# Abstract

In this thesis, we deal with techniques for query answering exploiting structural properties, with the integration of multiple data sources, and with the design and the implementation of a suite for process mining. The main contributions of this thesis are the following:

(1) A new algorithm that computes a hypertree decomposition of a query, by accounting for grouping operators and statistics on the data. .

(2) The study of advanced techniques and innovative methodologies for information integration systems and a prototype implementation of a knowledge based system for advanced information integration, by using computational logic and integrating research results on data acquisition and transformation.

(3) The study of techniques and algorithms for process mining and a suite implementing them.

## (1) Techniques for query evaluation

Answering queries is computationally very expensive, and many approaches have been proposed in the literature to face this fundamental problem. Some of them are based on optimization modules that exploit quantitative information on the database instance, while other approaches exploit structural properties of the query hypergraph.

Our efforts were carried on this last direction extending the notion of hypertree decomposition, which is currently the most powerful structural method. This new version, called query-oriented hypertree decomposition, is a suitable relaxation of hypertree decomposition designed for query optimization, and such that output variables and aggregate operators can be dealt with. Based on this notion, a hybrid optimizer is implemented, which can be used on top of available DBMSs to compute query plans. The prototype is also integrated into the well-known open-source DBMS PostgreSQL. Finally, we validate our proposal with a thorough experimental activity, conducted on PostgreSQL and on a commercial DBMS, which shows that both systems may significantly benefit from using hypertree decompositions for query optimization.

### (2) Techniques for data integration systems

Information integration is the problem of combining the data residing at different sources, and providing the user with a unified view of these data, called *global schema*. Our work was performed within of the INFOMIX project. Its principal goal was to provide advanced techniques and innovative methodologies for information integration systems. In a nutshell, the project developed a theory, comprising a comprehensive information model and information integration algorithms, and a prototype implementation of a knowledge based system for advanced information integration, by using computational logic and integrating research results on data acquisition and transformation. Special attention was devoted to the definition of declarative user-interaction mechanisms, and techniques for handling semi-structured data, and incomplete and inconsistent data sources.

### (3) Techniques for process mining

In the context of enterprise automation, *process mining* has recently emerged as a powerful approach to support the analysis and the design of complex business processes. In a typical process mining scenario, a set of traces registering the activities performed along several enactments of a transactional system is given to hand, and the goal is to (semi)automatically derive a model explaining all the episodes recorded in them.

We developed a novel Suite for Process Mining applications having an open and extendable architecture and introducing three innovative designing elements to meet the desiderata of flexibility and scalability arising in actual industrial scenarios.

- The concept of “flow of mining”, i.e., it allows to specify complex mining chains based on interconnecting elementary tasks

- Building interactive applications based on the possibility of customizing data types, algorithms, and graphical user interfaces used in the analysis.

- Ensuring scalability over large volumes of data.