Activity Report 2022

Team SHAMAN

Symbolic and Human-centric view of dAta MANagement

D7 – Data & Knowledge Management

1 Team composition

Researchers and faculty

Cheikh-Brahim El Vaigh, Enseignant-Chercheur contractuel (LRU), ENSSAT, from Sep. 20 to Aug. 22
François Goasdoué, Professor, ENSSAT, head of the team
Hélène Jaudoin, Associate Professor, ENSSAT
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Cyrielle Mallart, ATER, ENSSAT, from Dec. 21 to Aug. 22
Rahul Nath, Postdoc, ENSSAT, from Sep. 21 to Dec. 22
Pierre Nerzic, Associate Professor, IUT Lannion
Laurent d'Orazio, Professor, IUT Lannion
Olivier Pivert, Professor, ENSSAT
Grégory Smits, Associate Professor, HDR, IUT Lannion, up to Aug. 22
Virginie Thion, Associate Professor, HDR, ENSSAT

Research engineers, techical staff

Pierre Alain, Research engineer (PhD), ENSSAT (20%)

PhD students

Wafaa El Husseini, PhD student, ENSSAT, since Sep. 20
Yamen Haddad, PhD student, Inria Saclay & École Polytechnique, since Jan. 21
Vincent Lannurien, PhD student, ENSTA Bretagne, since Oct. 21
Van Long Nguyen Huu, PhD student ENSSAT Nokia, from Jan. 20 to Dec. 22
Vérone Yepmo, PhD student, ENSSAT, since Sep. 20

Administrative assistant

Angélique Le Pennec, team assistant, ENSSAT (20%)
Joëlle Thépault, team assistant, ENSSAT (20%)

2 Overall objectives

2.1 Overview

The overall goal pursued by Shaman is to improve the data management methods currently used in commercial systems, which suffer from a severe lack of flexibility in several respects. In particular, with the techniques currently available, it is difficult for a user to *i)* understand the data he/she has access to, and to *ii)* specify his/her information needs in an intuitive though sufficiently expressive way. Moreover, these systems/approaches have limited capabilities when it comes to handling imperfect data, in particular in a context where data come from different sources. Shaman addresses these shortcomings and strives to devise new tools with the objective of helping end users and/or database conceptors:

- *model* and *integrate* the data — possibly *heterogeneous* and/or *imperfect* — that are relevant in a given applicative context;
- *understand* the data (structure and semantics) that are accessible to them;
- *query* and *analyze* these data, taking into account their *preferences*, by means of a mechanism as *cooperative* as possible.

We favor *symbolic* approaches for the sake of intelligibility/ease of use (again, the objective is to define *human-centric* data management methods). Fuzzy set theory (and the closely related possibility theory) constitutes a natural and intuitive symbolic/numerical interface, between the symbolic aspect of a linguistic variable and the numerical nature of the corresponding characteristic function valued in the unit interval. Fuzzy set theory can be used to model preference queries, data summaries, and cooperative answering strategies, as well as to define a new data model and querying framework based on *clusters* instead of tables. On the other hand, possibility theory can serve as a basis to the modeling of uncertain databases where uncertainty is assumed to be of a *qualitative*, nonfrequential, nature.

Ontology-based data management is another central topic in Shaman inasmuch as ontologies *i)* are a powerful tool to make data more *intelligible* to users, and to *mediate* between data sources whose schemas differ, *ii)* make it possible to enhance data management systems with *reasoning capabilities*, thus to handle data in a more "intelligent" way.

A strong point of Shaman lies in its positioning at the junction between the Databases and Artificial Intelligence domains. Up to now, these two research communities have stayed much apart from each other, whereas we believe that data management should highly benefit from a cross-fertilization between DB technologies and AI approaches. Historically, the members of the team were always sensitized to this challenge, making use for instance of theoretical tools coming from fuzzy logic for making database querying more flexible. This trend also corresponds to an evolution of the data management landscape itself: the rise of the internet made it necessary to manage open and linked data, using methods that involve reasoning capabilities (i.e., what is called the Semantic Web).

2.2 Scientific foundations

2.2.1 Big Data management

Managing large volumes of data (with respect to the available resources) has been an important issue for decades. As an illustration, the first Very Large Data Bases (VLDB) conference was organized in 1975. Main contributions in the domain include parallel and distributed systems [DG92] with different approaches, in particular shared-nothing architectures [Sto86].

The deployment of large data centers consisting of thousand of commodity hardware-based nodes have led to massively parallel processing systems. In particular, large scale distributed file systems such as Google File System [GGL03], parallel processing paradigm/environment like MapReduce [DG08] have been the foundations of a new ecosystem with data management contributions in major conferences and journals on databases, such as VLDB, VLDBJ, SIGMOD, TODS, ICDE, IEEE DEB, ICDE and EDBT. Different (often open-source) systems have been provided such as Pig [ORS+08], Hive [TSJ+10] or more recently Spark [ZCD+12] and Flink [CKE+15], making it easier to use data center resources for managing big data.

2.2.2 Fuzzy logic applied to databases

Fuzzy sets were introduced by L.A. Zadeh in 1965 [Zad65] in order to model sets or classes whose boundaries are not sharp. This is particularly the case for many adjectives of the natural language which can be hardly defined in terms of usual sets (e.g., *high*, *young*, *small*, etc.), but are a matter of degree. A fuzzy (sub)set F of a universe X is defined

[DG92] D. J. DeWitt, J. Gray, "Parallel Database Systems: The Future of High Performance Database Systems", *Communications of the {ACM} 35*, 6, 1992, p. 85–98.

[Sto86] M. Stonebraker, "The Case for Shared Nothing", *IEEE Database Engineering Bulletin 9*, 1, 1986, p. 4–9.

[GGL03] S. Ghemawat, H. Gobioff, S.-T. Leung, "The Google file system", *in : Proceedings of the Symposium on Operating Systems Principles (SOSP)*, p. 29–43, Bolton Landing, NY, USA, 2003.

[DG08] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", *Communications of the ACM 51*, 1, 2008, p. 107–113.

[ORS+08] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins, "Pig latin: a not-so-foreign language for data processing", *in : Proceedings of the SIGMOD International Conference on Management of Data*, p. 1099–1110, Vancouver, BC, Canada, 2008.

[TSJ+10] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Anthony, H. Liu, R. Murthy, "Hive - a petabyte scale data warehouse using Hadoop", *in : Proceedings of the International Conference on Data Engineering ({ICDE})*, p. 996–1005, Long Beach, California, {USA}, 2010.

[ZCD+12] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, I. Stoica, "Resilient Distributed Datasets: {A} Fault-Tolerant Abstraction for In-Memory Cluster Computing", *in : Proceedings of the {USENIX} Symposium on Networked Systems Design and Implementation (NSDI)*, p. 15–28, San Jose, CA, USA, 2012.

[CKE+15] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, K. Tzoumas, "Apache Flink{\texttrademark}: Stream and Batch Processing in a Single Engine", *{IEEE} Data Engineering Bulletin 38*, 4, 2015, p. 28–38.

[Zad65] L. Zadeh, "Fuzzy sets", *Information and Control 8*, 1965, p. 338–353.

thanks to a membership function denoted by μ_F which maps every element x of X into a degree $\mu_F(x)$ in the unit interval $[0, 1]$. When the degree equals 0, x does not belong at all to F, if it is 1, x is a full member of F and the closer $\mu_F(x)$ to 1 (resp. 0), the more (resp. less) x belongs to F. Clearly, a regular set is a special case of a fuzzy set where the values taken by the membership function are restricted to the pair {0, 1}. Beyond the intrinsic values of the degrees, the membership function offers a convenient way for ordering the elements of X and it defines a symbolic-numeric interface.

Since Lotfi Zadeh introduced fuzzy set theory in 1965, many applications of fuzzy logic to various domains of computer science have been achieved. As far as databases are concerned, the potential interest of fuzzy sets in this area has been identified as early as 1977, by V. Tahani [Tah77] — then a Ph.D. student supervised by L.A. Zadeh — who proposed a simple fuzzy query language extending SEQUEL. This first attempt was then followed by many researchers who strove to exploit fuzzy logic for giving database languages more expressiveness and flexibility. Then, in 1978, Zadeh coined possibility theory [Zad78], a model for dealing with uncertain information in a qualitative way, which also opened new perspectives in the area of uncertain databases. The pioneering work by Prade and Testemale [PT84] has had a rich posterity and the issue of modeling/querying uncertain databases in the framework of possibility theory is still an active topic of research nowadays. Beside these two main research lines, several other ways of exploiting fuzzy logic have been proposed along the years for dealing with various other aspects of data management, for instance *fuzzy data summaries*. More recently, fuzzy logic has also been applied — notably by the Shaman team — to model and query non-relational databases such as RDF databases or graph databases.

2.2.3 Ontology-based data management

Till the end of the 20$^{\text{th}}$ century, there have been few interactions between these two research fields concerning data management, essentially because they were addressing it from different perspectives. KR was investigating data management according to human cognitive schemes for the sake of intelligibility, e.g. using *Conceptual Graphs* [CM08] or *Description Logics* [BCM+03], while DB was focusing on data management according to simple mathematical structures for the sake of efficiency, e.g. using the *relational model*

[Tah77] V. Tahani, "A Conceptual Framework for Fuzzy Query Processing — A Step Toward Very Intelligent Database Systems", *Information Processing and Management 13*, 5, 1977, p. 289–303.

[Zad78] L. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

[PT84] H. Prade, C. Testemale, "Generalizing database relational algebra for the treatment of incompleteuncertain information and vague queries", *Information Sciences 34*, 1984, p. 115–143.

[CM08] M. Chein, M.-L. Mugnier, *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*, Springer Publishing Company, Incorporated, 2008.

[BCM+03] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (editors), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2003.

[AHV95] or the *eXtensible Markup Language* [AMR+12].

In the beginning of the 21st century, these ideological stances have changed with the new era of *ontology-based data management* [Len11]. Roughly speaking, ontology-based data management brings data management one step closer to end-users, especially to those that are not computer scientists or engineers. It basically revisits the traditional architecture of database management systems by decoupling the models with which data is exposed to end-users from the models with which data is stored. Notably, ontology-based data management advocates the use of conceptual models from KR as human intelligible front-ends called *ontologies* [Gru09], relegating DB models to back-end storage.

The *World Wide Web Consortium* (W3C) has greatly contributed to ontology-based data management by providing *standards* for handling data through ontologies, the two *Semantic Web* data models. The first standard, the *Resource Description Framework* (RDF) [W3Ca], was introduced in 1998. It is a graph data model coming with a very simple ontology language, *RDF Schema*, strongly related to description logics. The second standard, the *Web Ontology Language* (OWL) [W3Cb], was introduced in 2004. It is actually a family of well-established description logics with varying expressivity/complexity tradeoffs.

The advent of RDF and OWL has rapidly focused the attention of academia and industry on *practical* ontology-based data management. The research community has undertaken this challenge at the highest level, leading to pioneering and compelling contributions in top venues on Artificial Intelligence (e.g. AAAI, ECAI, IJCAI, and KR), on Databases e.g. ICDT/EDBT, ICDE, SIGMOD/PODS, and VLDB), and on the Web (e.g. ESWC, ISWC, and WWW). Also, open-source and commercial software providers are releasing an ever-growing number of tools allowing effective RDF and OWL data management.

Last but not least, large societies have promptly adhered to RDF and OWL data management (e.g. library and information science, life science, and medicine), sustaining and begetting further efforts towards always more convenient, efficient, and scalable ontology-based data management techniques.

2.3 Application domains

We currently focus on the following application domains:

- Open data management. One of the challenges in web data management today is to define adequate tools allowing users to extract the data that are the most

[AHV95] S. Abiteboul, R. Hull, V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[AMR+12] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, P. Senellart, *Web Data Management*, Cambridge University Press, 2012.

[Len11] M. Lenzerini, "Ontology-based data management", 2011.

[Gru09] T. Gruber, "Ontology", *in: Encyclopedia of Database Systems*, Springer US, 2009, p. 1963–1965.

[W3Ca] W3C, "Resource Description Framework", *research report*.

[W3Cb] W3C, "Web Ontology Language", *research report*.

likely to fulfill all or part of their information needs, then to understand and automatically correlate these data in order to elaborate relevant answers or analyses. Open data may be of various levels of quality: they may be imprecise, incomplete, inconsistent and/or their reliability/freshness may be somewhat questionable. An appropriate data model and suitable querying tools must then be defined for dealing with the imperfection that may pervade data in this context. On the other hand, it is of prime importance to provide end-users with simple and flexible means to better understand and analyze open data. The standards of W3C offer popular languages for representing both open and structured data. Another objective is to propose analytical tools suited to these languages through the construction of RDF data warehouses, whereas fuzzy-set-based data summarization approaches should constitute an important step towards making open data more intelligible to non-expert users.

- Cybersecurity. Security monitoring is one subdomain of cybersecurity. It aims at guaranteeing the safety of systems, continuously monitoring unusual events by analyzing logs. The notion of a system in this context is very variable. It can actually be an information system in any organization or any device, like a laptop, a smartphone, a smartwatch, a vehicle (car, plane, etc.), a television, etc. Hence, the data to be managed with a high Velocity, are Voluminous with a high Variety. Security monitoring can thus be seen as a concrete use case of Big Data. Shaman is involved in several projects related to security monitoring, in particular SERBER funded by the Pôle d'Excellence Cyber. One of the main goals is to provide a Big Data platform applied to security monitoring. This makes it mandatory to address several issues like efficient big fuzzy joins, data management with new hardware (FPGA) or optimization on encrypted data.

- Maritime transportation of goods. Shaman participates in the project Sea Defender (2020 – 2024), founded by the DGA (Direction Générale de l'Armement), whose objective is to conceive a solution for automating the controls performed by financiary institutions related to the maritime transportation of goods (an important partner in the project is the banking company HSBC). These controls aim to check i) the coherence between the data contained in the documents describing the transaction and those related to the effective path and transportation mode of the goods; ii) the conformity of the transport wrt. the rules of international trade (embargoed countries, piracy, etc.). For doing so, it is necessary to i) aggregate the data provided by different sources: maritime transportation companies, sites devoted to ship tracking, sites specialized in risk detection and fraud management, maritime weather forecast information, customs, etc.); ii) correlate all these data according to precise business rules in order to detect suspicious activities. The approach advocated by Shaman involves two steps; First, one needs to model complex fuzzy concepts based on the combination of different dimensions (e.g., a batch of containers may be considered *suspicious* if its rotation frequency is *high*, the loading intervals are *long*, and if they come from a company *under surveillance*). Then one needs to conceive knowledge discovery tools working on a unified representation of the data in the form of linguistic summaries.

- Digital score libraries. *Sheet music scores* have been the traditional way to pre-

serve and disseminate western classical music works for centuries. Nowadays, their content can be encoded in digital formats that yield a very detailed representation of music content expressed in the language of *music notation*. These digitized music scores constitute, therefore, an invaluable asset for digital library services. Shaman studies the data management of digitized music score data, including the design of intuitive and effective querying process and the data quality management of such data.

3 Scientific achievements

3.1 Big data management

Participants: Laurent d'Orazio, Vincent Lannurien, Van Long Nguyen Huu, Chenxiao Wang, Michail Georgoulakis Misegiannis, Mohamed Handaoui.

- *Coalescing heuristic to replacement policy in MASCARA.* We introduced ModulAr Semantic CAching fRAmework (MASCARA) that deployed Semantic Caching (SC) to perform a fast query processing based on Field Programmable Gate Arrays (FPGAs) accelerators. In addition of the accelerators, cache management plays an important role to address coalescing strategy and replacement policy so as to maximize the performance of FPGA caching. Therefore, in [9], we presented a coalescing heuristic with a new replacement function that leverages advantages of traditional strategies and overcomes their drawbacks. The proposed heuristic reduces response time, improves data availability, and saves cache space with respect to the semantic locality of query workload.

- *SLA-Aware Cloud Query Processing with Reinforcement Learning-Based Multi-objective Re-optimization.* Query processing on cloud database systems is a challenging problem due to the dynamic cloud environment. In cloud database systems, besides query execution time, users also consider the monetary cost to be paid to the cloud provider for executing queries. Moreover, a Service Level Agreement (SLA) is signed between users and cloud providers before any service is provided. Thus, from the profit-oriented perspective for the cloud providers, query re-optimization is multi-objective optimization that minimizes not only query execution time and monetary cost but also SLA violations. In [12], we introduce ReOptRL and SLAReOptRL, two novel query re-optimization algorithms based on deep reinforcement learning. Experiments show that both algorithms improve query execution time and query execution monetary cost by 50.

- *Multi-objective query optimization in Spark SQL.* Query optimization is a challenging process of DBMSs. When tackling query optimization in the cloud, there exists a simultaneous need of providing an optimal physical query execution plan, as well as an optimal resource configuration among available ones. Cloud computing features like resource elasticity and pricing make the process of finding this optimal query plan a multi-objective problem, with the monetary cost being an equally important factor to query execution time. Apache Spark is a popular

choice for managing big data in the cloud. However, query optimization in its SQL module (Spark SQL) involves a number of limitations due to the rule-based nature of its optimizer, Catalyst. In [6], we propose a multi-objective cost model for the extension of the query optimizer of Apache Spark, aiming to minimize both objectives of query execution time and monetary cost, as well as a methodology for exploring the space of Pareto-optimal query plans and selecting one. The cost model is implemented and tuned, and an experimental study is conducted to validate its accuracy.

- *RISCLESS: A Reinforcement Learning Strategy to Guarantee SLA on Cloud Ephemeral and Stable Resources.* In [13], we propose RISCLESS, a Reinforcement Learning strategy to exploit unused Cloud resources. Our approach consists in using a small proportion of stable on-demand resources alongside the ephemeral ones in order to guarantee customers SLA and reduce the overall costs. The approach decides when and how much stable resources to allocate in order to fulfill customers' demands. RISCLESS improved the Cloud Providers (CPs)' profits by an average of 15.9% compared to past strategies. It also reduced the SLA violation time by 36.7% while increasing the amount of used ephemeral resources by 19.5%.

3.2 Flexible, cooperative and quality-aware data management

Participants: Ludovic Liétard, Pierre Nerzic, Olivier Pivert, Grégory Smits, Virginie Thion, Véronne Yepmo.

- *Anomaly detection and explanation.* Anomaly detection has been studied intensively by the data mining community for several years. As a result, many methods to detect anomalies have emerged, and others are still under development. But during the recent years, anomaly detection, just like a lot of machine learning tasks, is facing a wall. This wall, erected by the lack of trust of the final users, has slowed down the usage of these algorithms in the real-world situations for which they are designed. Having the best empirical accuracy is not enough anymore; there is a need for algorithms to explain their outputs to the users in order to increase their trust. Consequently, a new expression has emerged recently: eXplainable Artificial Intelligence (XAI). This expression, which gathers all the methods that provide explanations to the output of algorithms has gained popularity, especially with the outbreak of deep learning. A lot of work has been devoted to anomaly detection in the literature, but not as much to anomaly explanation. There is so much work on anomaly detection that several reviews can be found on the topic. In contrast, there does not exist any survey on anomaly explanation in particular, while there are a lot of surveys on XAI in general or on XAI for neural networks for example. In [2, 14], we provide a comprehensive review of the anomaly explanation field. After a brief recall of some important anomaly detection algorithms, the anomaly explanation methods from the literature are classified according to a taxonomy that we defined. This taxonomy stems from an analysis of what is really important when trying to explain anomalies. In

[11], we address the tasks of anomaly detection and explanation simultaneously, in the human-in-the-loop paradigm integrating the end-user expertise. The paper first proposes to exploit two complementary data representations to identify anomalies, namely the description induced by the raw features and the description induced by a user-defined vocabulary. These representations respectively lead to identify so-called data-driven and knowledge-driven anomalies. We then propose to confront these two sets of instances so as to improve the detection step and to dispose of tools towards anomaly explanations. Three cases are distinguished and discusses, and we underline how the two description spaces can benefit from one another, in terms of accuracy and interpretability.

- *Massive Data Exploration using Estimated Cardinalities.* In [8] and [7], we use linguistic summaries to provide personalized exploration functionalities on massive relational data. To ensure a fluid exploration of the data, cardinalities of the data properties described in the summaries are estimated from statistics about the data distribution. The proposed workflow also involves a vocabulary inference mechanism from these statistics and a sampling-based approach to consolidate the estimated cardinalities especially in the case of conjunctive summaries. The paper shows that soft computing techniques are particularly relevant to build concrete and functional business intelligence solutions.

- *Music score management.* In the context of Digital Score Libraries, we proposed a model that makes possible to model the musical content of digital score as graph data, which can be stored in a graph database management system, and an associated implementation of such an approach in a Neo4j database, and expressing searches and analyses through graph pattern queries with the query language [10, 1].

3.3 Ontology-based data management

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

- *Query optimization for knowledge bases.* Ontology-based data management (OBDM) consists in performing data management tasks on a knowledge base (KB), in particular consistency checking and query answering. A KB is made of a database on which a set of deductive constraints, called an ontology, holds. The main OBDM technique, called FOL-reducibility, reduces consistency checking and query answering on KBs to standard query answering on databases provided by database management systems. It has been studied for a variety of OBDM settings based on datalog±, description logics, existential rules, and RDF. In [4], we devise a novel, general optimization framework that applies to all the works on FOL-reducibility from the literature. In particular, (*i*) we revise the foundational definition of FOL-reducibility of query answering (and of consistency checking that reduces to it) to allow for better performance while retaining correctness, (*ii*) we provide a generic algorithm for revised FOL-reducibility of query answering that leverages any algorithm from the literature for standard FOL-reducibility of query

answering, and (*iii*) we experimentally evaluate our optimization framework in a setting that underpins the W3C's OWL2 QL standard for OBDM, and we show that it improves the performance of consistency checking and of query answering, significantly in general and up to several orders of magnitude.

- *OptiRef.* Ontology-based data management (OBDM) consists in performing data management tasks, e.g., consistency checking and query answering, on a knowledge base (KB); a KB is a database on which a set of deductive constraints, called an ontology, holds. The prominent OBDM technique is FOL-reducibility that (*i*) expresses the task to perform on a KB as a query, (*ii*) reformulates this query w.r.t. the ontology so that (*iii*) the evaluation of the query reformulation on the database, by a DBMS, solves the task. Alas, this technique suffers from performance issues when query reformulations are complex, hence are costly to evaluate by DBMSs. In [5], we describe and showcase the OptiRef system that implements a novel, general optimization framework of [4] that applies to all related work on FOL-reducibility. We demonstrate its effectiveness using DL-lite$_\mathcal{R}$ KBs; DL-lite$_\mathcal{R}$ underpins the W3C's OWL2 QL standard for OBDM. OptiRef significantly improves performance, up to several orders of magnitude.

- Query optimization for graph-relational analytics. Graph data is generally stored and processed using two main approaches: (*i*) extending existing relational database management systems (RDBMSs) with graph capabilities, and (*ii*) through native graph database management systems (GDBMSs). The advantage of leveraging RDBMSs is to benefit from the maturity of their query optimization and execution. Conversely, native GDBMSs treat complex graph structure as a first-class citizen, which may make them more efficient on complex structural queries. In [3], we consider the processing of *graph-relational queries*, that is, queries mixing graph and relational operators, on graph data. We take a purely relational approach, reorganizing the graph connectivity information using a novel CSR Optimised Schema (COS). Based on our storage model, incoming queries are reformulated to take into account the COS data organization, and can then be optimized and executed by an RDBMS. We have implemented our approach on top of PostgreSQL and we demonstrate that COS improves the performance for many graph-relational queries of the popular Social Network Benchmark.

4 Software development

4.1 FuzViz

Participants: Pierre Nerzic, Grégory Smits.

FuzViz turns two scientific contributions, [8] and [7], into an operational research prototype. It includes three fuzzy vocabulary elicitation methods based on the distribution of the data estimated from statistics, and a scalable linguistic summarization strategy. The goal of this prototype is to show how complementary our scientific contributions are and that they provide pragmatic solutions to concrete needs. In terms of functionalities, FuzViz provides fluid and intuitive exploration methods and interactive views of

massive relational data. We are currently collaborating with the SATT Ouest Valorisation company and Stratinnov to obtain a software maturation funding and to reach companies interested in such functionalities.

4.2 Musypher

Participants: Virginie Thion.

MUSYPHER is an application that makes it possible to transcript a music score, encoded in a XML dialect (MEI or MusicXML), into an attributed graph database hosted by a Neo4j database management system. Our goal is to illustrate the relevancy (expressiveness, efficiency) of managing music scores over a graph-based data model.

4.3 Smarten

Participants: Olivier Pivert, Virginie Thion.

SMARTEN is an application that allows extending a mind map by querying data stemming from graph databases. It implements a theoretical framework that uses fuzzy set theory in order to identify the graph databases concepts that could contribute to the extension of the mind map, and also to compute scores (a relevancy score and an originality acore) associated with each suggestion.

4.4 Sugar

Participants: Olivier Pivert, Virginie Thion.

SUGAR is a prototype, based on the Neo4j graph database management system, which allows querying graph databases — fuzzy or not — in a flexible way. It makes it possible to express preferences queries where preference criteria may concern i) the content of the vertices of the graph and ii) the structure of the graph (which may include weighted vertices and edges when the graph is fuzzy).

4.5 Tamari

Participants: Virginie Thion.

TAMARI is software add-on, based on the Neo4j graph database management system, which allows introducing data quality-awareness when querying a graph database. Based on quality annotations that denote quality problems appearing in data (the annotations typically result from collaborative practices in the context of open data usage like e.g. users' feedbacks), and on a user's profile defining usage-dependant quality requirements, the TAMARI prototype computes a quality level of each retrieved answer.

4.6 OptiRef

Participants: Pierre Alain, Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

OptiRef is a JAVA tool built on top of ontology-based data management systems in order to optimize them. It features a PHP/JSP/jQuery-based GUI in order to examine the performance it brings to off-the-shelf ontology-based data management systems.

4.7 FRESQUE

Participants: Hoang Van Tran, Laurent d'Orazio.

FRESQUE is a framework for secure range query processing, that enables a scalable consumption throughput while still maintaining strong privacy protection for outsourced data.

4.8 Time-Series Semantic Caching

Participants: Trung Dung Le, Laurent d'Orazio.

Time-Series Semantic Caching is a form-based semantic caching for Time Series Data (TSD) system. The approach reduces both query result storing based on semantic caching technique and the data transfer between clients and servers.protection for outsourced data.

4.9 MASCARA

Participants: Van Long Nguyen Huu, Laurent d'Orazio.

MASCARA is a FPGA-based semantic caching. The approach relies on hardware acceleration to improve performances (in particular response times and energy consumption) in big data processing.

4.10 OntoSQL

Participants: Maxime Buron, Cheikh-Brahim El Vaigh, François Goasdoué.

OntoSQL is a Java-based tool that provides two main functionalities: (i) loading RDF graphs (consisting of RDF assertions and possibly an RDF Schema) into a relational database; the data is integer-encoded and indexed; (ii) querying the loaded RDF graphs through conjunctive SPARQL queries, a.k.a. basic graph pattern queries. OntoSQL not only evaluates queries, it answers them, that is: its answers accounts for both the data explicitly present in the database, as well as the implicit data begotten by the ontology knowledge. To this aim, OntoSQL supports both materialization (aka saturation), and reformulation-based query answering.

4.11 Skrid

Participants: Virginie Thion, Pierre Alain.

The SKRID platform is a digital score library that makes available some Traditional Breton music scores. The SKRID platform is still under development but already makes available some music scores that were collected and encoded by Shaman. The platform will integrate the approaches developed in the *Music score management* research axis of Shaman (under development).
`https://shaman.enssat.fr/skrid/index`

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 DODAM

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, Francçois Goasdoué, Hélène Jaudoin.

The Stic-AmSud project DODAM (2022-2024) brings together experts from artificial intelligence and data management from Univ. Rennes and Univ. Sorbonnes Universités in France as well as from Univ. Adolfo Ibanez (Chile), Univ. Buenos Aires (Argentina), Univ. de la Republica (Uruguay) in South America. The goal of this project is to study how knowledge representation and reasoning can improve performance, interpretability and explanability of machine learning and data analytics.

5.2 National Initiatives

5.2.1 CQFD

Participants: Wafaa El Husseini, Cheikh-Brahim El Vaigh, François Goasdoué, Hélène Jaudoin.

The ANR project CQFD (2019-2024) brings together experts in automated reasoning, data management and knowledge representation from Inria, Telecom ParisTech, Univ. Bordeaux, Univ. Grenoble, Univ. Montpellier and Univ. Rennes 1. The aim of the project is to devise data management algorithms for distributed knowledge-based data management systems.

5.2.2 SeaDefender

Participants: Grégory Smits, Olivier Pivert, Pierre Nerzic.

Sea Defender is a project funded by the DGA that involves the Semsoft company (located in Rennes) and the SHAMAN team. The goal of this project is to provide a

novel anomaly detection workflow dedicated to the particular cases of under and upper pricing, which the main cause of money laundering in the world. To solve this issue, two scientific issues are addressed by the shaman team : the detection of contextual anomalies and the explanation of the found anomalies. This two tasks form the basis of research subjects studied by Véronne Yepmo (PhD) and Rahul Nath (research engineer).

5.3 Hardware acceleration, an application to big data analytics in security monitoring

Participants: Van Long Nguyen Huu, Laurent d'Orazio.

The project Think Cities, funded by a CIFRE grant aims at developing optimization techniques, namely semantic caches, on top of new hardware and more precisely FPGAs, with an application in Cyber Security and security monitoring. Apart from IRISA/Shaman, the other participant is Nokia/Alctal Lucent Bell Labs (Lannion).

5.4 International collaboration

From March to July 2022, Grégory Smits worked with Marek Reformat at the University of Alberta (Edmonton - Canada) as an invited researcher.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Selection

Member of Conference Program Committees

François Goasdoué served as a member of the following program committees:

- AAAI Conference on Artificial Intelligence (AAAI)
- European Conference on Artificial Intelligence (ECAI)
- Conference on Extraction et Gestion de Connaissances (EGC)
- International Joint Conference on Artificial Intelligence (IJCAI)

Laurent d'Orazio served as a member of the following program committees:

- International Conference on Big Data Analytics and Knowledge Discovery (DaWaK@DEXA)
- International Workshop on Data Engineering meets Intelligent Food and COoking Recipe (DECOR@ICDE)

- International Workshop on Intelligent Data - From Data to Knowledge (DOING@ADBIS)
- International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications (BPOD@BigData)
- Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)
- Conférence Extraction et Gestion des Connaissances (EGC), démonstrations

Olivier Pivert served as a member of the following program committees:

- ACM Symposium on Applied Computing (ACM SAC)
- International Symposium on Methodologies for Intelligent Systems (ISMIS)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)
- International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA)

Grégory Smits served as a member of the following program committees:

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA)
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)
- International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-

Virginie Thion served as a member of the following program committees:

- Journées Bases de Données Avancées, Research papers sessions (BDA)
- Journées Bases de Données Avancées, Thesis Award (BDA)
- Conference on Extraction et Gestion de Connaissances (EGC)

6.1.2 Journal

Member of the Editorial Boards

Olivier Pivert is a member of the following editorial boards:

- Journal of Intelligent Information Systems,
- Fuzzy Sets and Systems,
- International Journal of Fuzziness, Uncertainty and Knowledge-Based Systems,
- Revue Ouverte d'Ingénierie des Systèmes d'Information.

Reviewer - Reviewing Activities

François Goasdoué served as a reviewer for the following journals:

- Artificial Intelligence Journal (AIJ)
- Information Systems (IS)

Laurent d'Orazio served as a reviewer for the following journals:

- Information Sciences
- Applied Soft Computing Journal (IS)

Grégory Smits served as a reviewer for the following journal:

- IEEE Transactions on Network and Service Management
- Fuzzy Sets and Systems (FSS)

Virginie Thion served as a reviewer for the following journal:

- Fuzzy Sets and Systems (FSS)

6.1.3 Leadership within the Scientific Community

François Goasdoué is a member of the IJCAI Program Committee Board, from 2022 to 2024.

François Goasdoué is a member of the Steering Committee of "Communauté Francophone en Gestion de Données : Principes, Technologies et Applications" (BDA).

Olivier Pivert is a member of the permanent steering committees of

- the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA);
- the International Symposium on Methodologies for Intelligent Systems (ISMIS);
- the International Conference on Flexible Query-Answering Systems (FQAS).

6.1.4 Scientific Expertise

Olivier Pivert is an expert for the Czech Science Foundation.

6.1.5 Research Administration

François Goasdoué is a member of the Scientific Steering Committee of IRISA UMR 6074, since 2013.

François Goasdoué is a member of the Laboratory council of IRISA UMR 6074, since 2022.

François Goasdoué is the head of the Shaman team of IRISA, since 2019.

François Goasdoué is the head of the Lannion branch of IRISA, since 2020.

François Goasdoué is the head of the Scientific Committee of ENSSAT, since 2022.

Grégory Smits is a member of the Research steering committee at the IUT of Lannion.

6.2 Teaching, supervision

6.2.1 Teaching

Several members of the Shaman team give courses in the ENSSAT track of the Master's degree curriculum in Computer Science at University of Rennes 1: Olivier Pivert and Grégory Smits teach a course about *Advanced Databases*, François Goasdoué and Hélène Jaudoin teach a course on *Web Data Management*, and Laurent d'Orazio teaches a part of the course on *Data analysis and data mining*.

6.2.2 Supervision

- PhD in progress: Wafaa El Husseini, Efficient ontology-based data management, started in Oct. 20, François Goasdoué and Hélène Jaudoin
- PhD in progress: Yamen Haddad, Adaptative query planning and execution on heterogeneous data sources, started in Jan. 21, Angelos-Christos Anadiotis, François Goasdoué and Ioana Manolescu
- PhD in progress: Vincent Lannurien, Big data applications scheduling on heterogeneous Cloud resources, started in Oct. 21, Laurent d'Orazio, Jalil Boukhobza and Olivier Barais
- PhD: Van Long Nguyen Huu, Hardware acceleration, an application to big data analytics in security monitoring, defended in Dec. 22, Laurent d'Orazio, Emmanuel Casseau and Julien Lallet
- PhD in progress: Véronne Yepmo, Anomaly detection and explanation, started in Nov. 20, Grégory Smits and Olivier Piver

6.2.3 Juries

François Goasdoué

- PhD, referee, Théo Ducros, Université de Clermont-Auvergne
- PhD, referee and president, Ousmane Issa, Université de Clermont-Auvergne
- PhD, president, Maria Masri, Université de Rennes
- PhD, referee, Adrian Orozco Taboada, Université de Dijon-Bourgogne

Grégory Smits

- HDR, referee, Arnaud Calsteltort, Université de Montpellier

6.3 Popularization

François Goasdoué, with colleagues from INRIA Saclay and Univ. Montpellier, gave a full-day tutorial on "Heterogeneous Data Integration" at the French summer school on Massive Distributed Data.

7 Bibliography

M. Bienvenu, C. Bourgaux, F. Goasdoué, "Computing and Explaining Query Answers over Inconsistent DL-Lite Knowledge Bases", *Journal of Artificial Intelligence Research 64*, March 2019, p. 563–644, https://hal.inria.fr/hal-02066288.

M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier, "Ontology-Based RDF Integration of Heterogeneous Data", *in : EDBT/ICDT 2020 - 23rd International Conference on Extending Database Technology*, Copenhagen, Denmark, March 2020, https://hal.inria.fr/hal-02446427.

S. El Hassad, F. Goasdoué, H. Jaudoin, "Learning Commonalities in SPARQL", *in : International Semantic Web Conference (ISWC)*, Vienna, Austria, October 2017, https://hal.inria.fr/hal-01572691.

F. Goasdoué, P. Guzewicz, I. Manolescu, "RDF graph summarization for first-sight structure discovery", *The VLDB Journal 29*, 5, April 2020, p. 1191–1218, https://hal.inria.fr/hal-02530206.

M. Georgoulakis Misegiannis, L. D'orazio, V. Kantere, "From Cloud to Serverless: MOO in the new Cloud epoch", *in : International Conference on Extending Database Technology (EDBT)*, Virtual, United Kingdom, March 2022, https://hal.inria.fr/hal-03925696.

V. L. Nguyen Huu, J. Lallet, E. Casseau, L. D'Orazio, "MASCARA-FPGA cooperation model: Query Trimming through accelerators", *in : SSDBM 2021 - 33rd International Conference on Scientific and Statistical Database Management*, ACM, p. 203–208, Tampa, United States, July 2021, https://hal.inria.fr/hal-03503635.

O. Pivert, E. Scholly, G. Smits, V. Thion, "Fuzzy quality-aware queries to graph databases", *Information Sciences 521*, February 2020, p. 160–173, https://hal.inria.fr/hal-02484041.

O. Pivert, O. Slama, V. Thion, "Expression and efficient evaluation of fuzzy quantified structural queries to fuzzy graph databases", *Fuzzy Sets and Systems 366*, July 2019, p. 3–17, https://hal.inria.fr/hal-02444573.

G. Smits, O. Pivert, R. Yager, P. Nerzic, "A soft computing approach to big data summarization", *Fuzzy Sets and Systems 348*, October 2018, p. 4–20, https://hal.inria.fr/hal-01962961.

H. Van Tran, T. Allard, L. D'orazio, A. El Abbadi, "FRESQUE: A Scalable Ingestion Framework for Secure Range Query Processing on Clouds", *in : EDBT 2021 - 24th International Conference on Extending Database Technology*, Nicosia, Cyprus, March 2021, https://hal.inria.fr/hal-03198346.

V. Yepmo, G. Smits, O. Pivert, "Anomaly Explanation : A Review", *Data and Knowledge Engineering*, November 2021, https://hal.archives-ouvertes.fr/hal-03449887.

Articles in referred journals and book chapters

[1] P. Rigaux, V. Thion, "Exploration de partitions musicales modélisées sous forme de graphe", *Revue Ouverte Ingénierie des Systèmes d'Information*, à paraître 2022.

[2] V. Yepmo, G. Smits, O. Pivert, "Anomaly explanation: A review", *Data Knowl. Eng. 137*, 2022, p. 101946, https://doi.org/10.1016/j.datak.2021.101946.

Publications in Conferences and Workshops

[3] A. C. Anadiotis, F. Goasdoué, M. Y. Haddad, I. Manolescu, "Towards Speeding Up Graph-Relational Queries in RDBMSs", *in : BDA 2022 - 38èmes journées de la conférence BDA " Gestion de Données – Principes, Technologies et Applications*, Clermont-Ferrand, France, October 2022, https://hal.inria.fr/hal-03791272.

[4] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "Gestion de données efficace dans les bases de connaissances", *in : BDA 2022 - 38èmes journées de la conférence BDA "Gestion de Données – Principes, Technologies et Applications"*, Clermont-Ferrand, France, October 2022, https://hal.inria.fr/hal-03812670.

[5] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "OptiRef : optimisation pour la gestion de données dans les bases de connaissances", *in : BDA 2022 - 38èmes journées de la conférence BDA "Gestion de Données – Principes, Technologies et Applications"*, Clermont-Ferrand, France, October 2022, https://hal.inria.fr/hal-03812666.

[6] M. Georgoulakis, V. Kantere, L. D'orazio, "Multi-objective query optimization in Spark SQL", *in : IDEAS 2022 - International Database Engineered Applications Symposium*, Budapest, Hungary, August 2022, https://hal.inria.fr/hal-03925675.

[7] P. Nerzic, G. Smits, O. Pivert, M.-J. Lesot, "Exploration de données massives à l'aide d'estimations de cardinalités", *in : LFA 2022 - Rencontres francophones sur la*

logique floue et ses applications, Toulouse, France, October 2022, https://hal.inria.fr/hal-03777530.

[8] P. Nerzic, G. Smits, O. Pivert, M. Lesot, "Massive Data Exploration using Estimated Cardinalities", *in : IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padua, Italy, July 18-23, 2022*, IEEE, p. 1–8, 2022, https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882692.

[9] V. L. Nguyen Huu, J. Lallet, E. Casseau, L. d'Orazio, "Cache management in MASCARA-FPGA: from coalescing heuristic to replacement policy", *in : DaMoN 2022 - 18th International Workshop on Data Management on New Hardware*, ACM, p. 1–5, Philadelphia, United States, June 2022, https://hal.inria.fr/hal-03907912.

[10] P. Rigaux, V. Thion, "Towards a Graph-Oriented Perspective for Querying Music Scores", *in : Proceedings of the INFORSID conference*, Dijon, France, May 2022, https://hal.inria.fr/hal-03672224.

[11] G. Smits, M. Lesot, V. Y. Tchaghe, O. Pivert, "PANDA: Human-in-the-Loop Anomaly Detection and Explanation", *in : Information Processing and Management of Uncertainty in Knowledge-Based Systems - 19th International Conference, IPMU 2022, Milan, Italy, July 11-15, 2022, Proceedings, Part II*, D. Ciucci, I. Couso, J. Medina, D. Slezak, D. Petturiti, B. Bouchon-Meunier, R. R. Yager (editors), *Communications in Computer and Information Science*, *1602*, Springer, p. 720–732, 2022, https://doi.org/10.1007/978-3-031-08974-9_57.

[12] C. Wang, L. Gruenwald, L. d'Orazio, "SLA-Aware Cloud Query Processing with Reinforcement Learning-based Multi-Objective Re-Optimization", *in : DAWAK 2023 - International Conference on Data Warehousing and Knowledge Discovery*, Penang, Malaysia, August 2023, https://hal.inria.fr/hal-03925654.

[13] S. Yalles, M. Handaoui, J.-E. Dartois, O. Barais, L. d'Orazio, J. Boukhobza, "RISCLESS: A Reinforcement Learning Strategy to Guarantee SLA on Cloud Ephemeral and Stable Resources", *in : 2022 30th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, IEEE, p. 83–87, Valladolid, Spain, March 2022, https://hal.science/hal-03921309.

[14] V. Yepmo, "Anomaly Explanation", *in : Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L. D. Raedt (editor), ijcai.org, p. 5883–5884, 2022, https://doi.org/10.24963/ijcai.2022/844.