# Addressing Streaming and Historical Data in OBDA Systems: Optique's Approach (Statement of Interest)

Ian Horrocks[4], Thomas Hubauer[3], Ernesto Jimenez-Ruiz[2], Evgeny Kharlamov[2], Manolis Koubarakis[4], Ralf Möller[1], Konstantina Bereta[4], Christian Neuenstadt[1], Özgür Özçep[1], Mikhail Roshchin[3], Panayiotis Smeros[4], Dmitriy Zheleznyakov[2]

[1] Hamburg University of Technology, Germany
[2] Oxford University, UK
[3] Siemens Corporate Technology, Germany
[4] University of Athens, Greece

**Abstract.** In large companies such as Siemens and Statoil monitoring tasks are of great importance, e.g., Siemens does monitoring of turbines and Statoil of oil behaviour in wells. This tasks bring up importance of both streaming and historical (temporal) data in the Big Data challenge for industries. We present the Optique project that addresses this problem by developing an Ontology Based Data Access (OBDA) system that incorporates novel tools and methodologies for processing and analyses of temporal and streaming data. In particular, we advocate for modelling time time aware data by temporal RDF and reduce monitoring tasks to knowledge discovery and data mining.

## 1 Introduction

A typical problem that end-users face when dealing with Big Data is of data access, which arises due to the three dimensions (the so-called "3V") of Big Data: *volume*, since massive amounts of data have been accumulated over the decades, *velocity*, since the amounts may be rapidly increasing, and *variety*, since the data are spread over different formats. In this context accessing the *relevant* information is an increasingly difficult task and the Optique project [8] aims at providing solutions for it.

The project is focused around two demanding use cases that provide it with motivation, guidance, and realistic evaluation settings. The first use case is provided by Siemens[5] and encompasses several terabytes of temporal data coming from sensors, with a growth rate of about 30 gigabytes per day. Users need to query this data in combination with many gigabytes of other relational data that describe events. The second use case is provided by Statoil[6] and concerns more than one petabyte of geological data. The data are stored in multiple databases which have different schemata, and the users have to manually combine information from many databases, including temporal, in order to get the results for a single query. In general, in the oil and gas industry, IT-experts spend 30–70% of their time gathering and assessing the quality of data [7]. This

[5] `http://www.siemens.com`
[6] `http://www.statoil.com`

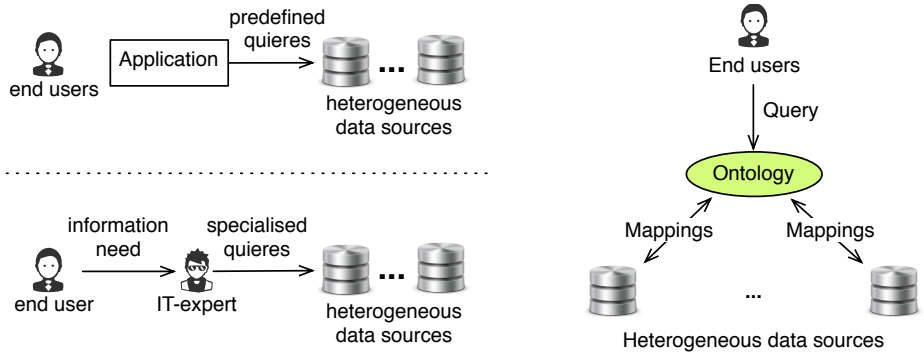


**Fig. 1.** Left: existing approaches to data access; Right: OBDA approach

is clearly very expensive in terms of both time and money. The Optique project aims at solutions that reduce the cost of data access dramatically. More precisely, Optique aims at automating the process of going from an information requirement to the retrieval of the relevant data, and to reduce the time needed for this process from days to hours, or even to minutes. A bigger goal of the project is to provide a platform with a generic architecture that can be adapted to any domain that requires scalable data access and efficient query execution.

The main bottleneck in the Optique's use cases is that data access is limited to a restricted set of predefined queries (cf. Figure 1, left, top). Thus, if an end-user needs data that current applications cannot provide, the help of an IT-expert is required to translate the information need of end-users to specialised queries and optimise them for efficient execution (cf. Figure 1, left, bottom). This process can take several days, and given the fact that in data-intensive industries engineers spend up to 80% of their time on data access problems [7] this incurs considerable cost.

The Semantic approach known as "Ontology-Based Data Access" (OBDA) [16, 6] has the potential to address the data access problem by automating the translation process from the information needs of users (cf. Figure 1, right) to data queries. The key idea is to use an ontology, that presents to users a semantically rich conceptual model of the problem domain. The user formulates their information requirements (that is, queries) in terms of the ontology, and then receives the answers in the same intelligible form. These requests should be executed over the data automatically, without an IT-expert's intervention. To this end, a set of mappings is maintained which describes the relationship between the terms in the ontology and the corresponding data source fields.

As discussed above, in the Siemens use case one has to deal with large amounts of streaming data, e.g., sensor and event data from many turbines and diagnostic centres in many different streams with the size of up to two kilobytes, in combination with historical, that is, temporal relational data sources. Thus, one has to provide Semantic technologies that enable modelling, e.g., with temporal RDF, and processing of both historical and streaming data which includes data mining and complex event processing. The data mining (and time series analysis) aspect is crucial for the Siemens use case as it sets the foundation for preventive diagnostics. More concretely, one of the diagnostic

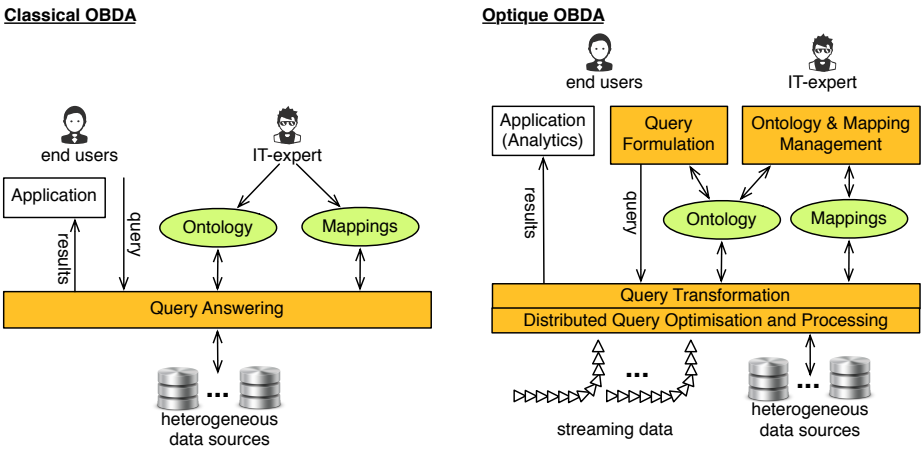


**Fig. 2.** Left: classical OBDA approach. Right: the Optique OBDA system

engineers' requirements is a means to find correlations between events like that of a can flame failure in a turbine and patterns in turbine's relevant timed stamped sensor data measuring temperature (or pressure etc.) If some correlation between an event and sensor data is detected in the historical data by some data mining procedure, then this correlation should be expressible by a continuous query that can be used on real-time data for preventive diagnostics. For example an error event in a turbine $T$ may be identified within the sensor data by at least five percent decrease of measured value TC255-Measurement in sensor TC255 w.r.t. the average in one hour followed by a statistically significant increase (here: more than the two times of the measured average in hour) of measured value TC256-Measurements. The continuous query may refer to a measurement ontology and a (rough) model of the turbine structure within the diagnostic engineer's ontology, that he has to set up in order to localize failures (up to some precision).

The classical OBDA systems (cf. Figure 2, left, shows a conceptual architecture of a classical OBDA system) fail to provide support for these tasks. In the Optique project, we aim at developing a next generation OBDA system (cf. Figure 2, right) that overcomes this limitations. More precisely, the project aims at a cost-effective approach that includes the development of tools and methodologies for processing and analytics of streaming and temporal data. These require to revise existing and develop new OBDA components, in particular, to support novel: *(i)* ontology and mapping management, *(ii)* user-friendly query formulation interface(s), *(iii)* automated query translation, and *(iv)* distributed query optimisation and execution in the Cloud. In this paper we will give a short overview of challenges that we encompass on the way to this goal.

The remainder of the paper is organised as follows. We discuss Optique's challenges in handling of streaming and historical data and present the general architecture of the Optique's OBDA solution in Section 2. Finally, we discuss related work (Section 3) and conclude (Section 4).

## 2 Stream Processing and Analytics in Optique's OBDA

A general requirement, especially motivated by the Siemens use case, is to support such a combination of the data, ontology, mapping, and query languages that is expressive enough for modelling machines, symptoms, and diagnoses, and guarantees a complete, correct, and feasible query answering over temporal and streaming data. We now overview challenges to be solved in achieving this goal: we discuss ontologies, query languages, query processing, and visualisation of answers.

We plan to model temporal data with some extension of RDF. This could be achieved, for example, by adding to RDF triples an extra fourth component: validity time. Thus, the first challenge to address is an understanding of the right temporal RDF data model.

The key component in the Optique OBDA solution is the domain ontology, since it enables users to understand the data at hand and formulate queries. Thus, the next challenge to address is how to model both data streams and temporal data via ontologies. In particular this will require to model time (at least in the query language). On the level of mappings, a homogeneous mapping language for static and streaming data has to be provided.

The query language that the system should provide to end users should combine

*(i)* *temporal operators*, that address the time dimension of data and allow to retrieve data which was true "always" in the past or "sometimes" in the last X months, etc.,
*(ii)* *time series analysis operators*, such as mean, variance, confidence intervals, standard deviation, as well as trends, regression, correlation, etc., and
*(iii)* *stream oriented operators*, such as sliding windows.

Besides, the query language should provide some means for intelligent query answering of queries on complex patterns in the data by, e.g., telling in the negative case what similar patterns exist (query relaxation) and in the positive case how the multitude of patterns can be further restricted (query refinement). Finally, the query language should support formulating queries based on results of explorative data analysis.

Given the query, mapping languages, and ontology, the Optique system should be able to translate queries into highly optimised executable code over the underlying temporal and streaming data. This requires techniques for automated query translation of one-time, continuous, temporal queries, and their combinations. Existing translation techniques are limited and they do not address query optimisation and distributed query processing. Thus, novel approaches should be developed.

Another set of challenges in Optique is related to handling answers to queries. One issue is visualisation of massive volumes of data formatted according to the domain ontology. To address this issue, in particular, Data Mining and Pattern Learning, techniques over ontological data should be developed to enable automatic identification of interesting patterns in the data. Another challenge is how to manage "dirty" data. The Optique system must provide basic means for data cleaning such as: automated identification, mapping and alignment of data types, including dates, time synchronisation, and data quality issues (outlier detection, noise, missing values, etc.).

To sum up this section, we will provide the general architecture of the Optique OBDA system.

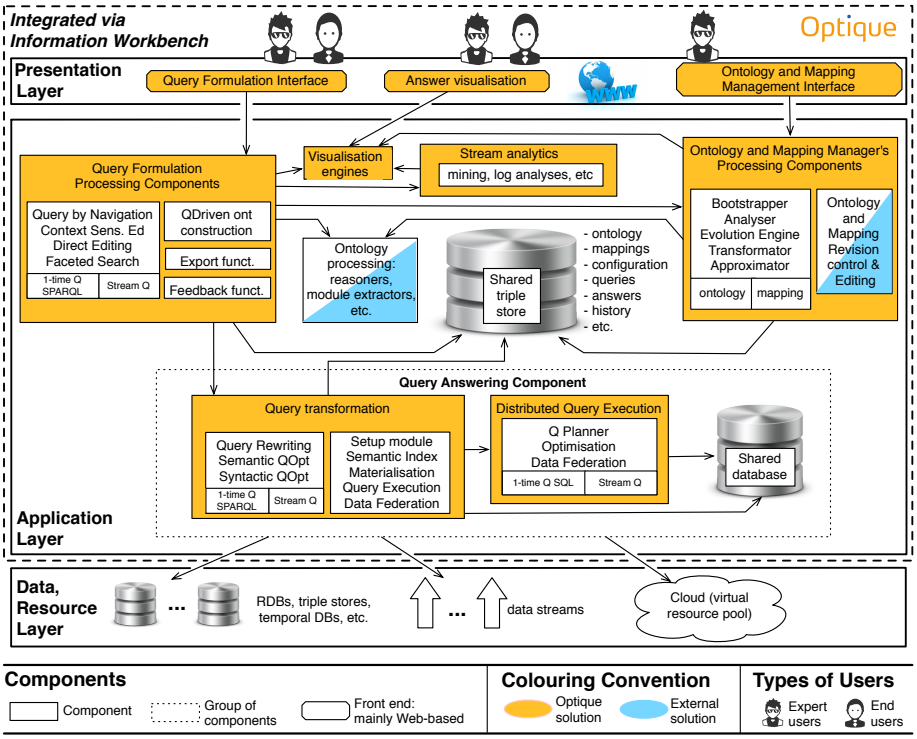


**Fig. 3.** The general architecture of the Optique OBDA system

*General architecture of Optique's OBDA solution.* Figure 3 gives an overview of the Optique's OBDA solution architecture and its components. The architecture is developed using the three-tier approach and has three layers:

- The *presentation layer* consists of three main user interfaces: to compose queries, to visualise answers to queries, and 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:
  - query formulation,
  - ontology and mapping management,
  - query answering, and
  - processing and analytics of streaming and temporal data.
- 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.

The entire Optique system will be integrated via the Information Workbench platform [11][7].

## 3 Related Work

Each of the approaches mentioned in the following deals with only one of the aspects that are relevant for the envisioned software component of the Optique query answering system: either the aspect of query answering over temporal data that can be described as historical; or the aspect of query answering over streamed data. Adhering to the requirements of the (Siemens) use case, the Optique approach favors a more integrative approach, that combines query answering over historical data and query answering over regularly updated temporal data stemming from many different streams.

There exist several approaches that address the problem of representing, inferring with, and querying temporal data within the general context of ontologies. As the Optique project will follow a weak temporalization of the OBDA paradigm, which will guarantee the conservation of so-called FOL rewritability (which essentially means a possibility to translate ontological queries into SQL queries over data sources), work on modal-style temporal ontology languages formalised via Description Logics [13] is of minor relevance; because of the bad complexity properties, this is even true for temporalized lightweight logics [2].

The approach in [10] introduces temporal RDF graphs, details out a sound and complete inference system, and gives a sketchy introduction to a possible temporal query language. A similar representation of temporal RDF graphs is adopted within the spatio-temporal RDF engine STRABON [12, 4][8], which also defines the spatio-temporal extension stSPARQL of the W3C recommendation query language SPARQL 1.1. Strabon is currently the only fully implemented spatio-temporal RDF store with rich functionality and very good performance as seen by the comparison in [12, 4]. For a similar temporal version of RDF, which is oriented at the temporal database language TSQL2, compare [9]. The authors of [17] favor a more conservative strategy by modeling time directly with language constructs within RDF and SPARQL—the resulting extensions of RDF and SPARQL being mere syntactic sugar. The logical approach of [14] follows ideas of [10] but shifts the discussion to the level of genuine ontology languages such as OWL; the semantics of the temporalized RDF and OWL languages are given by a translation to (a fragment of) first order logic. The temporalized SPARQL query language uses a careful separation of the time component and the thematic component that guarantees feasibility of query answering.

The concept of streaming relational data as well as the concepts underlying complex event processing are well understood and are the theoretical underpinnings for highly developed streaming engines used in industrial applications. The picture for stream processing within the OBDA paradigm is quite different; the few implemented streaming engines [5, 3, 15] are still under development and have been shown to lack one or other basic functionality [18]. Though all of the systems are intended to be used within the

[7] www.fluidops.com/information-workbench/

[8] www.strabon.di.uoa.gr

OBDA paradigm, only C-SPARQL [3] seems to have (minimal) capabilities for reasoning/query answering over ontologies. There is no agreement yet on how to extend SPARQL to work over streams; and so all of the mentioned systems have their own streamified version of SPARQL. However, the core of all extensions seems to be the addition of (sliding) window operators over streams, which are adapted from query languages over relational streams [1].

## 4 Conclusions

We presented motivations and challenges for the use of Ontology Based Data Access systems as a solution for Big Data access problem in industries. The important challenge in industries is knowledge discovery and data mining of temporal and streaming data, while the state of the art Semantic technologies that the OBDA systems rely on fail to address it adequately. The Optique project aims at providing this demanding technology which will be validated it in the two industrial use case: Statoil and Siemens. In particular we plan to *(i)* explore the possibility of modelling streaming data with temporal RDF and *(ii)* understand how knowledge discovery and data mining techniques developed for Linked Open Data could be adapted and extended in our setting.

## 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. Arasu, A., Babu, S., Widom, J.: The cql continuous query language: semantic foundations and query execution. The VLDB Journal 15, 121–142 (2006), 10.1007/s00778-004-0147-z
2. Artale, A., Kontchakov, R., Ryzhikov, V., Zakharyaschev, M.: Past and future of dl-lite. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10). AAAI Press (2010)
3. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: C-sparql: a continuous query language for rdf data streams. Int. J. Semantic Computing 4(1), 3–25 (2010)
4. Bereta, K., Smeros, P., Koubarakis, M.: Representation and querying of valid time of triples in linked geospatial data. In: ESWC 2013 (2013)
5. Calbimonte, J.P., Corcho, O., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. In: Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I. pp. 96–111. ISWC'10, Springer-Verlag, Berlin, Heidelberg (2010)
6. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Ruzzi, M., Savo, D.F.: The MASTRO System for Ontology-Based Data Access. Semantic Web 2(1), 43–53 (2011)
7. Crompton, J.: Keynote talk at the W3C Workshop on Semantic Web in Oil & Gas Industry: Houston, TX, USA, 9–10 December (2008), available from `http://www.w3.org/2008/12/ogws-slides/Crompton.pdf`

8. Giese, M., Calvanese, D., Haase, P., Horrocks, I., Ioannidis, Y., Kllapi, H., Koubarakis, M., Lenzerini, M., Möller, R., Özçep, 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)
9. Grandi, F.: T-sparql: a tsql2-like temporal query language for rdf. In: In International Workshop on on Querying Graph Structured Data. pp. 21–30 (2010)
10. Gutierrez, C., Hurtado, C., Vaisman, R.: Temporal rdf. In: In European Conference on the Semantic Web (ECSW' 05). pp. 93–107 (2005)
11. Haase, P., Schmidt, M., Schwarte, A.: The Information Workbench as a Self-Service Platform for Linked Data Applications. In: COLD (2011)
12. Kyzirakos, K., Karpathiotakis, M., Koubarakis, M.: Strabon: A Semantic Geospatial DBMS. In: International Semantic Web Conference. Boston, USA (Nov 2012)
13. Lutz, C., Wolter, F., Zakharyaschev, M.: Temporal description logics: A survey. In: Demri, S., Jensen, C.S. (eds.) 15th International Symposium on Temporal Representation and Reasoning (TIME-08). pp. 3–14 (2008)
14. Motik, B.: Representing and querying validity time in RDF and OWL: a logic-based approach. In: Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I. pp. 550–565. ISWC'10, Springer-Verlag, Berlin, Heidelberg (2010)
15. Phuoc, D.L., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N.F., Blomqvist, E. (eds.) 10th International Semantic Web Conference (ISWC 2011). pp. 370–388 (2011)
16. Rodriguez-Muro, M., Calvanese, D.: High Performance Query Answering over DL-Lite Ontologies. In: KR (2012)
17. Tappolet, J., Bernstein, A.: Applied temporal rdf: Efficient temporal querying of rdf data with sparql. In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. pp. 308–322. ESWC 2009 Heraklion, Springer-Verlag, Berlin, Heidelberg (2009)
18. Zhang, Y., Minh Duc, P., Corcho, O., Calbimonte, J.P.: Srbench: A Streaming RDF/SPARQL Benchmark. In: Proceedings of International Semantic Web Conference 2012 (November 2012)