Optique

Scalable End-User Access to Big Data

http://www.optique-project.eu/

UiO : University of Oslo

SAPIENZA
UNIVERSITÀ DI ROMA

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

1 Optique: Improving the competitiveness of European industry

For many advanced end users accessing the *relevant* data is becoming increasingly difficult due to the explosion in the *size* and *complexity* of data sets.

How much time do engineers in European industry spend searching for data?

Optique targets the key bottleneck limiting exploitation of "Big Data":

- Massive amounts of data are accumulated, in real time and over decades.
- Accessing relevant parts of the data requires in depth knowledge of the domain *and* of the organisation of data repositories
- End users' domain-specific applications limit data access to a restricted set of predefined queries.

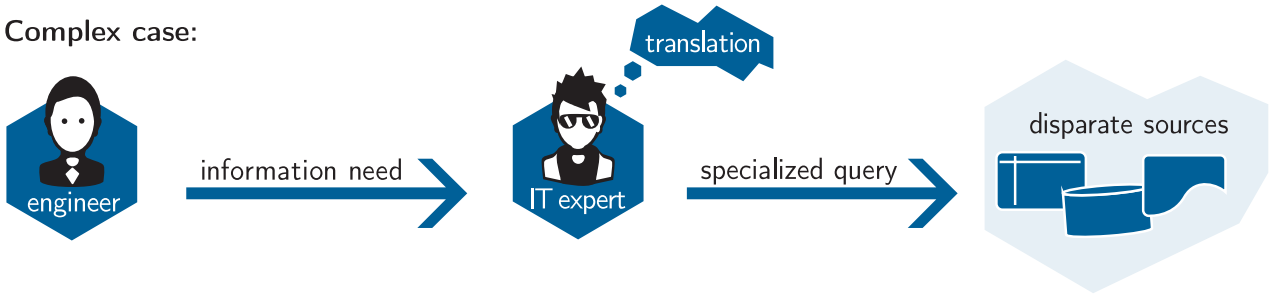
Simple case:

Maximally exploiting data requires flexible access—engineers need to *explore* the data in ways not supported by current applications. This typically requires an IT-expert to:

- write special purpose queries; and
- optimise queries for efficient execution.

How much value could they create in that time?

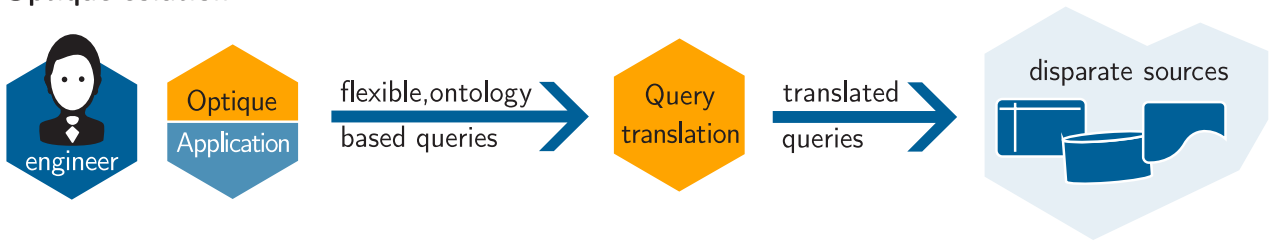
Complex case:



With this process, accessing the data can take several days. In data-intensive industries, engineers spend up to 80% of their time on data access. Apart from the enormous *direct cost*, freeing up expert time would lead to even greater *value creation* through deeper analysis and improved decision making.

The Optique platform will use an *ontology* to capture (possibly multiple) user conceptualisations, and declarative *mappings* to transform user queries into *complete, correct and highly optimised queries*:

Optique solution



Development of the Optique platform will be informed by and evaluated against the requirements of *complex real-world challenges*, with Siemens Energy Services and Statoil Exploration providing the project with comprehensive use cases. In both use cases, decision support is hampered by the data access problems that come with Big Data; and in both cases, the platform can be tested and exploited with minimal interfacing to the currently used decision support tools.

2 Technical approach

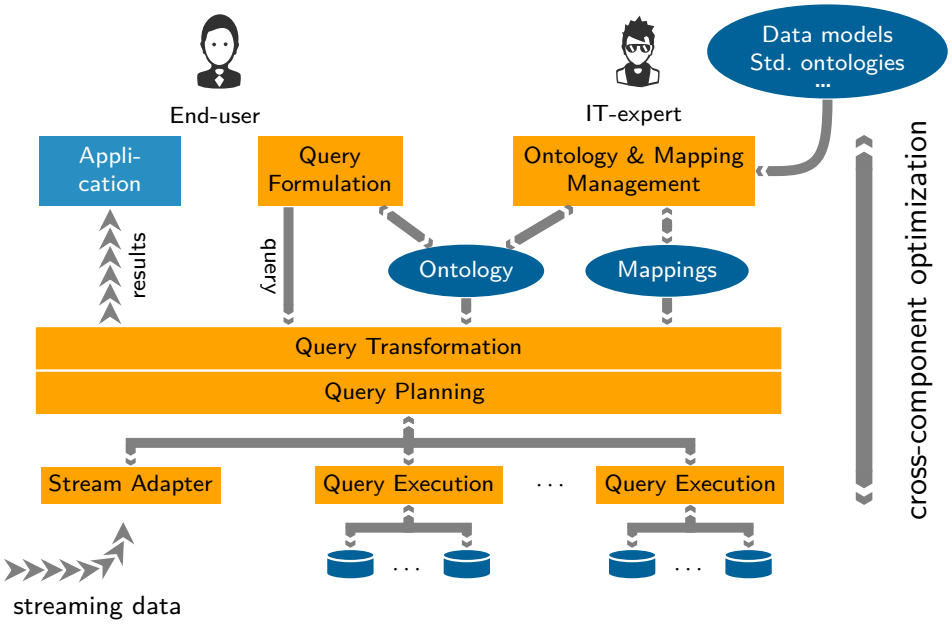


Figure 1: Optique platform architecture

The technical approach exploits recent ground-breaking European research on semantic technologies, in particular related to *query rewriting*, and combines this with techniques for scaling up query evaluation, in particular *massive parallelism*. These are integrated in a comprehensive and extensible platform that builds on open standards and protocols, enabling us to (i) focus our efforts on developing novel and performance-critical key modules; (ii) provide a prototype implementation at an early stage of the project; and (iii) maximise the reusability of the components developed in Optique.

The core architecture is illustrated in Figure 1. Front-end scalability is achieved by abstracting away from data sources, allowing users to construct queries in terms of an ontology that reflects their vocabulary. Further support for query construction is provided by a query formulation component that combines ontology based "query by navigation" with a context-sensitive query editor. Query results are presented to the user via existing tools, with interfaces to these provided by the platform.

Back-end scalability is achieved by optimising the transformation of user queries into queries over the data sources, which may include streams. A set of mappings relates the ontology to the data sources. Together, the ontology and mappings capture crucial expert knowledge that the query transformation component uses to rewrite end-user queries into queries over the data sources. The query planning component then devises an optimised plan for evaluating these queries, which are handled by a query execution or stream adaptor component for each autonomous source. Finally, a novel ontology & mapping management component supports the construction and maintenance of ontology and mappings. This includes support for semi-automatic "lazy" construction driven by the requirements of user queries, minimising the need for *a priori* ontology and mapping development.

The platform is non-invasive, in the sense that it runs on existing IT infrastructure and does not require data migration. The source data can be of a wide variety of types (including relational and semantic data formats as well as streams), provided that they can be accessed via structured query interfaces. However, in order to maximise back-end scalability, we also explore an approach in which data *is* migrated to an infrastructure that implements massive parallelisation.

Further scalability gains are achieved by taking a holistic approach to optimisation that goes beyond considering components individually, which is a distinguishing feature of Optique. Query transformation is optimised so as to produce queries that can be more effectively planned and executed, and query planning and execution are optimised for the kinds of queries that result from query transformation. These optimisations exploit features of typical (user) queries, ontologies and mappings, and the system is tuneable so as to maximise performance in specific applications and even for specific queries.

3 Objectives

The use cases from Siemens Energy Services and Statoil Exploration drive the development in Optique and this is reflected in the main objective of the project. Seven objectives derive from the main objective, the architecture in Figure 1, and Optique's commitment to dissemination and exploitation of the project results.

Main objective. To design and implement an end-to-end solution to the problem of providing comprehensive and timely access to large scale data sets, and to evaluate this solution in four use cases from two industrial scenarios.	
Objective 1.	Achieve the combination of intelligence, flexibility and scalability needed to meet the requirements of the use case challenges.
Objective 2.	Provide a front end that supports end-users and helps them to formulate queries.
Objective 3.	Extend query rewriting techniques with support for temporal queries and queries over streams.
Objective 4.	Integrate scalable methods for database management into an optimised Ontology-Based Data Access framework.
Objective 5.	Deploy and evaluate the Optique platform in the Siemens and Statoil use cases.
Objective 6.	Widely disseminate Optique scientific results and technologies.
Objective 7.	Deliver a business implementation and exploitation strategy.

4 Annual project phases

Optique follows a four phase work plan, annually integrating the use cases in an analysis, development, evaluation and re-analysis cycle. This provides for the early delivery and evaluation of prototypes, and enhance our ability to adjust to evolving technologies and user requirements.

In line with Optique's holistic perspective, there will be one major deliverable, covering all aspects of the platform, for each project phase. Each phase in the work plan consists of S&T development related to all parts of the platform, testing of platform installations by domain experts, and a concluding evaluation of test results against user requirements, possibly leading to a refined requirements analysis. Each phase also has its own scope and goals for the dissemination activities. The main R&D activities in each phase are outlined in this table:

Phase	Optique Platform	Use cases
1	Basic architecture Early prototype Combine existing components	Req. analysis, data collection for Ph. 2 Test prototype on simple tasks Make Optique known in Siemens/Statoil
2	Functional prototype – new prototypes from all technical WP's – treatment of time – stratigraphy	Req. analysis, data collection for Ph. 3 Siemens: reactive diagnosis Statoil: EPDS (central data store)
3	Fully functional platform – complete components from all tech. WP's – treatment of streams – integration of new sources	Req. analysis, data collection for Ph. 4 Siemens: predictive diagnosis Statoil: include project data
4	Optimisation – usability – performance – additional functionality	Cultivation of user groups – increase visibility of project – courses, tutorials

Each phase starts November 1 and ends October 31 the following calendar year.

5 Impact

The two use cases of Optique provide the project with an in-depth knowledge of what is required for successful deployment of the platform. This knowledge will be turned into a best practice methodology, describing how the platform can most effectively be integrated in industrial work flows, and experience from the use cases will be used to develop a comprehensive tutoring and training programme to complement the platform. The methodology and training programme will be designed so as to facilitate the widest possible uptake and deployment of the Optique platform.

Optique's impact generation strategy responds to fundamental market mechanisms. In the scope of the project, we will develop an ecosystem of companies around the Optique platform along the entire value chain. The Optique dissemination and exploitation strategy takes into account the variety of stakeholders, and we will introduce measures to ensure uptake and sustainability of the ecosystem also after the project has ended. Expanding beyond the Optique consortium, this Optique ecosystem will include the following actors:

- Large industrial enterprises playing a leading role in the energy and oil & gas sectors
- Ambitious IT vendors offering industrial-strength platforms for data management and analysis
- Specialised consulting companies providing expert services and technology to the energy industry
- Renowned universities promoting cutting-edge research in the areas of data management and semantic web

From the start of year 2014 the ecosystem vision is implemented in the form of the *Optique European Partner Programme.*

Expected impacts of Optique are, first, *reinforced ability for a wide range of innovators to tap data infrastructures and to add value beyond the original purpose of the data through data analysis* since

- the Optique Platform can be deployed on top of existing data infrastructures, allowing companies to integrate and query a wide range of existing sources of structured data, and use the results in existing analysis solutions;
- by efficiently evaluating expressive queries over integrated disparate data sources, information can be extracted that is practically inaccessible today.

Second, an expected impact is *reinforced ability to find, reuse and exploit data resources (collections, software components) created in one environment in very different, distant and unforeseen contexts.* The Optique platform allows users to

- find data by identifying data sets that are relevant to a question at hand through direct access to the semantics of the data, and by supporting exploration of the data through ad hoc queries;
- reuse data through a powerful and flexible query language along with support for coherent query formulation; as well as allowing to link disparate sources to a common vocabulary;
- exploit data by connecting end user applications directly to the data sources through Optique's end-to-end solution; the dramatic increase in data retrieval efficiency opens up higher quality exploratory data analysis, and thus better exploitation of the data;
- exploit data in different, distant and unforeseen contexts through support for building different conceptualisations into ontologies and mappings, thus allowing end users to explore data from their own perspective, independently of the original purpose and organisation of the data.

6 Project results in Year 1

The main focus of the technical work during the first year of the project has been on establishing a platform with a generic architecture that can be adapted to any domain that requires scalable data access and efficient query execution. An initial prototype of the Optique Platform has been developed based on fluid Operations' Information Workbench platform, and includes a detailed modular design with shared API's enabling a seamless integration and interaction of components, languages for ontologies, mappings and queries, and API's for accessing RDF data, managing mappings and ontologies, and accessing relational schemas.

As regards the implementation and integration of component subsystems, the main achievements in year 1 have been:

- A preliminary ontology-based visual query formulation subsystem has been implemented, based on multiple coordinated interaction paradigms, in particular graph navigation and facet refinement.

- A prototypical ontology and mapping management subsystem has been implemented; it provides basic and advanced bootstrapping wizards, and supports ontology bootstrapping from schema constraints, ontology integration (using LogMap) and novel ontology approximation techniques. Work has also begun on supporting ontology and mapping evolution, including the development of a novel technique for capturing instance level ontology evolution, and inconsistency tolerant semantics for OWL 2 QL.

- STARQL—a temporal and streaming extension of the SPARQL query language—has been developed, and an analysis of the architectural requirements for the time and streams subsystem has been carried out.

- A query rewriting subsystem has been implemented, based on the Ontop system brought into the project by the Free University of Bozen-Bolzano; it extends Ontop with support for OWL 2 QL and SPARQL queries, and improves performance by exploiting mappings and data dependencies. Techniques for metadata extraction from source databases have been developed in order to facilitate semi-automatic configuration and optimisation, and the subsystem has been tuned to adapt to different SQL dialects.

- A JDBC interface has been implemented for the ADP-based distributed query execution subsystem, and queries imported from use-cases have been used to benchmark the subsystem and its component algorithms.

The platform has already been successfully deployed in the use cases at Siemens and Statoil. Furthermore, relevant user interface components, such as those for visual query formulation and browsing, have been presented to a groups of end users at use-case workshops, and feedback suggests that the overall approach is promising and that end-users are able to use the system.

As regards impact, the project has worked toward increasing its visibility to the research community, the industry and the general public, with the following results:

- The project has produced 51 refereed publications, including one book chapter, 3 journal articles and 47 conference and workshop papers. We have given one keynote and 49 workshop and conference presentation.

- The project has given 24 presentations on-site for the industrial consortium partners Siemens, Statoil and DNV, many of these with high level executives in the audience. We have given 10 presentations on-site for external companies, 3 keynotes at industry events, and 10 workshop and conference presentations for industry. We chaired the Exhibition Committee of the *European Data Forum* conference.

- Towards the end of first year a demo, which includes all the main components, were presented at the ISWC conference. The demo will be further developed over the entire project period and

will in its final form be delivered in M48 as the project's Public showcase. The project web site had 5265 visits, and was by the end of Y1 integrated with Twitter (i.e., twitter messages tagged with optique-project are displayed in the website) and Facebook (e.g., like and share buttons). Towards the end of Year 1 a YouTube channel has been established to broadcast the Optique demo with 174 views in October. We have published 7 news papers and magazine articles and given 13 talks to the general public.

Optique has a stated ambition of producing results that can be exploited for real business value. In extensive interaction with industry partners and interested third parties, the project has had a strong focus on understanding the data access needs of industry, and how the Optique platform and system can provide solutions.

- The Initial Exploitation Plan (D11.1) presents an initial market analysis, outlines a strategy, and details how the Optique consortium will organise to produce implementable business and product plans.
- Primary attention is given to industry verticals Energy and Oil & Gas.
- Standardisation is crucial to the strategy, both in the sense of adherence to existing standards and in contributing results from Optique to advance the state of the art.
- The Optique Partner Program plays a vital role as the framework for collaboration between R&D and industry.