The Optique Project: Towards OBDA Systems for Industry (Short Paper)

D. Calvanese[3], M. Giese[10], P. Haase[2], I. Horrocks[5], T. Hubauer[7], Y. Ioannidis[9], E. Jiménez-Ruiz[5], E. Kharlamov[5], H. Kllapi[9], J. Klüwer[1], M. Koubarakis[9], S. Lamparter[7], R. Möller[4], C. Neuenstadt[4], T. Nordtveit[8], Ö. Özcep[4], M. Rodriguez-Muro[3], M. Roshchin[7], Marco Ruzzi[6], F. Savo[6], M. Schmidt[2], A. Soylu[10], A. Waaler[10], D. Zheleznyakov[5]

[1] Det Norske Veritas, [2] fluid Operations AG, [3] Free University of Bozen-Bolzano, [4] Hamburg University of Technology, [5] Oxford University, [6] Sapienza University of Rome, [7] Siemens Corporate Technology, [8] Statoil ASA, [9] University of Athens, [10] University of Oslo

Abstract. In this paper we present the EU Optique project that aims at developing an end-to-end OBDA system for managing Big Data in industries. We discuss limitations of state of the art OBDA systems and present the general architecture of the Optique's OBDA system that aims at overcoming these limitations.

Keywords: OBDA, ontologies, OWL 2, Big Data, System Architecture

1 Introduction

Accessing the *relevant* data in Big Data scenarios is increasingly difficult both for end-user and IT-experts, due to the *volume*, *variety*, *velocity*, and *complexity* dimensions of Big Data [1]. This brings a high cost overhead in data access for large enterprises. For instance, in the oil and gas industry, IT-experts spend 30–70% of their time gathering and assessing the quality of data [3]. The Optique project[1] [5] advocates for a next generation of the well known *Ontology-Based Data Access* (OBDA) approach to address the Big Data dimensions and in particular the data access problem. The project aims at solutions that reduce the cost of data access dramatically.

OBDA systems address the data access problem by presenting a general ontology-based and end-user oriented query interface over heterogeneous data sources (see Figure 1). The core elements in a classical OBDA systems are an *ontology*, describing the application domain, and a set of *mappings*, relating the ontological terms with the schemata of the underlying data sources. End-users formulate queries using the ontological terms and thus they are not required to understand the structure of the data sources. These queries are then automatically translated using the ontology and mappings into an executable code over the data sources.

State of the art OBDA systems based on the classical architecture, however, have shown among others the following limitations:

[1] `http://www.optique-project.eu/`

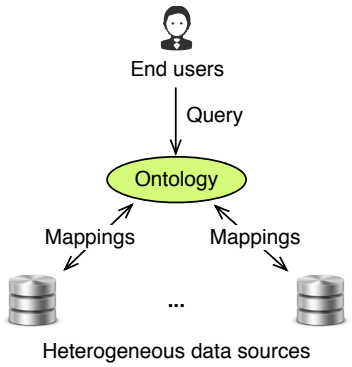

Fig. 1. A general view on Ontology Based Data Access

- The *usability* of OBDA systems is hampered by the need to use a formal query language which is difficult for end-users even if they know the ontological vocabulary.
- The *prerequisites* of OBDA, i.e., ontology and mappings, are in practice expensive to obtain. Additionally, they are not static artefacts and should evolve according to the new end-users' information requirements. In current OBDA systems, bootstrapping and maintenance of ontologies and mappings are in a premature.
- The *scope* of existing systems is too narrow. The chosen expressiveness of the ontology and mapping language are focused on very concrete solutions. Management of *streaming data* is essentially ignored despite their importance for industry.
- The *efficiency* of the translation process and the execution of the queries is too low.

2 Architecture

Figure 2 shows an overview of the Optique OBDA architecture which aims at overcoming the limitations above. The architecture is developed using the three-tier approach and has three layers:

- The *presentation layer* consists of three main user interfaces: *(i)* to compose queries, *(ii)* to visualise answers to queries, and *(iii)* to maintain the system by managing ontologies and mappings. The first two interfaces are for both end-users and IT-experts, while the third one is meant for IT-experts only.
- The *application layer* consists of several main components of the Optique's system, supports its machinery, and provides the following functionality: *(i)* query formulation [4], *(ii)* ontology and mapping management [6], *(iii)* query answering [2, 9], and *(iv)* processing and analytics of streaming data [8].
- The *data and resource layer* consists of the data sources that the system provides access to, that is, relational, semistructured, temporal databases and data streams. It also includes a cloud that provides a virtual resource pool.

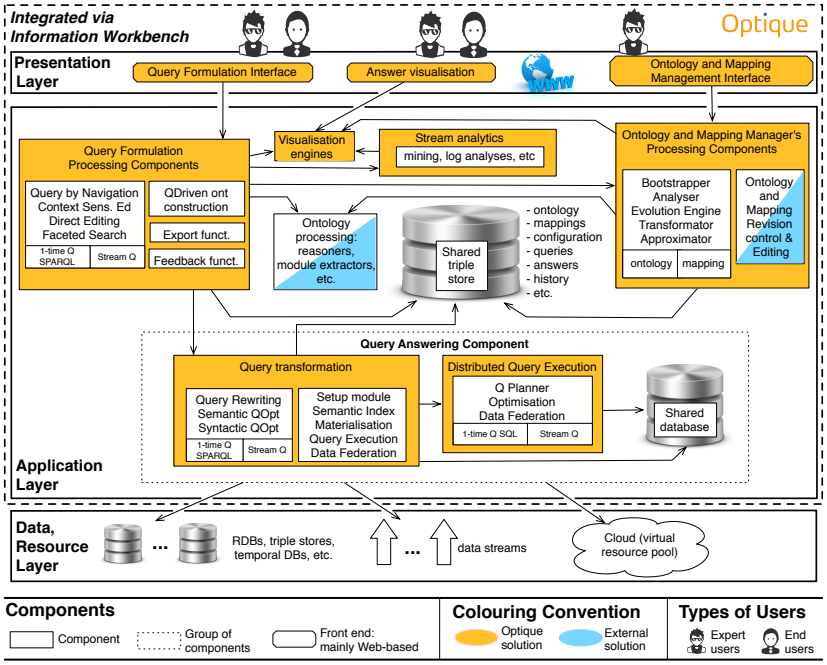

Fig. 2. The general architecture of the Optique OBDA system

The entire Optique system will be integrated via the Information Workbench platform [7].[2] We now briefly describe the four application layer components and the Information Workbench platform.

The query formulation component aims at providing a friendly interface for non-technical users combining multiple representation paradigms (query by navigation, faceted search, context sensitive editing, etc.). Furthermore, this component will also integrate a *query-driven ontology extension* subcomponent to insert new end-users' information requirements in the ontology.

The ontology and mapping management component will provide tools to *(i)* semi-automatically bootstrap an initial ontology and mappings and *(ii)* maintain the consistency between the evolving mappings and the evolving ontology.

The query answering component is compound of two large subcomponents: *(i)* query transformation subcomponent [2], and *(ii)* distributed query optimisation and processing component [9]. The transformation subcomponent is responsible for translating, usually referred to as *rewriting*, of queries received from the Query Formulation component, e.g., SPARQL queries, into an optimised executable code that should be evaluated over the data sources and streams in the data layer, e.g., into a set of SQL or

[2] `www.fluidops.com/information-workbench/`

sliding-window queries. In the nutshell, the transformation compiles the ontology in the input SPARQL query and then translates the result into an SQL query by means of mappings. Besides the compilation of ontology, the transformation component applies different query optimisation techniques, including syntactic and semantic query optimisation which may require ontology reasoning. Moreover, the transformation subcomponent creates and maintains a so-called semantic index that supports query optimisation. The Quest system [12] will be the core part of the Optique's query transformation, while we plan to develop novel rewriting and optimisation techniques to deal, e.g., with streaming data, and to employ other systems, such as PEGASUS [11].

Distributed query optimisation subcomponent provides query planning and execution. It distributes queries to individual servers and use massively parallelised (cloud) computing. The ADP [10, 13] system for complex dataflow processing in the cloud is going to be the core part of the Optique's distributed query processing. Its distinguished features that will guarantee efficiency of Optique's query processing are massive parallelism, i.e., running queries with the maximum amount of parallelism at each stage of execution, and elasticity, i.e., by allowing a flexibility to execute the same query with the use of resources that depends on the the resource availability for this particular query, and the execution time goals.

Processing and analytics for streaming data. This component is primarily motivated by the need of large industries. For example, Siemens[3] encompasses several terabytes of temporal data coming from sensors, with an increase rate of about 30 gigabytes per day. Addressing this challenge requires a number of techniques and tools which should be integrated in several modules of the Optique solution. For example, the query formulation module should support window queries and the query transformation module should support rewriting of such queries. It is also necessary to develop appropriate formalisms to support ontological modelling of streaming data. Besides that, analytical tools are required for stream processing.

The Information Workbench is a generic platform for semantic data management, which provides a central triple store for managing the OBDA system assets (such as ontologies, mappings, etc.), generic interfaces and APIs for semantic data management, and a flexible user interface that will be used for implementing the query formulation components. The user interface follows a semantic wiki approach, based on a rich, extensible pool of widgets for visualization, interaction, mashup, and collaboration, which can be flexibly integrated into semantic wiki pages, allowing developers to compose comprehensive, actionable user interfaces without any programming efforts.

3 Conclusions

The Optique system will provide an end-to-end OBDA solution for Big Data access which will address a number of important industry requirements. The technology and

[3] `http://www.siemens.com`

system will be developed in a close cooperation of six universities, two industrial partners, and two use cases: Statoil and Siemens. The system will be deployed and evaluated in our use cases. It will provide valuable insights for the application of semantic technologies to Big Data integration problems in industry.

Acknowledgements. The research presented in this paper was financed by the Seventh Framework Program (FP7) of the European Commission under Grant Agreement 318338, the Optique project.

References

1. Beyer, M.A., Lapkin, A., Gall, N., Feinberg, D., Sribar, V.T.: 'Big Data' is only the beginning of extreme information management. Gartner report G00211490 (April 2011)
2. Calvanese, D., Horrocks, I., Jiménez-Ruiz, E., Kharlamov, E., Meier, M., Rodriguez-Muro, M., Zheleznyakov, D.: On Rewriting and Answering Queries in OBDA Systems for Big Data (Short Paper). In: OWL Experiences and Directions Workshop (OWLED) (2013)
3. Crompton, J.: Keynote talk at the W3C Workshop on Sem. Web in Oil & Gas Industry (2008), available from `http://www.w3.org/2008/12/ogws-slides/Crompton.pdf`
4. Cuenca Grau, B., Giese, M., Horrocks, I., Hubauer, T., Jiménez-Ruiz, E., Kharlamov, E., Schmidt, M., Soylu, A., Zheleznyakov, D.: Towards Query Formulation and Query-Driven Ontology Extensions in OBDA. In: OWL Experiences and Directions Workshop (OWLED) (2013)
5. Giese, M., Calvanese, D., Haase, P., Horrocks, I., Ioannidis, Y., Kllapi, H., Koubarakis, M., Lenzerini, M., Möller, R., Özep, O., Rodriguez Muro, M., Rosati, R., Schlatte, R., Schmidt, M., Soylu, A., Waaler, A.: Scalable End-user Access to Big Data. In: Rajendra Akerkar: Big Data Computing. Florida : Chapman and Hall/CRC. To appear. (2013)
6. Haase, P., Horrocks, I., Hovland, D., Hubauer, T., Jiménez-Ruiz, E., Kharlamov, E., Klüwer, J., Pinkel, C., Rosati, R., Santarelli, V., Soylu, A., Zheleznyakov, D.: Optique System: Towards Ontology and Mapping Management in OBDA Solutions. In: Workshop on Debugging Ontologies and Ontology Mappings (WoDOOM) (2013)
7. Haase, P., Schmidt, M., Schwarte, A.: The information workbench as a self-service platform for linked data applications. In: COLD (2011)
8. Horrocks, I., Hubauer, T., Jiménez-Ruiz, E., Kharlamov, E., Koubarakis, M., Möller, R., Bereta, K., Neuenstadt, C., Özçep, O., Roshchin, M., Smeros, P., Zheleznyakov, D.: Addressing Streaming and Historical Data in OBDA Systems: Optique's Approach (Statement of Interest). In: Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD) (2013)
9. Kllapi, H., Bilidas, D., Horrocks, I., Ioannidis, Y., Jiménez-Ruiz, E., Kharlamov, E., Koubarakis, M., Zheleznyakov, D.: Distributed Query Processing on the Cloud: the Optique Point of View (Short Paper). In: OWL Experiences and Directions Workshop (OWLED) (2013)
10. Kllapi, H., Sitaridi, E., Tsangaris, M.M., Ioannidis, Y.E.: Schedule optimization for data processing flows on the cloud. In: Proc. of SIGMOD. pp. 289–300 (2011)
11. Meier, M.: The backchase revisited. Submitted for Publication (2013)
12. Rodriguez-Muro, M., Calvanese, D.: High Performance Query Answering over DL-Lite Ontologies. In: KR (2012)
13. Tsangaris, M.M., Kakaletris, G., Kllapi, H., Papanikos, G., Pentaris, F., Polydoras, P., Sitaridi, E., Stoumpos, V., Ioannidis, Y.E.: Dataflow processing and optimization on grid and cloud infrastructures. IEEE Data Eng. Bull. 32(1), 67–74 (2009)