Experimental Evaluation of a Dynamic Cubing system: workflow, metrics and prototype

Anne Tchounikine, Maryvonne Miquel, Usman Ahmed
LIRIS
CNRS UMR 5205, INSA-Université de Lyon, France

Motivations

- Motivé par des travaux précédents : Cube et OLAP dynamique
- Problématique de l'évaluation de la solution
  - Démonstration et qualification formelle
  - Démarche expérimentale
    - Prototypage
    - Appréciation des résultats
- **Démarche expérimentale**
  - Observer des comportements
  - Vérifier certaines intuitions
  - Mesurer des résultats

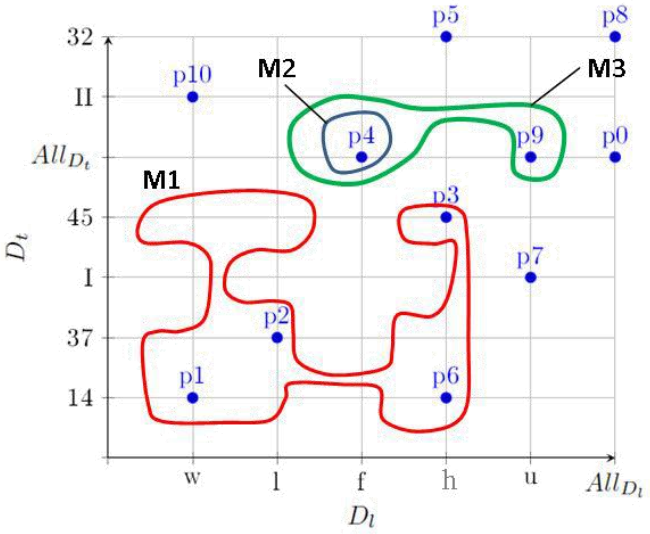
Objectifs et éléments de la contribution

- Résultats et **démarche** de l'évaluation expérimentale
- Un prototype et des expérimentations pour :
  - Montrer la faisabilité
  - Effectuer des tests fonctionnels
  - Ajuster les paramètres de la solution
  - Evaluer la performance
  - Etudier le comportement
  - Mener des études comparatives
- Définition d'un workflow pour l'expérimentation
  - Métriques
  - Caractérisation des éléments en entrée
  - Définition de scenarios d'exécution
  - Outils de monitoring

➔Appliqué à notre proposition de cube dynamique

Le modèle de cube

- Contexte : BI temps-réel
- Un espace multidimensionnel hiérarchique
  - Tous les membres d'une dimension sur le même axe
  - Membres non ordonnés
  - Axes construits dynamiquement
- Les **MBS** (Minimum Bounding Set) regroupent des points de même niveau hiérarchique



p4 ⊙ p2

M1 = {45, 37, 14} x {w, l, h}

|M1| = 9

||M1|| = 4

Let $\Delta = \{(x_{1,1}, x_{1,2}, ..., x_{1,n}), (x_{2,1}, x_{2,2}, ..., x_{2,n}), ..., (x_{m,1}, x_{m,2}, ..., x_{m,n})\}$ be a set of m points lying in a same n-dimensional hyper-plane i.e. $x_{j,i} \in domain(l_i^{k_i})$ for $1 \leq i \leq n$, $1 \leq j \leq m$ and k_i is a level in the hierarchy of dimension D_i. A minimum bounding space (MBS) constructed over Δ, denoted by M_Δ, is defined as:

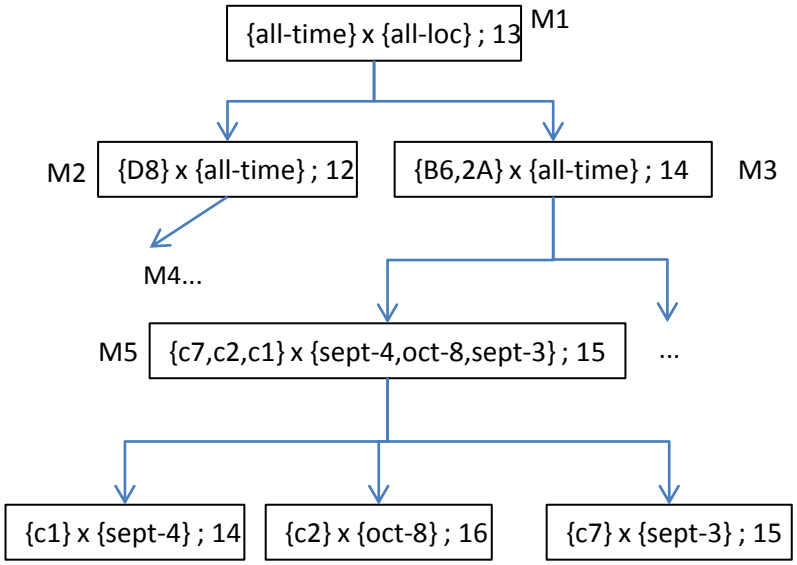
$$M_\Delta = \bigcup_{j=1}^{m} x_{j,1} \times \bigcup_{j=1}^{m} x_{j,2} \times ... \times \bigcup_{j=1}^{m} x_{j,n}$$

where each set $\bigcup_{j=1}^{m} x_{j,i}$ is called the $i^{th}(1 \leq i \leq n)$ dimension *edge* E_i of M_Δ

$|M_\Delta| = \prod_{i=1}^{n} |E_i|$ the number of points covered by the MBS M_Δ,
$||M_\Delta|| = |\triangle|$ the number of points enclosed in the MBS M_Δ.

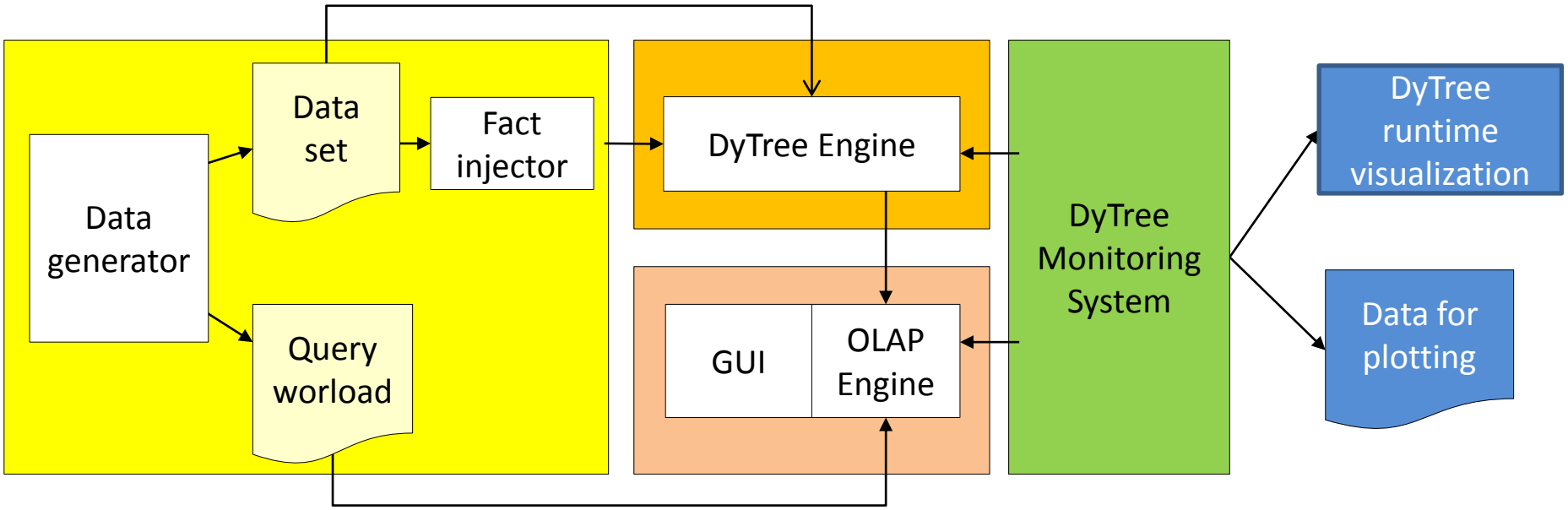
Le modèle d'arbre

- Le **DyTree** est le cube matérialisé
- Les points (faits) sont positionnés dans l'espace au fur et à mesure de leur arrivée. Ils sont les feuilles de l'arbre
- Les points sont regroupés dans un MBS. Lorsqu'un MBS est plein :
  - il est divisé (split)
  - un MBS de points couvrants est créé
  - La mesure agrégée est calculée

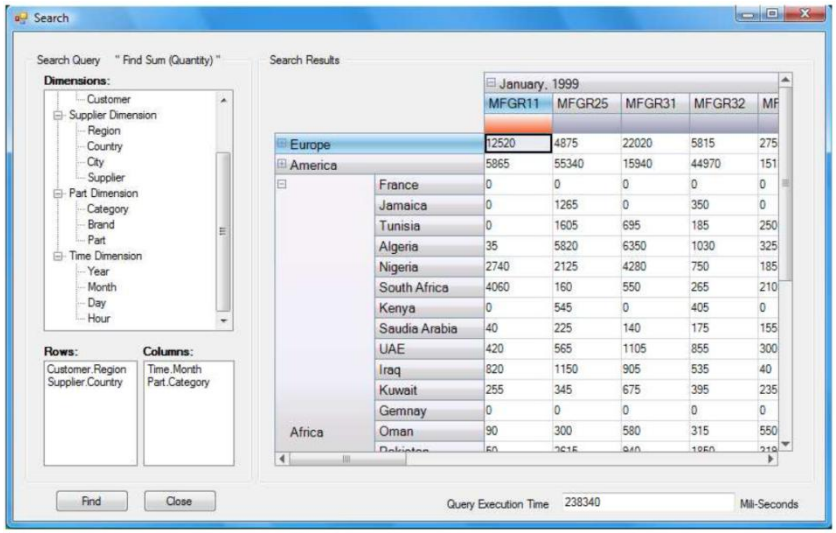


➔les MBS sont des morceaux de cuboïdes
➔les MBS sont de plus en plus détaillés
➔les MBS créés sont naturellement denses

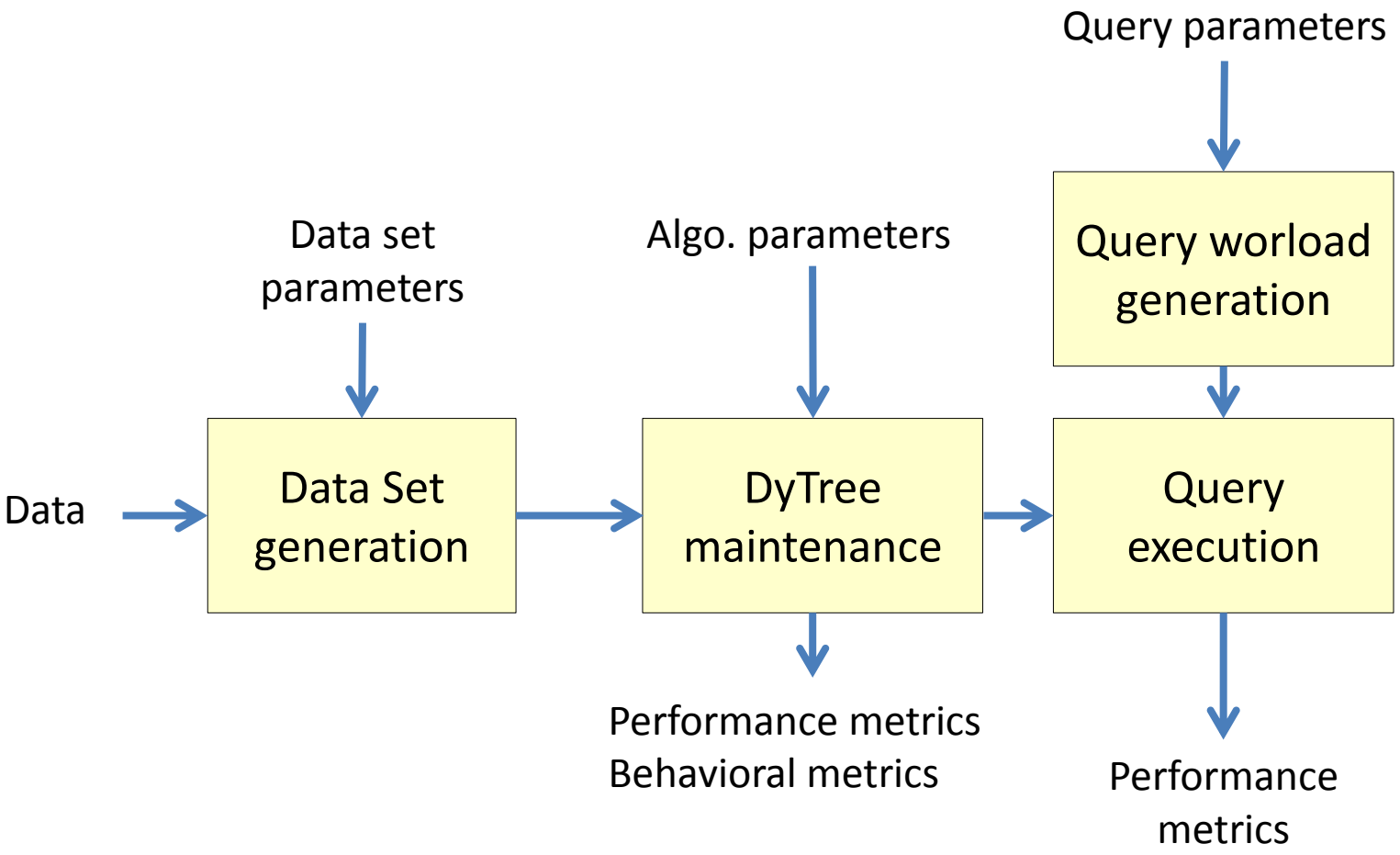
Le prototype



- Efficience du Dytree et des algorithmes associés ?
- Comportement en fonction des contextes d'exécution ?



Workflow de l'expérimentation



Métriques de performance

- Métriques de performance
  - Temps d'exécution
    - Temps cumulé pour l'insertion d'un ensemble de faits
    - Temps d'insertion atomique d'un fait
    - Temps de réponse aux requêtes
  - Taille de l'arbre en mémoire
    - Les feuilles (faits) sont stockés sur disque
    - Les nœuds (agrégats) en mémoire

Métriques de comportement

- Métriques sur l'arbre
  - Nombre de nœuds internes
  - Nombre de super nœuds
  - Profondeur
  - Largeur
- Métriques sur les nœuds
  - Taux de remplissage

$$fillRatio(node) = \frac{number\ of\ children(node)}{directory\ node\ capacity}$$

  - Volume et Densité
    - Calculé sur le résultat (N) d'un drill-down des coordonnées du MBS (M)
    - Nombre de points couverts et nombre de points inclus

$$volume(M) = |N| \text{ and } density(node) = \frac{||N||}{volume(M)}$$

Métriques vers indicateurs

- **Transformer les mesures en indicateurs (➔ signification, impact)**
- Métriques de performance
  - Temps d'exécution
    - Insertion d'un fait, temps cumulé, requête
  - Taille de l'arbre en mémoire
- Métriques de comportement
  - Métriques portant sur l'arbre
    - Nombre de nœuds internes
    - Nombre de super nœuds
    - Profondeur
    - Largeur
  - Métriques portant sur les nœuds
    - Taux de remplissage
    - Volume
    - Densité

➔ efficience de la solution dans un contexte BI "agile"

➔ nombre d'agrégats matérialisés

➔ granularité des agrégats

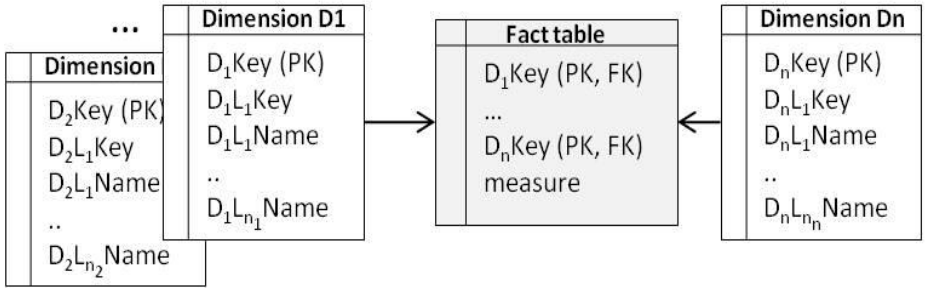
➔pertinence des agrégats

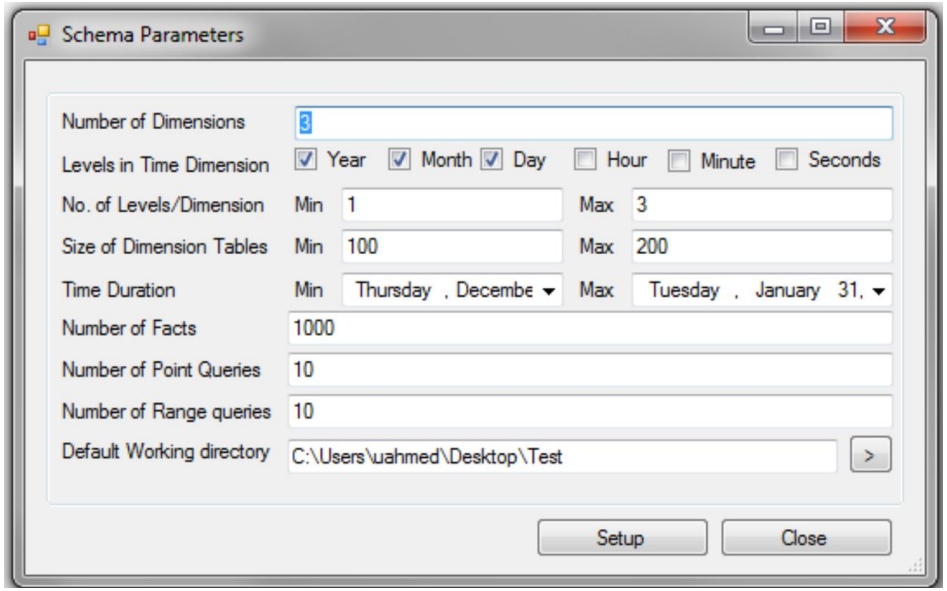
Eléments en entrée

- Lister et caractériser les inputs pour créer des scénarios d'exécution
- Paramétrage des algorithmes
  - Capacité des nœuds
  - Recouvrement (overlap)
- Paramétrage des data sets
  - Schéma
  - Taille (cardinalité des dimensions, volume de faits)
  - Densité
  - Ordre
  - Fréquence d'arrivée
- Paramétrage des requêtes
  - Nombre de requêtes
  - Types (point, range, group by)

Schéma des data sets

- Benchmark : Star Schema Benchmarks
- Données synthétiques
- Données réelles : SoQ4Home
  - Bâtiments intelligents
  - Capteurs de température
  - Déployé sur le campus LyonTech



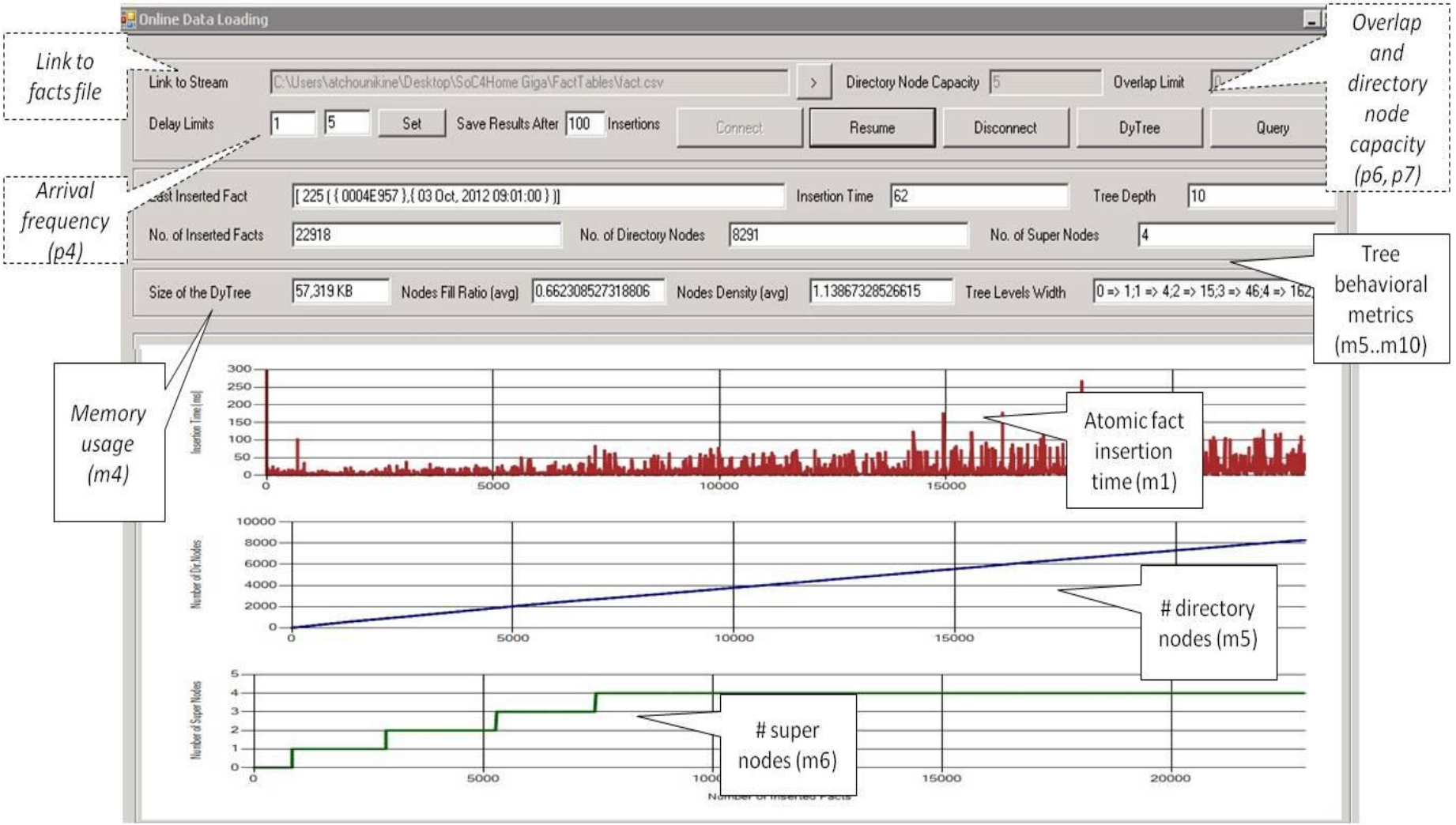


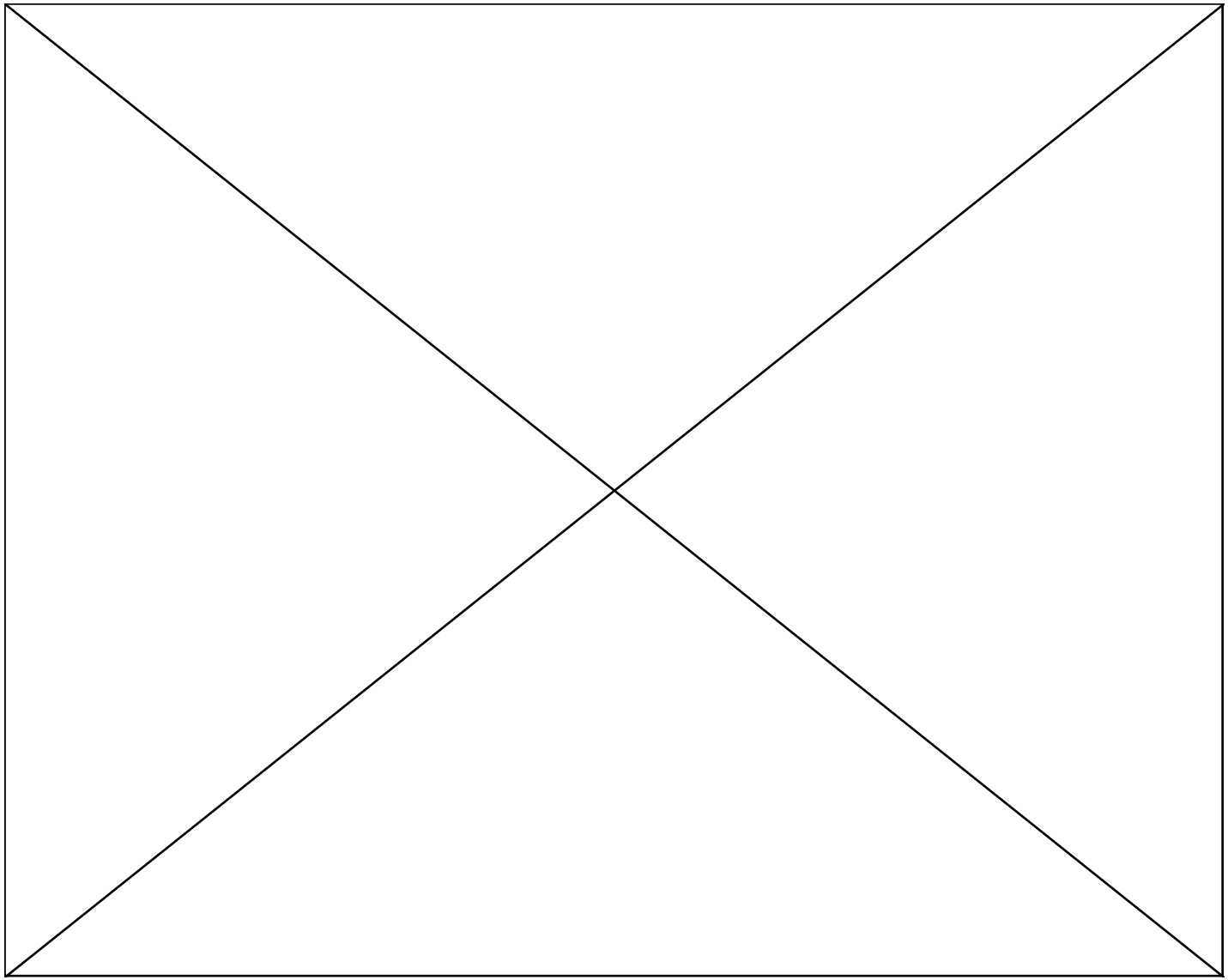
Constitutions de scenarios

- En faisant varier les paramètres en entrée

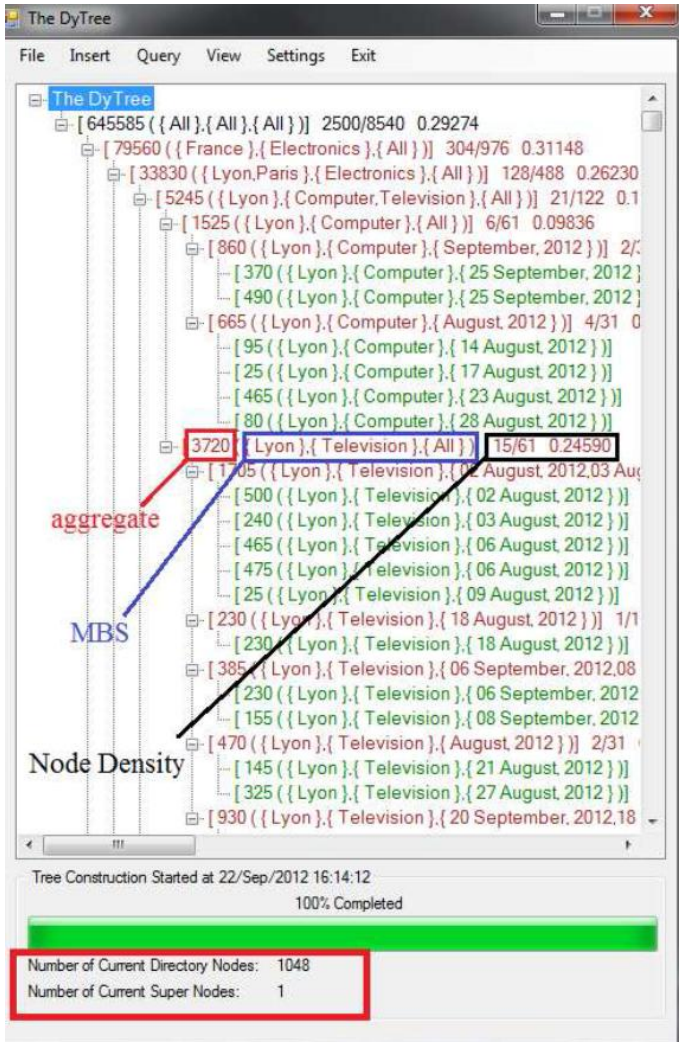
Scenario	Data Set	# dim	# fact	density	arrival freq	order	dir cap	overlap
Varying overlap	SSB	4	10M	calc.	na	na	15	0 to 15
Varying dir. cap.	SSB	4	10M	calc.	na	na	5 to 75	0
Scaling # facts	SSB	4	up to 10M	calc.	na	na	15	5
	synth	10	up to 100M	calc.	na	0%	15	5
	Soq4Home	2	na	na	na	na	15	5
Scaling # dim.	synth	2 to 30	up to 100M	calc.	1 to 5 ms	0%	15	5
Varying density	synth	10	calc.	0.2 to 0.6	1 to 5 ms	0%	15	5
Delayed arrival	synth	10	up to 100M	calc.	1 to 5 ms	5 to 80%	15	5

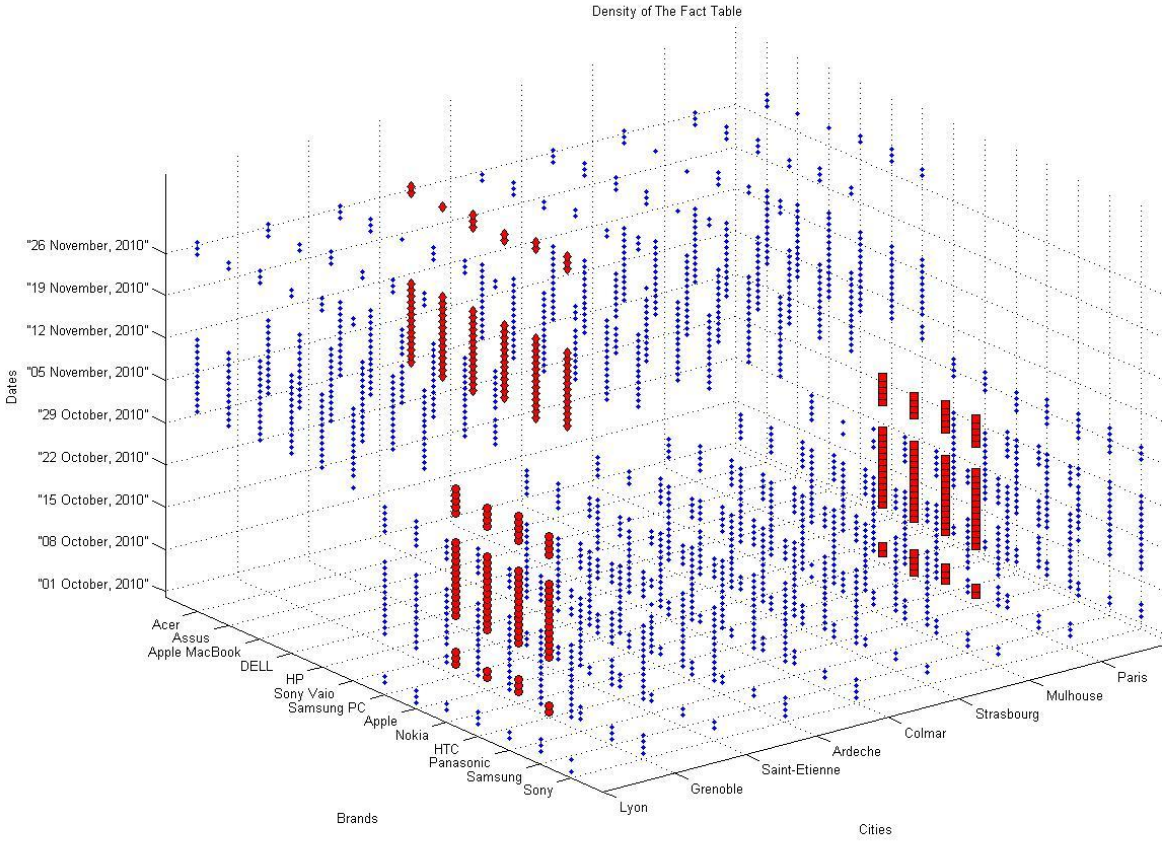
Outils de monitoring : en temps réel



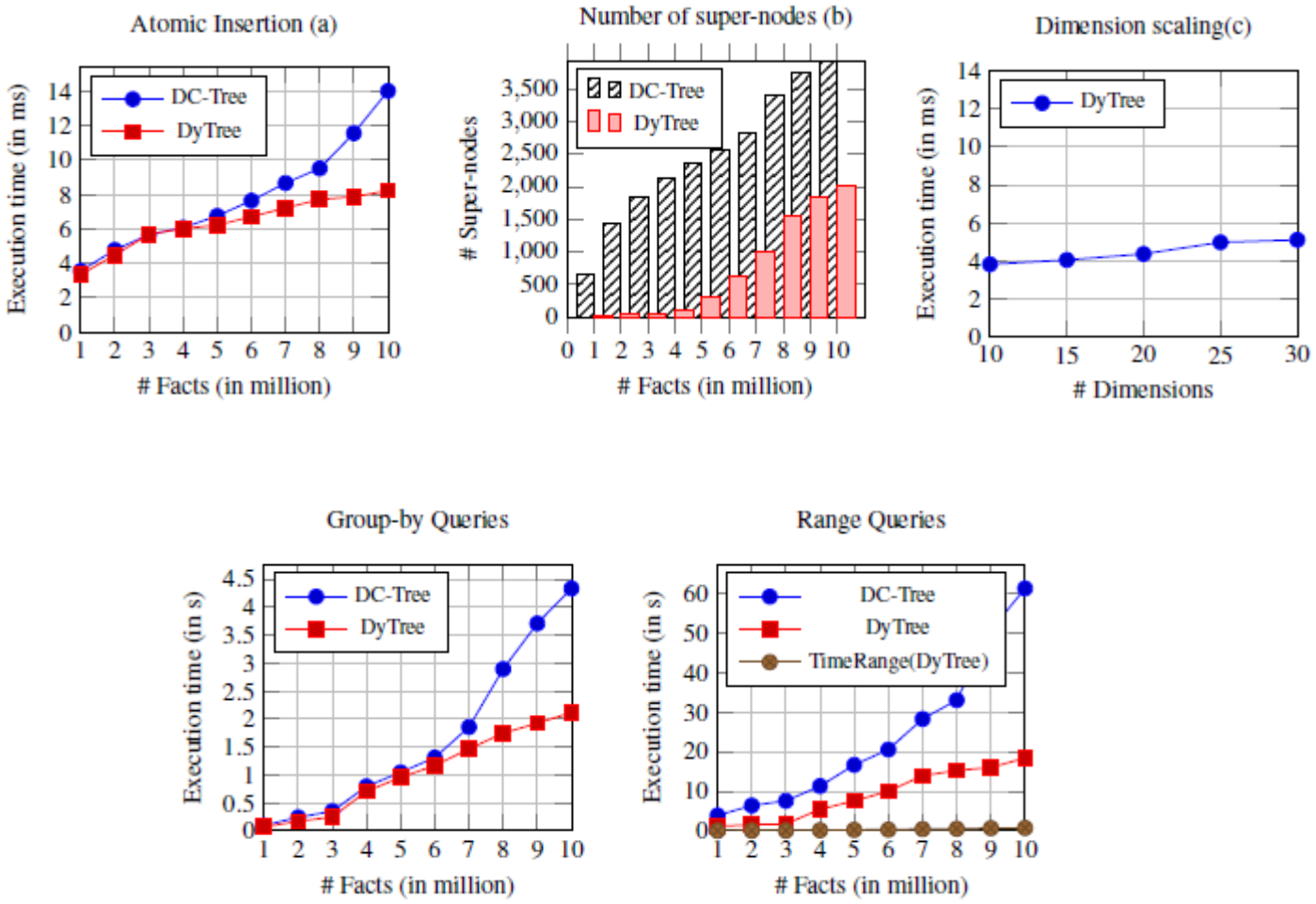


Outils de monitoring : des instantanés





Outils de monitoring : analyses a posteriori



Conclusions

- Ce prototypage nous a permis d'illustrer le modèle
- Une démarche expérimentale permet de valoriser le prototypage
  - Instancié sur notre exemple mais adaptable
- Approche complémentaire aux approches benchmark

- Perspectives
  - Les requêtes
    - Type de requêtes
    - Volume des résultats…
  - **Data set synthétiques**
    - **Génération de faits sous contraintes**