

Preface

Since the 1960s, database systems have been playing a relevant role in the information technology field. By the mid-1960s, several systems were also available for commercial purposes. Hierarchical and network database systems provided two different perspectives and data models to organize data collections. In 1970, E. Codd wrote a paper called *A Relational Model of Data for Large Shared Data Banks*, proposing a model relying on relational table structures. Relational databases became appealing for industries in the 1980s, and their wide adoption fostered new research and development activities toward advanced data models like object oriented or the extended relational. The online transaction processing (OLTP) support provided by the relational database systems was fundamental to make this data model successful. Even though the traditional operational systems were the best solution to manage transactions, new needs related to data analysis and decision support tasks led in the late 1980s to a new architectural model called data warehouse. It includes extraction transformation and loading (ETL) primitives and online analytical processing (OLAP) support to analyze data. From OLTP to OLAP, from transaction to analysis, from data to information, from the entity-relationship data model to a star/snowflake one, and from a customer-oriented perspective to a market-oriented one, data warehouses emerged as data repository architecture to perform data analysis and mining tasks. Relational, object-oriented, transactional, spatiotemporal, and multimedia data warehouses are some examples of database sources. Yet, the World Wide Web can be considered another fundamental and distributed data source (in the Web2.0 era it stores crucial information – from a market perspective – about user preferences, navigation, and access patterns).

Accessing and processing large amount of data distributed across several countries require a huge amount of computational power, storage, middleware services, specifications, and standards.

Since the 1990s, thanks to Ian Foster and Carl Kesselman, grid computing has emerged as a revolutionary paradigm to access and manage distributed, heterogeneous, and geographically spread resources, promising computer power as easy to access as an electric power grid. The term “resources” also includes the database,

yet successful attempts of grid database management research efforts started only after 2000. Later on, around 2007, a new paradigm named *Cloud Computing* brought the promise of providing easy and inexpensive access to remote hardware and storage resources. Exploiting pay per use models, virtualization for resource provisioning, cloud computing has been rapidly accepted and used by researchers, scientists, and industries.

Grid and cloud computing are exciting paradigms and how they deal with database management is the key topic of this book. By exploring current and future developments in this area, the book tries to provide a thorough understanding of the principles and techniques involved in these fields.

The idea of writing this book dates back to a tutorial on Grid Database Management that was organized at the 4th International Conference on Grid and Pervasive Computing (GPC 2009) held in Geneva (4–8 May 2009). Following up an initial idea from Ralf Gerstner (Springer Senior Editor Computer Science), we decided to act as editors of the book.

We invited internationally recognized experts asking them to contribute on challenging topics related to grid and cloud database management. After two review steps, 16 chapters have been accepted for publication.

Ultimately, the book provides the reader with a collection of chapters dealing with *Open standards and specifications* (Sect. 1), *Research efforts on grid database management* (Sect. 2), *Cloud data management* (Sect. 3), and some *Scientific case studies* (Sect. 4). The presented topics are well balanced, complementary, and range from well-known research projects and real case studies to standards and specifications as well as to nonfunctional aspects such as security, performance, and scalability, showing up how they can be effectively addressed in grid- and cloud-based environments.

Section 1 discusses the open standards and specifications related to grid and cloud data management. In particular, Chap. 1 presents an overview of the WS-DAI family of specifications, the motivation for defining them, and their relationships with other OGF and non-OGF standards. Conversely, Chap. 2 outlines the OCCI specifications and demonstrates (by presenting three interesting use cases) how they can be used in data management-related setups.

Section 2 presents three relevant research efforts on grid-database management systems. Chapter 3 provides a complete overview on the Grid Relational Catalog (GRelC) Project, a grid database research effort started in 2001. The project's main features, its interoperability with gLite-based production grids, and a relevant show-case in the environmental domain are also presented. Chapter 4 provides a complete overview about the OGSA-DAI framework, the main components for the distributed data management via workflows, the distributed query processing, and the most relevant security and performance aspects. Chapter 5 gives a detailed overview of the architecture and implementation of DASCOSA-DB. A complete description of novel features, developed to support typical data-intensive applications running on a grid system, is also presented.

Section 3 provides a wide overview on several cloud data management topics. Some of them (from Chaps. 6 to 8) specifically focus only on database aspects, whereas the remaining ones (from Chaps. 9 to 12) are wider in scope and address more general cloud data management issues. In this second case, the way these concepts apply to the database world is clarified through some practical examples or comments provided by the authors. In particular, Chap. 6 proposes a new security technique to measure the trustiness of the cloud resources. Through the use of the metadata of resources and access policies, the technique builds the privilege chains and binds authorization policies to compute the trustiness of cloud database management. Chapter 7 presents a method to manage the data with dirty data and obtain the query results with quality assurance in the dirty data. A dirty database storage structure for cloud databases is presented along with a multilevel index structure for query processing on dirty data. Chapter 8 examines column-oriented databases in virtual environments and provides evidence that they can benefit from virtualization in cloud and grid computing scenarios. Chapter 9 introduces a Windows Azure case study demonstrating the advantages of cloud computing and how the generic resources offered by cloud providers can be integrated to produce a large dynamic data store. Chapter 10 presents CloudMiner, which offers a cloud of data services running on a cloud service provider infrastructure. An example related to database management exploiting OGSA-DAI is also discussed. Chapter 11 defines the requirements of e-Science provenance systems and presents a novel solution (addressing these requirements) named the Vienna e-Science Provenance System (VePS). Chapter 12 examines the state of the art of workload management for data-intensive computing in clouds. A taxonomy is presented for workload management of data-intensive computing in the cloud and the use of the taxonomy to classify and evaluate current workload management mechanisms.

Section 4 presents a set of scientific use cases connected with Genomic, Health, Disaster monitoring, and Earth Science. In particular, Chap. 13 explores the implementation of an algorithm, often used to analyze microarray data, on top of an intelligent runtime that abstracts away the hard parts of file tracking and scheduling in a distributed system. This novel formulation is compared with a traditional method of expressing data parallel computations in a distributed environment using explicit message passing. Chapter 14 describes the use of Grid technologies for satellite data processing and management within the international disaster monitoring projects carried out by the Space Research Institute NASU-NSAU, Ukraine (SRI NASU-NSAU). Chapter 15 presents the CDM ActiveStorage infrastructure, a scalable and inexpensive transparent data cube for interactive analysis and high-resolution mapping of environmental and remote sensing data. Finally, Chap. 16 presents a mechanism for distributed storage of multidimensional EEG time series obtained from epilepsy patients on a cloud computing infrastructure (Hadoop cluster) using a column-oriented database (HBase).

The bibliography of the book covers the essential reference material. The aim is to convey any useful information to the interested readers, including researchers actively involved in the research field, students (both undergraduate and graduate), system designers, and programmers.

The book may serve as both an introduction and a technical reference for grid and cloud database management topics. Our desire and hope is that it will prove useful while exploring the main subject, as well as the research and industries efforts involved, and that it will contribute to new advances in this scientific field.

Lecce
February 2010

Sandro Fiore
Giovanni Aloisio



<http://www.springer.com/978-3-642-20044-1>

Grid and Cloud Database Management

Fiore, S.; Aloisio, G. (Eds.)

2011, X, 353 p., Hardcover

ISBN: 978-3-642-20044-1