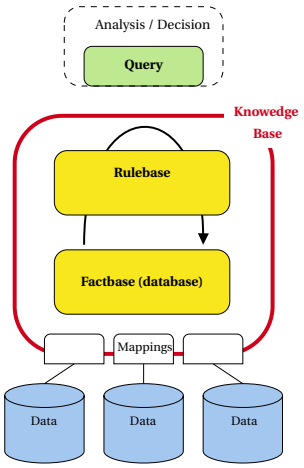
# Stage M2 Recherche

# Accès efficace aux données hétérogènes : étude d'approches alternatives à la matérialisation et à la saturation

Federico Ulliana (ulliana@lirmm.fr)

CONTEXTE ET MOTIVATION L'objectif du stage est d'étudier des techniques permettant l'*interrogation efficace de données hétérogènes.* Nous nous plaçons dans le cadre d'un système d'intégration de données à base de connaissances, illustré ci-dessous. Ceci est composé (1) d'un niveau des sources de données, (2) d'une base de connaissances, constituée à son tour par une base de faits et une base de règles, et (3) d'un niveau de mappings faisant le lien entre les données dans les sources et la base de connaissances (faits et règles).




La base de faits doit avant tout refléter la prise en compte des mappings. Celle ci peut être en principe *matérialisée* ou *virtuelle*. Dans le cas de la matérialisation, les données sont tout simplement transvasées des sources vers la base de faits. Dans le cas de la virtualisation, les données restent au niveau des sources, et cette fois toute interrogation sur la base de faits est traduite dans des requêtes de bas niveau utilisant le langage de sources de données.

De façon similaire au peuplement de la base de faits, les raisonnements permis par la base de règles sont pris en compte soit par la *saturation* des données (en faisant l'hypothèse que les données sont matérialisées) soit par la *reformulation* des requêtes (qui peuvent ensuite être soit évaluées sur les données matérialisées soit traduites une deuxième fois vers les sources).

L'inconvenient principal de l'approche qui combine matérialisation+saturation est qu'elle produit des bases volumineuses et qui nécessitent d'être périodiquement mises à jour. Pour pallier ce problème il est possible de mélanger les approches, et de (1) faire co-exister des parties virtuelles et matérialisées de la base de faits, ainsi que de (2) prendre en compte une partie des règles par saturation et l'autre par reformulation.

Les deux critères importants pour determiner une configuration possible du système sont (1) la faisabilité des raisonnements permis par les règles (en effet, certaines règles admettent exclusivement la saturation+matérialisation, d'autres permettent aussi la virtualisation) et bien évidemment (2) les performances de l'accès aux données du système résultant. À ce point, nous nous retrouvons avec un paysage de solutions possibles pour l'accès aux données - toutes avec des coûts différents.

Dans ce stage, deux questions pourront être considérées - de façon non mutuellement exclusive. De plus, l'accent pourra être mis sur des aspects plus théoriques ou pratiques en fonction des intérêts du candidat.

VIRTUALISATION VS MATÉRIALISATION Pour la base de faits du système :

(Q1) *Comment déterminer automatiquement la (meilleure) configuration en matérialisation et virtualisation des mappings ?*

SATURATION VS RÉÉCRITURE Pour les parties matérialisées de la base de faits du système :

(Q2) *Comment déterminer automatiquement la (meilleure) configuration en saturation et réformulation des règles ?*

RÈGLES EXISTENTIELLES. Pour les mappings et la base de connaissances, nous considérerons le langage des règles existentielles. Il s'agit de formules de la logique du premier ordre de la forme $\forall X, Y.\big(\varphi(X,Y) \rightarrow \exists Z.\psi(X,Z)\big)$, avec $\varphi$ et $\psi$ des conjonctions d'atomes sans symboles fonctionnels, permettant de faire abstraction sur une grande famille de tâches liées aux données et aux connaissances. Il sera en particulier possible de considérer des extensions des règles existentielles permettant plus de traitements de données (avec négation, fonctions, prédicats calculés, agrégations, etc) qui préservent néanmoins de bonnes propriétés computationnelles.

CAHIER DES CHARGES. Le stage pourra être articulé autour des questions énoncées ci-dessus. Dans les deux cas, le schéma de résolution est similaire, même si les défis techniques sont différents.

- La première étape consiste à caractériser formellement les configurations possibles donnant lieu à un système correct de réponse aux requêtes. Cette étape demande d'adapter des techniques standard d'analyse du graphe de dépendance des règles.
- La deuxième étape consiste à établir une fonction de coût permettant de comparer les différentes configurations (par exemple, cette fonction pourra essayer de minimiser le coût de stockage des données et inférences). Cela donne lieu à un espace de solutions qu'il est possible d'explorer avec un algorithme en style A*, mais dont il sera nécessaire néanmoins de définir des heuristiques et optimisations permettant de réduire le nombre d'états visités.
- La troisième étape prévoit l'implémentation de l'approche dans le logiciel Graal développé par l'équipe GraphIK du LIRMM, ainsi que l'évaluation expérimentale des approches résultantes.