Optique

Scalable End-User Access to Big Data

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

UiO : University of Oslo

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":

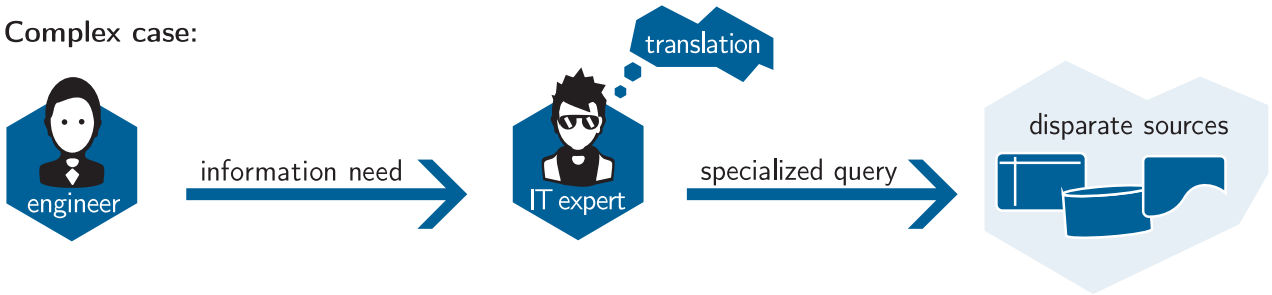
- Massive amounts of data are accumulated, in real time and over decades.
- Accessing relevant parts of the data requires in depth knowledge of the domain *and* of the organisation of data repositories.
- End users' domain-specific applications limit data access to a restricted set of predefined queries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

Simple case:

Maximally exploiting data requires flexible access – engineers need to *explore* the data in innovative ways not supported by current applications. This typically requires an IT-expert in order to write special purpose queries and optimise the queries for efficient execution (see Complex case figure below). 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.

How much value could they create in that time?

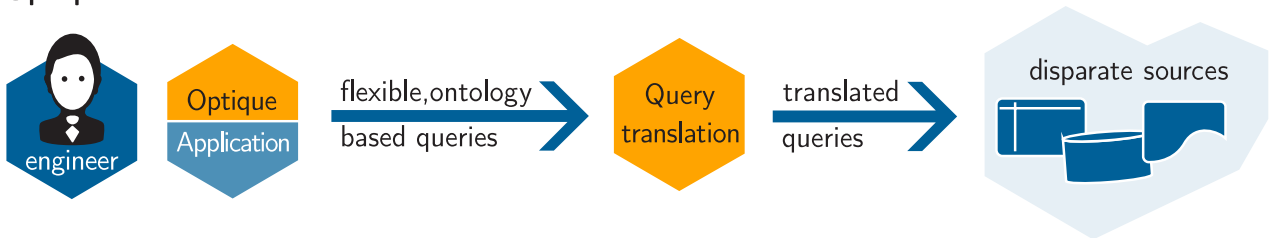
Complex case:



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 goal of Optique is to enable end users to formulate optimised special purpose queries on their own, without direct assistance from IT experts. In order to achieve this the **Optique platform** uses an *ontology* to capture (possibly multiple) user conceptualisations, so as to allow users to formulate queries using their own conceptualisations of the data. These user queries are then transformed into *complete, correct and highly optimised queries* over the data sources, which may include streams:

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.

2 The Optique Platform

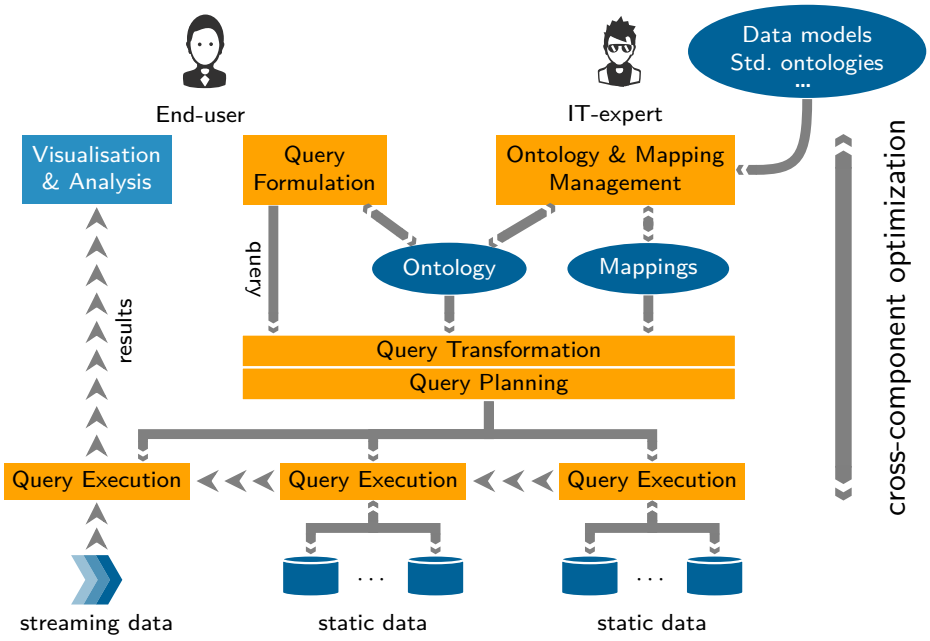


Figure 1: Optique platform architecture

The Optique platform 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.

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. Intelligent 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.

While the ontology captures crucial knowledge of the domain, the knowledge of the IT experts is captured by a set of declarative mappings that relate the ontology to the data sources. 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.

Back-end scalability is achieved by optimising the query transformation component that rewrites 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. 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.

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.

The platform runs against open data sources in a *Public showcase* installation, in addition to the two installations within the premises of Siemens and Statoil.

3 The Optique Partner Programme: A community of early adopters

Optique targets enterprises with big data challenges, and in particular challenges related not only to large *volumes* of data, but also to schemas and infrastructures of great *complexity*, new data coming in with high *velocity*, or data sources being structured according a large *variety* of schemas and formats.

Big data challeges are rooted deeply in organisations and their workflows; enterprises facing such challenges thus need to address not only new technologies, but also how well they know their data and their maturity for adopting new solutions. The Optique dissemination and exploitation strategy for targeted industry, summarised in Figure 2, assumes that a decision to adopt Optique will depend on a successful evaluation of a proof-of-concept implementation. This implementation should target a use case that is sufficiently complex, yet simple enough to be implemented and tested within a short timeframe and without requiring a major organizational effort.

Optique has identified five key phases of an enterprise project seeking to adopt the Optique platform, starting with exploration of introductory material, going through learning and assessment of business opportunities, and then to proof-of-concept implementation and evaluation. Optique delivers a comprehensive suite of resources in order to support the first three of these phases:

- Key personnel have to *observe* the features of Optique to be in position to decide on setting up a proof-of-concept project. Optique provides a proverbial *executive summary*, introductory presentations and *white papers* so as to enable staff to understand how Optique can meet their needs.
- In order to set up teams with the sufficient level of expertise for a proof-of-concept project, the enterprise has to go through a phase of *learning*. Optique delivers one-day *courses* to allow industry staff to explore how Optique can be applied to their specific needs and a *curriculum* with text-book character that can serve as reference material for implementation projects.
- A phase of *appraisal* is needed in order to identify suitable use cases. Optique provides analytic tools to support a focused work plan, leading up to an installed Optique system that delivers company data to expert users. Implementation guidelines and business model templates assist in choosing a use case, finding the right mix of people, and establishing an appropriate work process.

These resources will be freely available. Targeted events are offered to members of the community of early adopters, organised in the *Optique Partner Programme*. The Partner Programme is a framework for arranging binding, predictable agreements and a path to adoption that settles cost, responsibilities, and expectations. Ad hoc collaboration between a research programme and a commercial company is in general far from straightforward to set up. With the Optique Partner Programme, companies can sign up on a level appropriate to their interests, ranging from basic training on to use case evaluation and pilot project work.

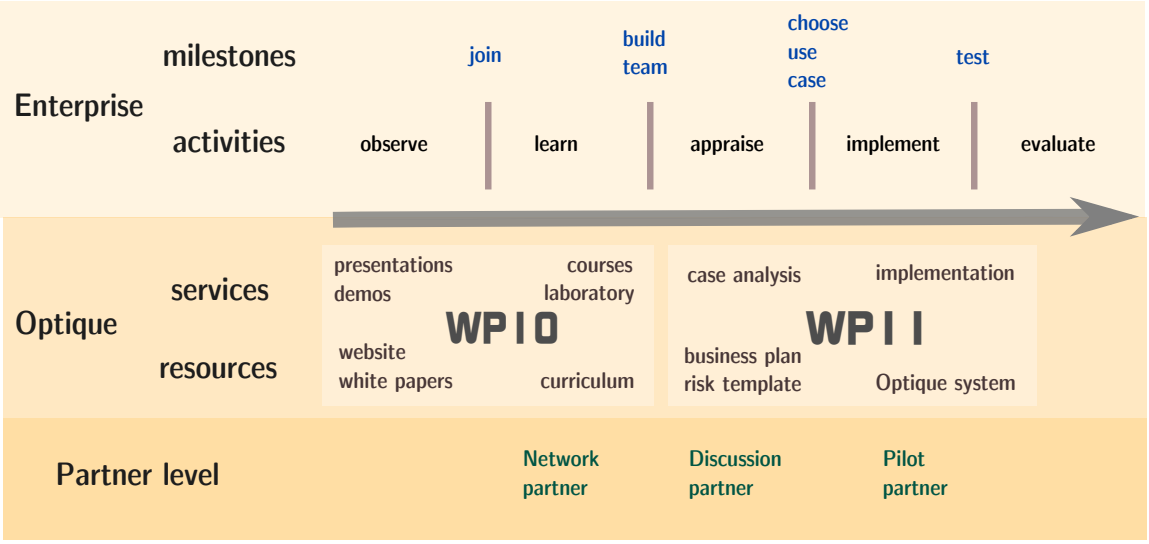


Figure 2: The Optique dissemination and exploitation strategy for targeted industry.

4 Project objectives

Use cases from Siemens Energy Services and Statoil Exploration drive the development of the Optique platform. Optique follows a four phase work plan; each phase starts November 1 and ends October 31 the following calendar year. The main priorities for each project year are outlined in Table 1. The overall project objectives are stated in Table 2.

Year	Optique Platform and Use Cases	Community Building
1	Early prototype incl. visual query interface – basic architecture – combine existing Background components Build initial Query Library Prototype processes simple queries	Make Optique widely known through: – publications, talks and demonstrations – targeted events in Siemens and Statoil – meetings with industry – development of fact sheets and web site
2	Functional prototype w/all components Build comprehensive Query Library Prototype processes realistic queries	Establish the Optique Partner Programme – targeted dissemination to network partners Release of first version of Public Showcase
3	Fully functional platform – new components fully integrated – Siemens: combine streams & static data – Statoil: federation of several sources	Extensive user training in Siemens/Statoil Partner Programme fully operational – deliver first courses – recruiting Discussion and Pilot Partners
4	Cross-component optimisation – increased usability & performance – additional functionality Most End User information needs covered	Broad showcasing in Siemens/Statoil Release of tutorials and Public Showcase Pilot Partners implementing demos Cultivation of user groups

Table 1: Phases of the Optique Work Plan

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.	
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.	Develop the Optique platform in response to requirements from use cases provided by Siemens and Statoil and deploy and evaluate the platform in these use cases.
Objective 6.	Widely disseminate Optique scientific results and technologies to the academic sector, to industry, and to the general public.
Objective 7.	Establish the Optique eco-system in order to disseminate and exploit the project results.

Table 2: The Optique project objectives

5 Project results up to Month 24

Main Objective. At Month 24, a fully functional prototype has been implemented and evaluated—with very promising results—in the two industrial use case scenarios.	
Objective 1.	• A generic architecture has been designed that can be adapted to any domain that requires scalable data access and efficient query execution. • Based on this architecture, an initial prototype of the Optique Platform has been developed. • The prototype has been deployed at the use case partners and is able to answer realistic user queries over large datasets (up to 1TB).
Objective 2.	• A stable prototype of the ontology-based visual query formulation subsystem has been implemented, based on multiple coordinated interaction paradigms, in particular graph navigation and facet refinement. Novel techniques to support query formulation have been developed. • Use case evaluations were carried out both with students and endusers from Statoil and Siemens with encouraging results showing that endusers can formulate realistic queries with mild training.
Objective 3.	• 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.
Objective 4.	• 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 and novel ontology approximation techniques. • A query rewriting subsystem has been implemented, based on the Ontop system; it supports OWL 2 QL, R2RML mappings, and most of the SPARQL query language, and improves performance by exploiting mappings and data dependencies. • 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.
Objective 5.	• Relevant user interface components, such as those for visual query formulation and browsing, have been presented to groups of end users at use-case workshops. • Feedback suggests that the overall approach is promising and that end-users are able to use the system.
Objective 6.	• Approx. 100 refereed publications, more than 100 presentations for the research community, more than 100 presentation for industry • The Optique Public Showcase, with a queryable data endpoint and sample queries and mappings. • An Executive White paper. • Several releases of the Query Transformation component as part of the open source ontop project.
Objective 7.	• The Optique Partner Program has been established. • An initial partner event with more than 20 companies have been arranged. • Partner events for 2015 have been scheduled.