UMR IRISA

Activity Report 2024

Team SHAMAN

Symbolic and Human-centric view of dAta MANagement

D7 – Data & Knowledge Management

Université de Rennes | ENS rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

1 Team composition

Researchers and faculty

Amine Boulhamel, Temporary teaching and research assistant, ENSSAT, since Nov. 1 24
Wafaa El Husseini, Temporary teaching and research assistant, ENSSAT, from Jan. 1 to Aug. 24
François Goasdoué, Professor, ENSSAT, head of the team
Hélène Jaudoin, Associate Professor, ENSSAT
Ludovic Liétard, Associate Professor, HDR, IUT Lannion
Pierre Nerzic, Associate Professor, IUT Lannion
Laurent d'Orazio, Professor, IUT Lannion
Olivier Pivert, Professor, ENSSAT
Virginie Thion, Associate Professor, HDR, ENSSAT
Vérone Yepmo, Temporary teaching and research assistant, ENSSAT, from Jan. 1 to Aug. 24

PhD and MSc students

Adel Aly, PhD student, ENSSAT, since Oct. 23
Vincent Lannurien, PhD student, ENSTA Bretagne, Oct. 21 - Dec. 24
Chanattan Sok, Master student, ENS Rennes, Sept. 23 - May 24

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 that was funded by the Pôle d'Excellence Cyber. One of the main goals was to provide a Big Data platform applied to security monitoring. The team is still investigated this direction with an informal collaboration with Thales. We are considering 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 preserve and disseminate western classical music works for centuries. Nowadays, their content can be encoded in digital formats that yield a very detailed representation of the music content expressed in the language of *music notation*. These digitized music scores constitute, therefore, an invaluable asset for digital library services. In this context, 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. This axis involves collaborations with the Dastum association, a cultural organization based in Rennes (Brittany, France), whose mission is to collect, protect and promote the cultural heritage of Brittany, and with teachers of the traditional music department of the Music Conservatory of Lannion Trégor.

3 Scientific achievements

3.1 Big data management

Participants: Laurent d'Orazio, Vincent Lannurien, Chanattan Sok.

- *GUESS* [13] This paper proposes a monitoring system called GUESS to compare the performance and energy consumption of join query processing on Spark in Serverless and Serverful environments. The system collects metrics on resource utilization, query execution times, and power usage through Prometheus, Grafana, Spark History Server, and Open- Manage Enterprise Power Manager. These metrics are visualized through an intuitive web dashboard to enable easy comparison between Serverless and Serverful Spark workloads. Experimental results using the TPC-H benchmark show that the Serverless environment consumes less energy than the Serverful environment due to on-demand resource allocation. However, the Serverful environment exhibits better query performance, especially for workloads with known resource requirements. GUESS provides insights into optimizing resource efficiency and query performance when deploying Spark analytic workloads.

- *BLOSSOM* [12] Big data's impact is driving research into efficient solutions for managing growing datasets, with a focus on distributed systems. Recent advancements in query processing, particularly the join operator, have been significant. WebAssembly (Wasm), known for its efficiency, is increasingly adopted. This project aims to evaluate the efficiency of Wasm in serverless environments, addressing challenges posed by serverless architectures and comparing Wasmbased joins against native in performance. We propose Blossom, a Rust-based experimental platform, to give insights into Wasmbased joins. Our initial results reveal trade-offs between Wasm performance and its generality.

- *HeroCache* [10] Intrusion Detection Systems (IDS) are time-sensitive applications that aim to classify potentially malicious network traffic. IDSs are part of a class of applications that rely on short-lived functions that can be run reactively

and, as such, could be deployed on edge resources, to offload processing from energy-constrained battery-backed devices. The serverless service model could fit the needs of such applications, given that the platform allows adequate levels of Quality of Service (QoS) for a variety of users, since the criticality of IDS applications depends on several parameters. Deploying serverless functions on unreserved edge resources requires to pay particular attention to (1) initialization delays that could be significant on low resources platforms, (2) inter-function communication between edge nodes, and (3) heterogeneous devices. In this paper, we propose both a storage-aware allocation and scheduling policy that seek to minimize task placement costs for service providers on edge devices while optimizing QoS for IDS users. To do so, we propose a caching and consolidation strategy that minimizes cold starts and inter-function communication delays while satisfying QoS by leveraging heterogeneous edge resources. We evaluated our platform in a simulation environment using characterization data from real-world IDS tasks and execution platforms and compared it with a vanilla Knative orchestrator and a storage-agnostic policy. Our strategy achieves 18

- *QRLIT* [1] Selecting indexes capable of reducing the cost of query processing in database systems is a challenging task, especially in large-scale applications. Quantum computing has been investigated with promising results in areas related to database management, such as query optimization, transaction scheduling, and index tuning. Promising results have also been seen when reinforcement learning is applied for database tuning in classical computing. However, there is no existing research with implementation details and experiment results for index tuning that takes advantage of both quantum computing and reinforcement learning. This paper proposes a new algorithm called QRLIT that uses the power of quantum computing and reinforcement learning for database index tuning. Experiments using the database TPC-H benchmark show that QRLIT exhibits superior performance and a faster convergence compared to its classical counterpart.

- *VG-Prefetcher Cache* [8] The demand for efficient and reliable cloud computing systems is increasing. However, effectively managing data workloads in edge cloud systems, especially for connected cars, can be challenging. To address this issue, we have developed a new cache management technique named VG-Prefetcher Cache that uses visibility graphs to handle time series data more effectively. Our approach involves predicting future data and prefetching it into the cache, which reduces retrieval time and improves system performance. VG-Prefetcher Cache presents a promising approach for overcoming challenges in managing data workloads, thus paving the way for a more efficient and reliable cloud computing system.

- *Form-based semantic caching for TSD system* [11] Time Series Databases Management System (TSMS) has been overcoming the Database Management Systems (DBMS) in storing vast amounts of data [35]. Nevertheless, TSMS only supports simple aggregate functions to analyze Time Series Data (TSD). Besides, to accelerate and save data transferring between clients and servers in the DBMS, semantic caching can be used. However, the semantic caching approach is not efficient because of not fully supporting aggregate functions in TSMS. Further-

more, the query result of TSD in the semantic caching technique could be huge for the in-memory database where the semantic caching technique is running on. A model-based compression can be used to compress data, reducing the data space in the in-memory database. In this paper, we present Form-based semantic caching for TSD system. The approach reduces both query result storing based on semantic caching technique and the data transfer between clients and servers. In particular, the approach accelerates up to 122 and 1.82 times the execution speed, comparing to the without cache and basic semantic caching approaches, respectively. On the public Reference Energy Disaggregation Data Set, the compression model ratio in the approach can be reached to 526.8:1.

3.2 Flexible, cooperative and quality-aware data management

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

- *Data Partioning and Anomaly Detection Based on an Isolation Forest.* Understanding why some points in a data set are considered anomalies cannot be done without taking into account the structure of the regular points. Whereas many machine learning methods are dedicated to the identification of anomalies on one side, or to the identification of the data inner-structure on the other side, a solution is introduced in [6, 14] to address these two tasks using a same data model, a variant of an isolation forest. The original Isolation Forest algorithm is revisited so as to preserve the data inner structure without affecting the efficiency of the outlier detection. Experiments conducted both on synthetic and real-world data sets show that, in addition to improving the detection of abnormal data points, the proposed variant of isolation forest allows for a reconstruction of the subspaces of high density. Therefore, the former can serve as a basis for a unified approach to detect global and local anomalies, which is a necessary condition to provide users with informative descriptions of the data.

- *Music Information Retrieval* Several data model were proposed in the litterature for modelling a musical score (in [7], we proposed a review of the litterature of the data models implemented in Digital Score Libraries). The existing models enable the storage of musical scores but are not well-suited for data querying. In [4, 5], we proposed a new approach for the modelling and the storage of music scores. The specificity of our approach relies on the fact that we model the musical content of the music scores in the form of a graph [4, 5]. This model allows to use graph pattern query to express musical searchings into data. The proposed graph-based model has led to the implementation of the Skrid platform (see Section 4.2), leveraging the Neo4j graoh DBMS for the implementation of the model and the Cypher query language to express graph pattern queries searching into the musical content of the scores.

3.3 Ontology-based data management

Participants: Wafaa El Husseini, François Goasdoué, Hélène Jaudoin.

- Ontology-mediated query answering (OMQA) consists in asking database queries on knowledge bases (KBs); a KB is a set of facts called the KB's database, which is described by domain knowledge called the KB's ontology. A widely-investigated OMQA technique is FO-rewriting: every query asked on a KB is reformulated w.r.t. the KB's ontology, so that its answers are computed by the relational evaluation of the query reformulation on the KB's database. Crucially, because FO-rewriting compiles the domain knowledge relevant to queries into their reformulations, query reformulations may be complex and their optimization is the crux of efficiency. In [9], we devise a novel optimization framework for a large set of OMQA settings that enjoy FO-rewriting: conjunctive queries, i.e., the core select-project-join queries, asked on KBs expressed using datalog+/-, description logics, existential rules, OWL, or RDFS. We optimize the query reformulations produced by state-of-the-art FO-rewriting algorithms by computing rapidly, with the help of a KB's database summary, simpler (contained) queries with the same answers that can be evaluated faster by RDBMSs. We show on a well-established OMQA benchmark that time performance is significantly improved by our optimization framework in general, up to three orders of magnitude.

4 Software development

4.1 FuzViz

Participants: Pierre Nerzic, Grégory Smits.

FuzViz 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 The SKRID platform

Participants: Adel Aly, Vincent Barreaud, Olivier Pivert, Virginie Thion.

The SKRID platform is a digital score library that makes available some Traditional Breton music scores. It is a collaborative effort with Dastum [1], an cultural organiza-

[1] https://www.dastum.bzh/association/

tion dedicated to preserving and disseminating the cultural heritage of Brittany. This platform is available at `https://shaman.enssat.fr/skrid/`

4.3 Maelis

Participants: Adel Aly, Olivier Pivert, Virginie Thion.

Maelis is a flexible querying module, implemented in the Skrid platform, which allows melodic pattern approximate searching in the graoh-based music score databases of Skrid.

4.4 Musypher

Participants: Adel Aly, 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. It allows illustrating the relevancy (expressiveness, efficiency) of managing music scores over a graph-based data model. Musypher is used as the populating tool of the Skrid platform.

4.5 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).

4.6 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.

4.7 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.

4.8 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.

4.9 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.

4.10 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.

4.11 HeROfake

Participants: Vincent Lannurien, Laurent d'Orazio.

HeROfake is a heterogeneity-aware serverless orchestrator for private clouds that consists of two components: the autoscaler allocates heterogeneous hardware resources (CPUs, GPUs, FPGAs) for function replicas, while the scheduler maps function executions to these replicas. Our objective is to guarantee function response time, while enabling the provider to reduce resource usage and energy consumption.

4.12 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.

4.13 HeROSim

Participants: Vincent Lannurien, Laurent d'Orazio.

HeROsim is an open source simulation environment for the evaluation of serverless resources allocation and tasks scheduling policies. Its main goal is to relieve researchers from the implementation of discrete-event simulation boilerplate in the context of the dynamic process of cloud orchestration.

4.14 QRLIT

Participants: Diogo Barbosa, Laurent d'Orazio.

QRLIT is an algorithm that uses the power of quantum computing and reinforcement learning for database index tuning.

4.15 BLOSSOM

Participants: Chanattan Sok, Laurent d'Orazio.

BLOSSOM is a Rust-based experimental platform, to give insights into Wasm-based joins.

4.16 GUESS

Participants: Chanattan Sok, Laurent d'Orazio.

GUESS (monitorinG join qUery Execution in Serverless and Serverful spark) is a monitoring system called GUESS to compare the performance and energy consumption of join query processing on Spark in serverless and serverful environments.

5 Contracts and collaborations

5.1 International Initiatives

5.1.1 DODAM

Participants: Wafaa El Husseini, 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.

5.1.2 Advancing OT Simulations in Cyber Ranges

Participants: Laurent d'Orazio.

Advancing OT Simulations in Cyber Ranges is a B-Monde project, funded by Conseil Régional de Bretagne. The Cyber Innovation Hub, in collaboration with the University of Rennes, Brittany, seeks Agile Cymru funding to recruit a student at Cardiff University to inte-grate and advance existing Operational Technology (OT) simulation environments into Cyber Range platforms. Both universities currently use the commercial Thales (Diateam) Cyber Range, and students will explore its capabilities to simulate near-realistic OT environments. Additionally, the project will investigate integrating these environments into Ludus Cyber Range, an open-source alternative that leverages Ansible roles and templates to create complex OT networks. This approach offers a versatile foundation for developing and customizing OT simula-tions. Depending on the project timeline, student may also enhance these simu-lations by developing user interfaces or adding complexity to bring them closer to realistic OT infrastructures, using Siemens industry processes as a model. Upon completion, the University of Rennes will deploy these projects within their infra-structure to assess their effectiveness. The University of Rennes also plans to apply for B-monde funding to hire additional students, ensuring the continuation and expansion of this pioneering work. This initiative targets the Brittany region and aligns perfectly with Agile Cymru's goals of fostering cross-border innovation, advancing cybersecurity, and developing essential digital skills.

5.1.3 IDENTITY: Big Data management in multimedia exploration in virtual reality

Participants: Laurent d'Orazio.

IDENTITY is a PHC Jules Verne (with Iceland) project.The Icelandic team has been developing a novel approach to multimedia exploration in Virtual Reality (VR), which leads to database queries that require significant query processing resources. The French team has been studying data caching methods, under the umbrella term of Semantic

Caching, that hold a promise to significantly improve the query processing cost of the approach. The French team has also been studying several multimedia applications, particularly in the medical domain, which could benefit from the new approach to multimedia exploration. The goal of the exchange is thus twofold: (a) in the short term, to apply the multimedia exploration methodology to novel applications, and (b) in the long term, to study and quantify the impact of Semantic Caching on VR-based multimedia exploration.

5.2 National Initiatives

5.2.1 CQFD

Participants: Wafaa El Husseini, 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 SKRID

Participants: Olivier Pivert, Virginie Thion.

SKRID (2024) was a project funded by the University of Rennes, France (in the *Défis scientifiques* ("Scientific challenges") program). The propose of this interdisciplinary project was twofold: (i) developing a first version of the Skrid platform (see Section 4.2) and (ii) enhancing the collaboration of the Shaman team with the Skrid's intented users (members of the Dastum association, musicologists specilized in traditional music of Brittany, and teachers of the traditional music department of the Music Conservatory of Lannion Trégor).

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 Semantic Web Conference (ESWC)
- Extraction et Gestion de Connaissances (EGC)

- International Conference on Principles of Knowledge Representation and Reasoning (KR)
- International Joint Conference on Artificial Intelligence (IJCAI)
- The ACM Web Conference (WWW)

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

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

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

- International Conference on Scientific and Statistical Database Management (SSDBM)
- International Conference on Computer Communications and Networks (ICCCN)
- International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)
- 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 Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS)
- Conférence Extraction et Gestion des Connaissances (EGC), démonstrations

Virginie Thion served as a member of the program committees of the Conférence Extraction et Gestion des Connaissances (EGC).

Reviewer

Hélène Jaudoin served as a reviewer for the International Conference on Scientific and Statistical Database Management 2024.

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
- Ingénierie des Systèmes d'Information.

Reviewer - Reviewing Activities Laurent d'Orazio served as a reviewer of the following journals:

- Information Sciences
- Expert Systems With Applications (ESWA)
- IEEE Internet Computing
- Transactions on Large-Scale Data and Knowledge-Centered Systems (TLDKS)
- IEEE Transactions on Parallel and Distributed Systems (TPDS)

6.1.3 Invited Talks

Laurent d'Orazio has been invited to give a keynote presentation at Museum Big Data (MBD), Athens, Greece 2024.

6.1.4 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).

François Goasdoué is a member of the project team for the new CNRS GDR MaDICS, from 2023 to 2024.

François Goasdoué is a member of the Steering Committee of the new CNRS GDR MaDICS, from 2024 to 2029.

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).

6.1.5 Scientific Expertise

François Goasdoué is an expert for the Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (Hcéres).

Laurent d'Orazio is an expert for the Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (Hcéres).

Laurent d'Orazio is an expert for the Direction générale de la recherche et de l'innovation (DGRI).

Olivier Pivert is an expert for the Czech Science Foundation.

6.1.6 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 council of ENSSAT, since 2022.

Laurent d'Orazio is the co-head of the International Relationship office of IRISA UMR 6074, since 2022.

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: Olivier Pivert teaches a course on *AI for Database Querying*, 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: Adel Aly, Storage and Querying of Musical Score Databases, advisor: Virginie Thion.
- PhD in progress: Vincent Lannurien, Big data applications scheduling on heterogeneous Cloud resources, advisors: Laurent d'Orazio, Jalil Boukhobza and Olivier Barais.

6.2.3 Juries

François Goasdoué

- HDR, referee, Mourad Ouziri, Université de Paris

Laurent d'Orazio

- PhD thesis, referee, Roxane Jouseau, Université Clermont Auvergne
- PhD thesis, referee, Léa El Ahdab, Université de Toulouse
- PhD thesis, referee, Sébastien Rivault, Université d'Orléans
- PhD thesis, referee, Qi Fan, Ecole Polytechnique
- PhD thesis, referee, Alexis Guyot, Université de Bourgogne

6.3 Popularization

Adel Aly was invited to present the Skrid platform querying process at the *Ton Air* conference, associated with the musical NoBorder festival. [2]

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.

W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "Query Optimization for Ontology-Mediated Query Answering", *in : The ACM Web Conference (WWW)*, Singapore, Singapore, May 2024, https://hal.science/hal-04470002.

[2] https://www.festivalnoborder.com/TON-AIR-A-BREST.html

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. Lannurien, C. Slimani, L. d'Orazio, O. Barais, S. Paquelet, J. Boukhobza, "HeROcache: Storage-Aware Scheduling in Heterogeneous Serverless Edge - The Case of IDS", *in : CCGrid 2024 - 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing*, p. 1–11, Philadelphia, United States, May 2024, https://hal.science/hal-04571484.

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.

P. Rigaux, V. Thion, "Topological Querying of Music Scores", *Data and Knowledge Engineering 153*, 2024, p. 102340, https://inria.hal.science/hal-04614440.

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] D. Barbosa, L. Gruenwald, L. D. Orazio, J. Bernardino, "QRLIT: Quantum Reinforcement Learning for Database Index Tuning", *Future internet 16*, 12, November 2024, https://cnrs.hal.science/hal-04837643.

[2] F. Foscarin, P. Rigaux, V. Thion, "Data Quality Assessment in Digital Score Libraries. The GioQoso Project", *International Journal on Digital Libraries 22*, 2, 2021, p. 159–173, https://hal.science/hal-03163156.

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

[4] P. Rigaux, V. Thion, "Exploration de partitions musicales modélisées sous forme de graphe", *Revue ouverte d'ingénierie des systèmes d'information 4*, 2, 2024, https://hal.science/hal-04464885.

[5] P. Rigaux, V. Thion, "Topological Querying of Music Scores", *Data and Knowledge Engineering 153*, 2024, p. 102340, https://inria.hal.science/hal-04614440.

[6] V. Yepmo, G. Smits, M.-J. Lesot, O. Pivert, "Leveraging an Isolation Forest to Anomaly Detection and Data Clustering", *Data and Knowledge Engineering 151*, March 2024, p. 102302, https://hal.science/hal-04516593.

Publications in Conferences and Workshops

[7] A. Aly, O. Pivert, V. Thion, "Database Approaches to the Modelling and Querying of Musical Scores: a Survey", *in : Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'24)*, Ljubljana, Slovenia, September 2024, https://hal.science/hal-04681804.

[8] A. Bensalem, L. d'Orazio, J. Lallet, A. Enrici, "VG-Prefetcher Cache: Towards Edge-Based Time Series Data Management Using Visibility Graph Prefetching", *in : International Conference on Scientific and Statistical Database Management (SSDBM)*, *35*, ACM, p. 1–4, Rennes France, France, July 2024, https://hal.science/hal-04722876.

[9] W. El Husseini, C. B. El Vaigh, F. Goasdoué, H. Jaudoin, "Query Optimization for Ontology-Mediated Query Answering", *in : Proceedings of the ACM Web Conference 2024*, Singapore, Singapore, May 2024, https://hal.science/hal-04470002.

[10] V. Lannurien, C. Slimani, L. d'Orazio, O. Barais, S. Paquelet, J. Boukhobza, "HeROcache: Storage-Aware Scheduling in Heterogeneous Serverless Edge - The Case of IDS", *in : CCGrid 2024 - 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing*, p. 1–11, Philadelphia, United States, May 2024, https://hal.science/hal-04571484.

[11] T.-D. Le, V. Kantere, L. d'Orazio, "Form-based semantic caching on time series", *in : International Conference on Computational Science and Its Applications (ICCSA)*, Hanoi (Vietnam), France, July 2024, https://hal.science/hal-04723263.

[12] C. Sok, L. d'Orazio, R. Tekin, D. Tombroff, "WebAssembly serverless join: A Study of its Application", *in : International Conference on Scientific and Statistical Database Management (SSDBM)*, *Lecture Notes in Computer Science*, *14813*, ACM, p. 1–4, Rennes France, France, July 2024, https://hal.science/hal-04722875.

[13] A.-T. Tran Phan, L. d'Orazio, T.-C. Phan, L. Gruenwald, "GUESS: monitorinG join qUery Execution in Serverless and Serverful spark", *in : International Conference on Database Systems for Advanced Applications (DASFAA)*, Gifu, Japan, July 2024, https://hal.science/hal-04544429.

[14] V. Yepmo, G. Smits, M.-J. Lesot, O. Pivert, "CADI: Contextual Anomaly Detection using an Isolation Forest", *in : The 39th ACM/SIGAPP Symposium On Applied Computing*, Avila, Spain, April 2024, https://hal.science/hal-04390676.