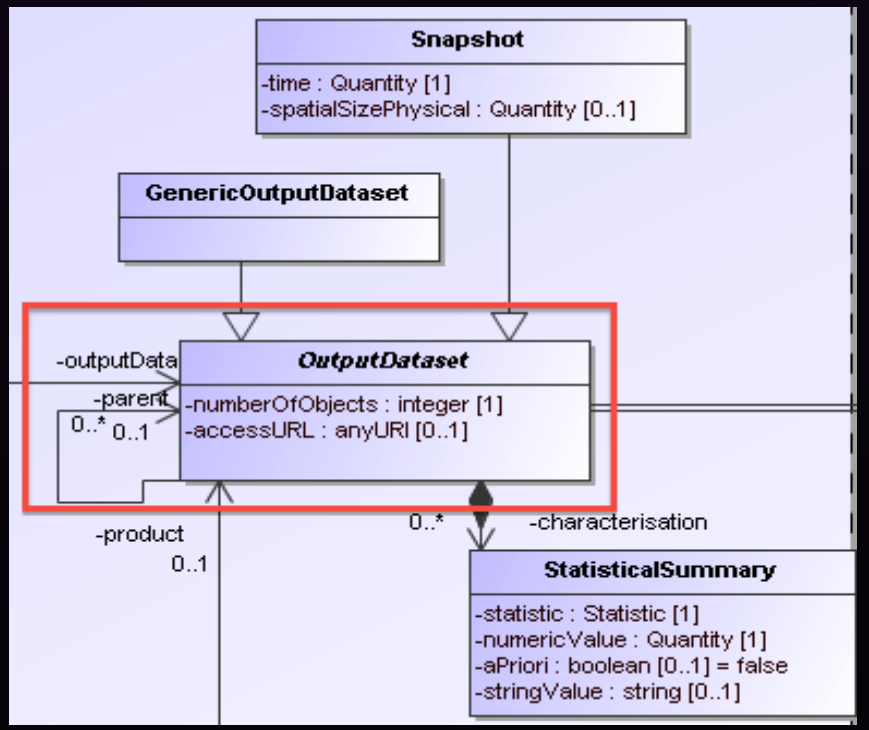
SimDM / SimDAL

Périmètre SimDM, SimDAL ?

Quel grain pour l'arborescence OutputDataset ?

Ex : Cosmologie





Utilisation des resultats

- Pipe entre plusieurs codes/services
- Preview (sous quelle forme ?)
- Navigation, fouille, analyse (ex : SkyServer)
- Accès direct, récupération (téléchargement ?)

Problème

- Résultats (SimDM Dataset) souvent TRES volumineux
  - Ex : Cosmologie

Comment faire ?

Existant non adapté

- Bases de données relationnelles (transactions)
- Conçu pour des utilisations spécifiques (finance)
- Datawarehouse ? Model étoile relationnel
- Clusters ? Ne fait que repousser les limites du relationnel
- Bases objets ? Orientées colonnes ? Le problème est déplacé

Et Pourtant

SkyServer fonctionne avec SqlServer (cluster), maintenant avec MonetDB (orienté colonne).

Est-ce encore extensible ?

Loi de Moore dirait oui

Problème

- Certains facteurs ne croissent pas automatiquement avec l'évolution des technologies
  - Manpower nécessaire
  - Disk i/o
  - Net bandwidth

Constats

- Pas besoin de transactionnel
- Il faut distinguer données brutes en sortie d'une experience et resultats extraits par le scientifique

- Besoin d'un preview (à différentes échelles)
- La science est faite en fouillant les resultats en ligne

Solutions à l'étude

- Étendre au maximum les capacités du relationnel lorsque c'est suffisant
  - bases plates (dénormalisation)
  - cluster, segmentation
- Nouvelles approches
  - Machine learning objets d'intéret et comportement utilisateur sur les ensembles de données
  - Indexation sémantique

Etude de cas

- Deuvo 2010
  - Simulations cosmologie
  - 4 000 000 halos
  - 140 milliards particules

  - Grain décrit par SimDM : cube ramses (= stats sur les halos)
  - Grain halos stocké dans des tables plates séparées (4 000 000 lignes) (stats sur les particules)
  - Grain particules en fichier binaire téléchargeable (cutout possible)

Besoin urgent d'une nouvelle approche

- Deuvo 2012
  - 200 000 000 halos.....

A term

- Est il pertinent de conserver TOUTES les données d'une simulation ?
  - Reproductible à moindre coût dans quelques années (mois) à partir des métadonnées
  - <> données observationnelles
- Approche “run à la demande” si loin que ça ?
  - Devons nous penser les standards dès maintenant ?
  - PDL