IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2013

Project-Team OAK

Database optimizations and architectures for complex large data

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team OAK

Keywords: Data Management, Reasoning, Semantics, Web, Cloud Computing, Distributed System

Creation of the Team: 2012 April 01*, updated into Project-Team:* 2013 January 01.

1. Members

Research Scientist

Ioana Manolescu [Team leader, Inria, Senior Researcher, HdR]

Faculty Members

Nicole Bidoit [UNIV. PARIS-SUD, Professor, HdR]
François Goasdoué [UNIV. PARIS-SUD, Associate Professor, until Aug 2013; Univ. Rennes I, Professor, from Sep 2013, HdR]
Melanie Herschel [UNIV. PARIS-SUD, Associate Professor]
Bogdan Cautis [UNIV. PARIS-SUD, Professor, from Sep 2013, HdR]
Dario Colazzo [Univ. Paris IX, Associate Professor, until Aug 2013; Univ. Paris IX (Dauphine), Professor, from Sep 2013, HdR]

External Collaborators

Sofoklis Floratos [Ecole Polytechnique, FP7 EIT ICT LABS GA project, from Apr 2013]
Konstantinos Karanasos [Post-doc, IBM Almaden, USA]

Engineers

Andrés Aranda Andujar [Inria, until Oct 2013]
Benjamin Djahandideh [Inria, from Oct 2013]
Tushar Indranath Ghosh [Inria, until Dec 2013]

PhD Students

Asterios Katsifodimos [Inria, PhD student, until Jul 2013]
Julien Leblay [UNIV. PARIS-SUD, until Sep 2013]
Raphael Bonaque [Inria, from Oct 2013]
Jesús Camacho Rodríguez [UNIV. PARIS-SUD]
Yifan Li [East China Normal University, in Shanghai, China, from Oct 2013]
Ioana-Alexandra Roatis [UNIV. PARIS-SUD]
Aikaterini Tzompanaki [UNIV. PARIS-SUD]
Stamatios Zampetakis [Inria]

Post-Doctoral Fellows

Zoi Kaoudi [Inria until Jul 2013, UNIV. PARIS-SUD from Aug to Oct 2013]
Francesca Bugiotti [Inria, from Nov 2013]
Soudip Roy Chowdhury [Inria, from Sep 2013]

Visiting Scientist

Alin Deutsch [UCSD, Professor, in Jul 2013]

Administrative Assistant

Maëva Jeannot [Inria]

Others

Damian Bursztyn [Inria, Intern from Univ. Buenos Aires, Argentina, Mar-Sep 2013]
Prachi Jain [Inria, undergraduate student at IIT Bombay, from May 2013 until Aug 2013]
Varun Malhotra [Inria, Master student at IIT Bombay, from May 2013 until Aug 2013]
Mathilde Verrier [UNIV. PARIS-SUD, undegraduate intern, from Jun 2013 until Jul 2013]

2. Overall Objectives

2.1. Highlights of the Year

The year has seen the finalization of five major research activities: XML materialized view maintenance [4], XML static type analysis [5], document management through semantic annotations [7], scalable dissemination of Web data [8], and XML type-based projection [13]. The respective five publications appeared in A* journals (according to the CORE ERA ranking).

3. Research Program

3.1. Scalable and Expressive Techniques for the Semantic Web

The Semantic Web vision of a world-wide interconnected database of *facts*, describing *resources* by means of *semantics*, is coming within reach as the W3C's RDF (Resource Description Format) data model is gaining traction. The W3C Linking Open Data initiative has boosted the publication and interlinkage of a large number of datasets on the semantic web resulting to the Linked Open Data Cloud. These datasets of billions of RDF triples have been created and published online. Moreover, numerous datasets and vocabularies from different application domains are published nowadays as RDF graphs in order to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. RDF storage, querying, and reasoning is now supported by a host of tools whose scalability and expressive power vary widely. Unsurprisingly, some of the most scalable tools draw upon the existing models and architecture for managing structured data. However, such tools often ignore the semantic aspects that make RDF interesting. For what concerns the semantics, a delicate balance must be found between expressive power and the efficiency of the resulting data management algorithms.

- The team works on identifying tractable dialects of RDF, amenable to highly efficient query answering algorithms, taking into account both data and semantics.
- Another line of research investigates the usage of RDF data and semantics to help structure, organize, and enrich other kinds of data, and in particular structured documents. The newly started DIGICOSME LabEx grant "Structured, Social and Semantic Search" is part of this research.
- Last but not least, we investigate novel models and algorithms for efficient Semantic Web data management, going beyond the existing standard languages. In particular, we study formal, flexible models for an all-RDF data analytics framework, combining the rich structure and semantics of RDF with the power of analysis tools previously developed for relational data, such as analytical schemas and queries. This work is related to our DIGITEO grant "Data Warehouses for RDF" (DW4RDF) and will continue as part of our recently started "Investissement d'Avenir" project Datalyse.

3.2. Massively Distributed Data Management Systems

Large and increasing data volumes have raised the need for distributed storage architectures. Among such architectures, computing in the cloud is an emerging paradigm massively adopted in many applications for the scalability, fault-tolerance and elasticity features it offers, which also allows for effortless deployment of distributed and parallel architectures. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment. For these reasons, cloud-based stores are an interesting avenue to explore for handling very large volumes of RDF data.

Our research aims at taking advantage of such widely available, large-scale distributed architectures to build scalable platforms for massively distributed management of complex data. We consider many different wide-scale distributed back-ends in this context, ranging from those provided by commercial cloud platforms to simple MapReduce and to more complex extensions thereof. Beyond these architectures that are characterized by a single master node (a single point of control and distribution), we also explored ad-hoc, peer-to-peer style data management, which is more suitable in certain contexts, in particular for disseminating high-velocity data based on the similarity of interests among peers.

This line of research is part of our participation to the Datalyse project previously mentioned, as well as the KIC EIT ICT Labs Europa activity, now in its third year, part of the "Computing in the Cloud" action line.

A recent development in this area is the start of our collaboration with social scientists from UNIV. PARIS-SUD, working on the management of innovation; we have started two collaborative research projects (ANR "Cloud-Based Organizational Design" and PEPS "Business Models for the Cloud") where we seek to build an interdisciplinary approach (both from a computing and from a business management perspective) on the adoption of cloud technologies within an enterprise.

3.3. Advanced Algorithms for Efficient XML processing

The development of Web technologies has led to a strong increase in the number and complexity of the applications which represent their data in Web formats, among which XML is used for structured documents. To manipulate very large volumes of XML data in a declarative fashion, the XQuery XML query language has been standardized by the W3C and is by now quite widely supported in industrial systems and research prototypes. The XQuery language allows expressing highly complex queries featuring complex navigation, joins, and nesting; the latest XQuery 3.0 has been extended with powerful grouping functionalities, too. For all these reasons, the *efficient* evaluation of XQuery queries and updates on large XML databases remains a challenge.

To address this challenge, the team specializes in two orthogonal performance enhancement techniques. The first one concerns the optimization of XML stores, in order to reduce as much as possible one of the main components of query evaluation cost, namely accessing the data. The second is static analysis of queries and updates, based on type systems; from a performance perspective, such static analysis techniques allow increasing parallelism, detecting operations whose results are not needed and thus whose evaluation can be omitted, etc.

3.4. Data Transformation Management

With the increasing complexity of data processing queries, for instance in applications such as relational data analysis or integration of Web data (e.g., XML or RDF) comes the need to better manage complex data transformations. This includes systematically verifying, maintaining, and testing the transformations an application relies on. In this context, Oak has focused on verifying the semantic correctness of a declarative program that specifies a data transformation query, e.g., an SQL query.

3.5. Social Data Management

While progress has been made in the area of personalized search in social applications, more remains to be done in order to address users' needs in practice. The social Web blurs today the distinction between search, recommendation, and advertising (three paradigms for information access that have been so far considered mostly in separation). Our research in this area strives to find better adapted and scalable ways to answer information needs in the social Web, often by techniques at the intersection of databases, information retrieval, and data mining.

4. Application Domains

4.1. Business Intelligence for Open Data

Research developed in the group helps publish, curate, and exploit open data, in particular the data produced by local or national administrations and which is returned to the general public under the form of applications (often Web-based, often mobile) which increase opportunities for business or leisure. This concerns in particular our work on Open Data entity resolution [36] and Open Data warehousing [33], [32]. This research is set to be deployed on real Open Data sets from the Grenoble urban area, within the industry-led Datalyse project (Section 7.2.1).

4.2. Social Data Management

We develop models and algorithms for efficiently exploiting, enhancing, and querying social network data, in particular based on structured content, semantic annotations, and user interaction networks. We pursue this research with many industrial partners within the ALICIA project (Section 7.2.1) as well as in the Structured, Social, and Semantic Search project (Section 7.2.2).

4.3. Data Journalism

Efficiently handling the deluge of news and other news-worthy electronic data being published today, requires powerful content management tools in order to handle news document structure, extract meaning from the text, connect pieces of information with each other, etc. To that purpose, we have built and experimented with FactMinder, a platform for gathering, enriching, annotating, storing, and querying news documents, with the help of existing ontologies that users may enrich and/or exploit next to their own [24]. Many more applications of our research are possible in this domain [38].

4.4. Data Transformation Debugging

All applications mentioned above, e.g., business intelligence, data integration, or data enrichment in social data management or data journalism take as input some data to be further manipulated and transformed. In many applications, including again business intelligence and data journalism, the correctness of the produced output data is crucial. It is thus important to verify the semantic correctness of a data transformation and to be able to trace back what has happened to the data within the transformation. We support this data transformation debugging based on provenance [31], [26].

5. Software and Platforms

5.1. Amada

Name: Amada (https://team.inria.fr/oak/amada/)

Contact: Jesús Camacho-Rodríguez (jesus.camacho-rodriguez[at]inria.fr))

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Dario Colazzo (dario.colazzo[at]dauphine.fr), François Goasdoué (fg[at]irisa.fr)

Presentation: A platform for Web data management in the Amazon cloud.

5.2. FactMinder

Name: FactMinder (http://tripleo.saclay.inria.fr/xr/demo/)

Contact: Julien Leblay (julien.leblay[at]inria.fr)

Other contacts: Stamatis Zampetakis (stamatis.zampetakis[at]inria.fr), François Goasdoué (fg[at]irisa.fr), Ioana Manolescu (ioana.manolescu[at]inria.fr)

Presentation: A system for archiving, annotating, and querying semantic-rich Web content.

5.3. Nautilus Analyzer

Name: Nautilus Analyzer (http://nautilus.saclay.inria.fr/)

Contact: Melanie Herschel (melanie.herschel[at]lri.fr)

Other contacts: n.a.

Presentation: A tool for analyzing and debugging SQL queries using why-provenance and why-not provenance.

5.4. RDFViewS

Name: RDFViewS (http://tripleo.saclay.inria.fr/rdfvs/)

Contact: Konstantinos Karanasos (kkaranasos[at]gmail.com)

Other contacts: François Goasdoué (fg[at]irisa.fr), Julien Leblay (julien.leblay[at]gmail.com), and Ioana Manolescu (ioana.manolescu[at]inria.fr)

Presentation: A storage tuning wizard for RDF applications.

5.5. ViP2P

Name: ViP2P (Views in Peer-to-Peer, http://vip2p.saclay.inria.fr)

Contact: Ioana Manolescu (ioana.manolescu[at]inria.fr)

Other contacts: Jesús Camacho-Rodríguez (jesus.camacho-rodriguez[at]inria.fr)

Presentation: A P2P platform for disseminating and querying XML and RDF data in large-scale distributed networks.

5.6. WARG

Name: WARG (https://team.inria.fr/oak/warg/)

Contact: Alexandra Roatiş (alexandra.roatis[at]lri.fr))

Other contacts: Ioana Manolescu (ioana.manolescu[at]inria.fr), Dario Colazzo (dario.colazzo[at]dauphine.fr), François Goasdoué (fg[at]irisa.fr)

Presentation: A platform for specifying and exploiting warehouses of RDF data.

5.7. XUpOp

Name: XUpOp (XML Update Optimization)

Contact: Dario Colazzo (dario.colazzo[at]dauphine.fr)

Other contacts: Nicole Bidoit (bidoit[at]lri.fr), Mohamed Amine Baazizi (baazizi[at]lri.fr)

Presentation: A general purpose type-based optimizer for XML updates.

5.8. XUpIn

Name: XUpIn (XML Update Independence)

Contact: Federico Ulliana (Federico.Ulliana[at]lri.fr)

Other contacts: Dario Colazzo (colazzo[at]lri.fr), Nicole Bidoit (bidoit[at]lri.fr)

Presentation: An XML query-update independence tester.

5.9. XUpTe

Name: XUpTe (XML Update for Temporal Documents)

Contact: Dario Colazzo (dario.colazzo[at]dauphine.fr)

Other contacts: Nicole Bidoit (bidoit[at]lri.fr), Mohamed-Amine Baazizi (amine.baazizi[at]gmail.com)

Presentation: A type-based optimizer for representing and updating XML temporal data.

5.10. XPUQ

Name: XPUQ (XML Partitioning for Updates and Queries)

Contact: Dario Colazzo (dario.colazzo[at]dauphine.fr)

Other contacts: Nicole Bidoit (bidoit[at]lri.fr)

Presentation: A static analyzer and partitioner for XML queries and updates.

6. New Results

6.1. Scalable and Expressive Techniques for the Semantic Web

The team has continued developing expressive models and scalable algorithms for exploiting Semantic Web data, in particular RDF graphs, as well as rich corpora consisting of Web documents with semantic annotations.

We have studied efficient algorithms for answering RDF queries in the presence of schema (or semantic) constraints such as described through the RDF Schema language. The difficulty here consists of efficiently taking into account the data that is implicitly present in the RDF database due to semantic constraints, and which needs to be reflected in query results. We have identified the expressive database fragment of RDF, which extends previously identified fragments of the RDF specification by allowing more expressive schema and queries, and provided novel efficient algorithms for answering Basic Graph Pattern queries (a popular dialect of the standard SPARQL query language) over RDF graphs pertaining to the RDF Database Fragment. Our query answering algorithms take advantage of the processing power of a relational database management system while also reflecting RDF semantics [25].

The ability to exploit large corpora of heterogeneous RDF data requires tools for analyzing RDF content through the lenses of a specific user perspective, or user need. Such tools are commonplace in the context of relational data management, where data warehousing is a well-developed area, but lack completely in the realm of RDF. We have proposed a novel framework for building and exploiting all-RDF data warehouses [33] and have implemented this framework in a proof-of-concept platform [32]. A main contribution of this work is to preserve RDF graph structure, heterogeneity, and rich semantics from the base data to the analytical schema and analytical schema instance. Thus, our proposal is the first to allow the analysis of rich Semantic Web (RDF) data while preserving its rich content and semantics. For more information on this project, see https://team.inria.fr/oak/warg/.

We have investigated the usage of semantics as a way to enrich, interconnect, and interpret rich corpora of Web data. In particular, within the XR project, we had proposed in prior work the XR (XML+RDF) data model which integrates XML documents and RDF triples treating both as first-class citizens. One particular use of XR is to annotate nodes in XML documents, by RDF triples which may for instance describe their properties or state how nodes are semantically related to some concept or to each other. In [18] we describe the data model and core query language, make a comprehensive analysis of query evaluation algorithms, and describe extensive experiments carried within a fully implemented platform, as part of the PhD thesis of J. Leblay [12]. The XR platform was put to task in an application context related to digital journalism, where an XR content warehouse is continuously enriched through document analysis and annotation. This scenario has lead to a software demonstration [24], [35] and a keynote tutorial [38]. In collaboration with A. Deutsch, we have extended the XR query language and provided query-view composition algorithms in [41].

6.2. Massively Distributed Data Management Systems

Our work on the AMADA platform has shown how the different sub-systems of a popular cloud platform (namely, Amazon Web Services, or AWS in short) can be harnessed to build scalable stores and query evaluation engines for XML and RDF data. In [23], we propose and compare several storage and indexing strategies within AWS, and show that they help reduce not only query evaluation time but also the monetary costs associated to the exploitation of the AWS-based store, since the index helps direct queries only to the subsets of the data likely to have results for the query. Thus, the total effort (and the costs charged by AWS) in relation to the processing of a given query are reduced. A similar study focused mostly on RDF data management appears as a book chapter [40]. More information can be found at http://cloak.saclay.inria.fr/research/amada/.

Semantic Web data collections, that is, RDF graphs, may be very voluminous since RDF natively enables connections between different RDF databases (which may have been produced independently and in ignorance of each other) through the usage of common URIs (resource identifiers) in two or more databases. To scale up to such large volumes, we have developed CliqueSquare, a novel platform for storing and querying RDF graphs in a MapReduce-based architecture such as Hadoop. We have described the storage and query algorithm in [34]. Our analysis of existing frameworks and algorithms for managing large RDF graphs in a highly distributed environment has lead to the tutorial [27].

Large-scale distributed processing of complex data was considered from a different perspective in our Delta project. Here, we considered the setting where one data source publishes new data items at a very high rate, and numerous clients subscribe to some of the updates by means of queries that must be matched by the published items. In this setting, the source may quickly become the bottleneck due to limitations in its capacity to match the published item against the subscription and/or to send the matching updates. We propose a fully automated approach for distributing the data dissemination effort across the network of subscribers, by identifying some which act as secondary data sources for others, in a peer-to-peer fashion. This distributed dissemination network is chosen so as to optimize a combination of overall dissemination costs and data propagation latency; since the space of options has daunting complexity, approximate algorithms involving Binary Integer Programming techniques were proposed in [20], [37], [42], and concluded in the PhD thesis of A. Katsifodimos [11].

6.3. Advanced Algorithms for Efficient XML processing

In 2013, several research works of the team focusing on advanced algorithms for processing XML data have been finalized and concluded through prestigious journal publications.

A first line of work concerned the usage of materialized views to speed up the evaluation of complex XML queries. In our previous work we had demonstrated that such views may bring up very significant speed-up factors of several orders of magnitude. However, materialized views need to be kept up to date when the underlying database changes. In [14] we have described efficient algorithms for updating materialized views expressed in a rich dialect of XQuery, the standard query language for XML.

A second class of work was concerned with XML static type analysis, in particular with the crucial problem of deciding XML type inclusion, that is: whether any XML tree of type τ_1 is also of type τ_2 where τ_1, τ_2 are XML types with interleaving and counting (currently adopted by main stream schema languages). For these types, inclusion is EXPSPACE-complete. We have defined and formally studied a quadratic subtype-checking algorithm for the case where the right-hand side type τ_2 meets some restrictions on symbol occurrences and the use of counting. These restrictions are often met by human-designed types, so our technique perfectly fits the needs of typical XML type-checking algorithms, which frequently require to check for inclusion a machine-generated subtype τ_1 against a human-defined supertype τ_2. Our approach has been validated by extensive experimental results [16]. In addition, we have devised and formally studied an alternative algorithm, still for the asymmetric case where τ_2 is restricted, based on structural, top-down analysis of types expression. This algorithm is almost linear: it has a linear-time backbone, and resorts to the above quadratic approach for some specific parts of the compared types. Our experiments show that this new algorithm is much faster than the

quadratic one and that it typically runs in linear time, hence it can be used as a building block for a practical type-checking compiler for XML programs and queries [15].

Third, we have completed and concluded our work on type-based document projection for efficient XML data management. The idea here is to restrict XML documents, prior to evaluating a query over them, to only those parts of the document that the query actually needs to consult. We provide algorithms for determining such document parts and experimentally demonstrate the benefits of such techniques, in [13].

Finally, we have devised a system that is able to process both queries and updates on very large XML documents [22]. As observed in recent works, such very large documents are generated and processed in several contexts, in particular in those involving scientific data and logs. Our system supports a large fragment of XQuery and XUF (XQuery Update Facility). The system exploits dynamic and static partitioning to distribute the processing load among the machines of a MapReduce cluster. The proposed technique applies when queries and updates are iterative, i.e., they iterate the same query/update operations on a sequence of subtrees of the input document. From our experience many real world queries and updates actually meet this property. Our partitioning technique is schema-less, as the presence of a user-supplied schema is not required; indeed, this technique only relies on path information extracted from the input query/update. Experiments conducted on a 8-machine Hadoop cluster have demonstrated that the system is able run both iterative queries and updates on quite large documents.

6.4. Data Transformation Management

With the increasing complexity of data processing queries, for instance in applications such as relational data analysis or integration of Web data (e.g., XML or RDF) comes the need to better manage complex data transformations. This includes systematically verifying, maintaining, and testing the transformations an application relies on. In this context, Oak has focused on verifying the semantic correctness of a declarative program that specifies a data transformation query, e.g., an SQL query. To this end, we have investigated how to leverage data provenance (the information of the origin of data and the query operators) for query debugging. More specifically, we developed and implemented novel algorithms to explain why data is missing from the result of a relational query. As opposed to our previous work, which produced explanations based on the available source data, our new algorithms return explanations based on query operators [31] or both [26].

6.5. Social Data Management

We considered top-k query answering in social tagging systems, also known as folksonomies, a problem that requires a significant departure from existing, socially agnostic techniques. In a network-aware context, one can and should exploit the social links, which can indicate how users relate to the seeker and how much weight their tagging actions should have in the result build-up. Our solutions addressed the main drawbacks of previous approaches. With respect to applicability and scalability, we avoid expensive and hardly updatable pre-computations of proximity values. With respect to efficiency, we show that our algorithm is instance optimal in the existing techniques. Our main results in this direction have been presented recently in [29], [28], [21].

7. Partnerships and Cooperations

7.1. Regional Initiatives

Data Warehousing for RDF (DW4RDF) is a 3-year project sponsored by the Digiteo foundation, between Inria and UNIV. PARIS-SUD. The project aims at defining and deploying a full framework for RDF data analytics, supporting its inherent structural heterogeneity and semantics, while at the same time providing powerful analytic tools for summarizing and analyzing the data. The project supports the PhD of Alexandra Roatiş.

7.2. National Initiatives

7.2.1. ANR

Apprentissage Adaptatif pour le Crowdsourcing Intelligent et l'Accès à l'Information (ALICIA) is a 3.5-year project, starting in February 2014, supported by the ANR CONTINT call. The project is coordinated by Bogdan Cautis, with Nicole Bidoit, Melanie Herschel, and Ioana Manolescu. Its goal is to study models, techniques, and the practical deployment of adaptive learning techniques in user-centric applications, such as social networks and crowdsourcing.

Cloud-Based Organizational Design (CBOD) is a 4-year project accepted by the ANR in 2013 and is currently under financial negotiation. The project is coordinated by prof. Ahmed Bounfour from UNIV. PARIS-SUD. Its goal is to study and model the ways in which cloud computing impacts the behavior and operation of companies and organizations, with a particular focus on the cloud-based management of data, a crucial asset in many companies.

Datalyse is funded for 3.5 years as part of the *Investissement d'Avenir - Cloud & Big Data* national program. The project is led by the Grenoble company Eolas, a subsidiary of Business & Decision. It is a collaboration with LIG Grenoble, U. Lille 1, U. Montpellier, and Inria Rhône-Alpes aiming at building scalable and expressive tools for Big Data analytics.

7.2.2. LabEx, IdEx

Structured, Social and Semantic Search is a 3-year project started in October 2013, financed by the *LabEx (Laboratoire d'Excellence)* DIGICOSME. The project aims at developing a data model for rich structured content enriched with semantic annotations and authored in a distributed setting, as well as efficient algorithms for top-k search on such content.

BizModel4Cloud is a one-year interdisciplinary research project funded under a *Projet Exploratoire Premier Soutien (PEPS)* call joint between the CNRS and the IdEx Paris Saclay. It reunites the same partners as the ANR CBOD project of which it is an initial, short version.

7.3. European Initiatives

7.3.1. Collaborations in European Programs, except FP7

Program: COST

Project acronym: Keystone

Project title: Semantic keyword-based search on structured data sources

Duration: Oct 2013 – Oct 2018

Coordinator: Francesco Guerra (U. Modena, Italy)

Other partners: The project involves 24 countries, see http://www.cost.eu/domains_actions/ict/Actions/IC1302?parties

Abstract: To build efficient and expressive keyword search tools, the action "semantic KEYword-based Search on sTructured data sOurcEs" (KEYSTONE) proposes to draw upon competencies from several disciplines, such as semantic data management, the semantic web, information retrieval, artificial intelligence, machine learning, user interaction, service science, service design, and natural language processing.

Program: KIC EIT ICT Labs "Computing in the Cloud" Action Line

Project acronym: Europa

Project title: Massively Parallel Data Management

Duration: Jan 2013 – Dec 2013

Coordinator: Volker Markl (TU Berlin, Germany)

Other partners: UNIV. PARIS-SUD (France), Aalto Univ (Finland), InternetMemory (France)

Abstract: Europa focuses on massively parallel algorithms and platforms for data management in the cloud. At TU Berlin, the Stratosphere open-source platform has been developed as part of this activity. At Inria and UNIV. PARIS-SUD, our work has focused on developing the AMADA platform for efficiently exploiting Web data in the Amazon cloud.

7.4. International Initiatives

7.4.1. Inria Associate Teams

7.4.1.1. OakSaD

Title: Languages and techniques for efficient large-scale Web data management

Inria principal investigator: Ioana Manolescu

International Partner (Institution - Laboratory - Researcher):

University of California San Diego (United States) - Computer Science and Engineering - Ioana Manolescu

Duration: 2013 - 2015

See also https://team.inria.fr/oak/oaksad/

Data on the Web is increasingly large and complex. The ways to process and share it have also evolved, from the classical scenario where users connect to a database, to today's complex processes whereas data is jointly produced on the Web, disseminated through streams, corroborated and enriched through annotations, and exploited through complex business processes, or workflows. The OAK and San Diego teams work together to devise expressive languages, efficient techniques and scalable platforms for such applications. The main areas on which our interest is shared are: semantic Web annotations; large-scale distributed data sharing; monitoring and verification of automated data processing workflows in the cloud.

7.4.2. Inria International Partners

7.4.2.1. Informal International Partners

We collaborate closely with TU Berlin within the Europa KIC EIT ICT Labs Europa project; A. Katsifodimos moved there for his post-doc after completing his PhD in OAK.

We have collaborated significantly with researchers from the University of Pisa and University of Basilicata [15], [16].

7.5. International Research Visitors

7.5.1. Visits of International Scientists

Participant: Stefano Ceri.

Date: September 2013

Institution: Politecnico di Milano, Italy.

Participant: Alin Deutsch.

Date: July-August 2013

Institution: UCSD, USA.

7.5.1.1. Internships

Damian Alexis Bursztyn

Subject: Take What You Need: Efficiently Querying Semantic Web Data

Date: from Mar 2013 until Aug 2013

Institution: University of Buenos Aires (Argentina).

Varun Malhotra

Subject: Task factorization for PACT programs on semistructured data

Date: from May 2013 until Aug 2013

Institution: IIT Delhi (India).

8. Dissemination

8.1. Scientific Animation

Journal editorial responsibilities

- I. Manolescu has been the editor-in-chief of the ACM SIGMOD Record and an associate editor of the ACM Transactions on Internet Technologies. She has also been a member of the editorial board of PVLDB. Further, she has been the guest editor of a special issue of the *Ingénierie et Science Informatique (ISI)* journal vol. 18, no. 4, 2013.

Program committees memberships

- Nicole Bidoit served in the program committee of ADBIS 2013, the 7th East-European Conf. on Advances in Databases and Information Systems, Genoa (Italy).
- Bogdan Cautis served as program committee member for the international conferences ACM SIGMOD 2013 and ACM CIKM 2013.
- Dario Colazzo served as program committee member for the national conference Bases de Données Avancées (BDA) 2013.
- F. Goasdoué has participated to the program committees of the International Joint Conference on Artificial Intelligence (IJCAI, senior PC), IEEE International Conference on Tools with Artificial Intelligence (ICTAI), and International Workshop on Open Data (WOD), 2013.
- M. Herschel served as program committee member at the following conferences: Bases de Données Avancées (BDA) 2013, Advances in Databases and Information Systems (ADBIS) 2013, Data Management Technologies and Applications (DATA) 2013, IEEE Big Data 2013, Extending Database Technology (EDBT) 2013, and Business, Technology, and Web (BTW) 2013.
- I. Manolescu has participated to the program committees of the ACM SIGMOD Conference 2013, EDBT 2013 demonstration track, International Workshop on the Web and Databases (WebDB) 2013, Workshop on Open Data (WOD) 2013, and the Workshop "Data Engineering Meets the Semantic Web (DESWEB)" associated with the IEEE ICDE Conference, 2013.

Participation to scientific bodies

- International:
  * I. Manolescu: member of the ACM SIGMOD Executive Committee, the EDBT Association, and the IEEE TCDE Awards Committee (since October 2013).
- National:
  * I. Manolescu has been the president of the *Bases de Données Avancées (BDA)* steering committee in Jan-Oct 2013.
  * N. Bidoit and I. Manolescu: members of the BDA committee.

- Faculty recruitment committees:
  * N. Bidoit has been a member of a Professor recruitment committee at UNIV. PARIS-SUD (IUT).
  * I. Manolescu has been a member of a Assistant Professor recruitment committees at U. Paris VI and U. Paris IX, and of a Professor committee at U. Paris IX.
- Other local committees and responsibilities:
  * Nicole Bidoit has been the head of the UNIV. PARIS-SUD Database team (until Jun 2013). In July, the team has fused with another Univ. one to form the new University team LaHDaK, of which N. Bidoit is the vice-head.
  * Nicole Bidoit is the head of the CS Doctoral School (ED 427) of UNIV. PARIS-SUD, a member of the HdR ad-hoc Committee, a member of the Inria SIF Scientific Committee, and is in charge of the CS Doctoral School in the EIT ICT Labs Doctoral Program. She is also a member of the LRI council, of the CCSU 27, and of the Telecom SudParis School Committee.
  * Nicole Bidoit and Ioana Manolescu are members of the steering committee of *Institut de la Société Numérique - ISN* within the Paris-Saclay IdEx.
  * Ioana Manolescu is a member of the executive committee of the DIGICOSME LabEx; she is also an elected member of Inria SIF personnel representative committee (since Oct 2013).

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Licence: François Goasdoué, Databases, 62.5 ETD, L3, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Relational database systems: application tuning, 21h ETD, M2, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Databases, 21 ETD, M2, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Database Management Systems, 36h ETD, Polytech, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Databases, 21 ETD, Polytech, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Advanced Databases, 31 ETD, Polytech, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Remedial Databases, 17 ETD, M1, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Web Data Management, 18 ETD, M1, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, Databases, 20 ETD, Polytech, UNIV. PARIS-SUD, France.

Master: Dario Colazzo, XML and Web programming, 18 ETD, M2, UNIV. PARIS-SUD, France.

Master: François Goasdoué, Semantic Web, 74 ETD, M2, UNIV. PARIS-SUD, France.

Master: François Goasdoué, Data and Knowledge for the Web, 7.5 ETD, M2, UNIV. PARIS-SUD, France.

Master: François Goasdoué, Distributed Reasoning, 4.5 ETD, M2, UNIV. PARIS-SUD, France.

Master: Ioana Manolescu, Big Data Algorithms and Techniques for the Web, 18 ETD, M2, UNIV. PARIS-SUD, France.

Master: Melanie Herschel, Data Warehouses and OLAP, 99.5 ETD, M2, UNIV. PARIS-SUD, France.

Master: Bogdan Cautis, Introduction to Databases, 30 ETD, M1, Telecom ParisTech, France.

Master: Bogdan Cautis, Advanced Databases, 60 ETD, M1, Telecom ParisTech, France.

Master: Bogdan Cautis, Modern Web Data Management, 45 ETD, Telecom ParisTech – Athens week, France.

Master: Bogdan Cautis, Massive Data Analysis, 18 ETD, Ecole Polytechnique – COMASIC Master, France.

8.2.2. Supervision

PhD: Julien Leblay, “Techniques d’optimisation pour des données semi-structurées du web sémantique”, UNIV. PARIS-SUD, September 2013, François Goasdoué and Ioana Manolescu.

PhD: Asterios Katsifodimos, “Scalable view-based techniques for web data : algorithms and systems”, UNIV. PARIS-SUD, July 2013, Ioana Manolescu.

Raphael Bonaque: “Structured, Social and Semantic Search”, since October 2013, Bogdan Cautis, François Goasdoué, and and Ioana Manolescu.

Jesús Camacho-Rodríguez : “Cloud-based Web Data Management”, since October 2011, Dario Colazzo and Ioana Manolescu.

Alexandra Roatiş: “Scalable Database Techniques for the Semantic Web”, since October 2011, Dario Colazzo, François Goasdoué, and Ioana Manolescu.

Aikaterini Tzompanaki: Foundations and Algorithms to Compute the Provenance of Missing Data, since November 2012, Melanie Herschel and Nicole Bidoit.

Stamatis Zampetakis: “Massively Parallel Algorithms for Semantic Web Data”, since October 2012, François Goasdoué and Ioana Manolescu.

Ioana Ileana: “Extracting and Archiving Web Data: a holistic approach”, since October 2011, Bogdan Cautis.

8.2.3. Juries

Ioana Manolescu has been a member of the PhD committee of Siarhei Bykau in April 2013, University of Trento, Italy.

François Goasdoué has been a member of the PhD committee of Anh Tuan Ly in December 2013, University Paris-Sud, France.

Bogdan Cautis has been a member of the PhD committee of Émilien Antoine (December 2013, University of Paris-Sud, France.

8.3. Popularization

- T. Ghosh, I. Manolescu, A. Tzompanaki, A. Roatiş and S. Zampetakis have presented a graph-based game to the audience of *Fête de la Science* in October 2013.
- Still part of *Fête de la Science* 2013, I. Manolescu has participated to a “speed dating” event between researchers and high-school students in Paris.

9. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M.-C. ROUSSET, P. SENELLART. , *Web Data Management*, Cambridge University Press, 2012, 456 p. , http://hal.inria.fr/hal-00677720

[2] N. BIDOIT, D. COLAZZO, F. ULLIANA. *Type-Based Detection of XML Query-Update Independence*, in "Proceedings of the VLDB Endowment", May 2012, http://hal.inria.fr/hal-00757544

[3] N. BIDOIT, M. HERSCHEL, A. TZOMPANAKI. *Query-Based Why-Not Provenance with NedExplain*, in "EDBT - 17th International Conference on Extending Database Technology, Athens, Greece", 2014, To appear

[4] A. BONIFATI, M. GOODFELLOW, I. MANOLESCU, D. SILEO. *Algebraic incremental maintenance of XML views*, in "ACM Transactions on Database Systems", April 2013, http://hal.inria.fr/hal-00816483

[5] D. COLAZZO, G. GHELLI, L. PARDINI, C. SARTIANI. *Almost-linear inclusion for XML regular expression types*, in "ACM Transactions on Database Systems", May 2013, http://hal.inria.fr/hal-00923690

[6] F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Efficient Query Answering against Dynamic RDF Databases*, in "EDBT - 16th International Conference on Extending Database Technology", Genoa, Italy, March 2013 [*DOI :* 10.1145/2452376.2452412], http://hal.inria.fr/hal-00804503

[7] F. GOASDOUÉ, K. KARANASOS, Y. KATSIS, J. LEBLAY, I. MANOLESCU, S. ZAMPETAKIS. *Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents*, in "VLDB Journal", 2013, vol. 22, n° 5, pp. 589-613, http://hal.inria.fr/hal-00828906

[8] K. KARANASOS, A. KATSIFODIMOS, I. MANOLESCU. *Delta: Scalable Data Dissemination under Capacity Constraints*, in "Proceedings of the VLDB Endowment", December 2013, vol. 7, n° 4, pp. 217-228, http://hal.inria.fr/hal-00930107

[9] A. KATSIFODIMOS, I. MANOLESCU, V. VASSALOS. *Materialized View Selection for XQuery Workloads*, in "SIGMOD - ACM SIGMOD International Conference on Management of Data 2012", Scottsdale, Arizona, United States, May 2012, http://hal.inria.fr/hal-00680365

[10] S. MANIU, B. CAUTIS. *Network-aware Search in Social Tagging Applications: Instance Optimality versus Efficiency*, in "ACM Conference on Information And Knowledge Management (CIKM)", San Francisco, United States, October 2013, http://hal.inria.fr/hal-00927308

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] A. KATSIFODIMOS. , *Scalable view-based techniques for web data : algorithms and systems*, Université Paris Sud - Paris XI, July 2013, http://hal.inria.fr/tel-00870456

[12] J. LEBLAY. , *Techniques d'optimisation pour des données semi-structurées du web sémantique*, Université Paris Sud - Paris XI, September 2013, http://hal.inria.fr/tel-00872883

Articles in International Peer-Reviewed Journals

[13] V. BENZAKEN, G. CASTAGNA, D. COLAZZO, K. NGUYỄN. *Optimizing XML Querying using Type-based Document Projection*, in "ACM Transactions on Database Systems", March 2013, vol. 38, n° 1, pp. 1-45, http://hal.inria.fr/hal-00798049

[14] A. BONIFATI, M. GOODFELLOW, I. MANOLESCU, D. SILEO. *Algebraic incremental maintenance of XML views*, in "ACM Transactions on Database Systems", April 2013, vol. 38, n° 3, pp. 1-45, http://hal.inria.fr/hal-00816483

[15] D. COLAZZO, G. GHELLI, L. PARDINI, C. SARTIANI. *Almost-linear inclusion for XML regular expression types*, in "ACM Transactions on Database Systems", May 2013, http://hal.inria.fr/hal-00923690

[16] D. COLAZZO, G. GHELLI, L. PARDINI, C. SARTIANI. *Efficient asymmetric inclusion of regular expressions with interleaving and counting for XML type-checking*, in "Theoretical Computer Science", May 2013, http://hal.inria.fr/hal-00923688

[17] C. COLLET, B. AMANN, N. BIDOIT, M. BOUGHANEM, M. BOUZEGHOUB, A. DOUCET, D. GROSS-AMBLARD, J.-M. PETIT, M.-S. HACID, G. VARGAS-SOLAR. *De la gestion de bases de données à la gestion de grands espaces de données*, in "Ingenierie des Systemes d Information", November 2013, vol. 18, n^{o} 4, pp. 11-31, RSTI série ISI: Personnalisation, contexte et mobilité - ISBN: 9782746246256 [*DOI :* 10.3166/ISI.18.4.11-31], http://isi.revuesonline.com/article.jsp?articleId=18839, http://hal.inria.fr/hal-00923534

[18] F. GOASDOUÉ, K. KARANASOS, Y. KATSIS, J. LEBLAY, I. MANOLESCU, S. ZAMPETAKIS. *Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents*, in "VLDB Journal", 2013, vol. 22, n^{o} 5, pp. 589-613, http://hal.inria.fr/hal-00828906

[19] F. GOASDOUÉ, M.-C. ROUSSET. *Robust Module-based Data Management*, in "IEEE Transactions on Knowledge and Data Engineering", March 2013, vol. 25, n^{o} 3, pp. 648-661 [*DOI :* 10.1109/TKDE.2011.255], http://hal.inria.fr/hal-00671004

[20] K. KARANASOS, A. KATSIFODIMOS, I. MANOLESCU. *Delta: Scalable Data Dissemination under Capacity Constraints*, in "Proceedings of the VLDB Endowment", December 2013, vol. 7, n^{o} 4, pp. 217-228, http://hal.inria.fr/hal-00930107

Articles in National Peer-Reviewed Journals

[21] B. CAUTIS, S. MANIU. *Recherche top-k dépendant d'un contexte à l'aide de vues*, in "Ingénierie des Systèmes d'Information", 2013, vol. 18, n^{o} 4, pp. 109-144, http://hal.inria.fr/hal-00926417

International Conferences with Proceedings

[22] N. BIDOIT, D. COLAZZO, N. MALLA, F. ULLIANA, M. NOLÉ, C. SARTIANI. *Processing XML queries and updates on map/reduce clusters*, in "International Conference on Extending Database Technology", Genova, Italy, March 2013, http://hal.inria.fr/hal-00923695

[23] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU. *Web Data Indexing in the Cloud: Efficiency and Cost Reductions*, in "EDBT - International Conference on Extending Database Technology", Genoa, Italy, March 2013, http://hal.inria.fr/hal-00803597

[24] F. GOASDOUÉ, K. KARANASOS, Y. KATSIS, J. LEBLAY, I. MANOLESCU, S. ZAMPETAKIS. *Fact Checking and Analyzing the Web*, in "SIGMOD - ACM International Conference on Management of Data", New York, United States, June 2013, http://hal.inria.fr/hal-00814285

[25] F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Efficient Query Answering against Dynamic RDF Databases*, in "EDBT - 16th International Conference on Extending Database Technology", Genoa, Italy, March 2013 [*DOI :* 10.1145/2452376.2452412], http://hal.inria.fr/hal-00804503

[26] M. HERSCHEL. *Wondering why data are missing from query results? Ask Conseil Why-Not*, in "International Conference on Information and Knowledge Management (CIKM)", Burlingname, United States, 2013, http://hal.inria.fr/hal-00909210

[27] Z. KAOUDI, I. MANOLESCU. *Triples in the clouds*, in "ICDE - 29th International Conference on Data Engineering", Brisbane, Australia, April 2013, http://hal.inria.fr/hal-00816942

[28] S. MANIU, B. CAUTIS. *Context-Aware Top-k Processing using Views*, in "ACM Conference on Information And Knowledge Management (CIKM)", San Francisco, United States, October 2013, http://hal.inria.fr/hal-00927307

[29] S. MANIU, B. CAUTIS. *Network-aware Search in Social Tagging Applications: Instance Optimality versus Efficiency*, in "ACM Conference on Information And Knowledge Management (CIKM)", San Francisco, United States, October 2013, http://hal.inria.fr/hal-00927308

[30] K. STEFANIDIS, V. EFTHYMIOU, M. HERSCHEL, V. CHRISTOPHIDES. *Entity Resolution in the Web of Data*, in "International Conference on Information and Knowledge Management (CIKM)", Burlingname, United States, 2013, http://hal.inria.fr/hal-00930159

Conferences without Proceedings

[31] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Answering Why-Not Questions*, in "Bases de Données Avancées (BDA)", Nantes, France, 2013, http://hal.inria.fr/hal-00909214

[32] D. COLAZZO, T. GHOSH, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *WaRG: Warehousing RDF Graphs*, in "Bases de Données Avancées", Nantes, France, October 2013, Demonstration, http://hal.inria.fr/hal-00868670

[33] D. COLAZZO, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Warehousing RDF Graphs*, in "Bases de Données Avancées", Nantes, France, October 2013, http://hal.inria.fr/hal-00868616

[34] F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare: efficient Hadoop-based RDF query processing*, in "BDA'13 - Journées de Bases de Données Avancées", Nantes, France, October 2013, http://hal.inria.fr/hal-00867728

[35] F. GOASDOUÉ, K. KARANASOS, Y. KATSIS, J. LEBLAY, I. MANOLESCU, S. ZAMPETAKIS. *Fact checking and Analyzing the web*, in "BDA'13 - Journées de Bases de Données Avancées", Nantes, France, October 2013, http://hal.inria.fr/hal-00867749

[36] M. HERSCHEL. *Entity Resolution*, in "International Workshop on Open Data (WOD)", Paris, France, 2013, http://hal.inria.fr/hal-00930154

[37] K. KARANASOS, A. KATSIFODIMOS, I. MANOLESCU. *Delta: Scalable Data Dissemination under Capacity Constraints*, in "Bases de Données Avancées", Nantes, France, October 2013, http://hal.inria.fr/hal-00867847

[38] I. MANOLESCU. *Fact-Checking the Web: Content Management Technologies for a Democratic Decision Process*, in "Workshop on Open Data", Paris, France, Vassilis Christophides and Dan Vodislav, May 2013, http://hal.inria.fr/hal-00930254

[39] N. PERNELLE, F. SAIS, B. SAFAR, M. KOUTRAKI, T. GHOSH. *N2R-Part: Identity Link Discovery using Partially Aligned Ontologies*, in "International Workshop on Open Data", Paris, France, June 2013, http://hal.inria.fr/hal-00832926

Scientific Books (or Scientific Book chapters)

[40] F. BUGIOTTI, J. CAMACHO-RODRÍGUEZ, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, S. ZAMPETAKIS. *SPARQL Query Processing in the Cloud*, in "Linked Data Management", A. HARTH, K. HOSE, R. SCHENKEL (editors), Emerging Directions in Database Systems and Applications, Chapman and Hall/CRC, April 2014, http://hal.inria.fr/hal-00909121

Research Reports

[41] A. DEUTSCH, F. GOASDOUÉ, J. LEBLAY, I. MANOLESCU. , *Extended query language and composition for the XR data model*, Inria, November 2013, n^{o} RR-8391, http://hal.inria.fr/hal-00879511

[42] K. KARANASOS, A. KATSIFODIMOS, I. MANOLESCU. , *Delta: Scalable Data Dissemination under Capacity Constraints*, Inria, October 2013, n^{o} RR-8385, 37 p. , http://hal.inria.fr/hal-00877758