

Titre : Gestion de données complexes pour la modélisation de niche écologique

Le phénomène du « Big Data » est de plus en plus perçu comme l'un des grands défis informatiques de la décennie en cours. De nombreux domaines font face à un 'déluge' de données sans précédent. La quantité des données produites augmente constamment et rend leur traitement de plus en plus difficile à gérer avec les outils actuels.

Par exemple Facebook enregistre plus de 800 millions d'utilisateurs actifs en moyenne par jour et ajoute quotidiennement plus de 500 TB de données. La base de données astrométrique européenne est complétée avec plus de 100 Go de données de nouvelles observations par jour (CNES : GAIA). En outre, selon une étude de Industrial Development Corporation (IDC) et EMC Corporation, la quantité de données générées en 2020 sera 44 fois supérieures par rapport à 2009 à 35 ZB par an.

L'accès, l'interrogation et l'analyse de ces nouvelles masses de données sont essentiels pour élargir les connaissances du domaine y afférent et font partie des défis majeurs du Big Data. Ceci est particulièrement crucial dans des domaines tels que les médias sociaux, la génomique, la climatologie, les réseaux énergétiques complexes, l'astronomie, l'écologie et la biodiversité.

La gestion de ces masses de données nécessite l'utilisation de nouveaux systèmes de gestion de données disposant de méthodes d'accès performantes et permettant d'exprimer des requêtes avec des langages de haut niveau. Un aspect particulièrement difficile à prendre en compte est le comportement très versatile des utilisateurs. Cela génère des demandes très fluctuantes : la charge est variable en nombre de requêtes, certaines données sont plus populaires (fréquemment demandées) que d'autres, la popularité est elle-même fluctuante (une donnée n'est populaire que pendant une durée limitée). Or les solutions actuelles n'ont pas été conçues pour s'adapter dynamiquement à ce type de situation.

Cette thèse se déroule dans le contexte du GBIF, initiative visant à fédérer et partager les données de biodiversité produites par de nombreux fournisseurs à l'échelle mondiale. Le GBIF propose actuellement des services pour interroger les données et les visualiser. Toutefois, avec un nombre croissant de fournisseurs qui ajoutent de nouvelles données et d'utilisateurs qui expriment de nouveaux besoins d'interrogation, l'accès aux données du GBIF pose un double problème d'expressivité et d'efficacité difficile à résoudre.

L'objectif principal de cette thèse, est de concevoir une solution qui offre un accès expressif et efficace à une très grande base de données, lorsque le nombre d'utilisateurs devient très grand. Dans cette thèse nous proposons une solution générale qui tient compte des spécificités réelles du cas d'usage du GBIF. Les résultats sont les suivants :

- 1) La conception d'une architecture décentralisée qui mutualise un grand nombre de ressources de stockage et de traitement de données. Partant du constat que les ressources disponibles sont très hétérogènes (disparité de leur capacité de stockage, traitement et communication) et réparties à large échelle. L'architecture est conçue pour agréger simplement et dynamiquement des nouvelles ressources, coordonner le stockage et l'interrogation entre les ressources. L'architecture proposée peut être couplée de manière non intrusive avec des sources de données existantes, tout en permettant de manipuler les données indépendamment des sources.

2) La définition d'une solution de répartition dynamique des données à la demande. Les données sont fragmentées en fonction des requêtes. Les prédicats pour la fragmentation portent sur les dimensions (attributs) hiérarchiques des données. Si nécessaire les fragments peuvent être répliqués. Le placement des fragments et leur éventuelle réplication s'adapte dynamiquement à la charge des requêtes tout en tenant compte de la taille maximale des ressources de stockage disponibles, à l'aide d'une fonction de coût.

3) Un modèle pour l'exécution répartie des requêtes dans cette architecture. Une requête est traitée en parallèle lorsque les données sont réparties sur plusieurs machines. L'exécution est décentralisée de telle sorte que toute machine peut recevoir les requêtes de plusieurs utilisateurs et gérer leur exécution. La réplication dynamique des données est contrôlée globalement.

4) Une méthode d'optimisation de requête qui considère plusieurs paramètres susceptibles d'impacter le coût de la requête : la fluctuation et la disparité des charges, la disparité des capacités de calcul et des liens de communication, la localisation des données impliquées, le schéma de fragmentation des données et les prédicats des requêtes. Ayant comme objectif de traiter la plupart des requêtes dans une durée impartie (contrainte de temps de réponse borné), la solution ajuste dynamiquement le schéma de placement et de réplication des données. Les fragments les plus sollicités sont détectés et sont répliqués afin de répartir les demandes sur plusieurs machines. Le choix de l'emplacement d'un fragment tend à favoriser la localité des requêtes en regroupant les fragments lus par une requête.

5) Une mise en œuvre de la solution proposée et sa validation expérimentale dans le cas réel de l'analyse de la biodiversité avec les données du GBIF et des requêtes d'experts du domaine. Les résultats obtenus, en utilisant un cluster de 200 cœurs, montrent la faisabilité de la solution décentralisée non-intrusive pour le traitement réparti de requêtes complexes. Ils montrent le bénéfice d'adapter dynamiquement le schéma de placement des fragments en fonction de leur popularité et de la charge des sites. Les performances de notre solution sont satisfaisantes : notre approche s'avère efficace pour garantir un temps de réponse borné et un débit acceptable. Les résultats de l'évaluation du passage à l'échelle sont prometteurs : le débit supporté croît presque linéairement avec le nombre de sites qui composent le système.