

Chapitre 6

De la construction d'entrepôts de données à l'extraction de connaissances sur grilles

L'évolution technologique en matière de moyens d'acquisition, de stockage et de puissance de calcul des machines a conduit à la prolifération de grandes masses de données dans différents domaines. Très souvent, ces données sont hétérogènes provenant de différentes sources géographiquement réparties, et disséminent des connaissances insoupçonnées. Les attentes d'interprétation, d'analyse, de recoupement de ces données représentent aujourd'hui, et plus encore dans le futur, un défi majeur.

Fournissant une infrastructure cohérente alliant puissance de calcul, dynamique, sécurité et capacité de stockage et de partage de gros volumes de données, les grilles informatiques apparaissent comme des candidates naturelles pour relever un tel défi. Néanmoins, l'exploitation d'une telle infrastructure de manière efficace n'est possible que si un certain nombre de verrous scientifiques et technologiques sont levés : organisation et stockage des données, optimisation des accès aux données, sécurité, etc. Ces verrous constituent autant de problématiques suscitant de nombreuses activités de recherche soutenues par les Actions Concertées Incitatives (ACI) du Ministère de la Recherche telles que *Masse de données*, *IMPBio*, *Grid*, *Sécurité*, etc.

Le but de ce chapitre est de traiter certaines de ces problématiques notamment celles portant sur la construction d'entrepôts de données distribués, l'optimisation des accès aux données et l'extraction de connaissances dans le contexte des grilles informatiques. Ces problématiques ont fait l'objet de trois articles ayant été sélectionnés dans le cadre du premier atelier sur

l'« Extraction et Gestion Parallèles Distribuées des Connaissances », organisé conjointement avec la conférence « Extraction et Gestion des Connaissances » (EGC2005).

- **Pascal Wehrle, Maryvonne Miquel, Anne Tchounikine. Entrepôts de données sur grilles de calcul.** La technologie des entrepôts de données et les outils OLAP permettent de recueillir et organiser de grandes masses de données et de naviguer sur des vues matérialisées au moyen d'opérateurs adaptés (opérateurs OLAP). De tradition, les entrepôts de données ont pour vocation la centralisation d'un ensemble de données multi-source, préalablement nettoyées et homogénéisées puis mises sous un schéma commun, le schéma de l'entrepôt. Cet article propose un modèle d'architecture distribuée pour la construction d'entrepôts de données sur grilles informatiques. Y sont abordés les problèmes de répartition des données, de leur indexation et leur échange entre nœuds de la grille.
- **Yonny Cardenas, Jean-Marc Mierzon, Lionel Brunie. Service de Cache pour les Grilles de Calcul.** L'accès aux données d'un entrepôt se traduit par l'exécution de requêtes, parfois complexes, sur des bases de données distribuées sur la grille. Les données manipulées sont toujours associées à des méta-données décrivant, entre autres, le contexte d'acquisition et l'interprétation de ces données. Cependant, la gestion de caches n'est que peu abordée en tenant compte de la sémantique des données, et donc des méta-données associées. Or, l'accès aux données n'est efficace que si l'accès et la gestion complète des méta-données sont prises en compte. Cet article présente une architecture de cache collaboratif sur grilles basée sur l'exploitation de méta-données. Celle-ci est constituée d'un ensemble de caches répartis et d'une couche de coopération permettant leur collaboration. Cette architecture est très intéressante pour les applications nécessitant un fort mouvement de données, telles que les applications d'extraction de connaissances.
- **Sébastien Cahon, Nordine Melab, El-Ghazali Talbi. Sélection d'attributs en fouille de données sur grilles.** Le développement d'applications, notamment d'extraction de connaissances, sur grilles informatiques nécessite de re-penser les algorithmes existants pour en assurer le passage à l'échelle. La complexité des grilles rend cet exercice souvent difficile. Ce qui a conduit à l'émergence d'intergiciels ou *middlewares* permettant de simplifier cette tâche en rendant l'accès au parallélisme et à la distribution transparents à l'utilisateur. Le but de cet article est de décrire une plate-forme logicielle ou *framework open source* d'aide au développement de méthodes d'optimisation approchées ou *méta-heuristiques* parallèles distribuées sur grilles. Cette plate-forme est appliquée à un problème d'extraction de connaissances à grande échelle dans le domaine de la spectroscopie proche infrarouge. Il s'agit de découvrir en utilisant une approche enveloppante des modèles de prédiction à partir d'un échantillon de spectres. L'application a été expérimentée sur une grille de 122 machines. Les résultats obtenus

montrent l'efficacité de l'approche utilisée tant en terme de précision des modèles extraits qu'en terme de performance à l'exécution.

Comité de programme et d'organisation :

Nordine Melab (LIFL - CNRS UMR 8022 - melab, @lifl.fr)

El-Ghazali Talbi (LIFL - CNRS UMR 8022 - talbi@lifl.fr)