A hybrid approach combining real-time and archived data for mobility analysis

Loic Salmon, Cyril Ray, Christophe Claramunt

A hybrid approach combining real-time and archived data for mobility analysis

Loic Salmon
Naval Academy Research Lab
29240 Brest Cedex 9, France
loic.salmon@ecole-navale.fr

Cyril Ray
Naval Academy Research Lab
29240 Brest Cedex 9, France
cyril.ray@ecole-navale.fr

Christophe Claramunt
Naval Academy Research Lab
29240 Brest Cedex 9, France
christophe.claramunt@ecole-navale.fr

ABSTRACT

Mobility analysis is involved in many areas such as urban planning, traffic monitoring, climatology, study of social and animal phenomena to mention a few examples. The emergence and proliferation of mobile and sensor-based systems generate a significant increase of spatial and temporal data in terms of volume and frequency of update. In particular, the storage, management and analysis of the large data sets generated become a non straightforward task. Current works related to the manipulation of mobility data have been directed towards either mining archived historical data or continuous processing of incoming data streams. Our research introduces a hybrid approach whose objective is to provide a combined processing of real-time data streams and archived data. The principles of our approach is to promote the distributed and parallelized processing of mobility data. The whole framework is currently applied to the real-time monitoring of maritime traffic.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications - Spatial databases and GIS; H.2.4 [**Database Management**]: Systems - *Query processing*, *Distributed databases*

General Terms

Theory, Design

Keywords

Moving object Database, Geostreaming, Maritime monitoring

1. INTRODUCTION

Over the past few years, the proliferation of sensors and devices recording positioning information regularly produces very large volumes of heterogeneous data. This leads to many research challenges as the storage, distribution, management, processing and analysis of the large mobility data repository generated that is far from being supported by most of current spatial databases.

Mobility analysis is a generic data manipulation functionality that relates to many scientific and application domains such as traffic monitoring, environmental and ecological modeling, urban planning, sociology and robotics. Often, one of the main challenges when analyzing mobility data is to infer patterns and outliers in the datasets generated, in order to infer some novel knowledge regarding the phenomena studied. For instance, in the context of traffic monitoring, regular behaviors and abnormal trajectories can be inferred [22]. Similarly, in sociology, emerging human mobility patterns can be inferred and studied, and this according to different social communities [11].

However, most current systems do not completely provide an efficient storage and analysis of mobility data [25]. Indeed, not only moving objects continuously recorded can produce huge amounts of trajectory data, but they also need to be continuously updated.

The emergence of new systems and paradigms to deal with huge amount of data such as map-reduce [6] provide some promising solutions but still do not completely deal with moving objects [8]. This motivate a few recent works such as the Map-reduce system specifically oriented to the spatial domain [2] and particularly moving objects [16]. But still these systems are oriented towards "a posteriori analysis" and can lack of efficiency to process data "on the fly". In most of current real-time application contexts, a successful processing system for mobility analysis have to be reactive enough and allow for anomalies detection in real-time, requiring to combine results extracted from historical data with incoming data.

The remainder of this paper is organized as follows. Section 2 provides an overview and discusses existing works oriented towards offline and online processing of mobility data. Section 3 gives a general presentation of our hybrid approach. In Section 4 we developed the principles of the online processing approach which is the focus of our hybrid-based proposal. Finally Section 5 summarizes and concludes this paper.

2. ONLINE AND OFFLINE APPROACHES TO HANDLE MOVING OBJECTS

Recent research works in the field of mobility analysis have been mainly oriented towards either an offline approach which stores all the data and process it on demand, or an online approach whose goal is to track and predict the trajectories of moving objects.

The mining or offline processing of historical data is characterized by a complete storage of the history of mobility data, while data is manipulated to generally retrospectively study and predict the next stages of the phenomena represented by such mobility data. In many application domains, and due to the large dataset volumes generated, response-time to any query or analysis is of crucial importance. Indeed, there is a need for some manipulation mechanisms (e.g., data structures, partitioning) to provide efficient access to the data, and in order to prevent continuous updates. Most of current works oriented to the manipulation of mobility data came from the moving object database domain [9]. Extended relational or object-oriented approaches have already integrated specialized data representation and manipulation extensions (e.g., complex data types, operators) to deal with moving objects [23], [5]. Usual database functions can be then applied to moving objects such as data mining techniques: extraction of outliers, aggregation, clustering [12]. Such manipulation functions allow for an identification of typical behaviors or outliers [22]. Moreover, it appears that mining techniques require the distribution of data and processing when the volume of data increases considerably [17]. While offline approaches can provide valuable solutions in many cases, they do not support specific real-time functionalities and for instance how to react to some specific events.

One of the main principles of the real-time or online processing of mobility data is to continuously track moving data, and to detect and predict some typical behaviors. Previous works have been oriented to the search for better response times at the memory level. For example, in [20], the authors address the problem of real-time analysis of mobility data by extending a data stream management system (DSMS) for handling spatio-temporal data. However this approach is still constrained by memory-based processing that implies to either sample or aggregate the data using thematic, spatial or temporal criteria, or incrementally process the data [14]. Another difficult aspect of the management of mobility data is that a given analysis should be performed while the considered moving object may change its location in the upcoming stream. Meanwhile, and as some continuous queries are for example processed, these queries should be re-evaluated continuously as well [19]. Some recent works such as Storm and S4 [21] suggest some distributed processing approach performed at multiple nodes, while some of those works have been extended to deal with mobility analysis [28], [15]. Although these systems distribute the process they do not appear to deal with query optimization and approximation issues. The limitations of an online approach is that it often leads to unsatisfactory situations. Indeed, even if the system is able to respond in sufficient time to a given query, accuracy is not guaranteed because of either limited amount of data considered, or alteration of the location of the moving object.

The problems and limitations mentioned above motivate our search for a hybrid solution whose objective will be to give query responses in sufficient time while maintaining appropriate accuracy and appropriateness regarding the dynamism of the system.

3. TOWARDS A HYBRID APPROACH

Let us introduce the main motivations and principles of our architecture inspired by a previous work that suggested a non distributed hybrid approach [4] in which three types of queries are distinguished: those on archived data, those related to the data received in real-time, and those so-called "hybrid" queries that require to combine real-time data and query results extracted from archived data.

More recently, Nathan Marz with his "Lambda Architecture" proposes a data management system taking into account both velocity and volume aspects with the constraint of low latency [18]. This architecture consists of three layers, a layer which corresponds to the data stored in a NOSQL database and pre-computed views related to frequently asked queries, a layer corresponding to real-time processing, and an intermediate layer that allows to merge the results of the previous two layers.

Existing hybrid systems can be classified as follows: DBMS-based systems, map-reduce-based systems and DSMS-based systems [13]. DBMS-based systems deal with high level queries (SQL-like) and encompass query planning optimization while map-reduce-based systems scale-out with fault tolerance and process huge amount of data. However, neither DBMS or map-reduce systems are real-time systems. DSMS-based systems is the only kind of system that supports real-time requirements. It enables real-time processing with a context given by the analysis of historical data, but still such systems are non easy to implement as they have to take into account the processing of large incoming real-time data with archived data. This is the kind of system retained by our study and that will be illustrated by an application to maritime traffic monitoring.

Maritime transportation is a domain of increasingly intense traffic (Figure 1). The monitoring and analysis of mobilities at sea is therefore crucial for safety and security reasons. For instance, these analysis of mobility and behavior should be designed to detect illegal or criminal activities, risks at sea (flow of illicit products, illegal immigration, overfishing, pollution by hazardous materials, piracy, accidents , etc.), and more generally any violation to regulations. Traffic monitoring is nowadays largely based on the continuous identification of vessel positions and trajectories and some additional functionalities such as pattern and abnormal behavior detections [22].

Practically, ships are fitted out with almost real-time position report systems whose objective is to identify and locate vessels at distance (Automatic Identification System (AIS) for example [1]). The multiplication of vessel positioning systems such as AIS but also Satellite AIS, Vessel Monitoring System (VMS) or Long Range Identification System (LRIT) contributes to the real-time availability of large traffic data at sea. The large datasets generated become difficult to manage and analyse due to the heterogeneity, large volumes and real-time components of the large datasets gen-

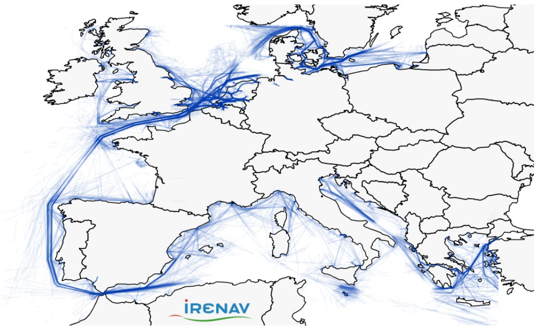

Figure 1: Ships' trajectories, density map in Europe during one month (AIS positions, December 2010)

erated. Detecting trends and abnormal behaviors at sea still require such large-scale continuous collection of vessel positions and the development of specific spatio-temporal analysis and knowledge extraction methods [27].

Amongst the requirements, not only such a system should be able to manipulate fresh and historical data separately and jointly but also appropriate functionalities to respond to any queries in almost real-time. This clearly motivates and supports our search for a hybrid system that needs to be stream-oriented (or DSMS-based), to respond in few seconds to minutes and not hours to any queries in order to allow a maritime agent to act immediately when any problem happens.

We thus propose a DSMS-based system for vessels mobility and traffic. The next section introduces the main principles of the proposed approach.

4. A DSMS-BASED APPROACH

In the system to develop, vessel location records are made via various data streams and should be managed on a real-time distributed system. The system must support real-time data analysis as incoming data flows. The following section introduces the main challenges and principles retained for the design of such a system.

4.1 Hybrid processing of moving object: goals and expectations

In order to manage and monitor maritime traffic, our system must respect some principles and requirements. Those principles raise some issues to still consider for processing of mobility data in real-time:

- *Using views for real-time analysis.* The system should give an answer within a short time period compared to a purely offline system and of better accuracy than a purely online one. However, a measure of the quality of response must be done to evaluate if the system can support execution of complicated queries and perform trajectory analysis in near real-time. To deal with real-time requirements such system will store online and offline views in main memory. The system should be able to take, translate and merge them with incoming data. Views should be the cornerstone of our architecture to share processing and results and then reduce the response time.

- *An adaptive system.* This system should be reactive and adapt itself gradually as the data and queries are received to be as efficient as possible and perform incoming queries. The framework must monitor the system performances and modify itself. For example, changing allocation of processes to the different nodes of the architecture to best fit with the queries already running in the system. Data that are no longer useful must be aggregated or transferred to the offline part to reduce the amount of data to process.

- *Query understanding to deal with mobility.* This system should handle a large variety of queries, this necessitates algorithms to decompose queries, analyze similarity between queries, use previous results of queries or views and merge it with incoming data. There is still a need to analyze and study the different kinds of queries related to moving objects and find similarities and common processing over them.

- *An autonomous and reactive processing system.* This system should handle processing and maintenance of views in accordance to queries frequently formulated by the agents. It should detect the emergence of new events and advise agents of those modifications. It might also reduce the intervention of human agents and process himself data and provide results only if necessary.

4.2 Hybrid processing of moving object: principles

Requirements and goals of our design having been formulated in the previous section. Let us describe the principles of our hybrid architecture whose objective is to handle moving objects.

The hybrid system suggested is structured with two main components: One relates to the offline processing while the other one is responsible for the online part (cf. Figure 2). Both components can run independently in pure offline or online way storing and processing data incoming in the system, but the goal of such a system is to exploit the advantages of both approaches in order to answer hybrid queries.

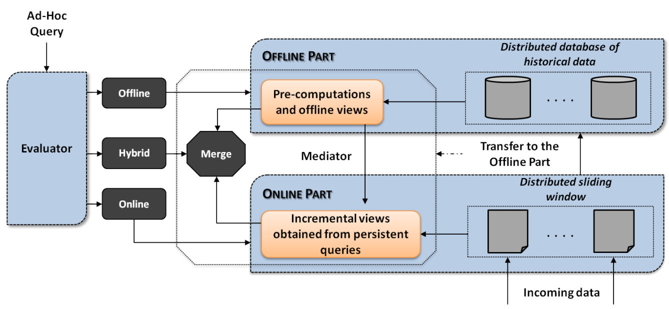

Figure 2: Architectural principles

Online processing is performed on a distributed sliding window whose size can be changed according to the amount of data collected in real-time on the respective coverage area. Online views on continuous queries are updated and incremented while incoming data stream is processed. When a user gives query for which no synthesized summary exists, necessary data to handle the query are accessible via the sliding window. When the temporal interval of the sliding window is exceeded, the data is transferred to the historical database to perform distributed processing offline. In order to have a reactive system, summaries are performed on historical data and updated upon arrival in the database and then transferred to online part to provide real-time answers. Both online and offline parts use a distributed processing schema that still needs to be defined to take advantage of the spatial and temporal distribution of positioning records.

Furthermore, two entities have the role (cf. Figure 2) of identifying the data to extract and process, and to manage the interactions between the historical database and real-time processing system. In other words, these are the major components of our hybrid system which allows us to merge the online and offline parts and to answer the query using the minimum data as possible in accordance to user's requirements.

The role of the Mediator is to manage the flows between components online and offline, preserve and store the associated views and to merge them to answer hybrid queries. The evaluator analyzes the input query and tries to infer the type of request, (i.e. online, offline or hybrid) to guide, based on the identified type of query, recovery of data and information needed in our architecture. It transmits the desired data to the mediator to deal with, then the mediator is responsible to answer the query taking, combining or performing processing on the sliding time window or on the archive following the query type.

4.3 Hybrid processing of moving object : Challenges involved in the online part

Our hybrid approach must be stream-oriented to deal with real-time requirements. Let us examine the challenges and difficulties involved in the design of a DSMS-based approach and those directly inherited from the DSMS part.

In a DSMS, a query consists of stream operators organized in a directed acyclic graph (DAG) or workflow [3]. Query operators are connected via queues and each stream operator has an in-memory state composed by the tuples needed to perform the operation whose is responsible for. The DAG uses a pipeline paradigm where each result produced by an operator is transmitted to the following operator(s) in the workflow. All data are processed in main memory and some tuples or results cached are shared between the different queries which run simultaneously.

Handling moving object in this context raises some additional issues that should be taken into account to design a distributed moving object stream management system. Let us sum up the challenges associated to the design of a distributed DSMS-based system for mobility analysis.

- *Distribution and optimization of the queries.* This system can be modeled as a dataflow where incoming data flows and queries should be re-evaluated accordingly [7]. Therefore, such system should create and delete novel processes when necessary to stay efficient as the system evolves. Online systems must also process multiple queries simultaneously, then incoming data associated to different queries must be brought and processed together in order to reduce number of operations. Finally this system must re-order the operators that process data and choose the best efficient query execution plan.

- *Evaluation and approximation.* An important problem to deal with in online systems is that data are processed in memory only. Therefore, the amount of data manipulated in memory should be reduced when processing the re-sampling, compressing load-shedding... [14]. Another approach is to design specific algorithms that give approximate solutions by a data evaluation performed at regular times. Data summaries or data views can be also used as a sort of incremental processing paradigm [10].

- *Distributed schema processing.* The two previous challenges must be extended in the context of a distributed system. Such distribution brings common additional issues such as fault tolerance, scalability and data distribution. In the context of our approach view placement, dynamic allocation of the operators, data transfer from one node to another must be also taken into account [24]. This implies to distribute both data and processing along the nodes of our architecture. Distribution can be done according to spatial coverage, temporal extent or by taking into account moving object movement patterns [26].

- *A hybrid stream-based approach.* In the context of hybrid processing stream-oriented, the online part must be the cornerstone of this system that can be combined with offline results. Therefore, this system must be able to merge query results of online and offline parts as well as making a difference between the online and offline parts of a hybrid query.

4.4 Hybrid processing of moving object : A stream processing approach

Considering both challenges for handling moving objects in a DSMS-based approach 4.3 and requirements of our system 4.1 to deal with operational context, let us propose the following architecture (cf. Figure 3) which is currently under development and falls into the conception of a stream-oriented hybrid system.

There are two different inputs, one related to the system which collects the data and processes it as it incomes, and one related to the user which formulates and adds his query to the pool of persistent queries or wants an answer to a hybrid query. The link between these two components is made by the Executor which is the distributed data flows processing engine whose role is to join moving objects with queries.

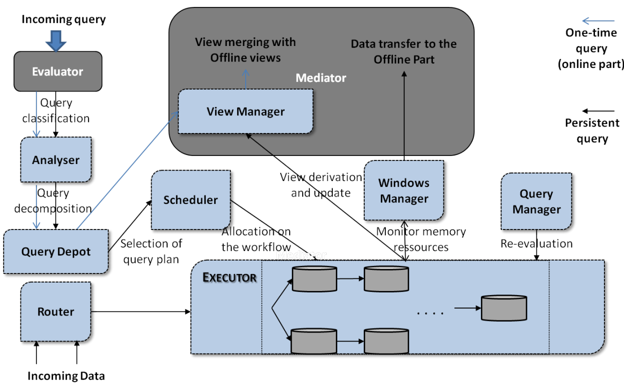

Figure 3: Online architectural principles

Incoming streams are received, pre-processed and supplied to the stream operators by the Router while new queries are analyzed and translated in the form of workflow (or query plan) by the Analyzer. As a new query is specified in the system, there are two different options following the result given by the Evaluator on the hybrid part. When the incoming query is persistent, then it is added to the system to run continuously whereas if it is a one-time query it will be executed once and processed mainly using views. The Analyser extracts the online part related to the query by desraggregating the query into online and offline sub-queries. The query plans associated to the online queries are then stored in a repository (Query Depot) and managed by the Query Manager that stores, indexes and ensures the periodic re-execution of continuous queries. The Executor, or dataflow based engine gives the link between the data sent by the Router and the queries which have been translated into scheduling processes and optimized by the Scheduler according to queries already running in the Executor. Finally, the Memory Manager is responsible for the management of the sliding windows and the transfer of data from the online part to offline one via the Mediator in case of obsolescence of data, while the View Manager is in charge of view maintenance obtained from successive executions of continuous queries.

When the incoming query is executed once then the system checks if the online sub-query(or a part of it) is already running in the pool of queries (in the Query Depot). If this is the case the system takes the view associated to the query via the View Manager, if not the system computes the result using the distributed sliding window and generates the associated view as it's done for a persistent query that run continuously.

4.5 Discussion

The research developed in this paper introduced the principles of a hybrid approach for mobility mining combining real-time analysis with information extracted from archived data. The system developed is mainly oriented to the online part of the architecture in order to deal with real-time requirements. We expect that most queries will be processed using only the online part and views associated to the analysis of archived data. Such system should be autonomous and process data and advise agents when necessary. The advantages of such an approach is its likely fast response-time and its reactivity to new events in a operational context, but can lack of efficiency for stream mining or running of complex queries.

5. CONCLUSION

Over the past few years the continuous increase of the large amounts of spatio-temporal data produced by vessel positioning systems requires an evaluation of the current methods, tools and decision support systems currently designed for monitoring and analysis maritime traffic. The research presented in this paper introduces the main principles of a hybrid approach for processing mobility data and that can avoid the common pitfalls due to higher volumes and velocities of data produced, that is, long processing times for purely offline models, and approximate data analysis for online systems. This work introduces a series of preliminary principles for the design of a hybrid approach for the combined processing of real-time streams and archived data. The hybrid architecture suggested is stream-based and deal with real-time requirements of operational contexts with a specific focus on maritime traffic. Ongoing work concerns the development and implementation of the different approaches and the evaluation of some specific algorithms for the stream mobility mining of maritime traffic.

6. REFERENCES

[1] *Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band.* Recommendation ITU-R M.1371-5 (02/2014), 2014.

[2] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz. Hadoop gis: A high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow.*, 6(11):1009–1020, Aug. 2013.

[3] A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. Datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom. Stream: The stanford data stream management system. Technical Report 2004-20, Stanford InfoLab, 2004.

[4] 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.

[5] 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.

[6] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[7] A. Deshpande, Z. Ives, and V. Raman. Adaptive query processing. *Found. Trends databases*, 1(1):1–140, Jan. 2007.

[8] A. Eldawy and M. F. Mokbel. The era of big spatial data. In *31st IEEE International Conference on Data*

Engineering Workshops, ICDE Workshops 2015, Seoul, South Korea, April 13-17, 2015, pages 42–49, 2015.

[9] L. Forlizzi, R. H. Güting, E. Nardelli, and M. Schneider. A data model and data structures for moving objects databases. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 319–330, New York, NY, USA, 2000. ACM.

[10] T. M. Ghanem, A. K. Elmagarmid, P. Larson, and W. G. Aref. Supporting views in data stream management systems. *ACM Trans. Database Syst.*, 35(1), 2010.

[11] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695–719, Oct. 2011.

[12] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 330–339. ACM, 2007.

[13] L. Golab and T. Johnson. Data stream warehousing. In *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 1290–1293, 2014.

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

[15] S. J. Kazemitabar, F. B. Kashani, and D. McLeod. Geostreaming in cloud. In *Proceedings of the 2011 ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2011, November 1, 2011, Chicago, IL, USA*, pages 3–9, 2011.

[16] J. Lu and R. H. Guting. Parallel secondo: Practical and efficient mobility data processing in the cloud. In *BigData Conference*, pages 17–25. IEEE, 2013.

[17] B. Y. Ma, Qiang, 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.

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

[19] M. F. Mokbel, X. Xiong, and W. G. Aref. SINA: scalable incremental processing of continuous queries in spatio-temporal databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, pages 623–634, 2004.

[20] 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.

[21] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 170–177. IEEE Computer Society, 2010.

[22] G. Pallotta, M. Vespe, and K. Bryan. Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, 15(6):2218–2245, 2013.

[23] 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.

[24] M. A. Shah, J. M. Hellerstein, and E. A. Brewer. Highly-available, fault-tolerant, parallel dataflows. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, pages 827–838, 2004.

[25] S. Shekhar, V. Gunturi, M. R. Evans, and K. Yang. Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '12, pages 1–6, New York, NY, USA, 2012. ACM.

[26] X. Sun, A. Yaagoub, G. Trajcevski, P. Scheuermann, H. Chen, and A. Kachhwaha. P^2est: parallelization philosophies for evaluating spatio-temporal queries. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, BigSpatial@SIGSPATIAL 2013, Nov 4th, 2013, Orlando, FL, USA*, pages 47–54, 2013.

[27] M. Vodas, N. Pelekis, Y. Theodoridis, C. Ray, V. Karkaletsis, S. Petridis, and A. Miliou. Efficient ais data processing for environmentally safe shipping. *SPOUDAI Journal of Economics and Business*, 63(3-4):181–190, 2013.

[28] Z. Yu, Y. Liu, X. Yu, and K. Q. Pu. Scalable distributed processing of K nearest neighbor queries over moving objects. *IEEE Trans. Knowl. Data Eng.*, 27(5):1383–1396, 2015.