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 Scientist

Patrick Valduriez [Team Leader, Research Director, INRIA, *Montpellier*, HdR]
Reza Akbarinia [Research Scientist, INRIA, *Nantes*]

Faculty Member

Marc Gelgon [Professor, University of Nantes, HdR]
Philippe Lamarre [Associate Professor, University of Nantes, HdR]
Esther Pacitti [Associate Professor, University of Nantes, until august, HdR]
Guillaume Raschia [Associate Professor, University of Nantes]
Gerson Sunyé [Associate Professor, University of Nantes]

External Collaborator

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

Technical Staff

Rabab Hayek [Engineer, Grid4All until june, *Nantes*]

PhD Student

Eduardo Almadeia [ATER University of Nantes until september]
Mounir Bechchi [ATER University of Nantes until september]
Pierrick Bruneau [ANR funding, *Nantes*]
Thomas Cerqueus [MENRT fellowship since october, *Nantes*]
William Kokou Dedzoe [CNRS fellowship, *Nantes*]
Vu Duc Trung [INRIA-Pays-de-la-Loire fellowship since november, *Nantes*]
Fadi Draidi [MAE-INRIA fellowship, *Montpellier*]
Ali El Attar [Pays-de-la-Loire fellowship, *Nantes*]
Manal El Dick [MENRT fellowship, *Nantes*]
Mohamed Jawad [MENRT fellowship, *Nantes*]
Jorge Manjarrez Sanchez [Conacyt fellowship until october, *Nantes*]
Wenceslao Palma [INRIA fellowship, *Nantes*]
Quang-Khai Pham [CNRS-University of New South Wales fellowship, *Nantes*]
Toufik Sarni [CNRS fellowship, *Nantes*]
Mounir Tlili [Pays-de-la-Loire fellowship, *Nantes*]
Anthony Ventresque [ATER University of Nantes until september]

Visiting Scientist

Victor Muntés Mulero [UPC, Barcelona, since october, *Montpellier*]

Administrative Assistant

Élodie Lizé [*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 now focus on data management in two large-scale distributed contexts: the web and P2P systems. In the context of the web, we consider information systems with autonomous participants (with heterogeneous data and different interests) and deal with the problems of data integration, data classification and data access. In the context of P2P systems, we capitalize on our experience in developping the APPA system, with various data management services (replication, caching, queries, clustering, privacy testing, etc.). We have also started to work on data management in a third context, that of multicore, because it promises to revolutionize basic data management tehcniques. In this context, we address the problem of real-time data access through transactional.

2.2. Highlights of the Year

The first highlight of the year is the bi-localization of the Atlas project-team between INRIA Rennes Bretagne - Atlantique, Nantes and INRIA Sophia Antipolis - Méditerrannée, Montpellier, where Patrick Valduriez and Esther Pacitti are now located. This year has been very productive in terms of industrial and international collaborations and research results with:

- the organization of VLDB2009 by P. Valduriez (general chair) in Lyon which had a record participation of 750 people, from 44 different countries.
- a new PREDIT project EPILOG on P2P data management for supply chain management in retail industry (with Euxenis SAS and RISC Solutions d'Assurances) for 18 months, starting in 2009;
- a new équipe associée (EA) SARAVÁ with UFRJ, Rio de Janeiro, to work on P2P data management for online comunities;
- a PICASSO research project, to start in 2010, Scaling GraphDB, with UPC, Barcelona, to work on very large graph database support;
- a very good publication record, including 3 papers published in top database journals (VLDBJ, DAPD);
- the paper [36] in the context of the SARAVÁ équipe associée was selected as best paper at the Colloquium of Computation Brazil / INRIA (COLIBRI 2009);
- 1 Habilitation (HDR) and 4 Ph.D. theses defended.

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 modelling, 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 neighbour 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 reasonnable 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 biais) 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 modelling tool. Mixtures have wide modelling 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 abovementionned 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 multdimensional 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 probabilites 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 form 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 naïve 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 neighbour 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, ressource description languages, reasoning engines...All these methods represent the fundations 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 corresondances 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 trafic 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.

Large scale distributed collaborative applications are getting common as a result of the progress of distributed technologies (GRID, P2P, and mobile computing). 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.

5. Software

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

Participants: Eduardo Almadeia, William Kokou Desdoe, Philippe Lamarre, Esther Pacitti, Gerson Sunyé, Mounir Tlili, Patrick Valduriez [contact].

URL: http://www.sciences.univ-nantes.fr/lina/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, SbQA, P2P-LTR and PeerUnit. The APPA services are used in several projects: Strep Grid4All, ANR RNTL Xwiki Concerto and ANR VERSO DataRing.

5.2. KTS (Key-based Timestamp Service)

Participants: William Kokou Desdoe, Esther Pacitti, Patrick Valduriez [contact].

URL: http://www.sciences.univ-nantes.fr/lina/gdd/appa/kts/

KTS (Key-based Timestamp Service) is a distributed service to manage timestamps in DHTs. It is useful to solve various DHT problems which need a total order on operations performed on each data. KTS has been initially proposed to support data currency in DHTs, i.e. the ability to return a current replica in a DHT despite peers leaving the network or concurrent updates. Experimental validation has shown that KTS incurs very little overhead in terms of communication cost. KTS is the basis for the P2P-LTR service. It has been implemented in Java on top of OpenChord.

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

Participants: William Kokou Desdoe, 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.

5.4. SbQA (Satisfaction-based Query Allocation Framework)

Participants: Philippe Lamarre [contact], Patrick Valduriez.

URL: http://www.sciences.univ-nantes.fr/lina/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.

5.5. PeerUnit (Peer-to-Peer Tester)

Participants: Eduardo Almadeia [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 (B-Tree) structure to avoid using a centralized testing coordination. Peerunit is implemented in Java and ha sbeen validated on two popular open-source P2P systems (FreePastry and OpenChord).

5.6. DBSum

Participants: Mounir Bechchi, Guillaume Raschia [contact], Amenel Voglozin.

URL: http://www.lina.univ-nantes.fr/

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.

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 have addressed the problem of integration of (possibly hierarchical) structures. Our focus is on integration of data descriptions, without resorting to raw data. We have also addressed the problem of efficient querying of database summaries.

6.1.1. Database Summaries

Participants: Guillaume Raschia, Mounir Bechchi, Quang-Khai Pham.

Our database summarization system DBSum provides multi-level summaries of tabular data stored in a centralized database. Summaries are computed online by means of a grid-based conceptual hierarchical clustering algorithm. Along this research area, we pursued two distinct directions : (i) approximate answering in large and distributed databases, and (ii) summarization of transaction DB.

In [13], we first proposed an efficient and effective algorithm coined Explore-Select-Rearrange Algorithm (ESRA), based on the DBSum model, to quickly provide users with hierarchical clustering schemas of their query results. Each node (or summary) of the hierarchy provided by ESRA describes a subset of the result set in a user-friendly form based on domain knowledge. The user then navigates through this hierarchy structure in a top-down fashion, exploring the summaries of interest while ignoring the rest. Experimental results show that the ESRA algorithm is efficient and provides well-formed (tight and clearly separated) and well-organized clusters of query results [47].

The ESRA algorithm assumes that the summary hierarchy of the queried data is already built using DBSum and available as input. However, DBSum requires full access to the data which is going to be summarized. This requirement severely limits the applicability of the ESRA algorithm in a distributed environment, where data is distributed across many sites and transmitting the data to a central site is not feasible or even desirable. Therefore, we proposed a solution for summarizing distributed data without a prior 'unification' of the data sources. We assume that the sources maintain their own summary hierarchies (local models), and we propose new algorithms for merging them into a single final one (global model). An experimental study shows that our merging algorithms result in high quality clustering schemas of the entire distributed data and are very efficient in terms of computational time.

As a second contribution to data reduction, we went one step further [39] into the definition of Time Sequence Summarization to support chronology-dependent applications on massive data sources. Time sequence summarization takes as input a sequence of events where each event is described by a set of descriptors. Time sequence summarization produces a concise time sequence that can be substituted for the original time sequence in chronology-dependent applications. We proposed an algorithm that achieves time sequence summarization based on a generalization, grouping and concept formation process. Generalization expresses event descriptors at higher levels of abstraction using taxonomies while grouping gathers similar events. Concept formation is responsible for reducing the size of the input time sequence of events by representing each group created by one concept. The process is performed in a way such that the overall chronology of events is preserved. The algorithm computes the summary incrementally and has reduced algorithmic complexity. The resulting output is a concise representation, yet, informative enough to directly support chronology-dependent applications. We validate our approach by summarizing one year of ?nancial news provided by Reuters.

6.1.2. Distributed Learning of Probabilistic Class Models

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

Learning a probabilistic model that describes the distribution of numerical features in a multidimensional continuous space, for supervised or unsupervised classification, is a fundamental and widely studied task. When data sources are distributed and dynamic, existing solutions must be reconsidered. We are indeed witnessing a strongly rising attention to classification and recognition from distributed data, supplied i.e. sensor networks or social networks.

Our proposal focuses on mixture aggregation, based on a probabilistic modelling over the parameters of the aggregated model and a variational Bayesian estimation procedure. To improve the model, we introduce a prior, based on a Poisson distribution, that favours grouping components coming from distinct models. While we showed that this generally improves both quality of the result and significantly speeds up computation, it has required the design of a new optimization scheme for the variational-Bayes EM algorithm [18], [25], [44]. We have recently extended this work to handle aggregation of models on different manifolds, i.e. variational-Bayes aggregation, at parameter-level, of mixtures of probabilistic PCA (paper submitted).

We are also extending this study to handle statistically robust clustering of distributed data. For this purpose, we have considered the counterpart of the above scheme, in its Student mixture model version. Student distributions may indeed be viewed as gamma-weighted infinite mixtures, thanks to which mixture model estimation can be made insensitive to some moderate amount of outlier data [31].

Finally, in cooperation with the COD team at LINA, we have proposed a scheme for interactive, semi-supervised clustering of a set of mixture models [19] and explored connections with biomimetic techniques for distributed clustering [26].

6.2. Data Access with Autonomous Participants

Taking into account the autonomy of participants holds an important part in the evolution of data management, systems and applications, especially considering open systems such as internet. Autonomy is some kind of freedom left to participants which can be managed differently from one participant to another. It can take very different forms [16]. For example, it may mean "enter or leave the system at will as in P2P systems. Intuitively, the more participants are autonomous, the easier it is for a participant to integrate the system, but, the harder it is to manage the system. In this context, we have focused on two problems: satisfaction-based query allocation and peer representation for query routing.

6.2.1. Satisfaction-based Query Allocation

Participants: Philippe Lamarre, Jorge Quiane Ruiz, Patrick Valduriez.

This work is related to supporting participants' objectives in joining a system. Intuitively, in the field of open systems, it is the hope of achieving some objectives which motivates a peer to participate in a system. Obviously, different participants may have different objectives. Consequently, to integrate as many participants as possible, a system should not assume that all of them are interested in the same normative objective. Instead, a system should enable participants to act accordingly to their own private objectives. This approach has been studied in the field of query allocation.

In the context of dynamic distributed systems, with large numbers of heterogeneous, autonomous consumers and providers (the participants) query allocation is challenging because participants' interests may be contradictory. For example, a consumer would desire to receive results from a given provider but this provider would not desire to perform the query of such a consumer. In [23], we defined a model to characterize the participants' satisfaction in the long run. We proposed a query allocation framework that takes into account participants' interests to satisfy them in the long-run. A particularity of our solution is that it dynamically makes the balance between providers' interests and consumers' interests taking into account their respective rewards. Experimental results show that our model enables a better evaluation of query allocation methods in these environments, and that our query allocation approach significantly outperforms baseline methods from both a satisfaction and performance point of view. This approach has been implemented and demonstrated [40].

We have also studied satisfaction in the context of query replication. Indeed, multiple allocation of the same query can be used to solve problems related to distributed systems integrating autonomous participants: First, to restrict the effects of a provider failure, as a preventive measure, the system can allocate the query to many of them; Second, to deal with Byzantine participants, a query initiator is sometimes interested in allocating its query to many participants in order to compare their results. In both cases, the major problem is to find a good compromise between the number of replicas and a possible degradation of the system (performances and participants departure). We propose to take advantage of the huge capacity of adaptation of a satisfaction based approach to adapt the number of replicas, according to the allocation policy and participantsÕ global satisfaction. This method takes care of participants intentions and thus automatically adapts query allocation to their individual situations. Obtained results are under submission.

6.2.2. Peer Representation for Query Routing

Participants: Philippe Lamarre, Anthony Ventresque, Patrick Valduriez.

We consider unstructured P2P systems, the most general form. When querying, it is important to be able to characterize an answer with respect to the set of all possible answers, especially when they are obtained using a scoring function and not an exact match, as for example, in document search. Current Top-k algorithms find the k best answers to a given query, but they are generally quite costly since they require to visit all the peers that can be reached. However, intuitively, it should be possible to avoid some peers which would not improve the results. We followed this line assuming a peer is able to represent its neighbors in such a way that it can determine if a neighbor is likely or not to improve the already obtained results. In [43] we studied the properties such a representation should satisfy 1 - to warranty the result of a Top-k query while avoiding some participants, and 2 - to be easily maintained. We then proposed a representation using semantic vectors and we proved it satisfies these properties. We presently work on algorithms to maintain such representation within a dynamic system and on related search algorithms.

6.3. 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 replicated data, data privacy and testing.

6.3.1. Data Replication in DHTs

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

Distributed Hash Tables (DHTs), e.g. CAN and Chord, provide an efficient solution for data location and lookup in large-scale P2P systems. One of the main characteristics of DHTs (and P2P systems) is the dynamic behavior of peers which can join and leave the system frequently, at anytime. When a peer gets offline, its data becomes unavailable. To improve data availability, most DHTs rely on data replication by storing (k, data) pairs at several peers, e.g. using several hash functions. If one peer is unavailable, its data can still be retrieved from the other peers that hold a replica. However, update management is difficult because of the dynamic behavior of peers and concurrent updates, and missed updates. One approach for update management is to stamp the updates with monotonically increasing timestamps, and send the updates and their timestamps to the replica holders. This gives us a total order on updates. To deal with missed updates, we can use timestamps which are continuous, i.e. without gap, so that a replica holder can detect the existence of missed updates by looking at the timestamps of the updates it has received. Examples of applications that can take advantage of continuous timestamping are the P2P collaborative text editing applications, e.g. P2P Wiki, which need to reconcile the updates done by collaborating users. However, the problem is how to generate such timestamps in a DHT.

In a recently submitted paper, 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. To perform updates on replicas, we developed a new protocol for CTRM that stamps the updates with timestamps in a distributed fashion. One of the main features of our protocol is that the updates' timestamps are not only monotonically increasing but also continuous. We take into account the peer failures that may happen during execution of the protocol, and show that our protocol works correctly under these failures.

The CTRM service is inspired by the P2P-LTR service (P2P Logging and Timestamping for Reconciliation) which we proposed in 2008. The objective of P2P-LTR is to perform distributed reconciliation over DHTs. It extends the Key-based Timestamping Service proposed in 2007 to support decentralized timestamping. While updating at collaborating peers, updates are timestamped and stored over a set of peers which are chosen using a set of hash functions. During reconciliation, these updates are retrieved in total order to enforce eventual consistency despite churn and failures. P2P-LTR was proposed in the context of XWiki Concerto and Grid4All projects. We finished the implementation of P2P-LTR this year.

6.3.2. Data Privacy

Participants: Mohamed Jawad, Patrick Valduriez.

Online peer-to-peer (P2P) communities such as professional communities (e.g., medical or research communities) are becoming popular due to increasing needs on data sharing. P2P environments offer valuable characteristics but limited guarantees when sharing sensitive or confidential data. They can be considered as hostile because data can be accessed by everyone (by potentially untrusted peers) and used for everything (e.g., for marketing or for activities against the owner's preferences or ethics).

Hippocratic databases provide mechanisms for enforcing *purpose-based* disclosure control, within a centralized datastore. This is achieved by using privacy metadata, i.e. privacy policies and privacy authorizations stored in tables. A privacy policy defines for each attribute, tuple or table the usage purpose(s), the potential users and retention period while privacy authorization defines which purposes each user is authorized to use. In the context of P2P systems, decentralized control makes it hard to enforce purpose-based privacy.

In addition to purpose-based data privacy, to prevent data misuse, it is necessary to trust participants. Trust management systems deal with unknown participants by testing their reputation. Reputation techniques verify the trustworthiness of peers by assigning them *trust levels*. A trust level is an assessment of the probability that a peer will not cheat.

In the context of P2P systems, few solutions for data privacy have been proposed and they focus on a small part of the general problem of data privacy, e.g. *anonymity* of uploaders/downloaders, *linkability* (correlation between uploaders and downloaders), *content deniability*, data encryption and authenticity. However, the major problem of data privacy violation due to data disclosure to malicious peers which misuse data, is not addressed.

In [35], we proposed a P2P data privacy model which combines the Hippocratic principles and trust. We proposed the algorithms of PriServ, a DHT-based P2P privacy service which supports this model and prevents data privacy violation. We also proposed three algorithms for trust level searching in PriServ. Our performance evaluation shows that PriServ introduces a small overhead.

In [34], [33], we extended PriServ functionalities. To improve availability, we give owners the choice to store locally their data or to distribute them on the system. Because distribution depends on the DHT, owners could see their private data stored on untrusted peers. To overcome this problem, before distribution, data is encrypted and decryption keys are stored and duplicated by owners. We also proposed a component-based architecture for PriServ. Several simulation results encourage our ideas and a prototype of PriServ is under development and will be tested on the Grid5000 platform.

6.3.3. Testing P2P Systems

Participants: Eduardo Almadeia, Gerson Sunyé, Patrick Valduriez.

Traditional architectures for testing, based on the Conformance Testing Methodology and Framework (CTMF) are not fully adapted to test large-scale distributed applications. Indeed, in this architecture, each node is tested by a Lower Tester (LT) and LTs are controlled by a centralized unit, the Lower Tester Control Function (LTCF). The LTCF establishes the synchronization among Lower Testers, ensuring for instance that a retrieve operation will only be executed after the insertion of data. Since the LTCF is centralized, it is a bottleneck when the number of nodes scales up. To address this problem, we proposed a distributed architecture for testing large-scale distributed data intensive applications.

This architecture was implemented in the PeerUnit software prototype. PeerUnit is based on the CTMF and implements both, a centralized and a distributed architecture to control Lower Testers [24], [14]. The distributed architecture showed a satisfactory performance when controlling more than a thousand nodes. We proposed an incremental methodology to deal with three aspects of P2P testing (functionality, volatility and scalability). The idea is to cover functionality first on a small system and then incrementally address the scalability and volatility aspects. The methodology was validated through the test of two popular open-source P2P systems, FreePastry and OpenChord.

6.4. P2P Query Support

We addressed three aspect related to efficient query support in P2P networks. by exploiting, in particular, DHTs and gossiping. First, we exploit DHTs and gossiping for improving the performance of join queries over data

streams. Second, we exploit DHTs and gossiping for improving content distribution. Third, we started a new research direction which considers uncertain data.

6.4.1. Join Queries over Data Streams

Participants: Reza Akbarinia, Esther Pacitti, Wenceslao Palma, Patrick Valduriez.

Recent years have witnessed the growth of a new class of data-intensive applications that do not fit the DBMS data model and querying 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. The unbounded nature of data streams makes it impossible to store the data entirely in bounded memory. However, approximate answers are often sufficient when the goal of a query is to understand trends and making decisions about measurements or utilizations patterns. One technique for producing an approximate answer to a continuous query is to execute the query over a window that maintains a restricted number of recent data items. In continuous query processing the join operator is one of the most important operators, which can be used to detect trends between different data streams. To emphasize access to recent data, the window conceptually slides over the input streams thereby giving rise to a type of join called sliding window join.

In [22], we addressed the problem of computing approximate answers to windowed stream joins over data streams. We propose a method, called DHTJoin, which combines hash-based placement of tuples in a Distributed Hash Table (DHT) and dissemination of queries by exploiting the embedded trees in the underlying DHT, thereby incuring little overhead. DHTJoin identifies, using query predicates, a subset of tuples in order to index the data required by the user's queries, thus reducing network traffic [37]. This is more efficient than the approaches based on structured P2P overlays, e.g. PIER and RJoin, which typically index all tuples in the network. We provided an analytical evaluation in [37] of the best number of nodes to obtain a certain degree of completeness given a continuous join query. DHTJoin tackles the dynamic behavior of DHT networks during query execution and dissemination of queries [22] [38]. When nodes fail during query dissemination, DHTJoin uses a gossip-based protocol that assures 100% of network coverage. When nodes fail during query execution, DHTJoin propagates messages to prevent nodes of sending intermediate results that do not contribute to join results, thereby reducing network traffic. Finally, DHTJoin provides an efficient solution to deal with overloaded nodes as a result of data skew [22][38]. The key idea is to distribute the tuples of an overloaded node to some underloaded nodes, called partners. When a node gets overloaded, DHTJoin discovers partners using information in the routing table and determines what tuples to send them using the concept of domain partitioning. We show that, in this case, DHTJoin incurs only one additional message per joined tuple produced, thus keeping response time low.

We evaluated the performance of DHTJoin through simulation. The results show the effectiveness of our solution compared with previous work.

6.4.2. P2P Content Distribution Network

Participants: Manal El Dick, Esther Pacitti.

P2P networks provide a very cost-effective alternative to build highly scalable infrastructures for content distribution. This is particularly useful for non-profit websites with a large user base (e.g., non-profit organizations) that cannot afford to distribute their popular content at large scales via commercial content distribution networks (CDN) like Akamai.

Our first contribution [29], [30] consists in building a P2P CDN, *Flower-CDN*, that enables any under-provisioned website to efficiently distribute its content, with the help of the non-profit community interested in its content. Our solution exhibits several unique characteristics that enable us to overcome all of the above mentioned challenges. It combines the strengths of both structured and unstructured P2P networks, exploiting DHT efficiency and gossip robustness. Flower-CDN introduces a novel DHT usage and management, called D-ring, that relies on a new locality- and interest-aware key service. It helps new peers to quickly find peers in the same network locality that are interested in the same website. We organize peers that share the same locality and are interested in the same website into unstructured overlay clusters (called *petals*). Within a petal,

peers use gossip protocols to exchange information about their content and contacts, allowing Flower-CDN to maintain accurate information despite dynamic changes in order to support eventual queries. We use this novel two-layered architecture consisting of a D-ring and petals to provide hybrid locality-aware query routing. The D-ring ensures a reliable access for new clients, while subsequent searches are performed within the petals. Thus, most of the query routing takes place within a local cluster leading to short query search and local data transfer. Our empirical analysis show that Flower-CDN can reduce lookup latency by a factor of 9 and the transfer distance by a factor of 2, compared to an existing P2P CDN (i.e. Squirrel). Moreover, Flower-CDN incurs very acceptable overhead in terms of gossip bandwidth, which can also be tuned according to hit ratio requirements and bandwidth availability.

Our second contribution [28] aims at providing our P2P CDN with high scalability and robustness under large scale and dynamic participation of peers. Thus, we propose *PetalUp-CDN*, which dynamically adapts Flower-CDN to increasing numbers of participants in order to avoid overload situations. In short, PetalUp-CDN enables D-ring to progressively expand to manage larger petals so that all the participants share the workload rather evenly. In addition, we maintain our P2P CDN in face of high churn and failures, by relying on low-cost gossip protocols. Our maintenance protocols preserve the locality and interest aware features of our achitecture and enables fast and efficient recovery. Based on extensive simulations, we show that our approach leverages larger scales to achieve higher improvements. Furthermore Flower-CDN can maintain an excellent performance under a highly dynamic participation of peers. This work was done in cooperation with Bettina Kemme (Mc Gill Univ.).

6.4.3. Uncertain Data Management

Participants: Reza Akbarinia, Esther Pacitti, Patrick Valduriez.

We are witnessing a rapid and important increase in the interest for uncertain data management. One of the main reasons is the emergence of many applications in which data uncertainty is unavoidable; e.g. data cleaning, sensor networks, information extraction, etc. In a recent work [36], we investigated the challenges of uncertain query processing in P2P online communities. In these environments the data are not 100 % certain, precise and correct, particularly when coming from peers with different levels of confidence. Query processing techniques designed for P2P systems should be revisited to deal with data uncertainty at all levels. Similarly, the recent extensions of DBMS that support data uncertainty should be revisited for P2P networks. We also addressed the problem of estimating the data confidence in P2P community information management systems. Since the data are not certain, we need to estimate the certainty degree (i.e. confidence) of the data. For this, we rely on the knowledge of all users of the systems, and use their feed-back to estimate the data confidence. We proposed a new data model, called feedback graph that models the relation between the users, their data and feedbacks. Based on this model, we developed a distributed approach for managing the feedback graph, and computing the data confidence based on a recursive formula.

In another work, we have started to study uncertain aggregate (aggr) queries which have been proven to be very useful for many uncertain data management applications. Examples of these applications are day-ahead energy market estimation, moving objects surveillance, mortgage default prediction, stock market prediction, etc. To evaluate aggr queries over uncertain data, we must firt provide a definition (semantics) of these queries in uncertain databases. In initial works, the aggr queries were defined based on the expected value semantics, i.e. expected value of aggregate attributes in uncertain tuples. However, recent works have shown that this semantics is not sufficient for many applications, and other semantics are needed. 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 in polynomial time. We plan to extend our algorithms to distributed systems, in particular, P2P systems. We should take into account the data distribution which makes the problem of uncertain aggr query processing much more complicated than that in centralized systems. Furthermore, we should deal with the dynamic behavior of peers that may leave the system or fail during query processing.

6.5. Transactional Memory in Multicore Systems

Participants: Toufiq Sarni, Patrick Valduriez.

We started a new, promising research direction on transactional memory in multicore systems, with the objective of levereging multicore technology for accessing shared data under real-time constraints. With the advent of multicore systems, the transactional memory (TM) concept has attracted much interest from both academy and industry as it eases programming and avoids the problems of lock-based methods. By supporting the ACI (Atomicity, Consistency and Isolation) properties of transactions, transactional memory relieves the programmer from dealing with locks to access resources. More important, it avoids the severe problems of lock-based methods such as deadlock situations and priority inversions. Furthermore, in the case of multicore systems, lock-based synchronization can reduce the data bandwidth by blocking several processes that try to access critical sections, thus reducing processors utilization. By contrast, TM allows several transactions to access resources in parallel. A transaction is either aborted when a conflict is detected, or committed in case of successful completion. Conflicts are handled with non-blocking synchronization which offers some guarantee of forward progress.

Existing solutions for transactional memory are implemented either in hardware (HTM) or software (STM) or both together (HyTM). STM researchers take care about performance issues on TM, and in the literature, several policies have been proposed to manage conflicts between transactions. The main challenge of these policies is to increase the system's throughput by increasing the number of committed transactions per unit of time.

However, the performance metrics used to study TM are rather intended for classical systems, without time constraints, and then not suitable for real-time applications [45]. Real-time means a predictable execution time of both tasks and their resource accesses, thus involving predictability at the transactions level.
Real-time applications are now becoming more and more concurrent and complex. For instance, telecommunication systems today, manage a huge and growing amount of shared data customers with time constraints on services. This challenge imposes to consider the interactions between the multiprocessor real-time scheduler of tasks and that of transactions embedded within TM. For this purpose, our first approach was to identify which STM among various kinds of non-blocking synchronization, is more suitable for the real-time context. For the best selected STM, the second step consisted in identifying the best multiprocessor real-time scheduling policy among both the partitionned, global, and fairness approaches [41]. Moreover, the scheduler of transactions takes a decision to abort or commit a transaction if a conflit occurs. This decision is usually based on the transaction time-stamp value or number the of transactions' retries. However, in real-time systems the task execution is based on its priority. For this reason, we proposed a new transaction model for STM that integrates deadline parameter of transactions. Based on this model, the scheduler of transactions commits first transactions that are close to their deadline, thus reducing the probability for a transaction to violate its time constraints. [41].

On the other hand, it is often claimed that the rollback times are the main cause of the execution time variation when using transactions in memory. However, the most recent STMs are based on a dynamic memory allocator. We carried out a directed real-time case study [42] to show that the rollbacks times are not the main reason for the execution time variation of transactions, and proved that a good memory allocator must also be provided to bound the execution time variation of transactions.

7. Contracts and Grants with Industry

7.1. STREP Grid4All (2006-2009)

Participants: William Kokou Dedzoe, Rabab Hayek, Philippe Lamarre, Esther Pacitti, Patrick Valduriez.

The project is with France Telecom R&D (leader), INRIA (Atlas, Grand-Large, Regal, and Sardes), Kungliga Tekniska Hoegskolan, Swedish Institute of Computer Science, ICCS (Greece), University of Piraeus Research Center, Universitat Politecnica de Catalunya and Rededia S.L. (Spain). Atlas and INRIA-Rennes are the INRIA representatives. The goal of Grid4All is to develop a grid infrastructure and middleware for the collaboration of dynamic, small virtual organizations such as communities, schools and families. The main technical innovation is to foster the combination of grid and P2P techniques to provide a light-weight, flexible solution. Atlas contributes to the definition of the P2P infrastructure (which is based on APPA) and to the development of two key services: resource discovery (using our mediation techniques) and optimistic replication (using our semantic reconciliation techniques). The project ended in september, after a successful review.

7.2. RNTL XWiki Concerto (2006-2008)

Participants: Esther Pacitti, Mounir Tlili, Patrick Valduriez.

The project involves XPertNet, ObjectWeb, INRIA, ENST, Mandriva, and EISTI. The goal of the project is to enable Xwiki, an open source second generation wiki product, to operate in a P2P environment and support mobile users. In this project, Atlas developed with the ECOO INRIA project-team the technologies for collaborative editing of wiki documents in P2P which has been integrated in a P2P wiki prototype. The project ended on april and obtained excellent reviews for P2P wiki.

7.3. ANR Safimage (2007-2010)

Participants: Marc Gelgon, Pierrick Bruneau.

This project involves Alcatel, IRCCyN, and IST. The project deals with inspection of data in high-speed routers for security purposes. The task devoted to Atlas is classification of multimedia data (examining how to scale up learning and recognition tasks with state-of-the-art classifiers in future routers).

7.4. PREDIT EPILOG (2009-2011)

Participants: Philippe Lamarre, Vu Duc Trung, 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 interesests, the modeling of information exchanged during transactions and query processing in the P2P network. The project was started in november.

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 Region-funded project 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.

8.2. National Actions

We are involved in the following project:

8.2.1. ANR VERSO DataRing(2008-2011)

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 Île 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 SARAVÁ with UFRJ, Rio de Janeiro (Marta Mattoso, Vanessa Braganholo, Alexandre Lima) to work on P2P data management for online comunities;
- the PICASSO project Scaling GraphDB, with UPC, Barcelona ((Josep Lluis Larriba Pey and Victor Muntès Mulero, visiting researcher in Atlas for 6 months) to work on very large graph database support;
- the STIC multimedia network between France and Morocco, with University Mohammed V of Rabat, EMI, ENSIAS and University of Fès.

Furthermore, we have regular scientific relationships with research laboratories in

- North America: Univ. of Waterloo (Tamer Özsu), McGill University (Bettina Kemme);
- Europe: Univ. of Madrid (Ricardo Jimenez-Periz), Univ. of Barcelona (Josep Lluis Larriba Pey), Univ. of Roskilde (Henrik Larsen), Nokia (Andreas Myka), ;
- Brazil: Univ. Federal of Rio de Janeiro (Marta Mattoso), PUC-Rio (Sergio Lifschitz), PUCPR, Curitiba, Brazil (Vidal Martins);
- International Univ. of Rabat (Noureddine Mouaddib, formely in the team).

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.

In 2009, P. Valduriez was the general chair of the VLDB conference in Lyon, with S. Abiteboul (Gemo project-team) as PC chair. VLDB 2009 broke the record of participation, with about 750 participants from 44 different countries.

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.
- Book series “Data Centric Systems and Applications” (Springer-Verlag): P. Valduriez.
- Ingénierie des Systèmes d’Information, Hermès : P. Valduriez.

Participation in conference program committees :

- Int. Conf. on VLDB 2009: E. Pacitti.
- ACM Int. Conf. on Information and Knowledge Management (CIKM) 2009: E. Pacitti.
- IEEE Int. Conf. on Distributed Computing Systems (ICDCS), Data management 2009: E. Pacitti; 2010: R. Akbarinia.
- Int. Conf. on High Performance Computing for Computational Science (VecPar) 2010: R. Akbarinia.
- IEEE/WIC/ACM Int. Conf. on Web Intelligence, 2009: M. Gelgon.
- Damap workshop on P2P data management, co-located with EDBT 2009: P. Lamarre, E. Pacitti.
- Int. Conf. on Extending DataBase Technologies (EDBT) 2009: E. Pacitti; 2010: P. Valduriez.
- EDBT Ph.D. Workshop 2009, 2010: P. Valduriez.
- ACM Symposium of Applied Computing (SAC), Privacy on the Web 2010: P. Valduriez.
- First International Workshop on MapReduce and its Applications (MAPREDUCE), 2010: P. Valduriez.
- Journées Bases de Données Avancées (BDA), 2009: P. Lamarre, G. Raschia.

9.3. Invited Talks

In April, E. Pacitti gave an invited talk at UFRJ, Rio de Janeiro, on Flower-CDN, a hybrid P2P overlay for efficient query processing in content distribution networks.

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.

Noureddine Mouaddib, on leave from Polytech’Nantes and a former member of the team, has been appointed the president of Rabat International University, Morocco, 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 (2nd edition) 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. A third edition is in progress and will be a major revision with much new material on replication, P2P, parallel systems and web data integration. A third edition is (at last) reaching completion.

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", vol. 10, n^{o} 3, April 2008, p. 385-392.

[6] E. PACITTI, P. VALDURIEZ, M. MATTOSO. *Grid Data Management: Open Problems and New Issues*, in "Journal of Grid Computing", vol. 5, n^{o} 3, 2007, 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", vol. 26, n^{o} 2-3, 2009, 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. 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", vol. 18, n^{o} 3, 2009, 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, 2nd edition*, Prentice Hall, 1999.

Year Publications

Doctoral Dissertations and Habilitation Theses

[13] M. BECHCHI. *Clustering-based Approximate Answering of Queries in Large and Distributed Databases*, Université de Nantes, 2009, Ph. D. Thesis.

[14] E. CUNHA DE ALMEIDA. *Test and Validation of P2P Systems*, Université de Nantes, 2009, Ph. D. Thesis.

[15] R. HAYEK. *Summary Management in P2P Systems*, Université de Nantes, 2009, Ph. D. Thesis.

[16] P. LAMARRE. *Contributions à la recherche d'information dans des systèmes distribués, ouverts, intégrant des participants autonomes*, Université de Nantes, 2009, Habilitation à Diriger des Recherches.

[17] J. MANJARREZ. *Parallel Content-based Retrieval in Image Databases*, Université de Nantes, 2009, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

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

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

[20] M. DIDONET DEL FABRO, P. VALDURIEZ. *Towards the Efficient Development of Model Transformations using Model Weaving and Matching Transformations*, in "Software and Systems Modeling (SoSyM)", vol. 8, n^{o} 3, 2009, p. 305-324.

[21] ALEXANDRE A. B. LIMA, C. FURTADO, P. VALDURIEZ, M. MATTOSO. *Parallel OLAP query processing in database clusters with data replication*, in "Distributed and Parallel Databases", vol. 25, n^{o} 1-2, 2009, p. 97-123 BR .

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

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

[24] E. C. DE ALMEIDA, G. SUNYÉ, Y. L. TRAON, P. VALDURIEZ. *Testing Peer-to-Peer Systems*, in "Empirical Software Engineering", 2009, to appear.

International Peer-Reviewed Conference/Proceedings

[25] P. BRUNEAU, M. GELGON, F. PICAROUGNE. *Parsimonious variational-Bayes mixture aggregation with a Poisson prior*, in "EURASIP European Signal Processing Conference (EUSIPCO'2009), Glasgow, U.K.", August 2009.

[26] P. BRUNEAU, F. PICAROUGNE, M. GELGON. *Incremental semi-supervised clustering in a data stream with a flock of agents*, in "IEEE Congress on Evolutionary Computing (CEC'2009), Trondheim, Norway.", May 2009.

[27] P. BRUNEAU, A. PIGEAU, M. GELGON, F. PICAROUGNE. *Geo-temporal structuring of a personal image database with two-level variational-Bayes mixture estimation*, in "Adaptive Multimedia Retrieval workshop (AMR'08), Berlin, Germany", LNCS, Springer, 2009, to appear.

[28] M. E. DICK, E. PACITTI, B. KEMME. *A Highly Robust P2P-CDN under Large-Scale and Dynamic Participation*, in "Int. Conf. on Advances in P2P Systems (AP2PS)", 2009, p. 180-185 CA .

[29] 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), Saint Petersburg, Russia", 2009, p. 427-438 CA .

[30] M. E. DICK, E. PACITTI, B. KEMME. *Un Réseau Pair-à-Pair de Distribution de Contenu Exploitant les Intérêts et les Localités des Pairs*, in "Journées Bases de Données Avancées (BDA), Namur, Belgium", 2009, p. 407-388 CA .

[31] A. EL ATTAR, A. PIGEAU, M. GELGON. *Fast aggregation of Student mixture models*, in "EURASIP European Signal Processing Conference (EUSIPCO'2009), Glasgow, U.K.", August 2009.

[32] M. EL DICK, E. PACITTI, P. VALDURIEZ. *Location-aware Index Caching and Searching for P2P Systems*, in "VLDB Int. Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P) - revised selected papers, Vienna, Austria", LNCS, Springer, 2009, to appear.

[33] M. JAWAD, P. SERRANO-ALVARADO, P. VALDURIEZ, S. DRAPEAU. *A Data Privacy Service for Structured P2P Systems*, in "Mexican International Conference on Computer Science (ENC), Mexico City, Mexico", IEEE Computer Society, 2009.

[34] M. JAWAD, P. SERRANO-ALVARADO, P. VALDURIEZ, S. DRAPEAU. *Data Privacy in Structured P2P systems with PriServ*, in "Journées Bases de Données Avancées (BDA), Namur, Belgium", 2009.

[35] M. JAWAD, P. SERRANO-ALVARADO, P. VALDURIEZ. *Protecting Data Privacy in Structured P2P Networks*, in "Second International Conference on Data Management in Grid and Peer-to-Peer Systems (Globe), Linz, Austria", LNCS 5697, Springer, 2009, p. 85-98.

[36] M. MATTOSO, E. PACITTI, P. VALDURIEZ, R. AKBARINIA, V. BRAGANHOLO, A. A. B. LIMA. *SARAVÁ: data sharing for online communities in P2P*, in "Colloquium of Computation: Brazil / INRIA, Cooperations, Advances and Challenges (COLIBRI)", SBC/INRIA, ISBN 978-85-7669-245-4, 2009, Best Paper Award BR .

[37] W. PALMA, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Distributed Processing of Continuous Join Queries using DHT Networks*, in "Int. Workshop on Data Management in Peer-to-Peer Systems (DaMaP), Saint Petersburg, Russia", ACM International Conference Proceeding Series, 2009.

[38] W. PALMA, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Processing of Continuous Join Queries in a P2P Environment*, in "Journées Bases de Données Avancées (BDA), Namur, Belgium", 2009.

[39] Q.-K. PHAM, G. RASCHIA, R. SAINT-PAUL, B. BENATALLAH, N. MOUADDIB. *Time Sequence Summarization to Scale Up Chronology-dependent Applications*, in "ACM Conf. on Information and Knowledge Management (CIKM), Hong-Kong, China", 2009, p. 1137-1146.

[40] J.-A. QUIANÉ-RUIZ, P. LAMARRE, P. VALDURIEZ. *SbQA: A Self-Adaptable Query Allocation Process*, in "IEEE Int. Conf. on Data Engineering (ICDE'09), Shangai, China", 2009, p. 1527-1530, (Demo paper).

[41] T. SARNI, A. QUEUDET, P. VALDURIEZ. *Real-Time Support for Software Transactional Memory*, in "IEEE Int. Conf. on Embedded and Real-Time Computing Systems and Applications, Beijing, China", August 2009, p. 477-485.

[42] T. SARNI, A. QUEUDET, P. VALDURIEZ. *Software Transactional Memory: Worst Case Execution Time Analysis*, in "Int. Conf. on Real-Time and Network Systems, Paris, France", INRIA, October 2009, p. 107-114.

[43] A. VENTRESQUE, P. LAMARRE, S. CAZALENS, P. VALDURIEZ. *Représentation optimiste de contenus dans les système P2P*, in "Journées Bases de Données Avancées (BDA), Namur, Belgium", 2009.

National Peer-Reviewed Conference/Proceedings

[44] 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.

Workshops without Proceedings

[45] T. SARNI, A. QUEUDET, P. VALDURIEZ. *Real-Time Scheduling of Transactions in Multicore Systems*, in "Workshop on Massively Multiprocessor and Multicore Computers, Rocquencourt, France", 2009.

Scientific Books (or Scientific Book chapters)

[46] R. AKBARINIA. *Data Access in Dynamic Distributed Systems, 1st edition*, VDM-Verlag, 2009.

[47] M. BECHCHI, G. RASCHIA, N. MOUADDIB. *Practical Approaches to the Many-Answers Problem*, in "Advanced Database Query Systems", IGI Global, 2009, to appear.

[48] R. HAYEK, G. RASCHIA, P. VALDURIEZ, N. MOUADDIB. *Data Localization and Description Through Summaries in P2P Collaborative Applications*, in "Handbook of Peer-to-Peer Networking", H. X. SHEN, J. YU, M. A. BUFORD (editors), Springer, 2009.

[49] E. PACITTI. *Parallel Query Processing*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer, 2009, p. 2038-2040.

[50] P. VALDURIEZ. *Parallel Data Placement*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer, 2009, p. 2024-2026.

[51] P. VALDURIEZ. *Parallel Database Management*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer, 2009, p. 2026-2029.