INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team ATLAS

Complex Data Management in Distributed Systems

Rennes - Bretagne-Atlantique, Sophia Antipolis - Méditerranée

Theme : Knowledge and Data Representation and Management

Table of contents

1. Team

Research Scientists

Patrick Valduriez [Team Leader, Senior Researcher, INRIA, *Montpellier*, HdR]
Reza Akbarinia [Junior Researcher, INRIA, *Nantes* until august, *Montpellier* since september]

Faculty Members

Rémi Coletta [Professor, University Montpellier 2, since september (INRIA delegation)]
Marc Gelgon [Professor, University of Nantes, HdR]
Philippe Lamarre [Associate Professor, University of Nantes, HdR]
Guillaume Raschia [Associate Professor, University of Nantes]
Gerson Sunyé [Associate Professor, University of Nantes until august]

External Collaborator

Esther Pacitti [Professor, University Montpellier 2, since september, HdR]

Technical Staff

Emmanuel Castanier [Engineer, ADT WebSmatch since october, *Montpellier*]
Guillaume Verger [Engineer, ANR DataRing since april, *Montpellier*]

PhD Students

Luis Barguno [UPC, Barcelona; INRIA COLOR funding (july-december), *Montpellier*]
Pierrick Bruneau [ANR funding, *Nantes*]
Thomas Cerqueus [MENRT fellowship, *Nantes*]
William Kokou Dedzoe [CNRS fellowship, *Nantes*]
Duc Trung Vu [INRIA-Pays-de-la-Loire fellowship since february, *Nantes*]
Fady Draidi [MAE-INRIA fellowship, *Montpellier*]
Ali El Attar [Pays-de-la-Loire fellowship, *Nantes*]
Mohamed Jawad [MENRT fellowship, *Nantes*]
Miguel Liroz [INRIA CORDIS fellowship since october, *Montpellier*]
Wenceslao Palma [INRIA fellowship until july, *Nantes*]
Gianvito Summa [Univ. of Basilicata, Italy; INRIA COLOR funding (june-november), *Montpellier*]
Toufik Sarni [CNRS fellowship, *Nantes*]
Mounir Tlili [Pays-de-la-Loire fellowship, *Nantes*]

Visiting Scientists

Angela Bonifati [CNR, Rende, Italy (april-june) *Montpellier*]
Victor Muntés Mulero [UPC, Barcelona, until march, *Montpellier*]

Administrative Assistants

Annie Aliaga [*Montpellier*]
Hanane Maaroufi [*Nantes*]

2. Overall Objectives

2.1. Introduction

Today's hard problems in data management go well beyond the traditional context of Database Management Systems (DBMS). These problems stem from significant evolutions of data, systems and applications. First, data have become much richer and more complex in formats (e.g., multimedia objects), structures (e.g., semi-structured documents), content (e.g., incomplete or imprecise data), size (e.g., very large volumes), and associated semantics (e.g., metadata, code). The management of such data makes it hard to develop data-intensive applications and creates hard performance problems. Second, data management systems need to scale up to support large-scale distributed systems and deal with both fixed and mobile clients. In a highly distributed context, data sources are typically in high number, autonomous and heterogeneous, thereby making

data management difficult. Third, this combined evolution of data and systems gives rise to new, typically complex, applications with ubiquitous, on-line data access: collaborative content management (e.g. Wiki), virtual libraries, virtual stores, global catalogs, services for personal content management, etc.

The general problem can be summarized as *complex data management in distributed systems*. The Atlas project-team addresses this problem with the objective of designing and validating new solutions with significant advantages in functionality and performance. To tackle this objective, we focus on data management in large-scale distributed contexts, including web, P2P systems and cloud. In the context of the web, we have continued our work on data classification and database summarization. In the context of P2P systems, we capitalized on our experience in developping the APPA system, with various data management services (replication, caching, queries, clustering, privacy, testing, etc.). In the context of cloud computing, with large-scale parallelism, we have started to work on data streaming, scientific workflow management and massive graph databases. We have also continued to work on data management in multicore systems, in particular, real-time data access through transactional memory.

The Atlas project-team was formally ended on 31 dec. 2010. However, the research initiated in Atlas does continue, both in Nantes and Montpellier, in various forms. Research in model engineering is continuing within the Atlanmod INRIA team in Nantes. Research in multimedia data management and P2P systems will continue at LINA, Nantes. Finally, P. Valduriez is starting a new INRIA team, called Zenith, at INRIA Sophia-Antipolis Méditerranée, located in Montpellier, with a focus on scientific data management.

2.2. Highlights of the Year

The third edition of the book *Principles of Distributed Database Systems* [12] has been completed (with Springer). This long awaited major revision is now about 850 pages. In addition to the fundamental principles of distributed data management, it now covers new hot topics such as web data management, peer-to-peer, data streaming, and cloud.

3. Scientific Foundations

3.1. Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMSs, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (application servers, document systems, directories, etc.).

The fundamental principle behind data management is *data independence*, which enables applications and users to deal with the data at a high conceptual level while ignoring implementation details. The relational model, by resting on a strong theory (set theory and first-order logic) to provide data independence, has revolutionized data management. The major innovation of relational DBMS has been to allow data manipulation through queries expressed in a high-level (declarative) language such as SQL. Queries can then be automatically translated into optimized query plans that take advantage of underlying access methods and indices. Many other advanced capabilities have been made possible by data independence : data and metadata modeling, schema management, consistency through integrity rules and triggers, transaction support, etc.

This data independence principle has also enabled DBMS to continuously integrate new advanced capabilities such as object and XML support and to adapt to all kinds of hardware/software platforms from very small smart devices (PDA, smart card, etc.) to very large computers (multiprocessor, cluster, etc.) in distributed environments.

Following the invention of the relational model, research in data management has continued with the elaboration of strong database theory (query languages, schema normalization, complexity of data management algorithms, transaction theory, etc.) and the design and implementation of DBMS. For a long time, the focus was on providing advanced database capabilities with good performance, for both transaction processing and decision support applications. And the main objective was to support all these capabilities within a single DBMS.

Today's hard problems in data management go well beyond the context of DBMS. These problems stem from the need to deal with data of all kinds, in particular, multimedia data and data streams, in highly distributed environments. Thus, we capitalize on scientific foundations in data reduction techniques, distributed data management, data stream management and semantic interoperability to address these problems. To deal with uncertain data, we rely on probabilistic databases which provide a powerful foundation for our work.

3.2. Data Reduction Techniques

With the explosion of the quantities of data to be analyzed, it is desirable to sacrifice the accuracy of the answers for response time. Particularly in the early, more exploratory, stages of data analysis, interactive response times are critical, while tolerance for approximation errors is quite high. In this context, data reduction is important to control the desired trade-off between answer accuracy and response time.

Data reduction is closely associated with aggregation. While histograms form the baseline approach and have been extensively used for query optimizers, a wealth of techniques have been proposed. In particular, cluster-based reduction of data, where each data item is identified by means of its cluster representative, leads to classical tree indexes, where data is partitioned recursively into buckets. The clusters may be data-driven, or independent from the data. With minimal augmentation, it becomes possible to answer queries approximately based upon an examination of only the top levels of an index tree. If these top levels are cached in memory, as is typically the case, then one can view these top levels of the tree as a reduced form of data suitable for approximate query answering.

To deal with large amounts of data, or high-dimensional data, much work has also been devoted to reducing the dimension of representations, by identifying lower dimension manifolds on which data essentially lies. Single-value decomposition or discrete wavelet transformations are two examples of such transform-based techniques. Among data reduction techniques, one may further distinguish parametric techniques (e.g. linear regression), that assume a model for the data, from non-parametric techniques. While the former offer generally more compression, automatically selecting the form of the model remains a difficult issue.

An important use of data reduction is for retrieval within collections of multimedia material, such as image, audio or video. For the purpose of comparing queries to target documents or for building an index, these documents are represented by features, i.e. multivariate attributes. These features may be used directly (e.g. nearest neighbourhood search among feature vectors, for image matching) or, often, through probabilistic models of their distribution, thereby capturing the variability of a given class. The design of these features requires a specific expertise for each media, to ensure a good trade-off between concision, ability to discriminate and invariance to certain imaging or acoustic conditions. This is typically handled by media-specific research communities.

Nearest neighbor queries are appropriate for multimedia information retrieval. Efficient multimedia feature vectors often span high dimensional spaces, where indexing structures classically used in database management systems (tree-based and hashing-based) are not effective, due to the dimensionality curse. Parallel databases may contribute to maintaining reasonable query processing time, but require the definition of data distribution strategies. Such strategies are one of the focuses of our work.

Among models, parametric probabilistic models build a very rich, well-founded and well-documented toolbox for representing the data distributions in a concise way, in association to statistical estimation techniques for determining the form of the model and values of its parameters. Together, they provide a strong share of existing solutions to multimedia data analysis problems (learning and recognition). Relating this to database summaries, seeking simple forms to *describe* the data (structure for efficient retrieval) and forms that *explain*

the data (structure for understanding, where parametric forms introduce the necessary inductive bias) are often very close goals, hence a growing number of techniques common to the database and machine learning communities. Among probabilistic models, generative mixture models consider the data to be a combination of several populations, whether this correspond to true variety of natures or whether is a only a modeling tool. Mixtures have wide modeling ability, like non-parametric methods, but retain the parsimony of parametric approaches. Hence, they have been much studied, extended and applied, in the contexts of both supervised and unsupervised learning. In the case of probabilistic models, Bayesian estimation supplies a principled solution to the above-mentioned model selection. This long remained either computation-intensive or very approximative, but nowadays, besides increasing computing power available, a corpus of efficient approximate inference mechanisms has been built, for a growing variety of graphical model structures. There remain questions which are receiving growing attention : how can such models be efficiently learned from dynamic distributed data sources ? How can a large set of probabilistic models be indexed ?

Among the broad range of reduction techniques, the database summarization paradigm has become an ubiquitous requirement for a variety of application environments, including corporate data warehouses,network-traffic monitoring and large socio-economic or demographic surveys. Besides, downsizing massive data sets allows to address some critical issues such as individual data obfuscation, optimization of the usage of system resources like storage space and network bandwidth, as well as effective approximate answers to queries. Depending on the application environment and the preferred goal of the approach, we distinguish three families of approaches concerned with database summarization. The first one focuses on aggregate computation and it is supported by statistical databases, OLAP cubes and multidimensional databases.The second class of approaches extends the previous one in that it tries to produce more compact representations of aggregates. The main challenge for such methods is to keep expressiveness of the provided access methods (aggregate queries) to the items without any need to uncompress the structure. Quotient cubes and linguistic summaries are two major contributions in that direction. The third family of approaches deals with intentional characterization of groups of individuals based on usual mining algorithms. Those categories are obviously not sharp and there are many orthogonal criteria that encompass such a classification. For instance, some of them share the same theoretical background (Zadeh's fuzzy set theory) and they use fuzzy partitions and linguistic variables to support a robust summarization process.

This database research field raises new challenges, in particular, to push more semantics into summaries while still remaining efficient in the context of database systems. Update of such metadata is also of major concern. Furthermore, traditional problems of data management such as query evaluation or data integration have to be revisited from the point of view of database summaries.

3.3. Probabilistic Databases

The generation of massive amounts of data with various levels of control and quality makes data uncertainty ubiquitous in many applications. Examples include web data cleaning, sensor networks, information extraction, data integration, RFID stream analysis, etc. Data uncertainty can be well captured by associating probabilities of data which is the basis for probabilistic databases. Thus, a probabilistic database management system (PDBMS) is a system that deals with storing and retrieving probabilistic data as well as supporting complex queries over the data. There are two important issues which any PDBMS should address: 1) how to represent a probabilistic database, i.e. data model; 2) how to answer queries using the chosen representation, i.e. query evaluation.

There are two main probabilistic data models which are the tuple level and attribute level models. With the tuple level model, each tuple t has an attribute that indicates the membership probability (also called existence probability) of t, i.e. the probability that the tuple appears in a random instance of the database. In the attribute level model, each tuple t has at least one uncertain attribute, e.g. a. The value of a in t is determined by a random variable whose probability density function (pdf) may be from a discrete or continuous domain. In both models, the tuples of the probabilistic database may be independent or correlated. Although the models that support correlation are more powerful than the others; they usually require exponential processing complexity.

Query evaluation is the hardest technical challenge in a PDBMS. A naive solution for evaluating probabilistic queries is to enumerate all possible worlds, i.e. all possible instances of the database, execute the query in each world, and return the possible answers together with their cumulative probabilities. However, this solution is not efficient due to the exponential number of possible worlds which a probabilistic database may have. Some queries can be evaluated on a probabilistic database by pushing the probabilistic computation inside the query plan. Thus, for these queries the output probabilities are computed inside the database engine, using the normal query processing. Queries for which this computation is possible are called safe queries, and the execution plan that computes the output probabilities is called a safe plan. However, there are many queries for which there is no safe plan, e.g. those containing self joins. For some complex queries, e.g. top-k and aggregate queries, we need to redefine the semantics of the query. For example, for top-k queries we should decide on how to take into account both tuple probabilities and scores in ranking the tuples. Although much research has been done in few last years on complex query evaluation in probabilistic databases, there remain many open problems in this domain.

Though difficult in centralized systems, the problem of query evaluation is more complicated in distributed systems, particularly because of new challenges in schema mapping and query routing. There may be some type of uncertainty in the defined schema mappings which should be considered in query reformulation, and in execution plans. Furthermore, the query must be routed to the nodes that involve relevant data with high probabilities.

3.4. Distributed Data Management

The Atlas project-team considers data management in the context of distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of distributed systems, in particular, large-scale distributed systems such as clusters, grid, and peer-to-peer (P2P) systems, to address issues in data replication and high availability, load balancing, and query processing.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL) [12]. Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system is a centralized server that supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases. Data integration systems extend the distributed database approach to access data sources on the Internet with a simpler query language in read-only mode.

Parallel database systems also extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

In contrast, peer-to-peer (P2P) systems adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. Popular examples of P2P systems such as Gnutella and Kaaza have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g. file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. To deal with the dynamic behavior of peers that can join and leave the system at any time, they rely on the fact that popular data get massively duplicated.

Initial research on P2P systems has focused on improving the performance of query routing in the unstructured systems which rely on flooding, whereby peers forward messages to their neighbors. This work led to structured solutions based on Distributed Hash Tables (DHT), e.g. CAN and CHORD, or hybrid solutions with super-peers that index subsets of peers. Another approach is to exploit gossiping protocols, also known as

epidemic protocols. Gossiping has been initially proposed to maintain the mutual consistency of replicated data by spreading replica updates to all nodes over the network. It has since been successfully used in P2P networks for data dissemination. Basic gossiping is simple. Each peer has a complete view of the network (i.e. a list of all peers' addresses) and chooses a node at random to spread the request. The main advantage of gossiping is robustness over node failures since, with very high probability, the request is eventually propagated to all nodes in the network. In large P2P networks, however, the basic gossiping model does not scale as maintaining the complete view of the network at each node would generate very heavy communication traffic. A solution to scalable gossiping is by having each peer with only a partial view of the network, e.g. a list of tens of neighbor peers. To gossip a request, a peer chooses at random a peer in its partial view to send it the request. In addition, the peers involved in a gossip exchange their partial views to reflect network changes in their own views. Thus, by continuously refreshing their partial views, nodes can self-organize into randomized overlays which scale up very well.

Other work has concentrated on supporting advanced applications which must deal with semantically rich data (e.g., XML documents, relational tables, etc.) using a high-level SQL-like query language. Such data management in P2P systems is quite challenging because of the scale of the network and the autonomy and unreliable nature of peers. Most techniques designed for distributed database systems which statically exploit schema and network information no longer apply. New techniques are needed which should be decentralized, dynamic and self-adaptive.

3.5. Data Stream Management

Recent years have witnessed major research interest in data stream management systems. A data stream is a continuous and unbounded sequence of data items. There are many applications that generate streams of data including financial applications, network monitoring, telecommunication data management, sensor networks, etc. Processing a query over a data stream involves running the query continuously over the data stream and generating a new answer each time a new data item arrives. Due to the unbounded nature of data streams, it is not possible to store the data entirely in a bounded memory. This makes difficult the processing of queries that need to compare each new arriving data with past ones. A common solution to the problem of processing join queries over data streams is to execute the query over a sliding window that maintains a restricted number of recent data items. This allows queries to be executed in a finite memory and in an incremental manner by generating new answers when a new data item arrives. Due to the continuous, often very fast, arrival of new data, it is impossible to produce exact answers to queries. Therefore, approximate answers are typically provided.

In real data settings, a data stream management system may process hundreds of user queries. Therefore, for most realistic distributed streaming applications the naive solution of collecting all the data at a single site is simply not viable. Therefore, we are interested in techniques for processing continuous queries over collections of distributed data streams. An example of such queries is join queries which are very important for many applications. A streaming join computation can be useful in understanding important trends and making decisions about measurements or utilization patterns.

3.6. Semantic Interoperability

Semantic interoperability ensures that the meaning of the information that is exchanged is automatically interpreted by the receiver of a message. In centralized systems, this property improves the relevance of query answers. In distributed heterogeneous systems, it is compulsory to enable autonomous heterogeneous sources understand each other to obtain relevant results.

To provide semantic interoperability within a system, much research has been conducted on semantic representations. The main idea is to use meta-information which eases the meaning understanding. This approach needs the definition of ontologies which describe the concepts and relations between them, for a given domain. During the last fifteen years, much effort has focused on formal methods to describe ontologies, resource description languages, reasoning engines...All these methods represent the foundations

of the semantic web. However, many works rely on the assumption that a single ontology is shared by all the participants of the system.

However, in distributed systems with autonomous participants, such as P2P systems, this assumption is not realistic anymore. On the contrary, one has to consider that the participants create their ontologies independently of each other. Thus, most often the ontologies differ. To tackle this problem, research on ontology matching proposes several techniques to define correspondences between entities of two ontologies. So, in some way, ontology matching highlights the shared parts of two ontologies. Thus it provides the basis for interoperability between heterogeneous participants and by "transitivity" in the whole system.

Although ontology matching and other semantic web techniques provide a basis for interoperability, the challenge is still to define a whole semantic infrastructure in which participants' search for information is both relevant and efficient. Considering semantics can be useful at different stages. First, semantic representation of queries and information may improve the relevance of the results. It can be used in place or in addition to usual request representation. Second, semantics can be used to represent participants, or groups of them, leading participants to better know each other. Such information can be useful for routing the requests to other participants in order to obtain the relevant answers within a short time and with a low traffic load. Third, this information can also be used to organize the network so as to improve efficiency. All these research directions have received partial answers but more work is needed on the interaction between all these elements and their impact on the efficiency of the global system.

4. Application Domains

4.1. Overview

Complex data management in distributed systems is quite generic and can apply to virtually any kind of data. Thus, we are potentially interested in many applications which help us demonstrate and validate our results in real-world settings. However, data management is a very mature field and there are well-established application scenarios, e.g., the On Line Transaction Processing (OLTP) and On Line Analytical Processing (OLAP) benchmarks from the Transaction Processing Council (TPC). We often use these benchmarks for experimentation as they are easy to deploy in our prototypes and foster comparison with competing projects.

However, there is no benchmark that can capture all the requirements of complex data management. Therefore, we also invest time in real-life applications when they exhibit specific requirements that bring new research problems. Examples of such applications are large-scale distributed collaborative applications, large decision-support applications or multimedia personal databases. This last year in Montpellier, we have started to study scientific applications.

Large scale distributed collaborative applications are getting common as a result of the progress of distributed technologies (GRID, P2P, and cloud). Consider a professional community whose members wish to elaborate, improve and maintain an on-line virtual document, e.g. reading or writing notes on classical literature, or common bibliography, supported by a P2P system. They should be able to read/write on the application data. An important aspect of large scale distributed collaborative applications is that user nodes may join and leave the network whenever they wish, thus hurting data availability. Other examples of collaborative applications we are interested in are social networks. In Atlas, we address the issues of data sharing for such applications assuming a P2P architecture (APPA) that is fully decentralized.

Large decision-support applications need to manipulate information from very large databases in a synthetic fashion. A widely used technique is to define various data aggregators and use them in a spreadsheet-like application. However, this technique requires the user to make strong assumptions on which aggregators are significant. We propose a new solution whereby the user can build a general summary of the database that allows more flexible data manipulation.

A major application of multimedia data management that we are dealing with is multimedia personal databases which can help retrieve and classify personal audio-visual material stored either locally on a PC/Settop-box, or a mobile handset. Content-based retrieval from distributed multimedia documents is also an important class of applications.

Scientific data management has become a major challenge for the data management research community. Modern science such as agronomy, bio-informatics, physics and environmental science must deal with overwhelming amounts of experimental data produced through empirical observation and simulation. Such data must be processed (cleaned, transformed, analyzed) in all kinds of ways in order to draw new conclusions, prove scientific theories and produce knowledge. However, constant progress in scientific observational instruments (e.g. satellites, sensors, large hadron collider) and simulation tools (that foster in silico experimentation, as opposed to traditional in situ or in vivo experimentation) creates a huge data overload. Scientific data is also very complex, in particular because of heterogeneous methods used for producing data and the inherently multi-scale nature (spatial scale, temporal scale) of many sciences, resulting in data with hundreds of attributes or dimensions. Processing and analyzing such massive sets of complex scientific data is therefore a major challenge since solutions must combine new data management techniques with large-scale parallelism in cluster, grid or cloud environments

5. Software

5.1. APPA (Atlas Peer-to-Peer Architecture)

Participants: Reza Akbarinia, William Kokou Dedzoe, Philippe Lamarre, Esther Pacitti, Gerson Sunyé, Mounir Tlili, Patrick Valduriez [contact].

URL: http://atlas.lina.univ-nantes.fr/gdd/appa/

APPA is a P2P data management system that provides scalability, availability and performance for applications which deal with semantically rich data (XML, relational, etc.). APPA provides advanced services such as queries, replication and load balancing. It is being implemented on top on various P2P networks such as JXTA, OpenChord and Pastry and tested on GRID5000 and PlanetLab. The current services of APPA are (see below): KTS, P2P-LTR, SbQA, PeerUnit and Priserv. The APPA services have been used in important projects: Strep Grid4All, ANR RNTL Xwiki Concerto, ANR VERSO DataRing and PREDIT Epilog.

5.2. P2P-LTR (P2P Logging and Timestamping for Reconciliation)

Participants: Reza Akbarinia, William Kokou Dedzoe, Esther Pacitti [contact], Mounir Tlili, Patrick Valduriez.

URL: http://p2pltr.gforge.inria.fr/

P2P-LTR provides two major functions: logging of user actions in a DHT and continuous, distributed timestamping of these actions. This is useful to perform reconciliation of replicated data. P2P-LTR extends KTS with continuous timestamping and logging of actions. To perform reconciliation using P2P-LTR, we use a simple reconciliation algorithm based on operational transforms, called SB, from the ECOO team at LORIA and readily available as Open Source Software. P2P-LTR has been implemented in Java on top of OpenChord. It has been validated in the Strep Grid4All and RNTL Xwiki Concerto projects to perform reconciliation of replicated documents in a P2P wiki system.

5.3. SbQA (Satisfaction-based Query Allocation Framework)

Participants: Philippe Lamarre [contact], Patrick Valduriez.

URL: http://atlas.lina.univ-nantes.fr/gdd/appa/sbqa/

SbQA is a Satisfaction-based Query Allocation framework for distributed environments where consumers and providers are autonomous and have special interests towards providers and queries, respectively. We experimentally demonstrated that it ensures good system performances while satisfying consumers and providers. Hence, SbQA can scale-up in these environments by preserving the total system capacity, i.e. the aggregate capacity of all providers. SbQA is used in the Strep Grid4All project as the basis to perform selection of services proposed by market-places as well as altruist contributors. SbQA is implemented in Java.

5.4. PeerUnit (Peer-to-Peer Tester)

Participants: Eduardo Almeida [contact], Gerson Sunyé, Patrick Valduriez.

URL: http://peerunit.gforge.inria.fr/

Peerunit is a testing framework for P2P systems. It is useful to developers who want to implement unit tests for a Java P2P system. The framework is based on two original aspects: (i) the individual control of peers volatility and (ii) a distributed testing architecture to cope with large numbers of peers. A distributed component, the tester, executes on peers, and controls their execution and their volatility, making them leave and join the system at any time, according to the needs of a test. Furthermore, testers communicate with each other across a balanced tree structure to avoid using a centralized testing coordination. Peerunit is implemented in Java and has been validated on two popular open-source P2P systems (FreePastry and OpenChord).

5.5. Priserv (Privacy Service)

Participants: Mohamed Jawad, Patrick Valduriez [contact].

URL: http://atlas.lina.univ-nantes.fr/gdd/appa/priserv/

PriServ is a privacy service for P2P data sharing that combines purpose-based access control, trust and encryption, for applications with sensitive data, e.g. medical data. The key feature is that owner peers (data publishers) keep full control over their private data and private keys. Data publishing in PriServ takes into account owner privacy preferences and does not reveal any private information about data (encrypted data or data references). PriServ uses a DHT to efficiently locate data. It is implemented in Java using the Service Component Architecture and Java RMI for peer communication. The implementation uses the Chord DHT but any other DHT could be used.

5.6. DBSum

Participants: Mounir Bechchi, Guillaume Raschia [contact].

URL: http://www.polytech.univ-nantes.fr/grim/doku.php?id=dbsum

DBSUM is a *Database Summary Management System* that provides various tools to support data reduction with query and analytical processing techniques on top of a DBMS. The current implementation has two parts: a summarization engine, namely SAINTETIQ, for building and updating database summaries; a full-feature user interface coined SEQT (*Summary Exploration and Querying Tool*) which provides languages, algorithms and views to query, search and browse into summaries. SAINTETIQ computes and maintains abstract and user-friendly views from very large databases. Aa an alternative to the win32 executable version of SAINTETIQ, SAINTETIQ is also exposed as a Web Service. SEQT is a new software component which provides efficient search algorithms to filter summaries and support flexible query processing and personalized queries.

5.7. WebSmatch (Web Schema Matching)

Participants: Emmanuel Castanier, Rémi Coletta, Patrick Valduriez [contact].

URL: http://websmatch.gforge.inria.fr/

WebSmatch is an Action de Développement Technologique (ADT) at INRIA Sophia-Antipolis Méditerranée started in october. The goal of this ADT is to develop, validate and promote WebSmatch, an environment for Web schema matching which is anticipated to play a crucial role in Web data integration applications. WebSmatch will capitalize on our experience gained in developing several schema matching tools at LIRMM and data integration systems in Atlas. We plan to develop WebSmatch as Eclipse plugins and deliver it through Web services, to be used directly by data integrators or other tools, in unanticipated situations.

6. New Results

6.1. Data Reduction and Classification

Data reduction and classification is needed to cluster large data sets in concise ways. We use two different formalisms for clustering data: grid-based conceptual hierarchies, for database summarization; and parametric probabilistic models, for continuous multivariate spaces typically encountered with multimedia data. To deal with distributed data sources, we addressed the problem of integration of (possibly hierarchical) structures. Our focus is on integration of data descriptions, without resorting to raw data.

6.1.1. Transaction Database Summaries

Participants: Guillaume Raschia, Quang-Khai Pham.

Transaction databases are sequences of item sets associated within a timestamp on the timeline. TSAR is a summarization technique to reduce the size of input sequences and represent them at higher levels of (semantic and temporal) presentation. TSAR transforms a time sequence into a concise summary sequence in three steps: (i) generalization, (ii) grouping and (iii) concept formation. Using domain specific taxonomies, input event descriptors are expressed at a higher level of abstraction. Then, the size of the sequence is reduced by grouping similar events while preserving the overall chronology of events. The concept formation phase then models each group of events by a single representative event. The resulting summary is a concise yet informative higher level representation of the transaction database (sequence of events). TSAR has been implemented as a web service and has been successfully applied within a dedicated methodology on Reuters' financial news archives.

Building upon textscTSaR, we proposed to make the process parameter-free. Thus, we reformulated the summarization problem into a new clustering problem with constraints, with a specific optimality criteria [40]. The objective function takes into account both the content similarity and the proximity of events on the timeline such that temporal dimension is considered as a first-class citizen in the problem. The problem is NP-hard and we proposed two greedy approaches called G-BUSS and GRASS to build an approximate solution.

We also explored and analyzed how time sequence summaries contribute to discovering Higher Order Knowledge. We analytically characterized the higher order patterns discovered from summaries and devised a methodology that uses the patterns discovered to uncover even more refined patterns. We evaluated and validated our summarization algorithms and our methodology by an extensive set of experiments on real world data extracted from Reuters' financial news archives [16].

6.1.2. Aggregating Probabilistic Models to Learn from Distributed Data

Participants: Pierrick Bruneau, Ali El Attar, Marc Gelgon.

Performing statistical pattern recognition and machine learning on distributed data is attracting rising attention. We have extended our previous focus on mixture aggregation of probabilistic mixture models [20], [47], in the following manner. We have proposed a variational Bayesian estimation procedure that conducts both model merging and dimensionality reduction jointly, rather than sequentially. The scheme operates over mixtures of probabilistic Principal Components Analysis (PCA) and only requires access to model parameters (rather than data), ensuring fast processing [31]. Extensions under progress are two fold : statistically robust clustering of models, in a decentralized fashion, through a joint process of identification and rejection of outlier probabilistic models ; dealing with the discrete-observation counterpart (Latent Dirichlet Allocation (LDA) models). Finally, in cooperation with the COD team at LINA, we proposed a scheme for interactive, semi-supervised clustering of a set of mixture models [21].

6.2. P2P Data Management

Data management in P2P systems offers new research opportunities since traditional distributed database techniques need to scale up while supporting autonomy, heterogeneity, and dynamicity of the data sources. In the context of the Atlas Peer-to-Peer Architecture (APPA) project, the main results this year are in the management of data replication, data privacy and testing.

6.2.1. Data Replication in DHTs

Participants: Reza Akbarinia, Esther Pacitti, Mounir Tlili, Patrick Valduriez.

Distributed Hash Tables (DHTs) provide an efficient solution for data location and lookup in large-scale P2P systems. However, it is up to the applications to deal with the availability of the data they store in the DHT, e.g. via replication. To improve data availability, most DHT applications rely on data replication. However, efficient replication management is quite challenging, in particular because of concurrent and missed updates in a very dynamic environment.

In [29], we gave an efficient solution to this problem. We proposed a service called Continuous Timestamp based Replication Management (CTRM) that deals with the efficient storage, retrieval and updating of replicas in DHTs. The CTRM service is inspired by the P2P-LTR service whose objective is to perform distributed reconciliation over DHTs [42] [41].In CTRM, the replicas are maintained by groups of peers which are determined dynamically using a hash function. To perform updates on replicas, we propose a new protocol that stamps the updates with timestamps that are generated in a distributed fashion using the dynamic groups. Timestamps are not only monotonically increasing but also continuous, i.e. without gap. The property of monotonically increasing allows applications to determine a total order on updates. The other property, i.e. continuity, enables applications to deal with missed updates. Through simulation and experimentation, we showed the effectiveness of our solution.

6.2.2. Data Privacy

Participants: Mohamed Jawad, Patrick Valduriez.

We continued our work on PriServ, a privacy service for P2P data sharing applications. Priserv could be used in many professional and public communities (e.g. medical or scientific applications) in order to control disclosure on sensitive data (e.g., medical records or research results) without violating privacy. Priserv combines the Hippocratic principles for enforcing *purpose-based* disclosure control, with trust and encryption. The key feature is that owner peers (data publishers) keep full control over their private data and private keys. Data publishing in PriServ takes into account owner privacy preferences and does not reveal any private information about data (encrypted data or data references). PriServ uses a DHT to efficiently locate data.

To test and validate PriServ, we implemented a Java prototype for collaborative data sharing applications. We consider a medical data sharing application where patients and doctors could share private medical records. In this context, we show how (a) PriMod (the privacy model of Priserv) is used to define privacy policies of data owners which are attached to data, and how (b) PriServ is used to publish and to request data according to owner privacy preferences. We also show how data requesters may access data provided if they have the necessary access rights and the sufficient trust levels.

6.2.3. Testing P2P Systems

Participants: Eduardo Almeida, Gerson Sunyé, Patrick Valduriez.

Typical testing architectures for distributed software rely on a centralized test controller, which decomposes test cases in sequence of *test steps* and deploy them across distributed testers. The controller guarantees the correct execution of test steps through synchronization messages. The distributed testers stimulate, locally to each nodes, the nodes that compose the system.

In the context of large-scale distributed systems, we implemented the PeerUnit software prototype, which is based on the CTMF architecture and implements both, a centralized and a distributed architecture to control the execution of test cases [44].

In [28], we presented a framework and a methodology for testing P2P applications. The framework is based on the individual control of nodes, allowing test cases to precisely control the volatility of nodes during their execution. We validated this framework through implementation and experimentation on an open-source P2P system. The correctness of the system under test is checked based on three dimensions: functionality, scalability and volatility. We proposed an incremental methodology to deal with these dimensions, which aims at covering functions first on a small system and then incrementally addressing the scalability and volatility issues. The experimentation tests the behavior of the system on different conditions of volatility and shows how the tests were able to detect complex implementation problems.

In [45], we presented a distributed architecture to synchronize the test execution sequence. The architecture organizes the testers in a tree, where messages are exchanged among parents and children. The experimental evaluation shows that the synchronization management overhead can be reduced by several orders of magnitude. The distributed architecture showed a satisfactory performance when controlling up to eight thousand nodes. We concluded that testing architectures should scale up along with the distributed system under test. The paper on PeerUnit [28] obtained an excellent review in ACM Computing Reviews from Gerald D. Everett, director of the American Software Testing Qualifications Board (ASTQB), who describes the work as a new paradigm of software testing. With Everett, we have started the process for obtaining the status of *ASTQB best practice* for PeerUnit.

6.3. P2P Query Support

We addressed three aspects related to efficient query support in unstructured P2P networks. First, we exploit gossiping for exchanging user profiles between peers, as a basis for a recommandation service. Second, we proposed a new solution to top-k query processing that avoids long wating times for users. Finally, we continued our work on semantic interoperability.

6.3.1. P2P Recommendation Using Gossiping

Participants: Fady Draidi, Esther Pacitti, Patrick Valduriez.

The popularity of P2P content sharing systems such as BitTorrent and eMule has translated into large amounts of documents being spread over high numbers of peers (and users). Although more information is available to more users, users tend to get overwhelmed with high numbers of documents returned as results of their queries. And it is hard for them to distinguish which are the most valuable and relevant documents. However, the users themselves, i.e. their interest in specific topics or their rankings of documents they have read, is simply ignored. In other words, what is missing in P2P content sharing systems is a recommendation service (RS) which can recommend high quality and valuable documents using user information.

In [36], we proposed a solution to this problem with P2Prec, a recommendation service for P2P content sharing systems that exploits users' social data. The key idea is to recommend to a user high quality documents in a specific topic using ratings of friends (or friends of friends) who are expert in that topic. To manage users' social data, we rely on Friend-Of-A-Friend (FOAF) descriptions. P2Prec has a hybrid P2P architecture to work on top of any P2P content sharing system. It combines efficient DHT indexing to manage the users' FOAF files with gossip robustness to disseminate the topics of expertise between friends. In our experimental evaluation, using the CiteSeer dataset, we show that P2Prec has the ability to get the maximum recall with very good performance. Furthermore, it increases recall and precision by a factor of 2 compared with centralized solutions.

6.3.2. ASAP Top-k Query Processing in Unstructured P2P Systems

Participants: Reza Akbarinia, William Kokou Dedzoe, Philippe Lamarre, Patrick Valduriez.

Top-k query processing techniques are useful in unstructured P2P systems to avoid overwhelming users with too many results and provide them with the best ones. However, existing approaches suffer from long waiting times, because top-k results are returned only when all queried peers have finished processing the query. As a result, response time is dominated by the slowest queried peer. We proposed to revisit the problem of top-k query processing.

In [33], we addressed this users' waiting time problem. For this, we revisited top-k query processing in P2P systems by introducing two novel notions in addition to response time: the stabilization time and the cumulative quality gap. Using these notions, we formally defined the as-soon-as-possible (ASAP) top-k processing problem. Then, we proposed a family of algorithms called ASAP to deal with this problem. We validated our solution through implementation and extensive experimentation. The results show that ASAP significantly outperforms baseline algorithms by returning final top-k result to users in much better times.

This work paves the way to a new class of algorithms, more efficient, which can take advantage of semantic representation of peers to route queries and which are able to balance load over time while limiting impact on performance.

6.3.3. Semantic Heterogeneity in Unstructured P2P Systems

Participants: Thomas Cerqueus, Philippe Lamarre, Patrick Valduriez.

Participants' autonomy in distributed systems may quickly lead to semantic heterogeneity which may induce many difficulties for semantic approaches. Indeed, each participant can use her own ontology or semantic context. In such a situation, it becomes a problem for a participant to understand the meaning of a request expressed by another participant using her own terms. If unsolved, this problem may have a deep impact on the quality of provided answers. Many solutions have been proposed to deal with heterogeneity, but their evaluation process is difficult because tools are missing. In particular, there is no way to quantify the difficulty of a particular semantically heterogeneous situation nor to understand what makes it difficult. We started to propose some answers to these quite difficult questions. In particular we introduced new measures of heterogeneity between two participants and, based on them, heterogenehity of a complex system. The first results are currently under submission. These results would be very useful to evaluate solutions proposed to improve semantic interoperabiliy and we plan to apply it to many of them and in particular to ExSI2D we proposed before. We also studied ExSI2D we initialy proposed to improve interoperability in semantic heterogeneous systems from the point of view of personalization. Indeed, according to ExSI2D approach, it is up to the document provider to adapt his representation to the requester's one and this leads to some kind of interoperability. First results has been published in [43] [34].

6.4. Cloud Data Management

Cloud computing is a natural evolution, and combination, of different computing models proposed for supporting applications over the web, in particular, cluster and grid computing. From a technical point of view, the grand challenge is to support in a cost-effective way, the very large scale of the infrastructure that has to manage lots of users and resources with high quality of service. Cloud computing can be very useful for scientific data-intensive applications. With this in mind, we have started new work on several related issues, all important for scientific applications: data streaming, scientific workflow management, massive graph management and uncertain data management.

6.4.1. StreamCloud

Participant: Patrick Valduriez.

Recent years have witnessed the growth of a new class of data-intensive applications that do not fit the DBMS query paradigm. Instead, the data arrive at high speeds taking the form of an unbounded sequence of values (data streams) and queries run continuously returning new results as new data arrive. Examples of data streams are sensor data (e.g. in environmental applications) or IP packets (e.g. in a network monitoring application). The unbounded nature of data streams makes it impossible to store the data entirely in bounded memory.

Current research efforts have mainly focused on scaling in the number of queries and/or query operators having overlooked the scalability with respect the stream volume. In [37], we proposed StreamCloud, a large scale data streaming system for processing large data stream volumes in a cloud. The focus is on how to parallelize continuous queries to attain a highly scalable data streaming infrastructure. StreamCloud goes beyond the state of the art by using a novel parallelization technique that splits queries into subqueries that are allocated to independent sets of nodes in a way that minimizes the distribution overhead. StreamCloud is

implemented as a middleware and is highly independent of the underlying data streaming engine. We explore and evaluate different strategies to parallelize data streaming and identify and tackle with the main bottlenecks and overheads to achieve large scalability. We presented the system design, implementation and a thorough evaluation of the scalability of the fully implemented system.

6.4.2. Scientific Workflow Management

Participants: Esther Pacitti, Patrick Valduriez.

Scientific Workflow Management Systems (SWfMS) are being used intensively to support large scale in silico experiments. In order to reduce execution time, current SWfMS have exploited workflow parallelization under the arising Many Tasks Computing (MTC) paradigm in homogeneous computing environments, such as multiprocessors, clusters and grids with centralized control. Although successful, this solution no longer applies to heterogeneous computing environments, such as hybrid clouds, which may combine users' own computing resources with multiple edge clouds. A promising approach to address this challenge is Peer-to-Peer (P2P) which relies on decentralized control to deal with scalability and dynamic behavior of resources.

In [39], we proposed a new P2P approach to apply MTC in scientific workflows. Through the results of simulation experiments, we showed that our approach is promising.

Large-scale scientific experiments are usually supported by scientific workflows that may demand high performance computing infrastructure. Within a given experiment, the same workflow may be explored with different sets of parameters. However, the parallelization of the workflow instances is hard to be accomplished mainly due to the heterogeneity of its activities. Many-Task computing paradigm seems to be a candidate approach to support workflow activity parallelism. However, scheduling a huge amount of workflow activities on large clusters may be susceptible to resource failures and overloading.

In [35], we proposed Heracles, an approach to apply consolidated P2P techniques to improve Many-Task computing of workflow activities on large clusters. We present a fault tolerance mechanism, a dynamic resource management and a hierarchical organization of computing nodes to handle workflow instances execution properly. We have evaluated Heracles by executing experimental analysis regarding the benefits of P2P techniques on the workflow execution time.

This work was done in the context of the Equipe Associée Sarava and the CNPq-INRIA project DatLuge.

6.4.3. Massive Graph Management

Participants: Victor Muntés Mulero, Esther Pacitti, Patrick Valduriez.

Traversing massive graphs as efficiently as possible is essential for many scientific applications. Many common operations on graphs, such as calculating the distance between two nodes, are based on the Breadth First Search (BFS) traversal. However, because of the exhaustive exploration of all the nodes and edges of the graph, this operation might be very time consuming. A possible solution is distributing the graph among the nodes of a shared-nothing parallel system. Nevertheless, this operation may generate a large amount of inter-node communication. In [38], we focused on reducing the inter-node communication during the parallelized execution of the BFS operation on massive graphs containing up to millions of vertices and edges. For this purpose, we assume graphs to be nonattributed since attributes are not relevant for the BFS operation. We studied the effect of different partitioning methods on the overall communication during the execution of this algorithm. For this, we proposed two new heuristic partitioning strategies which benefit from the fact that they work on compact collections of vertices and edges, such as those proposed for DEX [8]. We show that our techniques reduce communication from 74compared to previous work.

6.4.4. Uncertain Data Management

Participants: Reza Akbarinia, Esther Pacitti, Patrick Valduriez, Guillaume Verger.

Data uncertainty exists in many scientific applications because of several reasons: incomplete knowledge of the underlying system 2) inexact model parameters 3) inaccurate representation of the initial boundary conditions; 4) inaccuracy in equipments (e.g. sensors); etc. An example of application that needs to process uncertain data is the study of plant evolution, at INRA, Montpellier. In each plant there are several sensors monitoring different parameters of the plants, e.g. size and temperature. Every 15 minutes a new data from each sensor is sent to the application. However, sometimes the data are inconsistent, and even in contradiction. Therefore, in addition to the traditional services of a database system (storage/retrieval and query processing), an important requirement is to manage uncertainty rather than ignoring it, e.g. in the case of uncertainty among mutual exclusive data, the system should be able to return the most probable one.

We have started a prototype of a probabilistic database, in which we intend to embed specific functions needed by scientists. In an ongoing work, we studied uncertain aggregate (aggr) queries which have been proven to be very useful for many uncertain data management applications. To evaluate aggr queries over uncertain data, we must first provide a definition (semantics) of these queries in uncertain databases. In our work, in addition to taking into account the previously proposed semantics we proposed new semantics which are very useful for uncertain applications. The evaluation of aggr queries in both new and previously proposed semantics is quite challenging, particularly for SUM and AVG queries. Naïve algorithms, which are based on enumerating possible worlds, evaluate the aggr queries in exponential time. We developed new algorithms that in most cases execute aggr queries very efficiently compared to baseline solutions. First results of this work are under submission.

6.5. Transactional Memory in Multicore Systems

Participants: Toufik Sarni, Patrick Valduriez.

In 2009, we made the claim that the design of software real-time transactional memory (RT-STM) is possible, by considering both model of real-time tasks and transactions. We argued that lock-free algorithms are suitable for real-time using an helping mechanism between transactions. In general, this mechanism is based on heuristic rules and we proposed a new heuristic allowing a transaction to only help a higher priority one (i.e. transaction which is close to its deadline). In terms of deadline guarantee ratio of transactions, our experimental results have shown that RT-STM outperforms classical approaches.

However, this ratio does not represent the total number of conflicts. We analyzed the conflicts ratio using another platform for our experiments (eight 32-bit multicore architecture Intel-Xeon). And the new results show that the help mechanism only impacts 18% of the conflicts set. In other words, the other conflicts are not handled by our helping rules. This important proportion of conflicts is caused by the optimistic strategy which systematically - without taking into account the deadline of transaction - makes a transaction aborted when its objects are modified by others. This situation could be tolerated in a traditional system, but in soft real-time systems, the main issue is to minimize the number of transactions that miss their deadlines.

To tackle this problem, we proposed a novel real-time concurrency control protocol that prevents the conflicts before appearing (i.e. an early detection protocol). In this new protocol, each transaction tries to mark objects before opening them. As a result, a transaction cannot modify objects marked by a higher priority transaction, and thus avoids the situation caused by the optimistic strategy.
Furthermore, we designed this protocol to provide high throughput. It is based on lock-free synchronization which is a guarantee that at least one transaction is making progress. In particular, our protocol tries to make transactions which are close their deadlines.
For this purpose, our new protocol integrates a helping mechanism for serializability but also to fix priority among queued transactions as their deadlines are variables and directly depends on those of real-time tasks. Furthermore, all data-structures of classical STM are kept although the mark strategy involves simpler structures. We made this choice to relieve us from dealing with the performance impact due to these structures.

The main challenge of lock-free algorithms is how to efficiently implement them. In our case, the performance is guaranteed to be at least better than lock-based algorithms since our protocol enhances the best algorithm (Harris and Fraser) known today. We implemented our protocol. All synchronization procedures are lock-free and based on Compare and Swap operator with several memory barriers to ensure consistency while any shared memory access. Currently, we are testing this protocol, and trying to reduce the number of memory allocations because of its important impact on transactions rollbacks.

7. Contracts and Grants with Industry

7.1. ANR Safimage (2007-2010, 170Keuros)

Participants: Marc Gelgon, Pierrick Bruneau.

This project involved Alcatel-Lucent, IRCCyN, and IS2T, and ended in August 2010. The project deals with inspection of data in high-speed routers for security purposes. The task carried out by Atlas dealt with estimation of probabilistic modelling of distributed data.

7.2. OSEO/Région Pays-de-la-Loire EMAP (2010-2012, 75Keuros)

Participants: Guillaume Raschia, Marc Gelgon.

Besides our group, the project involves Human Connect and IM-Info, two small IT companies in Nantes, and the IRCCyN laboratory. It started in June 2010 and aims at developing techniques for accessing information over social networks. The Atlas contribution is two fold : matching and agregating profiles of the same user, appearing in various social networks ; identifying communities and providing recommendations through content-based and collaborative filtering.

7.3. PREDIT EPILOG (2009-2011, 60Keuros)

Participants: Philippe Lamarre, Duc Trung Vu, Patrick Valduriez.

The project EPILOG (Etude des technologies Pair-à-pair pour la collaboration Interentreprises dans la chaîne LOGistique) involves Euxenis SAS and RISC Solutions d'Assurances. The objective is to provide support for collaboration and supply chain management among partner enterprises in the retail industry The approach we plan to validate in the project is P2P. Atlas addresses the research issues associated with the definition of the P2P network for supply chain management, with autonomous partners with various interests, the modeling of information exchanged during transactions and query processing in the P2P network.

7.4. Data Publica (2010)

Participants: Emmanuel Castanier, Rémi Coletta, Patrick Valduriez.

Data Publica is an open data platform initially developped by Araok, Nexedi and Talend. In september, Origin and INRIA (Atlas and Leo) joined the Data Publica project. As a direct application of its WebSmatch ADT, Atlas contributes technologies for automatic schema extraction and matching from high numbers of public data sources. A first contribution has been the development of an Excel extraction component.

8. Other Grants and Activities

8.1. Regional Actions

We are involved in the following actions:

8.1.1. MILES (2007–2010)

MILES is the main project funded by Region Pays-de-Loire, Nantes, on information and communication technologies. Within the MILES project, M. Gelgon is in charge of a sub-project dealing with distributed multimedia systems, involving the Atlas project-team and IRCCyN (IVC group). This sub-project addresses, on one side, multimedia data learning and classification in a distributed computing and storage context and, on the other side, secure, distributed storage with involving techniques specific to multimedia data.

8.1.2. Pôle de compétitivité (2007-2010)

The ANR Safimage project (described above) is further supported by Pôle de Compétitivité Images & Réseaux, Rennes.

8.1.3. Labex NUMEV, Montpellier

In the context of the Excellence Initiative of the MENRT, we have participated to the proposal of the Laboratory of Excellence (labex) NUMEV (Digital and Hardware Solutions, Modelling for the Environment and Life Sciences) presented by University of Montpellier 2 in partnership with CNRS, University of Montpellier 1, and INRIA. NUMEV seeks to harmonize the approaches of hard sciences and life and environmental sciences in order to pave the way for an emerging interdisciplinary group with an international profile. The NUMEV project is decomposed in four complementary research themes: Modelling, Algorithms and computation, Scientific data (processing, integration, security), Model-Systems and measurements. Patrick Valduriez heads the theme on scientific data.

8.2. National Actions

We are involved in the following project:

8.2.1. ANR VERSO DataRing(2008-2011, 200Keuros)

Participants: Reza Akbarinia, Fady Draidi, Manal El Dick, Mohamed Jawad, Philippe Lamarre, Esther Pacitti, Patrick Valduriez.

The project, headed by P. Valduriez, involves the Gemo project-team (INRIA Saclay Ile de France), LIG, LIRMM and Telecom ParisTech. The objective is to address the problem of data sharing for online communities, such as social networks (e.g. sites like MySpace and Facebook) and professional communities (e.g. research communities, online technical support groups) which are becoming a major killer application of the web. The project addresses this problem by organizing community members in a peer-to-peer (P2P) network ring across distributed data source owners where each member can share data with the others through a P2P overlay network.

In this project, we study the following problems: query processing with data uncertainty, data indexing and caching, data privacy and trust. To validate our approach, we develop services in the context of the APPA prototype.

8.3. International actions

We are involved in the following international actions:

- the Equipe Associée SARAVA with UFRJ, Rio de Janeiro (Marta Mattoso, Vanessa Braganholo, Alexandre Lima) to work on P2P data management for online comunities;
- the CNPq-INRIA project DatLuge (Data & Task Management in Large Scale) with UFRJ (Marta Mattoso, Vanessa Braganholo, Alexandre Lima), LNCC, Rio de Janeiro (Fabio Porto), and UFPR, Curitiba (Eduardo Almeida) to work on large scale scientifc workflows;
- the PICASSO project Scaling GraphDB, with UPC, Barcelona ((Josep Lluis Larriba Pey and Victor Muntes Mulero, visiting researcher in Atlas for 6 months) to work on very large graph database support;
- the INRIA Sophia-Antipolis Méditerranée COLOR project VLGDB, with UPC, Barcelona to work on very large graph databases in cluster systems.

9. Dissemination

9.1. Animation of the Scientific Community

The members of the Atlas project-team have always been strongly involved in organizing the French database research community, in the context of the I3 GDR and the conference BDA.

P. Valduriez is the panel chair of the 2011 IEEE Int. Conf. on Data Engineering (ICDE).

9.2. Editorial Program Committees

Participation in the editorial board of scientific journals:

- Proceedings of the VLDB Endowment: P. Valduriez.
- Distributed and Parallel Database Systems, Kluwer Academic Publishers: P. Valduriez.
- Internet and Databases: Web Information Systems, Kluwer Academic Publishers: P. Valduriez.
- Journal of Information and Data Management, Brazilian Computer Society Special Interest Group on Databases: P. Valduriez.
- Book series "Data Centric Systems and Applications" (Springer-Verlag): P. Valduriez.
- Ingénierie des Systèmes d'Information, Hermès : P. Valduriez.

Participation in conference program committees :

- IEEE/ACM Int. Conf on Web Intelligence, 2010: M. Gelgon, G. Raschia
- Int. Conf. on Cloud Computing (CloudCom)2010: P. Valduriez
- PersDB Workshop co-located with VLDB 2010: G. Raschia
- IEEE Int. Conf. on Networking, Architecture and Storage (NAS) 2010: E. Pacitti.
- IEEE Int. Conf. on Distributed Computing Systems (ICDCS), Data management 2010: R. Akbarinia.
- Int. Conf. on High Performance Computing for Computational Science (VecPar) 2010: R. Akbarinia.
- Int. Conf. on Extending DataBase Technologies (EDBT) 2010: P. Valduriez; 2011: E. Pacitti
- EDBT Ph.D. Workshop 2010: P. Valduriez.
- ACM Symposium of Applied Computing (SAC), Privacy on the Web 2010: P. Valduriez.
- International Workshop on MapReduce and its Applications (MAPREDUCE), 2010, 2011: P. Valduriez.
- Journées Bases de Données Avancées (BDA), 2010: E. Pacitti, G. Raschia
- Ecole Bases de Données Avancées (BDA), 2010: E. Pacitti
- IEEE Int. Conf. on Data Engineering (ICDE) 2011: P. Valduriez
- ACM SIGMOD Int. Conf. 2011: P. Valduriez
- Int. Conf. on VLDB 2011: P. Valduriez
- Int. Conf. on Current Trends in Theory and Practice of Computer Science (SoftSem)2011: P. Valduriez

9.3. Invited Talks

Patrick Valduriez gave an invited conference on the "social internet" (or Web 2.0) at HiPhis (Séminaire inter-universitaire d'histoire et philosophie des sciences) at Univ. Montpellier 2 in february. The conference is online at http://www.paroledechercheurs.net/spip.php?article768.

Esther Pacitti and Patrick Valduriez gave an invited talk on “Cloud data management: open issues and research directions”at UFRJ in july and at the DNAC (De Nouvelles Architectures pour les Communications) Congress in Paris in november.

Esther Pacitti gave an invited talk on “Large-scale data sharing by exploiting gossiping” at the 1st Gossple workshop on Social networking at INRIA-Rennes Bretagne Atlantique in december.

9.4. Teaching

All the members of the Atlas project-team teach database management, multimedia, and software engineering at the Bs, Ms and Ph.D. degree levels at the University of Nantes, and, since sept. 2009, at University Montpellier 2.

Guillaume Raschia is heading a new final-year engineering degree track on Digital Content Management at Polytech'Nantes. He is also in charge of an Advanced Database course in the Erasmus Mundus Data Management and Knowledge Discovery (DMKM) Master. Marc Gelgon is in charge (with IRCCyN) of a future master degree on Multimedia and Data Management, targeting international students and validated by AERES in July 2010.

Noureddine Mouaddib, on leave from Polytech'Nantes and a former member of the team, is the president of Rabat International University, Morocco, inaugurated in 2010, which aims at training high-level engineers and managers for Africa.

The book Principles of Distributed Database Systems, co-authored with professor Tamer Özsu, U. Waterloo, published by Prentice Hall in 1991 et 1999, and by Springer in 2011 has become the standard book for teaching distributed databases all over the world. Our Web site features course material, exercises, and direct communication with professors. The third edition is a major revision (800 pages!) with much new material on data integration, replication, P2P, parallel systems, web data management, data stream management and cloud.

10. Bibliography

Major publications by the team in recent years

[1] R. AKBARINIA, V. MARTINS, E. PACITTI, P. VALDURIEZ. *Design and Implementation of Atlas P2P Architecture*, in "Global Data Management", R. BALDONI, G. CORTESE, F. DAVIDE (editors), IOS Press, 2006.

[2] R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Best Position Algorithms for Top-k Queries*, in "Int. Conf. on Very Large Data Bases (VDLB)", Vienna, Austria, 2007, p. 495-506.

[3] R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Data currency in replicated DHTs*, in "ACM SIGMOD Int. Conf. on Management of Data (SIGMOD", Beijing, China, 2007, p. 211-222.

[4] M. E. DICK, E. PACITTI, B. KEMME. *Flower-CDN: a hybrid P2P overlay for efficient query processing in CDN*, in "Int. Conf. on Extending Database Technology (EDBT)", 2009, p. 427-438.

[5] A. NIKSERESHT, M. GELGON. *Gossip-based Computation of a Gaussian Mixture Model for Distributed Multimedia Indexing*, in "IEEE Transactions on Multimedia", April 2008, vol. 10, n^{o} 3, p. 385-392.

[6] E. PACITTI, P. VALDURIEZ, M. MATTOSO. *Grid Data Management: Open Problems and New Issues*, in "Journal of Grid Computing", 2007, vol. 5, n^{o} 3, p. 273-281.

[7] W. PALMA, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *DHTJoin: processing continuous join queries using DHT networks*, in "Distributed and Parallel Databases", 2009, vol. 26, n° 2-3, p. 291-317.

[8] A. PIGEAU, M. GELGON. *Building and Tracking Hierarchical Partitions of Image Collections on Mobile Devices*, in "ACM Multimedia Conf.", Singapore, 2005, p. 141-150.

[9] J.-A. QUIANÉ-RUIZ, P. LAMARRE, P. VALDURIEZ. *SQLB: A Query Allocation Framework for Autonomous Consumers and Providers*, in "Int. Conf. on Very Large Data Bases (VDLB)", Vienna, Austria, 2007, p. 974-985.

[10] J.-A. QUIANÉ-RUIZ, P. LAMARRE, P. VALDURIEZ. *A Self-Adaptable Query Allocation Framework for Distributed Information Systems*, in "The VLDB Journal", 2009, vol. 18, n° 3, p. 649-674.

[11] R. SAINT-PAUL, G. RASCHIA, N. MOUADDIB. *General Purpose Database Summarization*, in "Int. Conf. on Very Large Databases (VLDB)", Trondheim, Norway, 2005, p. 733-744.

[12] T. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, 3rd edition*, Springer, 2011.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] P. BRUNEAU. *Contributions en classification automatique : agrégation bayésienne de mélanges de lois et visualisation interactive*, Université de Nantes, 2010.

[14] M. ELDICK. *P2P Infrastructure for Content Distribution Networks*, Université de Nantes, 2010.

[15] W. PALMA. *Processing Continuous Join Queries in Structured P2P Systems*, Université de Nantes, 2010.

[16] Q.-K. PHAM. *Time Sequence Summarization: Theory and Applications*, University of Nantes, 2010.

Articles in International Peer-Reviewed Journal

[17] R. AKBARINIA, M. TLILI, E. PACITTI, P. VALDURIEZ, A. A. B. LIMA. *Replication in DHTs using Dynamic Groups*, in "Journal of Transactions on Large-Scale Data and Knowledge-Centered Systems", 2011, Selected among the (two) best papers of Globe'2010 as extended version. To appear.

[18] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Parsimonious reduction of Gaussian mixture models with a variational-Bayes approach*, in "Pattern Recognition, Elsevier", 2010, vol. 43, n° 3, p. 850-858.

[19] P. BRUNEAU, F. PICAROUGNE, M. GELGON. *Interactive unsupervised classification and visualization for browsing an image collection*, in "Pattern Recognition, Elsevier", 2010, vol. 43, n° 2, p. 485-493.

[20] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Parsimonious reduction of Gaussian mixture models with a variational-Bayes approach*, in "Pattern Recognition", 2010, vol. 43, p. 850-858.

[21] P. BRUNEAU, F. PICAROUGNE, M. GELGON. *Interactive unsupervised classication and visualization for browsing an image collection*, in "Pattern Recognition", 2010, vol. 43, p. 485-493.

[22] M. ELDICK, E. PACITTI, R. AKBARINIA, B. KEMME. *Building a Content Distribution Network over a Peer-to-Peer Network*, in "Information Systems Journal", 2011, vol. 36, n^o 2, p. 222-247.

[23] A. A. B. LIMA, M. MATTOSO, P. VALDURIEZ. *Adaptive Virtual Partitioning for OLAP Query Processing in a Database Cluster*, in "Journal of Information and Data Management (JIDM)", 2010, vol. 1, n^o 1, p. 75-88.

[24] A. A. B. LIMA, M. MATTOSO, P. VALDURIEZ. *Adaptive Virtual Partitioning: Further Developments*, in "Journal of Information and Data Management (JIDM)", 2010, vol. 1, n^o 1, p. 89-92.

[25] K. PARK, H. CHOO, P. VALDURIEZ. *A Scalable Energy-efficient Continuous Nearest Neighbor Search in Wireless Broadcast Systems*, in "Wireless Networks", 2010, vol. 16, n^o 4, p. 1011-1031.

[26] K. PARK, P. VALDURIEZ. *Energy Efficient Data Access in Mobile P2P Networks*, in "IEEE Trans. on Knowledge and Data Engineering (TKDE)", 2011, To appear.

[27] G. A. VOUROS, A. PAPASALOUROS, K. TZONAS, A. G. VALARAKOS, K. KOTIS, J.-A. QUIANÉ-RUIZ, P. LAMARRE, P. VALDURIEZ. *A Semantic Information System for Services and Traded Resources in Grid e-Markets*, in "Future Generation Computer Systems", 2010, vol. 26, n^o 7, p. 916-933.

[28] E. C. DE ALMEIDA, G. SUNYÉ, Y. LE TRAON, P. VALDURIEZ. *Testing peer-to-peer systems*, in "Empirical Software Engineering", 2010, vol. 15, n^o 4, p. 346-379.

International Peer-Reviewed Conference/Proceedings

[29] R. AKBARINIA, M. TLILI, E. PACITTI, P. VALDURIEZ, A. A. B. LIMA. *Continuous Timestamping for Efficient Replication Management in DHTs*, in "Int. Conf. on Data Management in Grid and P2P Systems (Globe)", LNCS, Springer, 2010, vol. 6265, p. 38-49.

[30] P. BRUNEAU, A. PIGEAU, M. GELGON, F. PICAROUGNE. *Geo-temporal structuring of a personal image database with two-level variational-Bayes mixture estimation*, in "Proc. of Adaptive Multimedia Retrieval workshop (AMR'08-AMR'09)", Berlin, Germany, LNCS, Springer, 2010, vol. 5811, p. 127-139.

[31] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Aggregation of probabilistic PCA mixtures with a variational-Bayes technique over parameters*, in "IAPR Int. Conf. on Pattern Recognition (ICPR)", IEEE Computer Society, 2010, p. 340-345.

[32] P. BRUNEAU, A. PIGEAU, M. GELGON, F. PICAROUGNE. *Geo-temporal structuring of a personal image database with two-level variational-Bayes mixture estimation*, in "Revised selected papers from Adaptive Multimedia Retrieval workshops (AMR'08-AMR'09)", LNCS, Springer, 2010, vol. 5811, p. 127-139.

[33] W. K. DEDZOE, P. LAMARRE, R. AKBARINIA, P. VALDURIEZ. *ASAP Top-k Query Processing in Unstructured P2P Systems*, in "IEEE Int. Conf. on Peer-to-Peer Computing (P2P)", 2010, p. 1-10.

[34] W. K. DEDZOE, P. LAMARRE, R. AKBARINIA, P. VALDURIEZ. *Reducing User Waiting Time for Top-k Queries in Unstructured P2P Systems*, in "Journées Bases de Données Avancées (BDA)", 2010.

[35] J. DIAS, E. OGASAWARA, D. OLIVEIRA, E. PACITTI, M. MATTOSO. *Improving Many-Task Computing in Scientific Workflows Using P2P Techniques*, in "IEEE Workshop on Many-Task Computing on Grids and Su-

percomputers (MTAGS) co-located with ACM/IEEE SC10 (International Conference for High Performance, Networking, Storage and Analysis)", 2010.

[36] F. DRAIDI, E. PACITTI, P. VALDURIEZ. *P2Prec: a Recommendation Service for P2P Content Sharing Systems*, in "Journées Bases de Données Avancées (BDA)", 2010.

[37] V. GULISANO, R. JIMÉNEZ-PERIS, M. PATIÑO-MARTÍNEZ, P. VALDURIEZ. *StreamCloud: A Large Scale Data Streaming System*, in "IEEE Int. Conf. on Distributed Computing Systems (ICDCS)", 2010, p. 126-137.

[38] V. MUNTES-MULERO, N. MARTINEZ-BAZAN, J.-L. LARRIBA-PEY, E. PACITTI, P. VALDURIEZ. . *Graph Partitioning Strategies for Efficient BFS in Shared-Nothing Parallel Systems*, in "Int. Workshop on Graph Databases (IWGD) in conjunction with International Conference on Web-Age Information Management (WAIM)", LNCS, Springer, 2010, vol. 6185, p. 13-24.

[39] E. OGASAWARA, J. DIAS, D. OLIVEIRA, C. RODRIGUES, C. PIVOTTO, R. ANTAS, V. BRAGANHOLO, P. VALDURIEZ, M. MATTOSO. *A P2P Approach to Many Tasks Computing for Scientific Workflows*, in "Int. Conf. on High Performance Computing for Computational Science (VecPar)", 2010.

[40] Q.-K. PHAM, G. RASCHIA, B. BENATALLAH, N. MOUADDIB. *Le résumé de séquence d'événements: un nouveau problème de classification*, in "Journées Bases de Données Avancées (BDA)", 2010.

[41] M. TLILI, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *A Fault-Tolerant Infrastructure for P2P Collaborative Text Edition*, in "Journées Bases de Données Avancées (BDA)", 2010.

[42] M. TLILI, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Scalable P2P Reconciliation Infrastructure for Collaborative Text Editing*, in "Int. Conf. on Advances in Databases, Knowledge, and Data Applications (DBKDA)", IEEE Computer Society, 2010, p. 155-164.

[43] A. VENTRESQUE, S. CAZALENS, T. CERQUEUS, P. LAMARRE, G. PASI. *Personalization through Query Explanation and Document Adaptation.*, in "Int. Workshop on Personalized Access, Profile Management, and Context Awareness in Databases (PeersDB) in conjunction with VLDB Conference", 2010.

[44] E. C. DE ALMEIDA, J. E. MARYNOWSKI, G. SUNYÉ, P. VALDURIEZ. *PeerUnit: a Framework for Testing Peer-to-peer Systems*, in "IEEE/ACM Int. Conf. on Automated Software Engineering (ASE)", 2010, p. 169-170.

[45] E. C. DE ALMEIDA, J. E. MARYNOWSKI, G. SUNYÉ, Y. LE TRAON, P. VALDURIEZ. *Efficient Distributed Test Architectures for Large-Scale Systems*, in "Int. Conf. on Testing Software and Systems (ICTSS)", LNCS, Springer, 2010, vol. 6435, p. 174-187.

National Peer-Reviewed Conference/Proceedings

[46] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Fusion de mélanges gaussiens par une approche variationnelle*, in "Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA)", Caen, France, 2010.

[47] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Fusion bayésienne de mélanges de gaussiennes par une approche variationnelle*, in "Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA)", 2010, p. 170-179, Nominated in the top five papers.

[48] A. VENTRESQUE, T. CERQUEUS, L.-A. CELTON, G. HERVOUET, D. LEVIN, P. LAMARRE, S. CAZALENS. *Mysins : make your semantic INformation system*, in "Extraction et Gestion des Connaissances (EGC)", 2010, p. 629-630.

Workshops without Proceedings

[49] P. VALDURIEZ, E. PACITTI. *Cloud Data Management: open issues and research directions*, in "24ème Congrès DNAC : De Nouvelles Architectures pour les Communications: Réseaux et Cloud", 2010.

Scientific Books (or Scientific Book chapters)

[50] R. HAYEK, G. RASCHIA, P. VALDURIEZ, N. MOUADDIB. *Data Sharing in P2P Systems*, in "Handbook of Peer-to-Peer Networking", Springer, 2010, p. 531-570.

[51] R. HAYEK, G. RASCHIA, P. VALDURIEZ, N. MOUADDIB. *Managing Linguistic Data Summaries in Advanced P2P Applications*, in "Handbook of Peer-to-Peer Networking", Springer, 2010, p. 571-600.