Activity Report 2017

Team CEDAR

Rich Data Exploration at Cloud Scale

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Team CEDAR

Creation of the Team: 2016 January 01

Keywords:

Computer Science and Digital Science:

A3.1.1. - Modeling, representation
A3.1.2. - Data management, quering and storage
A3.1.3. - Distributed data
A3.1.6. - Query optimization
A3.1.7. - Open data
A3.1.8. - Big data (production, storage, transfer)
A3.1.9. - Database
A3.2.1. - Knowledge bases
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.2.5. - Ontologies
A3.3.1. - On-line analytical processing
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A9.1. - Knowledge
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1.6. - Genomics
B8.5.1. - Participative democracy
B9.4.5. - Data science
B9.7.2. - Open data

1. Personnel

Research Scientists

Ioana Manolescu [Team leader, Inria, Senior Researcher, HDR]
Yanlei Diao [Ecole Polytechnique, Senior Researcher, HDR]
Michael Thomazo [Inria, Researcher]

Post-Doctoral Fellow

Fei Song [Inria]

PhD Students

Maxime Buron [Inria, from Oct 2017]
Tien Duc Cao [Inria]
Sejla Cebiric [Inria]
Luciano Di Palma [Ecole Polytechnique, from Sep 2017]

Felix Raimundo [Ecole Polytechnique, from Sep 2017]
Alexandre Sevin [Ecole Polytechnique, from Oct 2017]
Khaled Zaouk [Ecole Polytechnique, from Oct 2017]
Enhui Huang [Ecole Polytechnique]

Technical staff

Oscar Santiago Mendoza Rivera [Inria, until Aug 2017]
Tayeb Merabti [Inria, from Sep 2017]

Interns

Emna Kamoun [Inria, from May 2017 until Aug 2017]
Francesco Pierri [Ecole Polytechnique, from Sep 2017]
Shu Shang [Ecole Polytechnique, from Mar 2017 until Sep 2017]
Quang-Duy Tran [Inria, from Apr 2017 until Jun 2017]

Administrative Assistant

Maeva Jeannot [Inria]

External Collaborators

Ahmed Abdelkafi [Ecole Polytechnique]
Lars Kegel [University of Dresden, from Sep 2017]
Swen Ribeiro [CNRS, until May 2017]
Xavier Tannier [CNRS]
Stamatios Zampetakis [Orchestra Networks, from Sep 2017]

2. Overall Objectives

2.1. Overall Objectives

Our research aims at **models, algorithms and tools for highly efficient, easy-to-use data and knowledge management**; throughout our research, **performance at scale** is a core concern, which we address, among other techniques, by designing algorithms for a **cloud (massively parallel)** setting. Our scientific contributions fall in three interconnected areas:

Expressive models for new applications As data and knowledge applications keep extending to novel application areas, we work to devise appropriate data and knowledge models, endowed with formal semantics, to capture such applications' needs. This work mostly concerns the domains of data journalism and journalistic fact checking;

Optimization and performance at scale This topic is at the heart of Y. Diao's ERC project "Big and Fast Data", which aims at optimization with performance guarantees for real-time data processing in the cloud. Machine learning techniques and multi-objectives optimization are leveraged to build performance models for data analytics the cloud. The same boal is shared by our work on efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration Today's Big Data is complex; understanding and exploiting it is difficult. To help users, we explore: compact summaries of knowledge bases to abstrac their structure and help users formulate queries; interactive exploration of large relational databases; techniques for automatically discovering interesting information in knowledge bases; and keyword search techniques over Big Data sources.

3. Research Program

3.1. Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited "as they are", with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2. Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3. Multi-Model Querying

As the world's affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4. Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example".

3.5. Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs.

3.6. Representative Semantic Query Answering

Top-k search is a classical topic, studied in relational databases, semantic web, recommandation systems,... It is extremely useful, among other, when a human user face a large number of query results, allowing the user to reformulate the query if necessary. However, we argue that top-k search incurs a bias on the perception of the set of results which is out of the control of the user. Our goal is to provide the user with k answers as well which are chosen so as to represent the diversity of the answer set. We will first consider this problem in the setting of relational or RDF databases. We will then extend to more heterogeneous sources, including in particular plain text.

4. Application Domains

4.1. Cloud Computing

Cloud computing services are strongly developing and more and more companies and institutions resort to running their computations in the cloud, in order to avoid the hassle of running their own infrastructure. Today's cloud service providers guarantee machine availabilities in their Service Level Agreement (SLA), without any guarantees on performance measures according to a specific cost budget. Running analytics on big data systems require the user not to only reserve the suitable cloud instances over which the big data system will be running, but also setting many system parameters like the degree of parallelism and granularity of scheduling. Chosing values for these parameters, and chosing cloud instances need to meet user objectives regarding latency, throughput and cost measures, which is a complex task if it's done manually by the user. Hence, we need need to transform cloud service models from availabily to user performance objective rises and leads to the problem of multi-objective optimization. Research carried out in the team within the ERC project "Big and Fast Data Analytics" aims to develop a novel optimization framework for providing guarantees on the performance while controlling the cost of data processing in the cloud.

4.2. Computational Journalism

Modern journalism increasingly relies on content management technologies in order to represent, store, and query source data and media objects themselves. Writing news articles increasingly requires consulting several sources, interpreting their findings in context, and crossing links between related sources of information. CEDARresearch results directly applicable to this area provide techniques and tools for rich Web content warehouse management. Within the ANR ContentCheck project, and also as part of our international collaboration with the AIST institute from Japan, we work on one hand, to lay down foundations for computational data journalism and fact checking, and also work to devise concrete algorithms and platforms to help journalists perform their work better and/or faster. This work is carried in collaboration with Le Monde's "Les Décodeurs".

On a related topic, heterogeneous data integration under a virtual graph abstract model is studied within the ICODA Inria project which has started in September 2017. There, we collaborate with Les Décodeurs as well as with Ouest France and Agence France Presse (AFP). The data and knowledge integration framework resulting from this work will support journalists' effort to organize and analyze their knowledge and exploit it in order to produce new content.

4.3. Open Data Intelligence

The Web is a vast source of information, to which more is added every day either in unstructured form (Web pages) or, increasingly, as partially structured sources of information, in particular as Open Data sets, which can be seen as connected graphs of data, most frequently described in the RDF data format recommended by the W3C. Further, RDF data is also the most appropriate format for representing structured information extracted automatically from Web pages, such as the DBPedia database extracted from Wikipedia or Google's InfoBoxes. We work on this topic within the 4-year project ODIN started in 2014.

4.4. Genomics

One particular case of area where the increase in data production is the more consequent is genomic data, indeed the amount of data produced doubles every 7 months. Thus we want to bring the expertise from the database and big data community to help both scale the existing algorithms and design new algorithms that are scalable from the ground up.

5. Highlights of the Year

5.1. Program Committee Chair

Yanlei Diao has been the PC chair of the IEEE International Conference on Data Engineering (ICDE) 2017.

5.2. Strong recruitment of PhD students

The team has started work on many new projects, particularly ; six new PhD thesis starting this year (M. Buron, L. Di Palma, L. Duroyon, F. Raimundo, A. Sevin and K. Zaouk) have rejoined the three more senior students (D. Cao, S. Cebiric, E. Huang). These recruitments boost our efforts on core topics of the team, namely: data exploration, fact checking and data journalism, and performance optimization in the cloud.

5.3. Keynotes

Y. Diao gave a distinguished talk at TU Darmstadt; I. Manolescu gave two keynotes at the international conferences DEXA 2017 and iiWAS 2017.

6. New Software and Platforms

6.1. RDF-Commons

KEYWORDS: Data management - RDF
FUNCTIONAL DESCRIPTION: RDF-Commons is a set of modules providing the abilities to: - load and store RDF data in a DBMS - parse RDF conjunctive queries - encode URIs and literals into integers - encode RDF conjunctive queries - build statistics on RDF data - estimate the cost of the evaluation of a conjunctive query - saturate the RDF data, with respect to an RDF Schema - reformulate a conjunctive query with respect to an RDF Schema - propose algebraic plans

- Contact: Ioana Manolescu

6.2. RDFSummary

FUNCTIONAL DESCRIPTION: RDF Summary is a standalone Java software capable of building summaries of RDF graphs. Summaries are compact graphs (typically several orders of magnitude smaller than the original graph), which can be used to get acquainted quickly with a given graph, they can also be used to perform static query analysis, infer certain things about the answer of a query on a graph, just by considering the query and the summary.

- Contact: Sejla Cebiric

6.3. Tatooine

KEYWORDS: Data integration - Databases - Knowledge database - JSon - RDF - Polystore

FUNCTIONAL DESCRIPTION: Tatooine allows to jointly query data sources of heterogeneous formats and data models (relations, RDF graphs, JSON documents etc.) under a single interface. It is capable of evaluating conjunctive queries over several such data sources, distributing computations between the underlying single-data model systems and a Java-based integration layer based on nested tuples.

- Participants: François Goasdoué, Ioana Manolescu, Javier Letelier Ruiz, Michaël Thomazo, Oscar Santiago Mendoza Rivera, Raphael Bonaque, Swen Ribeiro, Tien Duc Cao and Xavier Tannier
- Contact: Ioana Manolescu

7. New Results

7.1. Semantic Query Answering

Building upon last year's work on regular path queries, we studied the complexity of answering conjunctive regular path queries under linear existential rules and under guarded existential rules. These queries generalized conjunctive queries by their ability to check for a path between two individuals which is labeled by a word belonging to a given regular language. Linear and guarded rules are widely recognized as two important classes of existential rules, that among other generalizes most popular Horn description logics. The results are quite positive, in the sense that the complexity is as good as we could hope for: we provided matching upper-bound that correspond to much less expressive query or ontology languages (i.e., they come from RPQs over linear rules or CQs over guarded rules). These results have been published at IJCAI'17 [13].

7.2. Representative Semantic Query Answers

The availability of large knowledge bases such as Yago or DBPedia allows theoretically anybody to tap in their resources through structured and semantics queries. Thi is still not as widespread as it could be, and we postulate this is mainly for two reasons. First, it is complex to write queries in such a setting. Second, the value added of such querying is improvable. We focused on the second point, with the rationale that increasing the value aded may motivate more easily users to spend the time and energy necessary to learn to write SPARQL queries. More specifically, the internship of M. Buron [22] explored the possibility of exploiting the reasoning performed to find a tuple as an answer to cluster answers in a semantic (and explainable) way.

7.3. Interactive Data Exploration at Scale

To respond to increasing user information needs in the era of Big Data, we aim to build interactive data exploration as a new database service, using an approach called "explore-by-example". In particular, we cast the "explore-by-example" problem in a principled "active learning" framework, and bring the properties of important classes of database queries to bear on the design of new algorithms and optimizations for active learning based database exploration. We introduce a dual-space (data and version space) model for convex pattern queries, leverage the factorized dual-space model and online feature selection to handle high dimensional exploration, and design a new active learning algorithm based on version space reduction. These new techniques allow the database system to not only gain improved accuracy but also overcome fundamental limitations of traditional active learning, in particular, the slow convergence problem. Evaluation results using real-world datasets and user interest patterns show that our new system significantly outperforms state-of-the-art active learning techniques and data exploration systems in accuracy while achieving desired efficiency for interactive performance. In addition, we will extend current data exploration system to handle more complex inputs, such as pictures, by adding a active representation learning phase via neural networks to the existing system. Part of this work was explored during the M2 internship of Alexandre Sevin [25].

7.4. A Quotient Framework for Summarizing RDF Graphs

RDF is the data model of choice for Semantic Web applications. RDF graphs are often large and heterogeneous, thus users may have a hard time determining whether a graph is useful for a certain application. We consider answering such questions by inspecting a *graph summary*, a compact structure conveying as much information as possible about the input graph. A summary is *representative* of a graph if it represents both its explicit and implicit triples, the latter resulting from RDF Schema constraints. To ensure representativeness, we defined a novel RDF-specific summarization framework based on *RDF node equivalence* and graph *quotients*; our framework can be instantiated with many different RDF node equivalence relations. We have shown that our summaries are representative, and establish a *sufficient condition* on the RDF equivalence relation to ensure that a graph can be *efficiently summarized*, without materializing its implicit triples. We illustrate our framework on *bisimulation* equivalence relations between graph nodes, and demonstrate the performance benefits of our efficient summarization method through a set of experiments. These results appeared in [17] and are extended in [20], [19].

7.5. Exploring RDF Graphs through Aggregation

RDF graphs may be large and their structure is heterogeneous and complex, making them very hard to explore and understand. To help users discover valuable insights from RDF graph, we have developed Dagger, a tool which automatically recommends *interesting aggregation queries* over the RDF graphs; Dagger evaluates the queries and graphically shows their results to the user, in the ranked order of their interestingness. To specify aggregate RDF queries, we rely on a dialect of SPARQL 1.1, the standard Semantic Web query language, which has been recently enhanced with the capability to specify aggregation; for the interestingness measure, we relied on variance (or second statistic moment). Dagger was developed as part of the M2 internship of Shu Shang [26] and was demonstrated at the International Semantic Web Conference [15]. A short video of our demo appears online at: https://team.inria.fr/cedar/projects/dagger.

7.6. Models and Algorithms for Fact-Checking and Data Journalism

We have advanced toward a generic definition of a computational fact-checking platform, and identified the set of core functionalities it should support: (i) extraction of a claim from a larger document (typically a text published online in some media, social network etc.); this may require identifying the time and space context in which the claim is supposed to hold; (ii) checking the accuracy of the claim against a set of reference data sources; (iii) putting the claim into perspective by checking its significance in a broader context, for instance by checking if the claim still holds after some minor modification of its temporal, spatial or numeric parameters. Checking a claim is not possible in the absence of a set of reference sources, containing data we consider to be true; thus reference source construction, refinement and selection are also central tasks in such an architecture. We have carried this work as part of the ANR ContentCheck project (Section 8.1.1) and also within our associated team with AIST Japan (Section 8.2.1.1). The architecture of the generic platform we envision has been presented in the Paris DB Day event in May 2017, in an ERCIM News [21] and in a keynote [24].

Within this architecture, an important task is to construct reference data sources and to make them more accessible. Toward this goal, we have devised an approach to extract Linked Open Data (RDF graphs) from Excel tables published by INSEE, the French national statistics institute [14]; the resulting data has been published online. Another ongoing line of work explored within the PhD of Ludivine Duroyon concerns establishing new models for temporal beliefs and statements, allowing journalists to increase the value of refernece sources on which to check who said what when.

7.7. Design and optimization for population genomics

As mentioned above, the area of genomics experiences a massive increase in the amount of data to be processed. Furthermore the data generated can sometimes hard to interpret (in particular NGS data for CNV detection).

We investigate new means to discover Copy Number Variation in the human population using methods from the deep learning community. Indeed, great success has been achieved in that area within projects such as DeepVariant; such projects managed to considerably lower the latency for getting results (about 10 fold) but at a higher computational cost. Such methods are currently attracting significant attention in the biology / bioinformatics community, as witnesed by an editorial in Cell Systems (December 2017) [1].

As the area of population genomics is fairly new, we hope to help design a complete framework allowing for better optimisations and integration with database tools. This work is carried by Yanlei Diao and Felix Raimundo, together with Dr. Avinash Abhyankar at the New York Genome Center (NYGC) who co-advises the PhD of F. Raimundo and Dr. Toby Bloom (head of informatics at NYGC).

7.8. Performance Modeling and Multi-Objective Optimization For the Cloud

We study cloud service models based on attaining user's performance objectives; these immediately lead to problems of multi-objective optimization.

Given different cost models, we consider the optimizer will search a multi-dimensional space, compute execution plans that are not dominated by others (known as Pareto plans) and explore meaningful tradeoffs between different objectives to find the optimal plan for each analytical task. We focused on analytical tasks encoded as dataflow programs as in Hadoop and Spark systems. When such dataflow programs are submitted to the cloud, we aim to provide a multi-objective optimizer that can automatically find an optimal execution plan of the dataflow program, which meets specific user performance objectives. Developing an optimizer for dataflow programs in the cloud raises two major challenges: The optimizer needs cost models for running complex dataflow programs in the cloud, and, it further needs a new algorithmic foundation for multi-objective optimization across user-specific objectives.

We have worked to develop a performance model for the optimizer in order to build the skylines for the user-objectives. We found that deep learning offers an incremental prediction framework (using embedding architecture) or online prediction framework (using auto-encoder along with a gradient boosting regressor) that are not available in a baseline regressor approach. However, there is a tradeoff between using the online prediction framework and having good performance, since of course retraining improves results. That said, the online prediction framework gave us acceptable generalization power over unseen jobs. This work has been carried in the M2 internship of Khaled Zaouk [27], and it continues through his PhD.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

- AIDE ("A New Database Service for Interactive Exploration on Big Data") is an ANR "Young Researcher" project led by Y. Diao, started at the end of 2016.
- CBOD ("Cloud-Based Organizational Design") is a 4-year ANR started in 2014, coordinated by prof. Ahmed Bounfour from UPS. 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.
- ContentCheck (2015-2018) is an ANR project in collaboration with U. Rennes 1 (F. Goasdoué), INSA Lyon (P. Lamarre), the LIMSI lab from U. Paris Sud, and the Le Monde newspaper, in particular their fact-checking team Les Décodeurs. Its aim is to investigate content management models and tools for journalistic fact-checking.

8.1.2. LabEx, IdEx

[1] http://www.cell.com/cell-systems/fulltext/S2405-4712(17)30554-9

- CloudSelect is a three-years project started in October 2015. It is financed by the *Institut de la Société Numérique* (ISN) of the IDEX Paris-Saclay; it funds the PhD scholarship of S. Cebiric. The project is a collaboration with A. Bounfour from the economics department of Université Paris Sud. The project aims at exploring technical and business-oriented aspects of data mobility across clo ud services, and from the cloud to outside the cloud.

8.1.3. Others

- ODIN is a four-year project started in 2014, funded by the Direction Générale de l'Armement, between the SemSoft company, IRISA Rennes and Cedar. The project aims to develop a complete framework for analytics on Web data, in particular taking into account uncertainty, based on Semantic Web technologies such as RDF.
- The goal of the iCODA project is to develop the scientific and technological foundations for knowledge- mediated user-in-the-loop collaborative data analytics on heterogenous information sources, and to demonstrate the effectiveness of the approach in realistic, high-visibility use-cases. The project stands at the crossroad of multiple research fields—content analysis, data management, knowledge represen- tation, visualization—that span multiple Inria themes, and counts on a club of major press partners to define usage scenarios, provide data and demonstrate achievements. This is a project funded directly by Inria ("Inria Project Lab"), and is in collaboration with GraphIK, ILDA, LINKMEDIA (coordinator), as well as the press partners AFP, Le Monde (Les Décodeurs) and Ouest-France.

8.2. International Initiatives

8.2.1. Inria Associate Teams Not Involved in an Inria International Labs

8.2.1.1. WebClaimExplain

Title: Mining for explanations to claims published on the Web

International Partner (Institution - Laboratory - Researcher):

AIST (Japan) - Julien Leblay

Start year: 2017

See also: https://team.inria.fr/cedar/projects/webclaimexplain/

The goal of this research is to create tools to find explanations for facts and verify claims made online. While this process cannot be fully automated, the main focus of our work will be explanation finding via trusted sources, based on the observation that one can only trust a statement if he/she can explain it through rules and proofs that can themselves be trusted.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

8.3.1.1. Internships

Lars Kegel, a PhD student at the university of Dresden, is visiting the team since September 2017. He is working on the systematic description of time series with features that capture the global, structural characteristics of a series in a lower dimensional space.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Selection

9.1.1.1. Chair of Conference Program Committees

Y. Diao has been the PC co-chair of IEEE International Conference of Data Engineering (ICDE) 2017.

I. Manolescu has been a track chair for the ACM SIGMOD (Special Interest Group on the Management of Data) 2017 conference.

9.1.1.2. Member of the Conference Program Committees

I. Manolescu has been member of the PC of the DASFAA Conference 2017, of the Semantic Big Data Workshop (SBD) and of the WebDB (Web and Databases) Workshops in conjunction with the SIGMOD Conference 2017.

M. Thomazo has been member of the PC of BDA 2017, and IJCAI 2017.

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

Y. Diao is the editor-in-chief of ACM SIGMOD Record.

I. Manolescu has been an Associate Editor for PVLDB (Proceedings of Very Large Databases) 2017.

9.1.2.2. Reviewer - Reviewing Activities

M. Thomazo has been a reviewer for JODS and TOCL.

9.1.3. Invited Talks

- Y. Diao gave a distinguished lecture at the Technische Universitaet Darmstadt, in November 2017.
- I. Manolescu gave a keynote talk at the DEXA Conference in August 2017 [23].
- I. Manolescu gave a keynote talk at the iiWAS Conference in December 2017 [24]

9.1.4. Leadership within the Scientific Community

Y. Diao is a member of the ACM SIGMOD Executive Committee, and also a member of the PVLDB Endowment.

I. Manolescu is a member of the PVLDB Endowment, of the ACM SIGMOD "Jim Gray" PhD Award Committee, and of the steering committee (*Comité de Pilotage*) of "Bases de Données Avancées" (BDA), the informal association organizing the database research community in France and french-speaking countries.

9.1.5. Research Administration

Y. Diao is on the advisory board of the Data Science Initiative (DSI), a joint center between the applied mathematics and computer science departments of Ecole Polytechnique.

I. Manolescu is responsible of the "Massive Data Processing" axis of the Inria partnership with DGA (Direction Générale de l'Armement).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Master: Y. Diao is a Professor at Ecole Polytechnique, where she teaches "System for Big Data" in M1; she also teaches "Systems for Big Data Analytics" in M2 in the Data Science Master Program of Université Paris Saclay, and organizes a data science research project of M1 (third-year) students at École Polytechnique.
- Master: I. Manolescu, Architectures for Massive Data Management, 12h, M2, Université Paris-Saclay.
- Master: I. Manolescu, Database Management Systems, 52h, M1, École Polytechnique.

9.2.2. Supervision

PhD in progress: Maxime Buron: "Raisonnement efficace sur des grands graphes hétérogènes ", since October 2017, François Goasdoué, Ioana Manolescu and Marie-Laure Mugnier (GraphIK Inria team in Montpellier)

PhD in progress: Tien Duc Cao: "Extraction et interconnexion de connaissances appliquée aux données journalistiques", since October 2016, Ioana Manolescu and Xavier Tannier (LIMSI/CNRS and Université de Paris Sud)

PhD in progress: Sejla Čebirić: "CloudSelect: Data Mobility Within, Across and Outside Clouds", since September 2015, François Goasdoué Goasdoué and Ioana Manolescu.

PhD in progress: Ludivine Duroyon: "Data management models, algorithms & tools for fact-checking", since October 2017, François Goasdoué and Ioana Manolescu (Ludivine is in the Shaman team of U. Rennes 1 and IRISA, in Lannion)

PhD in progress: Enhui Huang: "Interactive Data Exploration at Scale", since October 2016, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA)

PhD in progress: Luciano di Palma, "New sampling algorithms and optimizations for interactive exploration in Big Data", since October 2017, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA)

PhD in progress: Felix Raimundo: "Nouveaux algorithmes et optimisations pour l'analyse profonde du génome à l'échelle de la population", since October 2017, Yanlei Diao and Avinash Abhyankar (New York Genome Center, USA)

PhD in progress: Alexandre Sevin: "Exploration interactive de données sur de grandes sources de données hétérogènes", since October 2017, Yanlei Diao and Peter Haas (U. Massachussets at Amherst, USA)

PhD in progress: Khaled Zaouk: "Performance Modeling and Multi-Objective Optimization for Data Analytics in the Cloud", since October 2017, Yanlei Diao

9.2.3. Juries

Y. Diao has been a member of the PhD committee of Julien Pilourdault (defended in September 2017, at LIG, Grenoble). The thesis was titled "Scalable Algorithms for Monitoring Activity Traces".

I. Manolescu has been a member of the PhD committee of Olivier Wang (defended in June 2017, at LIX), the thesis was titled "Adaptive Rule Models: Active Learning for Rule-Based Systems" and also of the PhD committee of Maria Rossi (defended in November 2017, at LIX). The thesis was titled "Graph Mining for Influence Maximization in Social Networks".

I. Manolescu has been a reviewer and a member of the HDR committee of Vicent Leroy (defended in September 2017, at LIG, Grenoble). The thesis was titled "Data Analysis at Scale: Systems, Algorithms and Information".

9.3. Popularization

- M. Buron, I. Manolescu, F. Raimundo and T. Merabti animated a booth at the "Fête de la Science 2017" at Inria Saclay, in October 2017.
- I. Manolescu published an article on computational fact-checking titled "La vérité, rien que la vérité" in the Binaire blog of Le Monde (http://binaire.blog.lemonde.fr/2017/04/05/la-verite-rien-que-la-verite/)
- I. Manolescu participated to a panel on Data Journalism at the Web2Day, a 3000-strong IT and digital conference in Nantes, in June 2017 (https://web2day.co/en/speakers/ioana-manolescu/)
- I. Manolescu gave a talk at the Inria Alumni Jam Session " Fausses informations, post vérité : allons aux faits ! ", presenting our current work on data management for data journalism (http://www.inria-alumni.fr/evenement/session-inria-alumni-fausses-informations-post-verite-allons-aux-faits-25-octobre-2017-cnam-paris/)

10. Bibliography

Major publications by the team in recent years

[1] J.-F. BAGET, M. BIENVENU, M.-L. MUGNIER, M. THOMAZO. *Answering Conjunctive Regular Path Queries over Guarded Existential Rules*, in "IJCAI: International Joint Conference on Artificial Intelligence", Melbourne, Australia, August 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01632224

[2] R. BONAQUE, B. CAUTIS, F. GOASDOUÉ, I. MANOLESCU. *Toward Social, Structured and Semantic Search*, in "Surfacing the Deep and the Social Web (SDSW)", Riva del Garda, Trentino, Italy, Proceedings of the Workshop on Surfacing the Deep and the Social Web co-located with the 13th International Semantic Web Conference (ISWC 2014), COST Action KEYSTONE, October 2014, vol. 1310, https://hal.inria.fr/hal-01109123

[3] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Teaching an RDBMS about ontological constraints*, in "Proceedings of the Very Large Databases (PVLDB)", 2016, vol. 9, n^{o} 12, pp. 1161–1172, http://www.vldb.org/pvldb/vol9/p1161-bursztyn.pdf

[4] D. COLAZZO, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *RDF Analytics: Lenses over Semantic Graphs*, in "23rd International World Wide Web Conference", Seoul, South Korea, April 2014 [*DOI :* 10.1145/2566486.2567982], https://hal.inria.fr/hal-00960609

[5] Y. DIAO, I. MANOLESCU, S. SHANG. *Dagger: Digging for Interesting Aggregates in RDF Graphs*, in "International Semantic Web Conference (ISWC)", Vienna, Austria, October 2017, https://hal.inria.fr/hal-01577464

[6] F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J.-A. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare: Flat Plans for Massively Parallel RDF Queries*, in "International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01108705

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

[8] M. KÖNIG, M. LECLÈRE, M. MUGNIER, M. THOMAZO. *Sound, complete and minimal UCQ-rewriting for existential rules*, in "Semantic Web", 2015, vol. 6, n^{o} 5, pp. 451–475, http://dx.doi.org/10.3233/SW-140153

[9] A. ROY, Y. DIAO, U. EVANI, A. ABHYANKAR, C. HOWARTH, R. LE PRIOL, T. BLOOM. *Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study*, in "SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Dat", Chicago, Illinois, United States, SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Data, ACM, May 2017, pp. 187-202 [*DOI :* 10.1145/3035918.3064048], https://hal.inria.fr/hal-01683398

[10] H. ZHANG, Y. DIAO, N. IMMERMAN. *On complexity and optimization of expensive queries in complex event processing*, in "International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014", 2014, pp. 217–228, http://doi.acm.org/10.1145/2588555.2593671

Publications of the year

Articles in International Peer-Reviewed Journals

[11] M. KRÖTZSCH, T. MASOPUST, M. THOMAZO. *Complexity of universality and related problems for partially ordered NFAs*, in "Information and Computation", 2017, vol. 255, pp. 177 - 192 [*DOI :* 10.1016/J.IC.2017.06.004], https://hal.inria.fr/hal-01571398

[12] T. MASOPUST, M. THOMAZO. *On Boolean Combinations forming Piecewise Testable Languages*, in "Theoretical Computer Science", June 2017, vol. 682, https://hal.inria.fr/hal-01637057

International Conferences with Proceedings

[13] J.-F. BAGET, M. BIENVENU, M.-L. MUGNIER, M. THOMAZO. *Answering Conjunctive Regular Path Queries over Guarded Existential Rules*, in "IJCAI: International Joint Conference on Artificial Intelligence", Melbourne, Australia, August 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01632224

[14] T. D. CAO, I. MANOLESCU, X. TANNIER. *Extracting linked data from statistic spreadsheets*, in "International Workshop on Semantic Big Data", Chicago, United States, International Workshop on Semantic Big Data, May 2017, pp. 1 - 5 [*DOI :* 10.1145/3066911.3066914], https://hal.inria.fr/hal-01583975

[15] Y. DIAO, I. MANOLESCU, S. SHANG. *Dagger: Digging for Interesting Aggregates in RDF Graphs*, in "International Semantic Web Conference (ISWC)", Vienna, Austria, October 2017, https://hal.inria.fr/hal-01577464

[16] A. ROY, Y. DIAO, U. EVANI, A. ABHYANKAR, C. HOWARTH, R. LE PRIOL, T. BLOOM. *Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study*, in "SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Dat", Chicago, Illinois, United States, ACM, May 2017, pp. 187-202 [*DOI :* 10.1145/3035918.3064048], https://hal.inria.fr/hal-01683398

[17] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *A Framework for Efficient Representative Summarization of RDF Graphs*, in "International Semantic Web Conference (ISWC)", Vienna, Austria, October 2017, https://hal.inria.fr/hal-01577778

Research Reports

[18] T. D. CAO, I. MANOLESCU, X. TANNIER. *Extracting Linked Data from statistic spreadsheets*, Inria Saclay Ile de France, March 2017, https://hal.inria.fr/hal-01496700

[19] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *A Framework for Efficient Representative Summarization of RDF Graphs*, Inria Saclay Ile de France ; Ecole Polytechnique ; Université de Rennes 1 [UR1], August 2017, n° RR-9090, 11 p. , https://hal.inria.fr/hal-01577431

[20] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, Inria Saclay ; Université Rennes 1, June 2017, n° RR-8920, https://hal.inria.fr/hal-01325900

Scientific Popularization

[21] I. MANOLESCU. *ContentCheck: Content Management Techniques and Tools for Fact-checking*, in "ERCIM News", October 2017, https://hal.inria.fr/hal-01596563

Other Publications

[22] M. BURON. *Grouping Answers in Ontology-Based Query Answering*, Inria Saclay, September 2017, 20 p. , https://hal.inria.fr/hal-01622564

[23] I. MANOLESCU. *Data Discovery in RDF Graphs*, August 2017, pp. 1-63, DEXA 2017 - 28th International Conference on Database and Expert System Applications, https://hal.inria.fr/hal-01657144

[24] I. MANOLESCU. *Data integration for journalism: goals, tools, and architectures (Keynote)*, December 2017, pp. 1-46, iiWAS 2017 - 19th International Conference on Information Integration and Web-based Applications & Services, https://hal.inria.fr/hal-01657152

[25] A. SEVIN. *Creation of a smart representation of pictures for interactive data exploration*, ENSAE ParisTech, September 2017, https://hal.inria.fr/hal-01643077

[26] S. SHANG. *Exploratory Analytics for RDF Graphs*, Université de Paris Saclay, September 2017, https://hal.inria.fr/hal-01657163

[27] K. ZAOUK. *Performance Modeling and Multi-Objective Optimization for Data Analytics in the Cloud*, Ecole Polytechnique (Palaiseau, France) ; Telecom ParisTech, September 2017, https://hal.inria.fr/hal-01647208