Efficient ontology-based integration of heterogeneous data sources

François Goasdoué (Univ. Rennes)
Ioana Manolescu (Inria & Institut Polytechnique de Paris)

1 Context

Heterogeneous data management is a longstanding research topic. The *database (DB)* community made strong contributions to it by devising popular data processing architectures ranging from federated databases [SL90] to data warehouses [Jar03] and mediators [Wie92], and more recently polystores [DES+15, ABD+19]. The *knowledge representation & reasoning (KRR)* community also contributed by revisiting such architectures in order to use ontologies on top of them; ontologies allow performing data management tasks (queries, updates, etc) through a more abstract, conceptual representation of the application domain. Crucially, in ontology-based data management systems, typical data management tasks require AI-style reasoning.

The PhD project described below targets the efficient ontology-based data management of heterogeneous data sources within the ANR project CQFD, a research project funded by the French government from 2020 to 2023. This project gathers DB and KRR teams from several French academic research organizations (CNRS, INRIA, and universities) that are used to disseminate results on ontology-based data management in highly-visible international venues.

2 Expected contributions

The goal of this PhD is to study the *efficient* management of heterogeneous data in both *centralized* and *distributed* architectures, with a focus on cost-based optimization, and novel query answering mechanisms and algorithms to be devised.

Centralized architecture. We recently proposed RDF integration systems [BGMM20] that allow querying heterogeneous data sources (SQL, JSON, etc) through an RDF Schema ontology either by materializing the source data relevant to the application into an *RDF warehouse*, or by rewriting queries into distributed ones over the relevant data sources in a *mediator* fashion. In both cases, an incoming query is transformed into a query that may be syntactically complex (due to the reasoning incurred by the ontology and possibly by the source accesses). The evaluation of such queries may be difficult even for modern query engines.

In this centralized, warehouse scenario, we propose to study cost-based optimization for query answering, building on two of our prior contributions. First, we proposed in [BGM15] a query answering technique that reformulates every incoming query into a set of equivalent reformulated queries that take into account the ontology, among which we can pick one with lowest estimated evaluation cost. Second, we are working on a novel storage layout which leads to robust query answering performance [BGM+20]. A first expected contribution C_1 will be to devise a *cost model for the evaluation of reformulated queries on this novel data layout.*

Distributed architecture. In this setting, the two main classes of systems we consider are: mediators (where user queries target a single integrated schema), and polystores (where an integration language is used to explicitly specify the subqueries to be evaluated by each data source).

1. In a *mediator* scenario, a query is transformed in a single distributed over the heterogeneous data sources. A second expected contribution C_2 will be to transpose the query answering technique of [BGM15] in order to produce a set of equivalent queries, involving several sources, and to select one with lowest estimated evaluation costs. The cost model should be devised building on previous techniques from mediator systems [OV11]. Ideally, it should be able to take into account the possibility of pushing operations (selection, projection, joins, etc) into the data sources so as to avoid performing them within the mediator query engine.

2. The core expected contributions of this PhD concern query answering in *ontology-based polystores*. A polystore is system holding a federation of heterogeneous data sources, which can be jointly queried through a hybrid query language, i.e., which combines query languages for different data models. Such a language has been introduced in [ABD$^+$19]. Within a polystore, queries are evaluated on the sources in a distributed fashion.

   A third expected and main contribution C_3 will be to devise ontology-based polystores, i.e., by incorporating ontologies in the polystore architecture. For instance, many options could be considered like putting an ontology on top of every sources, mappings between these ontologies could also be considered in the spirit of ontology-based peer-to-peer data management systems, or a global ontology could also be put on top of the ontologies of the data sources, etc. For the defined ontology-based polystores, query answers will be defined as well as techniques to compute them. In particular, a fourth expected contribution C_4 is to study cost-based query optimization for such polystores. A starting point could be the above contribution C_2 as a mediator query distributed over the data sources is similar to a polystore hybrid query.

3 PhD prerequisites, location and supervision

We are looking for PhD candidates with curricula in computer science, showing excellent skills in databases and knowledge representation (or mathematical logic). Candidates must also have good programming skills. Knowledge about Semantic Web formalisms (RDF and OWL) and technologies would be a plus.

The 3-years PhD position is located at Inria Saclay (30 minutes/kilometers from Paris) on the Ecole Polytechnique's campus, within the Inria Cedar team. The PhD should ideally start in October 2020. In the current sanitary circumstances, Inria encourages work from home whenever possible; the PhD could start (and continue) remotely. Inria will supply a computer and other necessary hardware (e.g., external screen, headset etc.)

The PhD candidate will be jointly supervised by Ioana Manolescu[1] (Senior Inria Researcher), head of the CEDAR team[2] and François Goasdoué[3] (Full Professor at IRISA), head of the SHAMAN team[4]. They have been working together for years on the topic of ontology-based data management, on which they obtained highly-visible results for databases [GMR13, BGM15, BGM16, BGMM19], data wharehouses [GKLM11, CGMR14] and mediators [BGMM20].

[1] `http://pages.saclay.inria.fr/ioana.manolescu/`
[2] `https://team.inria.fr/cedar/`
[3] `http://people.irisa.fr/Francois.Goasdoue/`
[4] `http://www-shaman.irisa.fr/`

4 How to apply?

The PhD candidate must contact first by email both Ioana Manolescu (`ioana.manolescu@inria.fr`) and François Goasdoué (`fg@irisa.fr`), to send them: a detailed CV, university grades (to assess the required skills), as well as either recommendation letters or coordinates supporting the application that we can contact.

References

[ABD+19] Rana Alotaibi, Damian Bursztyn, Alin Deutsch, Ioana Manolescu, and Stamatis Zampetakis. Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue. In *SIGMOD*, June 2019.

[BGM15] Damian Bursztyn, François Goasdoué, and Ioana Manolescu. Optimizing reformulation-based query answering in RDF. In *International Conference on Extending Database Technology, EDBT*, 2015.

[BGM16] Damian Bursztyn, François Goasdoué, and Ioana Manolescu. Teaching an RDBMS about ontological constraints. *PVLDB*, 9(12), 2016.

[BGM+20] Maxime Buron, François Goasdoué, Ioana Manolescu, Tayeb Merabti, and Marie-Laure Mugnier. Revisiting RDF storage layouts for efficient query answering. Preprint, 2020.

[BGMM19] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Reformulation-based query answering for RDF graphs with RDFS ontologies. In *The Semantic Web (ESWC)*, 2019.

[BGMM20] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Ontology-based RDF integration of heterogeneous data. In *International Conference on Extending Database Technology (EDBT)*, 2020.

[CGMR14] Dario Colazzo, François Goasdoué, Ioana Manolescu, and Alexandra Roatis. RDF analytics: lenses over semantic graphs. In *International World Wide Web Conference (WWW)*. ACM, 2014.

[DES+15] Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magdalena Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stanley B. Zdonik. The bigdawg polystore system. *SIGMOD Rec.*, 44(2):11–16, 2015.

[GKLM11] François Goasdoué, Konstantinos Karanasos, Julien Leblay, and Ioana Manolescu. View selection in semantic web databases. *PVLDB*, 5(2), 2011.

[GMR13] François Goasdoué, Ioana Manolescu, and Alexandra Roatis. Efficient query answering against dynamic RDF databases. In Giovanna Guerrini and Norman W. Paton, editors, *International Conference on Extending Database Technology, EDBT*, 2013.

[Jar03] Matthias Jarke. *Fundamentals of data warehouses, 2nd Edition*. Springer, 2003.

[OV11] M. Tamer Ozsu and Patrick Valduriez. *Principles of Distributed Database Systems*. Springer Publishing Company, Incorporated, 3rd edition, 2011.

[SL90] Amit P. Sheth and James A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.*, 22(3):183–236, 1990.

[Wie92] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.