1 Motivations and Challenges

Accessing the relevant data in Big Data scenarios is increasingly difficult both for end-user and IT-experts, due to the volume, variety, and velocity dimensions of Big Data. 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 [1]. The Optique project (http://www.optique-project.eu/) advocates 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. 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 source. OBDA is natural for addressing the 3V of Big Data: the ontology covers the variety of data sources, on the fly data access for query evaluation allows to obtain fresh data regardless the velocity of its changes, and the virtual nature of data integration allows to manage large volumes of data.

The important limitations of the state of the art OBDA systems are as follows:

- 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 of ontologies and mappings are in a premature stage at the best.
- 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.

* This research was financed by the Optique project with the grant agreement FP7-318338.
Some of these items were addressed by the Semantic Web community, but the solutions are limited and there is no unified approach to deal with all these aspects in one system.

2 Optique Solution

Figure 1 shows an overview of the Optique architecture and its components which aim at overcoming the limitations above. We now briefly introduce the main components:

- **The query formulation component** aims at providing a friendly interface for non-technical users combining multiple representation paradigms for ontologies (query by navigation, faceted search, context sensitive editing, etc.). *Query-driven ontology extension* subcomponent will allow to extend the ontology on the fly.

- **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 decomposed into several layers in order to achieve efficiency: transformation, planning and execution. The transformation layer will exploit the mappings and the ontology and will apply optimised query rewriting techniques. The planning and execution layers will distribute queries to individual servers and use massively parallelised (cloud) computing.

To sum up, the Optique system will provide a novel end-to-end OBDA solution for Big Data access which will be integrated in the Information Workbench platform.
Optique: OBDA Solution for Big Data and address a number of important industry requirements. The technology and system will be developed in a close cooperation of six universities, two industrial partners, and 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 in industry.

Reference