Chapter 1
Advanced Query Processing: An Introduction

Barbara Catania and Lakhmi Jain

Abstract. Traditional query processing techniques have played a major role in the success of relational Database Management Systems over the last decade. However, they do not obviously extend to much more challenging, unorganized and unpredictable data providers, typical of emerging data intensive applications and novel processing environments. For them, advanced query processing and data integration approaches have been proposed with the aim of still guaranteeing an effective and efficient data access in such more complex data management scenarios. The aim of this chapter is to present the main issues and trends arising in advanced query processing and to relate them to the various parts of this book. For each part, a brief description of the background concepts and of the presented contributions is also provided.

1.1 Introduction

One of the main reasons for the success of Database Management Systems (DBMSs) and of the main key concepts upon which traditional query processing techniques have been developed is logical data independence. With logical data independence we mean the neat separation between the specification of ‘what’ we are searching for from ‘how’ these searches, specified in terms of queries, are processed. The system is responsible for transforming declarative queries into execution plans, statically determined before the processing starts. The result obtained by processing...
the query according to the chosen execution plan is the set of items which exactly satisfy the specified query conditions.

Since the late 1970s and the introduction of System R \cite{17}, this approach has proved to be highly efficient and effective and it has played a major role in the relational DBMSs success over the last decades. Efficiency is guaranteed by the usage of several sophisticated optimization techniques. These techniques heavily rely on the existence of metadata information about the data which have to be processed, such as the distribution of values and the selectivity of the relational operators. Effectiveness is guaranteed by the usage of simple declarative languages which perfectly adhered to the needs of traditional database domains and applications, characterized by data which have a completely known structure, are executed in stable environments, and for which a reasonable set of statistical information on data is usually available.

The query processing language-to-engine stack of classical DBMSs embodies advanced and fundamentally elegant concepts and ideas, that however do not obviously extend to much more challenging, unorganized and unpredictable data providers, typical of emerging data intensive applications and novel processing environments. New applications and environments include, to mention only a few, data integration applications, web services, Future Web, sensors databases and networks, P2P, cloud computing, and hosting. They are characterized by: high network connectivity and resource sharing (as in data integration applications, Future Web, cloud computing, and hosting); new types of data availability (data can be stored or produced as a stream as in sensor databases); high data heterogeneity and incompleteness (as in data integration applications, P2P, and cloud computing); extremely high variability and unpredictability of data characteristics during the processing (because transactions could be long-running or the data schema may change during computation); limited user knowledge about the data which have to be processed and limited resources with respect to the data volumes under processing (the response time should be low also in presence of high volume of data as in cloud computing or sensor databases and the space may not be sufficient to store all of the data, which are sometimes unbounded, as in data streams).

All the characteristics discussed above make traditional query processing and data integration approaches not feasible for most of the new processing environments and lead to a radical modification of query processing requirements. As a consequence, new query processing approaches have been defined, that we call advanced. Two main innovative aspects are taken into account by advanced query processing techniques. A first issue concerns approximation. Data characteristics (e.g., heterogeneity, incompleteness, and uncertainty), resource limitations, huge data volumes, and volatility, typical of the new applications and processing environments, suggest it may be preferred to relax the query definition, using Query Relaxation (QR) techniques, or to generate an approximate result set, with quality guarantees, using Approximate Query Processing (ApQP) techniques, instead of getting an unsatisfactory answer. An answer can be unsatisfactory because either the user has