Bourse M2 Challenge 2/LabExIMobS3

Title : **Accommodating Trajectory Data Variety and Volume by a Multimodel Star Schema: application to autonomous robots planning in the agricultural context.**

Abstract

Nowadays, more and more trajectory data is collected from new acquisition systems (smartphones, vehicles, etc.). A trajectory is described by temporal and spatial data, and it is accompanied by contextual data (such as field, markets, meteo, etc.). Then, we can consider trajectory data as Big Data presenting 3Vs features: Velocity, Variety and Volume. *In particular in the context of the I-Site CAP2025 SupeRob project that aims to provide an information system for the planning and monitoring of autonomous robots planning in the agricultural context a big data set of trajectory data is generated*.

Recent approaches adopt multimodel databases (MMDBs) to natively handle the variety and volume issues arising from the increasing amounts of heterogeneous data (structured, semi-structured, graph based, etc.) made available. However, when it comes to analyzing these data, traditional data Warehouses (DWs) and OLAP systems fall short because they rely on relational DBMSs for storage and querying, thus constraining data variety into the rigidity of a structured schema. DW and OLAP systems allow the online analysis of huge datasets with simple and userfriendly user interfaces.

This project will provide a preliminary investigation of the performance of MMDBs when used to store multidimensional trajectory Big Data for OLAP analysis. *The proposals will be applied to data generated in the context of the SupeRob project to handle robots experts to visually analyze their datasets*.

Responsable du stage : Sandro Bimonte
Courriel / Téléphone : sandro.bimonte@inrae.fr/0473440666
Unité de rattachement : TSCF
Equipe : COPAIN
Etablissement de rattachement : Irstea

Co-encadrant : Roland Lenain
Unité de rattachement : TSCF
Etablissement de rattachement : Inrae

Work plan

A. **Analysis of spatio-temporal functionalities of multimodel databases for the identification of a reference platform.**
B. **Proposal of a logical model for trajectory DW**
C. **Integration of Variety in the logical model for trajectory DW**
D. **Design of a multimodel trajectory DW using data issued from the SupeRob project**

Location: INRAE, Campus Cezeaux
Date : 6 months from March-April 2021

Context. With the advanced techniques of acquisition of geographical positions (sensors, objects connected, etc.) huge trajectory data has been generated. These trajectory data are one of the most important sources of information for many applications in different areas such as for example, mobility (travel behavior, mobility, etc.), the environment, marketing, agriculture (fleet management, tractors, precision farming via sensors, etc.), etc, which characterize data and applications of the I-Site challenges 1 and 2.

In particular, in the agricultural context, today more and more work is being done to set up autonomous robots at service to reduce the costs of manual labor for farmers, to improve their quality of life, as well as to promote agro-ecology with agricultural practices having less significant impacts on the agricultural ecosystem. Autonomous robots move on plots to perform technical tasks such as plowing or weeding or mechanical weeding. They are programmed to perform these tasks by minimizing movement on plots, via trajectories planned, while avoiding potential fixed obstacles (such as a rut or a pole) or mobile (human, animals, or vehicle) requiring a deviation to the trajectory predefined. To analyze the experiments of algorithms for calculating trajectories, a possible solution is to set up a Data Warehouse. This data set represents Trajectory Big Data. This is the context of the *in the context of the I-Site CAP2025 SupeRob project that aims to provide an information system for the planning and monitoring of autonomous robots planning in the agricultural context a big data set of trajectory data is generated*.

Big Data are notoriously characterized by (at least) the 3 V's: volume, velocity, and variety. To handle velocity and volume, some distributed file system-based storage (such as Hadoop) and new Database Management Systems (DBMSs) have been proposed. In particular, four main categories of NoSQL databases have been proposed [1]: key-value, extensible record, graph-based, and document-based. Although NoSQL DBMSs have successfully proved to support the volume and velocity features, variety is still a challenge [19]. Indeed, several practical applications ask for collecting and analyzing data of different types: structured (e.g., relational tables), semi-structured (e.g., XML and Json), and unstructured (such as text, images, etc.). Using the right DBMS for the right data type is essential to grant good storage and analysis performance. Traditionally, each DBMS has been conceived for handling a specific data type, for example, relational DBMSs for structured data, document-based DBMSs for semi structured data, etc. Therefore, when an application requires different data types, two solutions are actually possible: (i) integrating all data into a single DBMS, or (ii) using two or more DBMSs together. The former solution presents serious drawbacks: first of all, some types of data cannot be stored and analyzed (e.g., the pure relational model does not support the storage of images and XML arrays [27]); besides, even when data can be converted and stored in the target DBMS, querying performances could be unsatisfactory. The latter approach (known as polyglot persistence [15]) presents important challenges as well, namely, technically managing more DBMSs, complex query languages, inadequate performance optimization, etc. Multimodel databases (MMDBs) have recently been proposed to overcome these issues. A MMDB is a DBMS that natively supports different data types under a single query language to grant performance, scalability, and fault tolerance [19]. Remarkably, using a single platform for multimodel data promises to deliver several benefits to users besides that of providing a unified query interface; namely, it will simplify query operations, reduce development and maintenance issues, speed up development, and eliminate migration problems [19]. Examples of MMDBs are Postgres and ArangoDB.

Related work. Handling variety while granting at the same time volume and velocity is even more complex in Data Warehouses (DWs) and OLAP systems. Indeed, warehoused data result from the integration of huge volumes of heterogeneous data, and OLAP requires very good performances for data-intensive analytical queries [18]. Traditional DW architectures rely on a single, relational DBMS for storage and querying. To offer better support to volume while maintaining velocity, some recent works propose the usage of NoSQL DBMSs; for example, [7] relies on a document-based DBMS, and [4] on a column-based DBMS. NoSQL proposals for DWs are based on a single data model, and all data are transformed to fit with that model (document, graph, etc.). Overall, although these approaches offer interesting results in terms of volume and velocity, they have been mainly conceived and tested for structured data,

without taking into account variety. Furthermore, to facilitate OLAP querying, DWs are normally based on the multidimensional model, which introduces the concepts of facts, dimensions, and measures to analyze data, so source data must be forcibly transformed to fit a multidimensional logical schema. Since this is not always painless because of the schemaless nature of some source data, some recent work (such as [11]) propose to directly rewrite OLAP queries over document stores that are not organized according to the multidimensional model, following a schema-on-read approach.

Contribution. However, even this approach relies on a single DBMS An interesting direction towards a solution for effectively handling the 3 V's in DW and OLAP systems is represented by MMDBs. A multimodel data warehouse (MMDW) can store data according to the multidimensional model and, at the same time, let each of its elements be natively represented through the most appropriate model. It will support OLAP querying over large volumes of multimodel and multidimensional data, thus ensuring support to both volume, velocity, and variety.*

We will apply our proposals to the dataset used in the context of the SupeRob project in order to provide robots experts with an easy visual interface to explore their data by means of simple and graphical reporting tools.

References

[1] Paolo Atzeni, Francesca Bugiotti, and Luca Rossi. 2014. Uniform access to NoSQL systems. Inf. Syst. 43 (2014), 117–133.

[2] Nabila Berkani, Ladjel Bellatreche, Selma Khouri, and Carlos Ordonez. 2019. Value-driven Approach for Designing Extended Data Warehouses. In Proc. DOLAP@EDBT/ICDT. Lisbon, Portugal.

[3] Doulkifli Boukraâ, Mohammed Amin Bouchoukh, and Omar Boussaïd. 2015. Efficient Compression and Storage of XML OLAP Cubes. IJDWM 11, 3 (2015), 1–25.

[4] Mohamed Boussahoua, Omar Boussaid, and Fadila Bentayeb. 2017. Logical Schema for Data Warehouse on Column-Oriented NoSQL Databases. In Proc. DEXA. Lyon, France, 247–256.

[5] Arnaud Castelltort and Anne Laurent. 2014. NoSQL Graph-based OLAP Analysis. In Proc. KDIR. Rome, Italy, 217–224.

[6] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2015. Implementation of Multidimensional Databases in ColumnOriented NoSQL Systems. In Proc. ADBIS. Poitiers, France, 79–91.

[7] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2015. Implementation of Multidimensional Databases with Document-Oriented NoSQL. In Proc. DaWaK. Valencia, Spain, 379–390.

[8] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2016. Document-Oriented Data Warehouses: Complex Hierarchies and Summarizability. In Proc. UNet. Casablanca, Morocco, 671–683.

[9] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2016. Document-oriented data warehouses: Models and extended cuboids, extended cuboids in oriented document. In Proc. RCIS. Grenoble, France, 1–11.

[10] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2016. Document-oriented Models for Data Warehouses – NoSQL Document-oriented for Data Warehouses. In Proc. ICEIS. Rome, Italy, 142–149.

[11] Mohamed Lamine Chouder, Stefano Rizzi, and Rachid Chalal. 2019. EXODuS: Exploratory OLAP over Document Stores. Inf. Syst. 79 (2019), 44–57.

[12] Khaled Dehdouh. 2016. Building OLAP Cubes from Columnar NoSQL Data Warehouses. In Proc. MEDI. Almería, Spain.
[13] Ibtisam Ferrahi, Sandro Bimonte, and Kamel Boukhalfa. 2017. A Model & DBMS Independent Benchmark for Data Warehouses. In Proc. EDA. Lyon, France, 101–110.
[14] Ibtisam Ferrahi, Sandro Bimonte, Myoung-Ah Kang, and Kamel Boukhalfa. 2017. Design and Implementation of Falling Star - A Non-Redundant SpatioMultidimensional Logical Model for Document Stores. In Proc. ICEIS. Porto, Portugal, 343–350.
[15] Vijay Gadepally, Peinan Chen, Jennie Duggan, Aaron J. Elmore, Brandon Haynes, Jeremy Kepner, Samuel Madden, Tim Mattson, and Michael Stonebraker. 2016. The BigDAWG polystore system and architecture. In Proc. HPEC. Waltham, MA, USA, 1–6.
[16] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. 2019. Approximate OLAP of document-oriented databases: A variety-aware approach. Inf. Syst. 85 (2019), 114–130.
[17] Matteo Golfarelli and Stefano Rizzi. 2009. Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, Inc., New York, NY, USA.
[18] Ralph Kimball and Margy Ross. 2002. The data warehouse toolkit: the complete guide to dimensional modeling, 2nd Edition. Wiley.
[19] Jiaheng Lu and Irena Holubová. 2019. Multi-model Databases: A New Journey to Handle the Variety of Data. ACM Comput. Surv. 52, 3 (2019), 55:1–55:38.
[20] Mohammed El Malki, Arlind Kopliku, Essaid Sabir, and Olivier Teste. 2018. Benchmarking Big Data OLAP NoSQL Databases. In Proc. UNet. Hammamet, Tunisia, 82–94.
[21] Patrick E. O'Neil, Elizabeth J. O'Neil, Xuedong Chen, and Stephen Revilak. 2009. The Star Schema Benchmark and Augmented Fact Table Indexing. In Proc. TPCTC. Lyon, France, 237–252.
[22] Zoubir Ouaret, Rachid Chalal, and Omar Boussaid. 2013. An overview of XML warehouse design approaches and techniques. IJICoT 2, 2/3 (2013), 140–170.
[23] Franck Ravat and Yan Zhao. 2019. Data Lakes: Trends and Perspectives. In Proc. DEXA. Linz, Austria, 304–313.
[24] Oscar Romero and Alberto Abelló. 2009. A Survey of Multidimensional Modeling Methodologies. IJDWM 5, 2 (2009), 1–23. DOI:http://dx.doi.org/10. 4018/jdwm.2009040101
[25] Stefanie Scherzinger, Meike Klettke, and Uta Störl. 2013. Managing Schema Evolution in NoSQL Data Stores. In Proc. DBPL. Riva del Garda, Italy.
[26] Amal Sellami, Ahlem Nabli, and Faïez Gargouri. 2018. Transformation of Data Warehouse Schema to NoSQL Graph Data Base. In Proc. ISDA. Vellore, India, 410–420.
[27] Takeyuki Shimura, Masatoshi Yoshikawa, and Shunsuke Uemura. 1999. Storage and Retrieval of XML Documents Using Object-Relational Databases. In Proc. DEXA. Florence, Italy, 206–217.
[28] Chao Zhang, Jiaheng Lu, Pengfei Xu, and Yuxing Chen. 2018. UniBench: A Benchmark for Multi-model Database Management Systems. In Proc. TPCTC. Rio de Janeiro, Brazil, 7–23.