Une approche holistique combinant flux temps-réel et données archivées pour la gestion et le traitement d'objets mobiles

Loic Salmon, Cyril Ray, Christophe Claramunt

Une approche holistique combinant flux temps-réel et données archivées pour la gestion et le traitement d'objets mobiles

Loic Salmon *
Institut de Recherche de l'Ecole Navale
29240 BREST Cedex 9 - FRANCE
loic.salmon@ecole-navale.fr

Cyril Ray †
Institut de Recherche de l'Ecole Navale
29240 BREST Cedex 9 - FRANCE
cyril.ray@ecole-navale.fr

Christophe Claramunt
Institut de Recherche de l'Ecole Navale
29240 BREST Cedex 9 - FRANCE
christophe.claramunt@ecole-navale.fr

ABSTRACT

La numérisation de nos espaces de vie et de mobilité s'est largement accentuée durant la dernière décennie. La multiplication des capteurs de toute nature permettant de percevoir et de mesurer notre espace physique en est le levier principal. L'ensemble de ces systèmes produit aujourd'hui de grands volumes de données hétérogènes sans cesse croissants ce qui soulève de nombreux enjeux scientifiques et d'ingénierie en termes de stockage et de traitement pour la gestion et l'analyse de mobilités. Les travaux dans le domaine d'analyse des données spatio-temporelles ont largement été orientés soit vers la fouille de données historiques archivées, soit vers le traitement continu. Afin d'éviter les écueils de plus en plus prégnants dûs à l'augmentation des volumes de données et de leur vélocité (temps de traitement trop long, modèles conceptuellement plus adaptés, analyse des données approximative), nous proposons la conception d'une approche hybride distribuée permettant le traitement combiné de flux temps-réel et de données archivées. L'objectif de cette thèse est donc de développer un système nouveau de gestion et de traitement distribué pour l'analyse des mobilités maritimes.

Keywords

Base de données spatio-temporelles, objets mobiles, traitement temps-réel, système distribué

1. INTRODUCTION

L'analyse de mobilités intervient dans de nombreux domaines tels que l'aménagement urbain, la surveillance du trafic, la climatologie, l'étude des phénomènes sociaux ou la zoologie. L'émergence et la multiplication de systèmes mobiles et des capteurs véhiculant des informations provoquent une explosion du volume de données spatiales et temporelles. Ce gisement de données qui n'a évidemment pas encore atteint sa pleine mesure devient de plus en plus difficile à traiter et soulève de nombreux enjeux scientifiques et d'ingénierie en termes de stockage et de traitement des objets mobiles.

L'analyse de mobilités est un domaine spécifique qui met en difficulté les systèmes de bases de données relationnelles dans la mesure où les objets mobiles reportent leur position en continu (**V**élocité) ce qui produit rapidement une masse de données conséquente (**V**olume) ce qui nécessitera de mettre en place une solution dite *Big Data*. Enfin, bien que moins déterminant par rapport aux deux facteurs précédents, il faut prendre en compte le fait que les objets mobiles peuvent être de toute sortes : points, polylignes, surfaces dont la taille et la forme peuvent fortement varier (**V**ariété) dans l'espace et le temps nécessitant l'usage d'index particuliers.

Aux trois "V" traditionnels s'ajoutent d'autres problèmes plus spécifiques concernant les données spatio-temporelles. Une distribution équilibrée des données sur l'ensemble des nœuds du système par rapport à leur couverture spatiale, spatio-temporelle ou sémantique est plus difficile à mettre en œuvre car les phénomènes et déplacements observés se répartissent dans l'espace et le temps à différents niveaux de densité. Ceci a également une incidence sur l'échelle de représentation choisie et le volume de données manipulées. En effet, si on restreint trop le volume temporel ou spatial de données à analyser l'information extraite peut être biaisée ou erronée. A contrario s'il est trop grand, l'information obtenue peut être lissée et peu représentative car certaines particularités locales (spatiales, temporelles ou spatio-temporelles), auront affecté les résultats observés, sans avoir été détectées. Enfin, les traitements et opérateurs spatiaux font interagir des objets de nature et de taille différentes ce qui peut faire intervenir de nombreuses jointures et des calculs plus complexes que pour des données usuelles (opérateurs topologiques, comparaison de trajectoires ...).

2. TRAITEMENT DISTRIBUÉ DE DONNÉES SPATIO-TEMPORELLES

L'objectif de cette thèse sera donc de développer un sys-

*L. Salmon, corresponding author
†

BDA 14 octobre 2014, Grenoble-Autrans, France.

tème nouveau de traitement distribué et parallélisé, tenant compte de ces spécificités, afin de favoriser la gestion et le traitement de données spatio-temporelles dans un contexte *Big Data*.

2.1 Traitement on-line vs. off-line

Les travaux dans le domaine de l'analyses de mobilité ont largement été orientés soit vers la fouille de données historiques archivées, soit vers le traitement temps-réel.

La fouille de données historiques ou traitement *off-line* se caractérise par le stockage de la totalité de l'historique des mouvements des entités mobiles pour pouvoir étudier à posteriori les phénomènes du passé et éventuellement inférer le comportement futur d'un objet donné. Au vu des forts volumes de données à manipuler, le temps de réponse est important et certains mécanismes sont nécessaires pour accéder plus vite aux données (index, partitionnement) empêchant des mises à jour en continu. Les techniques actuelles de collecte, de stockage et d'interrogation des mobilités sont issues des travaux sur les bases de données pour objets mobiles (Moving Object Database ; MOD) [3]. Ces dernières sont presque exclusivement basées sur un modèle relationnel et intègrent ou exploitent des extensions pour la gestion de ces mobiles (types et opérateurs spatiaux, notion de temps intégrée, index associés aux objets mobiles) comme Hermes [9] ou Secondo [2]. Ces données d'objets mobiles stockées et archivées peuvent être exploitées à l'aide de différentes techniques de fouille de données : extraction, agrégation, *clustering*, fusion et permettre notamment l'identification de comportements type et d'anomalies. Seulement ces techniques de fouilles nécessitent la distribution des données et des traitements lorsque le volume de données augmente considérablement [5].

Le traitement temps-réel ou approche *on-line* s'intéresse au maintien continu des informations sur la position actuelle de l'entité pour pouvoir détecter des événements se produisant en temps-réel et éventuellement prédire une future position proche. Divers travaux ont été réalisés concernant ce type de traitement qui se caractérise par un temps de réponse rapide car effectué en mémoire. Par exemple, dans [8] les auteurs tentent de répondre à la problématique d'analyse de mobilité temps-réel en étendant un système de gestion des flux temps-réel au contexte spatio-temporel. Cependant cette approche peut fournir une réponse de moins bonne qualité à cause du traitement mémoire imposant de supprimer des données, de faire de l'échantillonnage, d'utiliser des fenêtres temporelles ou d'agréger certaines données et résultats intermédiaires par un traitement incrémental des flux [4]. L'analyse se fait alors en même temps que l'objet mobile évolue et les requêtes sur les données ne s'exécutent plus une seule fois comme en *off-line* mais en continu au gré du flux de données entrant [7].

L'évaluation des requêtes est un compromis entre temps d'exécution et précision ou qualité de la réponse. L'approche base de données historiques a donc pour précepte de préférer la qualité au temps de calcul et inversement en ce qui concerne les systèmes temps-réels.

2.2 Proposition d'une architecture hybride

Afin d'éviter les écueils de plus en plus prégnants dûs à l'augmentation des volumes de données et de leur vélocité (temps de traitement trop long, modèles conceptuellement plus adaptés, analyse des données approximative), nous proposons une approche hybride distribuée permettant le traitement combiné de flux temps-réel et de données archivées qui permettra de fournir une réponse satisfaisante en un temps acceptable (Figure 1).

Cette architecture est inspirée de l'approche hybride non distribuée de [1] dans laquelle trois types de requêtes sont distinguées : celles portant sur les données archivées, celles portant sur les données reçues en temps-réel et enfin les requêtes dites "hybrides" nécessitant de combiner les données arrivant en temps-réel et les informations extraites des données historiques. Plus récemment, Nathan Marz propose avec son architecture lambda un système de gestion de données prenant en compte aussi bien les aspects vélocité, volumétrie que la contrainte de faible latence [6]. L'architecture se compose de trois couches, une couche qui correspond aux données archivées dans une base de données NOSQL et pré-calcule des vues relatives à des requêtes souvent posées, une couche qui correspond au traitement temps-réel et une couche intermédiaire qui permet de fusionner facilement les résultats obtenus des deux couches précédentes.

Dans notre système, les reports de position s'effectuent via différents flux de données qui seront gérés sur un système temps-réel distribué. Au niveau de la gestion des données, on distingue le composant relatif au traitement *off-line* et celui relatif au traitement *on-line*.

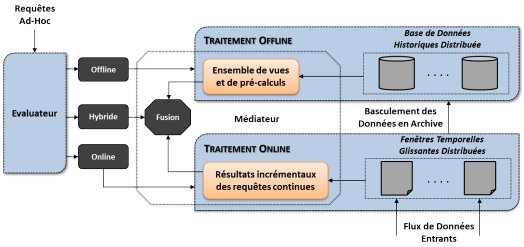

Figure 1: Principe architectural

La gestion des traitements en mémoire est faite sur une fenêtre glissante distribuée dont la taille pourra être modifiée selon le nombre de données collectées en temps-réel sur la zone de couverture concernée. Des vues *on-line* sur les requêtes continues sont mises à jour et incrémentées au gré du flux entrant de données. Si l'utilisateur exprime une requête portant sur des données n'étant pas synthétisées par le traitement continu, les données nécessaires sont accessibles via la fenêtre glissante. Une fois, que la période temporelle dédiée à la fenêtre glissante est dépassée, les données sont déplacées vers la base de données historiques distribuée pour effectuer les traitements *off-line*. Afin d'avoir un système réactif, des pré-calculs sont effectués sur les données historiques et mis à jour au fur et à mesure des arrivées en base de données.

Au niveau des requêtes deux entités sont utilisées pour identifier les données à extraire et traiter, ainsi que pour gérer les interactions entre la base de données historiques et le système de traitement temps-réel. Une de ces entités est le médiateur dont le rôle est de gérer les flux entre les composants *on-line* et *off-line*, de conserver et stocker les vues associées et de pouvoir les fusionner pour permettre de répondre aux requêtes hybrides. L'évaluateur analyse la

requête en entrée et essaie d'inférer le type de la requête, à savoir *on-line*, *off-line* ou hybride pour orienter, en fonction du type de requête identifiée, la récupération des données et des informations nécessaires dans notre architecture. Il transmet au médiateur les données désirées à traiter et ce dernier se charge de prendre, combiner ou d'effectuer des traitements sur la fenêtre temporelle glissante ou l'archive suivant la demande de l'évaluateur.

3. CONCLUSIONS

L'objectif principal de ce travail concerne la mise en place d'une architecture hybride pour la gestion et le traitement d'objets mobiles. Nous nous concentrerons en premier lieu sur la gestion des mécanismes de médiation ainsi que la distribution des données et des traitements. Le cas d'application de cette thèse, débutée en novembre 2013 (encadrée par Cyril Ray et dirigée par Christophe Claramunt), sera l'étude des positions et trajectoires de navires issues du système de positionnement AIS (Automatic Identification System). Le but final étant de traiter, stocker et analyser les positions de navires qui permettront d'obtenir des vues analytiques (multidimensionnelles) du trafic maritime et l'identification de comportements types (eg. trajectoire anormale) en temps-réel.

4. REFERENCES

[1] S. Chandrasekaran and M. Franklin. Remembrance of streams past : Overload-sensitive management of archived streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, VLDB '04, pages 348–359, 2004.

[2] V. T. de Almeida, R. H. Guting, and T. Behr. Querying moving objects in secondo. In *Proceedings of the 7th International Conference on Mobile Data Management*, MDM '06, pages 47–52. IEEE Computer Society, 2006.

[3] L. Forlizzi, R. H. Güting, E. Nardelli, and M. Schneider. A data model and data structures for moving objects databases. pages 319–330, 1999.

[4] L. Golab and M. T. Özsu. Issues in data stream management. *SIGMOD Rec.*, pages 5–14, 2003.

[5] Q. Ma, B. Y. 0002, W. Qian, and A. Zhou. Query processing of massive trajectory data based on mapreduce. In X. Meng, H. Wang, and Y. Chen, editors, *CloudDb*, pages 9–16. ACM, 2009.

[6] N. Marz. *Big data : principles and best practices of scalable realtime data systems*. O'Reilly Media, [S.l.], 2013.

[7] M. F. Mokbel, X. Xiong, M. A. Hammad, and W. G. Aref. Continuous query processing of spatio-temporal data streams in place. *Geoinformatica*, pages 343–365, 2005.

[8] K. Patroumpas. Multi-scale window specification over streaming trajectories. *J. Spatial Information Science*, pages 45–75, 2013.

[9] N. Pelekis, Y. Theodoridis, S. Vosinakis, and T. Panayiotopoulos. Hermes - a framework for location-based data management. In *In Proceedings of EDBT*, pages 1130–1134, 2006.