IN PARTNERSHIP WITH:
CNRS
Université de Montpellier

Activity Report 2016

Project-Team ZENITH

Scientific Data Management

IN COLLABORATION WITH: Laboratoire d'informatique, de robotique et de microélectronique de Montpellier (LIRMM)

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ZENITH

Creation of the Team: 2011 January 01, updated into Project-Team: 2012 January 01

Keywords:

Computer Science and Digital Science:

1. - Architectures, systems and networks
1.1. - Architectures
1.1.6. - Cloud
1.1.7. - Peer to peer
3.1. - Data
3.3. - Data and knowledge analysis
3.5. - Social networks
3.5.2. - Recommendation systems
4. - Security and privacy
4.8. - Privacy-enhancing technologies
5. - Interaction, multimedia and robotics
5.4. - Computer vision
5.4.3. - Content retrieval

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.9. - Bioinformatics
1.2. - Ecology
1.2.1. - Biodiversity
6. - IT and telecom
6.5. - Information systems

1. Members

Research Scientists

Reza Akbarinia [Inria, Researcher]
Alexis Joly [Inria, Researcher]
Florent Masseglia [Inria, Researcher, HDR]
Didier Parigot [Inria, Researcher, HDR]
Patrick Valduriez [Team leader, Inria, Senior Researcher, HDR]

Faculty Members

Esther Pacitti [Associate Team Leader, Univ. Montpellier, Professor, HDR]
Dennis Shasha [Inria international chair, NYU, Professor, HDR]

Engineers

Julien Champ [Inria and Inra, Pl@ntNet and ARCAD projects, until may 2016]
Boyan Kolev [Inria, FP7 CoherentPaaS project, until oct. 2016]
Pierre Larmande [IRD, collaborator]
Jean-Christophe Lombardo [Inria]
Oleksandra Levchenko [Inria, FP7 CoherentPaaS project]

Antoine Affouard [Inria, PIA Floris'tic project]
Benjamin Billet [Inria, Triton I-labt]
Sen Wang [Inria, PIA Floris'tic project]
Valentin Leveau [Inria, PIA Floris'tic project, since dec. 2016]

PhD Students

Christophe Botella [INRA-Inria fellowship, since oct. 2016]
Carlyna Bondiombouy [Congo fellowship]
Valentin Leveau [INA CIFRE, until nov. 2016]
Titouan Lorieul [LIRMM fellowship, since oct. 2016]
Ji Liu [Inria-MSR until nov. 2016]
David Fernandez [Univ. Nice, Triton e-lab, until june 2016]
Daniel Gaspar [LNCC, Rio de Janeiro, until aug. 2016]
Sakina Maboubi [Averroes fellowship]
Khadidja Meguelati [Averroes fellowship]
Djamel-Edine Yagoubi [Averroes fellowship]
Mehdi Zitouni [University of Tunis, Tunisia]

Post-Doctoral Fellows

Maximilien Servajean [Inria, FP7 CoherentPaaS project, until aug. 2016]
Ji Liu [Inria, H2020 HPC4E project, since dec. 2016]

2. Overall Objectives

2.1. Overall Objectives

Data-intensive science such as agronomy, astronomy, biology 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) creates a huge data overload. For example, climate modeling data are growing so fast that they will lead to collections of hundreds of exabytes by 2020.

Scientific data is very complex, in particular because of heterogeneous methods used for producing data, the uncertainty of captured data, the inherently multi-scale nature (spatial scale, temporal scale) of many sciences and the growing use of imaging (e.g. molecular imaging), resulting in data with hundreds of attributes, dimensions or descriptors. Modern science research is also highly collaborative, involving scientists from different disciplines (e.g. biologists, soil scientists, and geologists working on an environmental project), in some cases from different organizations in different countries. Each discipline or organization tends to produce and manage its own data, in specific formats, with its own processes. Thus, integrating such distributed data gets difficult as the amounts of heterogeneous data grow.

Despite their variety, we can identify common features of scientific data: big data; manipulated through complex, distributed workflows; typically complex, e.g. multidimensional or graph-based; with uncertainty in the data values, e.g., to reflect data capture or observation; important metadata about experiments and their provenance; and mostly append-only (with rare updates).

Relational DBMSs, which have proved effective in many application domains (e.g. business transactions, business intelligence), are not efficient at dealing with scientific data or big data, which is typically unstructured. In particular, they have been criticized for their "one size fits all" approach. As an alternative , more specialized solutions are being developped such as NoSQL/NewSQL DBMSs and data processing frameworks (e.g. Spark) on top of distributed file systems (e.g. HDFS).

The three main challenges of scientific data management can be summarized by: (1) scale (big data, big applications); (2) complexity (uncertain, multi-scale data with lots of dimensions), (3) heterogeneity (in particular, data semantics heterogeneity). These challenges are also those of data science, with the goal of making sense out of data by combining data management, machine learning, statistics and other disciplines. The overall goal of Zenith is to address these challenges, by proposing innovative solutions with significant advantages in terms of scalability, functionality, ease of use, and performance. To produce generic results, these solutions are in terms of architectures, models and algorithms that can be implemented in terms of components or services in specific computing environments, e.g. cloud. We design and validate our solutions by working closely with our scientific application partners such as INRA and IRD in France, or the National Research Institute on e-medicine (MACC) in Brazil. To further validate our solutions and extend the scope of our results, we also foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

Our approach is to capitalize on the principles of distributed and parallel data management. In particular, we exploit: high-level languages as the basis for data independence and automatic optimization; data semantics to improve information retrieval and automate data integration; declarative languages to manipulate data and workflows; and highly distributed and parallel environments such as P2P, cluster and cloud. We also exploit machine learning and statistics for data analytics and data search. To reflect our approach, we organize our research program in five complementary themes:

- Data integration, including data capture and cleaning;
- Data management, in particular, indexing and privacy;
- Scientific workflows, in particular, in grid and cloud;
- Data analytics, including data mining and statistics;
- Data search, including machine learning and content-based image retrieval.

3. Research Program

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 DBMS, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (search engines, application servers, document systems, 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 (smart phone, PDA, smart card, etc.) to very large computers (multiprocessor, cluster, etc.) in distributed environments. For a long time, the research 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.

The problems of scientific data management (massive scale, complexity and heterogeneity) go well beyond the traditional context of DBMS. To address them, we capitalize on scientific foundations in closely related domains: distributed data management, cloud data management, big data, big data integration, scientific workflows, data analytics and search.

3.2. Distributed Data Management

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud.

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 to be 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.

Parallel database systems 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 [8] 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 BitTorrent 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. A P2P solution is well-suited to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources. But for very-large scale scientific data analysis, we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the best of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.3. Cloud Data Management

Cloud computing encompasses on demand, reliable services provided over the Internet (typically represented as a cloud) with easy access to virtually infinite computing, storage and networking resources. Through very simple Web interfaces and at small incremental cost, users can outsource complex tasks, such as data storage, system administration, or application deployment, to very large data centers operated by cloud providers. However, cloud computing has some drawbacks and not all applications are good candidates for being "cloudified". The major concern is w.r.t. data security and privacy, and trust in the provider (which may use no so trustful providers to operate). One earlier criticism of cloud computing was that customers get locked in proprietary clouds. It is true that most clouds are proprietary and there are no standards for cloud interoperability.

There is much more variety in cloud data than in scientific data since there are many different kinds of customers (individuals, samll companies, large corporations, etc.). However, we can identify common features. Cloud data can be very large, unstructured (e.g. text-based) or semi-structured, and typically append-only (with rare updates). And cloud users and application developers may be in high numbers, but not DBMS experts.

3.4. Big Data

Big data (like its relative, data science) has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte of Hard Disk Drive for: 1M$ in 1982, 1K$ in 1995, 0.02$ in 2015), making it affordable to keep more data around. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's such as: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Many different big data management solutions have been designed, primarily for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility, hence the new term Data-Intensive Scalable Computing (DISC). Examples of DISC systems include data processing frameworks (e.g. Hadoop MapReduce, Apache Spark, Pregel), file systems (e.g. Google GFS, HDFS), NoSQL systems (Google BigTable, Hbase, MongoDB), NewSQL systems (Google F1, CockroachDB, LeanXcale). In Zenith, we exploit or extend DISC technologies to fit our needs for scientific workflow management and scalable data analysis.

3.5. Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are integrated and materialized in a data warehouse. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

In both cases, to understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the autonomy of the underlying data sources, which leads to a large variety of models and formats. Thus, it is necessary to identify semantic correspondences between the metadata of the related data sources.

This requires the matching of the heterogeneous metadata, by discovering semantic correspondences between ontologies, and the annotation of data sources using ontologies. In Zenith, we rely on semantic web techniques (e.g. RDF and SparkQL) to perform these tasks and deal with high numbers of data sources.

Scientific workflow management systems (SWfMS) are also useful for data integration. They allow scientists to describe and execute complex scientific activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data. This requires using distributed and parallel execution environments. However, existing workflow management systems have limited support for data parallelism. In Zenith, we use an algebraic approach to describe data-intensive workflows and exploit parallelism.

3.6. Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management. Data mining provides methods to discover new and useful patterns from very large datasets. These patterns may take different forms, depending on the end-user's request, such as:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. This problem was first motivated by commercial and marketing purposes (*e.g.* discovering frequent correlations between items bought in a shop, which could help selling more). A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset discovery but considering the order between. In the smart building example, a frequent sequence could say that "in 40% of rooms, lights are on at time i, the room is empty at time i+j and the door is closed at time i+j+k". Discovering frequent sequences has become critical in marketing, as well as in security (e.g. detecting network intrusions), in web usage analysis and any domain where data come in a specific order, typically given by timestamps.
- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

One main problem in data analytics is to deal with data streams. Existing methods have been designed for very large data sets where complex algorithms from artificial intelligence were not efficient because of data size. However, we now must deal with data streams, sequences of data events arriving at high rate, where traditional data analytics techniques cannot complete in real-time, given the infinite data size. In order to extract knowledge from data streams, the data mining community has investigated approximation methods that could yield good result quality.

3.7. Data Search

Technologies for searching information in scientific data have relied on relational DBMS or text-based indexing methods. However, content-based information retrieval has progressed much in the last decade, with much impact on search engines. Rather than restricting the search to the use of metadata, content-based methods index, search and browse digital objects by means of signatures that describe their content. Such methods have been intensively studied in the multimedia community to allow searching massive amounts of multimedia documents that are created every day (e.g. 99% of web data are audio-visual content with very sparse metadata). Scalable content-based methods have been proposed for searching objects in large image collections or detecting copies in huge video archives. Besides multimedia contents, content-based information retrieval methods have expanded their scope to deal with more diverse data such as medical images, 3D models or even molecular data. Potential applications in scientific data management are numerous. First, to allow searching within huge collections of scientific images (earth observation, medical images, botanical images, biology images, etc.) or browsing large datasets of experimental data (e.g. multisensor data, molecular data or instrumental data). However, scalability remains a major issue, involving complex algorithms (such as similarity search, clustering or supervised retrieval), in high dimensional spaces (up to millions of dimensions) with complex metrics (Lp, Kernels, sets intersections, edit distances, etc.). Most of these algorithms have linear, quadratic or even cubic complexities so that their use at large scale is not affordable without major breakthroughs. In Zenith, we investigate the following challenges:

- **High-dimensional similarity search**. Whereas many indexing methods were designed in the last 20 years to efficiently retrieve multidimensional data with relatively small dimensions, high-dimensional data are challenged by the well-known curse of dimensionality . Only recently have some methods appeared that allow approximate Nearest Neighbors queries in sub-linear time, in particular, Locality Sensitive Hashing methods that offer new theoretical insights in high-dimensional Euclidean spaces and random projections. But there are still challenging issues such as efficient similarity search in any kernel or metric spaces, efficient construction of k-nearest neighbor graphs (k-NNG) or relational similarity queries.
- **Large-scale supervised retrieval**. Supervised retrieval aims at retrieving relevant objects in a dataset by providing some positive and/or negative training samples. Toward this goal, Support Vector Machines (SVM) offer the possibility to construct generalized, non-linear predictors in high-dimensional spaces using small training sets. The prediction time complexity of these methods is usually linear in dataset size. Allowing hyperplane similarity queries in sub-linear time is for example a challenging research issue. A symmetric problem in supervised retrieval consists in retrieving the most relevant object categories that might contain a given query object, providing huge labeled datasets (up to millions of classes and billions of objects) and very few objects per category (from 1 to 100 objects). SVM methods that are formulated as quadratic programming with cubic training time complexity and quadratic space complexity are clearly not usable. Promising solutions include hybrid supervised-unsupervised methods and supervised hashing methods.
- **Distributed content-based retrieval**. Distributed content-based retrieval methods appear as a promising solution to manage masses of data distributed over large networks, in particular when the data cannot be centralized for privacy or cost reasons, which is often the case in scientific social networks. However, current methods are limited to very simple similarity search paradigms. In Zenith, we consider more advanced distributed content-based retrieval and mining methods such as k-NNG construction, large-scale supervised retrieval or multi-source clustering.

4. Application Domains

4.1. Data-intensive Scientific Applications

The application domains covered by Zenith are very wide and diverse, as they concern data-intensive scientific applications, i.e., most scientific applications. Since the interaction with scientists is crucial to identify and tackle data management problems, we are dealing primarily with application domains for which Montpellier has an excellent track record, i.e., agronomy, environmental science, life science, with scientific partners like INRA, IRD and CIRAD. However, we are also addressing other scientific domains (e.g. astronomy, oil extraction) through our international collaborations (e.g. in Brazil).

Let us briefly illustrate some representative examples of scientific applications on which we have been working on.

- **Management of astronomical catalogs**. An example of data-intensive scientific applications is the management of astronomical catalogs generated by the Dark Energy Survey (DES) project on which we are collaborating with researchers from Brazil. In this project, huge tables with billions of tuples and hundreds of attributes (corresponding to dimensions, mainly double precision real numbers) store the collected sky data. Data are appended to the catalog database as new observations are performed and the resulting database size is estimated to reach 100TB very soon. Scientists around the globe can query the database with queries that may contain a considerable number of attributes. The volume of data that this application holds poses important challenges for data management. In particular, efficient solutions are needed to partition and distribute the data in several servers. An efficient partitioning scheme should try to minimize the number of fragments accessed in the execution of a query, thus reducing the overhead associated to handle the distributed execution.

- **Personal health data analysis and privacy** Today, it is possible to acquire data on many domains related to personal data. For instance, one can collect data on her daily activities, habits or health. It is also possible to measure performance in sports. This can be done thanks to sensors, communicating devices or even connected glasses. Such data, once acquired, can lead to valuable knowledge for these domains. For people having a specific disease, it might be important to know if they belong to a specific category that needs particular care. For an individual, it can be interesting to find a category that corresponds to her performances in a specific sport and then adapt her training with an adequate program. Meanwhile, for privacy reasons, people will be reluctant to share their personal data and make them public. Therefore, it is important to provide them solutions that can extract such knowledge from everybody's data, while guaranteeing that their private data won't be disclosed to anyone.
- **Botanical data sharing**. Botanical data is highly decentralized and heterogeneous. Each actor has its own expertise domain, hosts its own data, and describes them in a specific format. Furthermore, botanical data is complex. A single plant's observation might include many structured and unstructured tags, several images of different organs, some empirical measurements and a few other contextual data (time, location, author, etc.). A noticeable consequence is that simply identifying plant species is often a very difficult task; even for the botanists themselves (the so-called taxonomic gap). Botanical data sharing should thus speed up the integration of raw observation data, while providing users an easy and efficient access to integrated data. This requires to deal with social-based data integration and sharing, massive data analysis and scalable content-based information retrieval. We address this application in the context of the French initiative Pl@ntNet, with CIRAD and IRD.
- **Biology data integration and analysis**.

  Biology and its applications, from medicine to agronomy and ecology, are now producing massive data, which is revolutionizing the way life scientists work. For instance, using plant phenotyping platforms such as PhenoDyn at INRA Montpellier, quantitative genetic methods allow to identify genes involved in phenotypic variation in response to environmental conditions. These methods produce large amounts of data at different time intervals (minutes to days), at different sites and at different scales ranging from small tissue samples until the entire plant. Analyzing such big data creates new challenges for data management and data integration.

These application examples illustrate the diversity of requirements and issues which we are addressing with our scientific application partners. To further validate our solutions and extend the scope of our results, we also want to foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

5. Highlights of the Year

5.1. Highlights of the Year

- The Pl@ntNet application, developed by Zenith and its partners, is enjoying a huge success: more than 2.7M downloads as in November 2016 in 150 countries; the number of users doubles every 6 months; tens of thousands of users each day, 12% being professionnals in agriculture or education.
- Alexis Joly and his collaborators of the Pl@ntNet project have been awarded the prize “La Recherche 2016” organized by the French magazine La Recherche for the article [2].

6. New Software and Platforms

6.1. Pl@ntNet

Participants: Antoine Affouard, Jean-Christophe Lombardo, Hervé Goëau, Alexis Joly [contact].

Pl@ntNet is an image sharing and retrieval application for the identification of plants. It is developed in the context of the Floris'tic project that involves Inria, CIRAD, INRA, IRD and Tela Botanica. The key feature of the iOS and Android front ends is to help identifying plant species from photographs, through a server-side visual search engine. Since its first release in march 2013 on the apple store, the application has been downloaded by 3M users in more than 170 countries, with between 15,000 and 50,000 active users daily. The collaborative training set that allows the content-based identification is continuously enriched by the users of the application and the members of Tela Botanica social network. At the time of writing, it includes about 300K images covering more than 10K species in the world (and about 60% of the West European flora).

6.2. The Plant Game: crowdsourced plants identification

Participants: Maximilien Servajean, Alexis Joly [contact], Antoine Affouard.

URL: http://theplantgame.com/

The Plant Game is a participatory game whose purpose is the production of large masses of taxonomic data to improve our knowledge of biodiversity. The objective is to learn botany with fun and participate to a large citizen sciences project in biodiversity. The game relies on consistent research contributions in scalable crowdsourcing models and algorithms that can deal with thousands of complex classes such as plant species. One major contribution is the active training of the users based on innovative sub-task creation and assignment processes that are adaptive to the increasing skills of the user. The first public version of the game was released in july 2015. As of today, about 22K players are registered and produce hundreds of new validated plant observations per day. The accuracy of the produced taxonomic tags is about 94%, which is quite impressive considering the fact that a majority of users are beginners when they start playing.

6.3. Smart'Flore

Participants: Antoine Affouard [contact], Alexis Joly, Hervé Goëau.

URL: http://otmedia.lirmm.fr/

Smart'Flore is an Android mobile application for the discovery of the surrounding vegetal biodiversity. It has three main features: (i) the geo-based exploration of the world's largest repository of biodiversity occurrences (GBIF, http://www.gbif.org/), (ii) the exploration of virtual botanical trails (created offline through a dedicated web application hosted by TelaBotanica NGO) and (iii) the access to a variety of information about the plants. Smart'Flore is the first mobile app in the world making use of the GBIF web services which makes it a remarkable and possibly highly visible realization. The first public version of the application was released in may 2016. Since then, it has been downloaded by more than 22K users and the daily number of sessions is about 250.

6.4. Snoop & SnoopIm

Participants: Alexis Joly, Julien Champ, Jean-Christophe Lombardo.

URL: http://otmedia.lirmm.fr/

Snoop is a generalist C++ library dedicated to high-dimensional data management and efficient similarity search. Its main features are dimension reduction, high-dimensional feature vector hashing, approximate k-nearest neighbors search and Hamming embedding. Snoop is a refactoring of a previous library called PMH developed jointly with INA. SnoopIm is a content-based image search engine built on top of Snoop that allows retrieving small visual patterns or objects in large collections of pictures. The software is used as the visual search engine of the Pl@ntNet applications and it is used in several other contexts, including a logo retrieval application in collaboration with INA, a whale's individuals matching application in collaboration with the CetaMada NGO, and a hieroglyph recognition application in collaboration with the Egyptology department of University Montpellier 3.

6.5. MultiSite-Rec

Participants: Mohamed Reda Bouadjenek, Florent Masseglia, Esther Pacitti.

Recommender systems are used as a mean to supply users with content that may be of interest to them. They have become a popular research topic, where many aspects and dimensions have been studied to make them more accurate and effective. In practice, recommender systems suffer from cold-start problems. However, users use many online services, which can provide information about their interest and the content of items (e.g. Google search engine, Facebook, Twitter, etc.). These services may be valuable data sources, which supply information to help a recommender system in modeling users and items' preferences, and thus, make the recommender system more precise. Moreover, these data sources are distributed, and geographically distant from each other, which raise many research problems and challenges to design a distributed recommendation algorithm. MultiSite-Rec is a distributed collaborative filtering algorithm, which exploits and combine these multiple and heterogeneous data sources to improve the recommendation quality.

6.6. Chiaroscuro

Participants: Tristan Allard, Florent Masseglia, Esther Pacitti.

URL: http://people.irisa.fr/Tristan.Allard/chiaroscuro/

Chiaroscuro is a software developped in the context of a research contract with EDF. It aims at clustering time series with privacy preserving guarantees. It is a distributed system, working in a P2P environment. It is used by the team for experiments and by EDF as a proof-of-concept. Chiaroscuro is the first software for that purpose. It is written in Java. The distributed algorithm implemented in Chiaroscuro has been filed by EDF in a patent (with Inria and University of Montpellier)

6.7. LogMagnet

Participant: Florent Masseglia.

URL: https://team.inria.fr/zenith/software/LogMagnet

LogMagnet is a software for analyzing streaming data, and in particular log data. Log data usually arrive in the form of lines containing activities of human or machines. In the case of human activities, it may be the behavior on a Web site or the usage of an application. In the case of machines, such log may contain the activities of software and hardware components (say, for each node of a computing cluster, the calls to system functions or some hardware alerts). Analyzing such data is often difficult and crucial in the meanwhile. LogMagnet allows to summarize this data, and to provide a first analysis as a clustering. This summary may also be exploited as easily as the original data.

6.8. FP-Hadoop

Participants: Reza Akbarinia, Patrick Valduriez.

https://gforge.inria.fr/plugins/mediawiki/wiki/fp-hadoop

FP-Hadoop is an extension of Hadoop that efficiently deals with the problem of data skew in MapReduce jobs. In FP-Hadoop, there is a new phase, called intermediate reduce (IR), in which blocks of intermediate values, constructed dynamically, are processed by intermediate reduce workers in parallel, by using a scheduling strategy.

6.9. CloudMdsQL Compiler

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

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

The CloudMdsQL (Cloud Multi-datastore Query Language) compiler transforms queries expressed in a common SQL-like query language into an optimized query execution plan to be executed over multiple cloud data stores (SQL, NoSQL, HDFS, etc.) through a query engine. The compiler/optimizer is implemented in C++ and uses the Boost.Spirit framework for parsing context-free grammars. CloudMdsQL has been validated on relational, document and graph data stores, as well as Spark/HDF in the context of the CoherentPaaS European project.

6.10. AgroLD

Participants: Pierre Larmande, Patrick Valduriez.

URL: http://www.agrold.org

Agronomic Linked Data (AgroLD) is a portal to help bioinformatics and domain experts exploiting the homogenized data models towards efficiently generating research hypotheses. AgroLD is an RDF knowledge base that is designed to integrate data from various publicly available plant centric data sources and ontologies, using Web Ontology Language (OWL) and the SPARQL Query Language (SPARQL).

6.11. SciFloware

Participants: Benjamin Billet, Didier Parigot.

URL: http://www-sop.inria.fr/members/Didier.Parigot/pmwiki/Scifloware

SciFloware is an action of technology development (ADT Inria) with the goal of developing a middleware for the execution of scientific workflows in a distributed and parallel way. It capitalizes on our experience with SON and an innovative algebraic approach to the management of scientific workflows. SciFloware provides a development environment and a runtime environment for scientific workflows, interoperable with existing systems. We validate SciFloware with workflows for analyzing biological data provided by our partners CIRAD, INRA and IRD.

7. New Results

7.1. Data Integration

7.1.1. CloudMdsQL, a query language for heterogeneous data stores

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

In the context of the CoherentPaaS European project, we have developped the Cloud Multi-datastore Query Language (CloudMdsQL), and its query engine. CloudMdsQL is a functional SQL-like language, capable of querying multiple heterogeneous data stores, e.g. relational, NoSQL or HDFS) [21]. The major innovation is that a CloudMdsQL query can exploit the full power of the local data stores, by simply allowing some local data store native queries to be called as functions, and at the same time be optimized. In [42], we demonstrate CloudMdsQL on two use cases each involving four diverse data stores (graph, document, relational, and key-value) with its corresponding CloudMdsQL queries. The query execution flows are visualized by an embedded real-time monitoring subsystem. In [17], we extend CloudMdsQL to allowing the ad-hoc usage of user defined map/filter/reduce operators in combination with traditional SQL statements, to integrate relational data and big data stored in HDFS and accessed by a data processing framework like Spark. Our experimental validation with several different data stores and representative queries [43] demonstrates the usability of the query language and the benefits from query optimization.

7.1.2. *Agronomic Linked Data*

Participant: Pierre Larmande.

Agronomic Linked Data (AgroLD) [30], [55], [54] is a knowledge system that exploits Semantic Web technology to integrate information on plant species widely studied by the agronomic research community. The objective is to provide the community with a platform for domain specific knowledge, capable of answering complex biological questions and thus facilitating the formulation of new hypotheses. The conceptual framework is based on well-established ontologies in plant sciences such as Gene Ontology, Sequence Ontology, Plant Ontology and Plant Environment Ontology. AgroLD version 1 consists of 50 million knowledge statements (i.e. RDF triples), which will grow in the subsequent versions to provide the required critical mass for hypotheses generation.

AgroLD relyes on AgroPortal [40], a reference ontology repository for the agronomi domain that features ontology hosting and search visualization with services for semantically annotating data with the ontologies. We used the AgroPortal Annotator web service to annotate more than 50 datasets and produced 22% additional triples validated manually. We also developed a dedicated AgroLD vocabulary that bridges the gap between these references ontologies and formalizes their mappings.

7.2. Data Management

7.2.1. *Scalable Query Processing with Big Data*

Participants: Reza Akbarinia, Patrick Valduriez.

In [22], we extend the popular Hadoop framework to deal efficiently with skewed MapReduce jobs. We extend the MapReduce programming model to allow the collaboration of reduce workers on processing the values of an intermediate key, without affecting the correctness of the final results. In FP-Hadoop, the reduce function is replaced by two functions: intermediate reduce and final reduce. There are three phases, each phase corresponding to one of the functions: map, intermediate reduce and final reduce phases. In the intermediate reduce phase, the function, which usually includes the main load of reducing in MapReduce jobs, is executed by reduce workers in a collaborative way, even if all values belong to only one intermediate key. This allows performing a big part of the reducing work by using the computing resources of all workers, even in case of highly skewed data. We implemented a prototype of FP-Hadoop by modifying Hadoop's code, and conducted extensive experiments over synthetic and real datasets. The results show that FP-Hadoop makes MapReduce job processing much faster and more parallel, and can efficiently deal with skewed data. We achieve excellent performance gains compared to native Hadoop, e.g. more than 10 times in reduce time and 5 times in total execution time.

7.2.2. *Management of Simulation Data*

Participant: Patrick Valduriez.

Supported by increasingly efficient HPC infrastructures , numerical simulations are rapidly expanding to fields such as oil and gas, medicine and meteorology. As simulations become more precise and cover longer periods of time, they may produce files with terabytes of data that need to be efficiently analyzed. In [24], we investigate techniques for managing such data using an array DBMS. We take advantage of multidimensional arrays that nicely models the dimensions and variables used in numerical simulations.We propose efficient techniques to map coordinate values in numerical simulations to evenly distributed cells in array chunks with the use of equi-depth histograms and space-filling curves. We implemented our techniques in SciDB and, through experiments over real-world data, compared them with two other approaches: row-store and column-store DBMS. The results indicate that multidimensional arrays and column-stores are much faster than a traditional row-store system for queries over a larger amount of simulation data. They also help identifying the scenarios where array DBMSs are most efficient, and those where they are outperformed by column-stores.

7.3. Scientific Workflows

7.3.1. A Scientific Workflow Infrastructure for Plant Phenomics

Participants: Didier Parigot, Patrick Valduriez.

Plant phenotyping consists in the observation of physical and biochemical traits of plant genotypes in response to environmental conditions. There are many challenges, in particular in the context of climate change and food security. High-throughput platforms have been introduced to observe the dynamic growth of a large number of plants in different environmental conditions. Instead of considering a few genotypes at a time (as it is the case when phenomic traits are measured manually), such platforms make it possible to use completely new kinds of approaches. However, the datasets produced by such widely instrumented platforms are huge, constantly augmenting and produced by increasingly complex experiments, reaching a point where distributed computation is mandatory to extract knowledge from data.

In[25], we introduce InfraPhenoGrid,an infrastructure to efficiently manage datasets produced by the PhenoArch plant phenomics platform in the context of the French Phenome Project. Our solution consists in deploying scientific workflows on a grid using a middle-ware to pilot workflow executions. Our approach is user-friendly in the sense that despite the intrinsic complexity of the infrastructure, running scientific workflows and understanding results obtained (using provenance information) is kept as simple as possible for end-users.

7.3.2. Managing Scientific Workflows in Multisite Cloud

Participants: Ji Liu, Esther Pacitti, Patrick Valduriez.

A cloud is typically made of several sites (or data centers), each with its own resources and data. Thus, it becomes important to be able to execute big scientific workflows at multiple cloud sites because of geographical distribution of data or available resources. Therefore, a major problem is how to execute a SWf in a multisite cloud, while reducing execution time and monetary cost. In [23], we propose a general solution based on multi-objective scheduling in order to execute SWfs in a multisite cloud. The solution includes a multi-objective cost model with execution time and monetary cost, a Single Site Virtual Machine (VM) Provisioning approach (SSVP) and ActGreedy, a multisite scheduling approach. We present an experimental evaluation, based on the execution of the SciEvol workflow in Microsoft Azure cloud. The results reveal that our scheduling approach significantly outperforms two adapted baseline algorithms and the scheduling time is reasonable compared with genetic and brute-force algorithms.

In [46], we present a hybrid decentralized/distributed model for handling frequently accessed metadata in a multisite cloud. We couple our model with a scientific workflow management system (SWfMS) to validate and tune its applicability to different real-life scientific scenarios. We show that efficient management of hot metadata improves the performance of SWfMS, reducing the workflow execution time up to 50% for highly parallel jobs and avoiding unnecessary cold metadata operations.

7.3.3. Online Input Data Reduction in Scientific Workflows

Participant: Patrick Valduriez.

Many scientific workflows are data-intensive and must be iteratively executed for large input sets of data elements. Reducing input data is a powerful way to reduce overall execution time in such workflows. When this is accomplished online (i.e., without requiring users to stop execution to reduce the data and resume execution), it can save much time and user interactions can integrate within workflow execution. Then, a major problem is to determine which subset of the input data should be removed. Other related problems include guaranteeing that the workflow system will maintain execution and data consistent after reduction, and keeping track of how users interacted with execution. In [48], we adopt the approach "human-in-the-loop" for scientific workflows by enabling users to steer the workflow execution and reduce input elements from datasets at runtime. We propose an adaptive monitoring approach that combines workflow provenance monitoring and computational steering to support users in analyzing the evolution of key parameters and determining which subset of the data should be removed. We also extend a provenance data model to keep track of user interactions when users reduce data at runtime. In our experimental validation, we develop a test case from the oil and gas industry, using a 936-cores cluster. The results on our parameter sweep test case show that the user interactions for online data reduction yield a 37% reduction of execution time.

7.4. Data Analytics

7.4.1. Parallel Mining of Maximally Informative k-Itemsets

Participants: Saber Salah, Reza Akbarinia, Florent Masseglia.

The discovery of informative itemsets is a fundamental building block in data analytics and information retrieval. While the problem has been widely studied, only few solutions scale. This is particularly the case when the dataset is massive, or the length K of the informative itemset to be discovered is high.

In [26], [52], we address the problem of parallel mining of maximally informative k-itemsets (miki) based on joint entropy. We propose PHIKS (Parallel Highly Informative K-itemSets) a highly scalable, parallel mining algorithm. PHIKS renders the mining process of large scale databases (up to terabytes of data) succinct and effective. Its mining process is made up of only two compact, yet efficient parallel jobs. PHIKS uses a clever heuristic approach to efficiently estimate the joint entropies of miki having different sizes with very low upper bound error rate, which dramatically reduces the runtime process. PHIKS has been extensively evaluated using massive, real-world datasets. Our experimental results confirm the effectiveness of our approach by the significant scale-up obtained with high featuresets length and hundreds of millions of objects.

7.4.2. Chiaroscuro

Participants: Tristan Allard, Florent Masseglia, Esther Pacitti.

New personal data fields are currently emerging due to the proliferation of on-body/at-home sensors connected to personal devices. However, strong privacy concerns prevent individuals to benefit from large-scale analytics that could be performed on this fine-grain highly sensitive wealth of data. In [32] we propose a demonstration of Chiaroscuro, a complete solution for clustering massively-distributed sensitive personal data while guaranteeing their privacy. The demonstration scenario highlights the affordability of the *privacy vs. quality* and *privacy vs. performance* tradeoffs by dissecting the inner working of Chiaroscuro, exposing the results obtained by the individuals participating in the clustering process, and illustrating possible uses.

7.5. Data Search

7.5.1. Spatially Localized Visual Dictionary Learning

Participants: Valentin Leveau, Alexis Joly, Patrick Valduriez.

In [44], we devise new representation learning algorithms that overcome the lack of interpretability of classical visual models. We introduce a new recursive visual patch selection technique built on top of a Shared Nearest Neighbors embedding method. The main contribution is to drastically reduce the high-dimensionality of such over-complete representation using a recursive feature elimination method. We show that the number of spatial atoms of the representation can be reduced by up to two orders of magnitude without degrading much the encoded information. The resulting representations are shown to provide competitive image classification performance with the state-of-the-art while enabling to learn highly interpretable visual models. This contribution was the last one in Valentin Leveau's PhD on Nearest Neighbor Representations [13].

7.5.2. Crowdsourcing Biodiversity Monitoring

Participants: Alexis Joly, Julien Champ, Herve Goeau, Jean-Christophe Lombardo.

Large scale biodiversity monitoring is essential for sustainable development (earth stewardship). With the recent advances in computer vision, we see the emergence of more and more effective identification tools, thus allowing large-scale data collection platforms such as the popular Pl@ntNet initiative to reuse interaction data. Although it covers only a fraction of the world flora, this platform has been used by more than 300K people who produce tens of thousands of validated plant observations each year. This explicitly shared and validated data is only the tip of the iceberg. The real potential relies on the millions of raw image queries submitted by the users of the mobile application for which there is no human validation. People make such requests to get information on a plant along a hike or something they find in their garden but do not know anything about. Allowing the exploitation of such contents in a fully automatic way could scale up the world-wide collection of implicit plant observations by several orders of magnitude, thus complementing the explicit monitoring efforts.

In [37], we first survey existing automated plant identification systems through a five-year synthesis of the PlantCLEF benchmark and an impact study of the Pl@ntNet platform. We then focus on the implicit monitoring scenario and discuss related research challenges at the frontier of computer science and biodiversity studies. Finally, we discuss the results of a preliminary study focused on implicit monitoring of invasive species in mobile search logs. We show that the results are promising while there is room for improvement before being able to automatically share implicit observations within international platforms.

7.5.3. Unsupervised Individual Whales Identification

Participants: Alexis Joly, Jean-Christophe Lombardo.

Identifying organisms is critical in accessing information related to the ecology of species. Unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly identify and record living organisms. To bridge this gap, a lot of work has been done on the development of automated species identification tools such as image-based plant identification or audio recordings-based bird identification. Yet, for some groups, it is preferable to monitor the organisms at the individual level rather than at the species level. The automation of this problem has received much less attention than species identification.

In [39], we address the specific scenario of discovering humpack whale individuals in a large collection of pictures collected by nature observers. The process is initiated from scratch, without any knowledge on the number of individuals and without any training samples of these individuals. Thus, the problem is entirely unsupervised. To address it, we set up and experimented a scalable fine-grained matching system, which allows discovering small rigid visual patterns in highly cluttered backgrounds. The evaluation was conducted in blind in the context of the LifeCLEF evaluation campaign. Results show that the proposed system provides very promising results with regard to the difficulty of the task but that there is still room for improvements to reach higher recall and precision in the future. This work was done in collaboration with the Cetamada NGO.

7.5.4. Evaluation of Biodiversity Identification and Search Techniques

Participants: Alexis Joly, Herve Goeau, Jean-Christophe Lombardo.

We ran a new edition of the LifeCLEF evaluation campaign in the context of the CLEF international research forum. We did share a new subset of the data produced by the Pl@ntNet platform and set up three new challenges: one related to the identification of plant images in open-world data streams, one related to bird sounds identification in soundscapes and one related to the visual-based identification of fish species and whales individuals. More than 150 research groups registered to at least one of the challenges and about 15 of them crossed the finish lines by running their system on the final test data. A synthesis of the results is published in the LifeCLEF 2016 overview paper [38] and more detailed analyses are provided in research reports for the plant task [35] and the bird task [36].

7.5.5. *Crowdsourcing Thousands of Specialized Labels using a Bayesian Approach*

Participants: Maximilien Servajean, Alexis Joly, Dennis Shasha, Julien Champ, Esther Pacitti.

Large-scale annotated corpora are often at the basis of huge performance gaps in machine learning based content analysis. However, the availability of such datasets has only been made possible thanks to the great amount of human labeling efforts leveraged by popular crowdsourcing and social media platforms. When the labels correspond to well known concepts, it is straightforward to train the annotators by giving a few examples with known answers. It is also straightforward to judge the quality of their labels. But neither is true with thousands of complex domain specific labels. Training on all labels is infeasible and the quality of an annotator's judgements may be vastly different for some subsets of labels than for others. This paper proposes a set of data-driven algorithms to (i) train annotators on how to disambiguate automatically labelled images, (ii) evaluate the quality of annotators' answers on new test items and (iii) weight predictions. The algorithms adapt to the skills of each annotator both in the questions asked and the weights given to their answers. The underlying judgements are Bayesian, based on adaptive priors. We measure the benefits of these algorithms by a live user experiment related to image-based plant identification involving around 1,000 people [47] (at the origin of ThePlantGame, see Software section). The proposed methods yield huge gains in annotation accuracy. While a standard user could correctly label around 2% of our data, this goes up to 80% with machine learning assisted training and almost 90% when doing a weighted combination of several annotators' labels.

8. Bilateral Contracts and Grants with Industry

8.1. Microsoft ZcloudFlow (2013-2017)

Participants: Jalexis Joly, Ji Liu, Esther Pacitti, Patrick Valduriez.

ZcloudFlow is a project in collaboration with the Kerdata team in the context of the Joint Inria–Microsoft Research Centre. It addresses the problem of advanced data storage and processing for supporting scientific workflows in the cloud. The goal is to design and implement a framework for the efficient processing of scientific workflows in clouds. The validation is performed using synthetic benchmarks and real-life applications from bioinformatics on the Microsoft Azure cloud with multiple sites.

8.2. Triton I-lab (2014-2016)

Participants: Benjamin Billet, Didier Parigot.

Triton is a common Inria lab (i-lab) between Zenith and Beepeers (http://beepeers.com/) to work on a scalable platform for developing social networks in mobile/Web environments. The main objective of this project is to design and implement a new architecture for Beepeers applications to scale up to high numbers of participants. The new platform relyes on our SON middleware and NoSQL graph databases.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. *Labex NUMEV, Montpellier*

URL: http://www.lirmm.fr/numev

We participate in the Laboratory of Excellence (labex) NUMEV (Digital and Hardware Solutions, Modelling for the Environment and Life Sciences) headed by University of Montpellier in partnership with CNRS, 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 project is decomposed in four complementary research themes: Modeling, Algorithms and computation, Scientific data (processing, integration, security), Model-Systems and measurements. Florent Masseglia co-heads the theme on scientific data.

9.1.2. Institute of Computational Biology (IBC), Montpellier

URL: http://www.ibc-montpellier.fr

IBC is a 5 year project (2012-2017) with a funding of 2Meuros by the MENRT (PIA program) to develop innovative methods and software to integrate and analyze biological data at large scale in health, agronomy and environment. Patrick Valduriez heads the workpackage on integration of biological data and knowledge.

9.2. National Initiatives

9.2.1. PIA (Projets Investissements d'Avenir

9.2.1.1. PIA Floris'Tic (2015-2018), 430Keuro.

Participants: Julien Champ, Alexis Joly.

Floris'tic aims at promoting the scientific and technical culture of plant sciences through innovative pedagogic methods, including participatory initiatives and the use of IT tools such as the one built within the Pl@ntNet project. A. Joly heads the work package on the development of the IT tools. This is a joint project with the AMAP laboratory, the TelaBotanica social network and the Agropolis foundation.

9.2.2. Others

9.2.2.1. CIFRE INA/Inria (2013-2016), 100Keuros

Participants: Alexis Joly, Valentin Leveau, Patrick Valduriez.

This contract with INA allows funding a 3-years PhD (Valentin Leveau). This PhD addresses research challenges related to large-scale supervised content-based retrieval in distributed environments.

9.2.2.2. INRA/Inria PhD program, 100Keuros

Participant: Alexis Joly.

This contract between INRA and Inria allows funding a 3-years PhD student (Christophe Botella). The addressed challenge is the large-scale analysis of Pl@ntNet data with the objective to model species distribution (a big data approach to species distribution modeling). The PhD student is supervised by Alexis Joly with François Munoz (ecologist, IRD) and Pascal Monestiez (statistician, INRA).

9.3. European Initiatives

9.3.1. FP7 Projects

9.3.1.1. CoherentPaaS

Participants: Carlyna Bondiombouy, Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez.

Project title: A Coherent and Rich Platform as a Service with a Common Programming Model
Instrument: Integrated Project
Duration: 2013 - 2016
Total funding: 5 Meuros (Zenith: 500Keuros)
Coordinator: U. Madrid, Spain
Partner: FORTH (Greece), ICCS (Greece), INESC (Portugal) and the companies MonetDB (Netherlands), QuartetFS (France), Sparsity (Spain), Neurocom (Greece), Portugal Telecom (Portugal).
Inria contact: Patrick Valduriez

CoherentPaaS has been developing a PaaS that incorporates a rich and diverse set of cloud data management technologies, including NoSQL data stores, such as key-value data stores and graph databases, SQL data stores, such as in-memory and column-oriented databases, hybrid systems, such as SQL engines on top on key-value data stores, and complex event processing data management systems. It uses a common query language to unify the programming models of all systems under a single paradigm and provides holistic coherence across data stores using a scalable, transactional management system. CoherentPaaS will dramatically reduce the effort required to build and the quality of the resulting cloud applications using multiple cloud data management technologies via a single query language, a uniform programming model, and ACID-based global transactional semantics. CoherentPaaS will design and build a working prototype and will validate the proposed technology with real-life use cases. In this project, Zenith is in charge of designing the CloudMdsQL language and implementing its compiler/optimizer and query engine.

9.3.1.2. HPC4E

Participants: Reza Akbarinia, Florent Masseglia, Esther Pacitti, Patrick Valduriez.

Project title: High Performance Computing for Energy
Instrument: H2020
Duration: 2015 - 2017
Total funding: 2 Meuros
Coordinator: Barcelona Supercomputing Center (BSC), Spain
Partner: Europe: Inria, Lancaster University, Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, Repsol S.A., Iberdrola Renovables Energía S.A., Total S.A. Brazil: COPPE/Universidade Federal de Rio de Janeiro, LNCC, Instituto Tecnológico de Aeronáutica (ITA), Universidade Federal do Rio Grande do Sul, Universidade Federal de Pernambuco, Petrobras.
Inria contact: Patrick Valduriez

The main objective is to develop high performance simulation tools that can help the energy industry to respond future energy demands and also to carbon-related environmental issues using HPC systems. The project also aims at improving the usage of energy using HPC tools by acting at many levels of the energy chain for different energy sources. Another objective is to improve the cooperation between energy industries from EU and Brazil. The project includes relevant energy industrial partners from Brazil (Petrobras) and EU (Repsol and Total as O&G industries), which benefit from the project's results. A last objective is to improve the cooperation between the leading research centres in EU and Brazil in HPC applied to energy. This includes sharing supercomputing infrastructures between Brazil and EU. In this project, Zenith is working on Big Data management and analysis of numerical simulations.

9.3.1.3. CloudDBAppliance

Participants: Reza Akbarinia, Boyan Kolev, Florent Masseglia, Esther Pacitti, Patrick Valduriez.

Project title: CloudDBAppliance
Instrument: H2020
Duration: 2016 - 2019
Total funding: 5 Meuros (Zenith: 500Keuros)
Coordinator: Bull/Atos, France
Partner: Europe: Inria Zenith, U. Madrid, INESC and the companies LeanXcale, QuartetFS, Nordea, BTO, H3G, IKEA, CloudBiz, and Singular Logic.
Inria contact: Florent Masseglia, Patrick Valduriez

The project aims at producing a European Cloud Database Appliance for providing a Database as a Service able to match the predictable performance, robustness and trustworthiness of on premise architectures such as those based on mainframes. The cloud database appliance features: (1) a scalable operational database able to process high update workloads such as the ones processed by banks or telcos, combined with a fast analytical engine able to answer analytical queries in an online manner; (2) an operational Hadoop data lake that integrates an operational database with Hadoop, so operational data is stored in Hadoop that will cover the needs from companies on big data; (3) a cloud hardware appliance leveraging the next generation of hardware to be produced by Bull, the main European hardware provider. This hardware is a scale-up hardware similar to the one of mainframes but with a more modern architecture. Both the operational database and the in-memory analytics engine will be optimized to fully exploit this hardware and deliver predictable performance. Additionally, CloudDBAppliance will tolerate catastrophic cloud data centres failures (e.g. a fire or natural disaster) providing data redundancy across cloud data centres. In this project, Zenith is in charge of designing and implementing the components for analytics and parallel query processing.

9.4. International Initiatives

9.4.1. MUSIC

Title: MUltiSite Cloud (MUSIC) data management
Inria principal investigator: Esther Pacitti
International Partner):

Laboratorio Nacional de Computaçao Cientifica, Petropolis (Brazil) - Fabio Porto

Universidade Federal do Rio de Janeiro (Brazil) - Alvaro Coutinho and Marta Mattoso

Universidade Federal Fluminense, Niteroi (Brazil) - Daniel Oliveira

Centro Federal de Educa cao Tecnologica, Rio de Janeiro (Brazil) - Eduardo Ogasawara

Duration: 2014 - 2016
See also: https://team.inria.fr/zenith/projects/international-projects/music/

By centralizing all data in a large-scale data center, the cloud significantly simplifies the task of system administration. But for scientific data, where different organizations may have their own data centers, a distributed (multisite) cloud model where each site is visible from outside, is needed. The main objective of this research and scientific collaboration is to develop a multisite cloud architecture for managing and analyzing scientific data, including support for heterogeneous data; distributed scientific workflows, and complex big data analysis. The resulting architecture will enable scalable data management infrastructures that can be used to host a variety of scientific applications that benefit from computing, storage, and networking resources that span multiple data centers.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

We have regular scientific relationships with research laboratories in

- North America: Univ. of Waterloo (Tamer Özsu), UCSB Santa Barbara (Divy Agrawal and Amr El Abbadi)
- Asia: National Univ. of Singapore (Beng Chin Ooi, Stéphane Bressan), Wonkwang University, Korea (Kwangjin Park)
- Europe: Univ. of Madrid (Ricardo Jiménez-Periz), UPC Barcelona (Josep Lluis Larriba Pey), HES-SO (Henning Müller), University of Catania (Concetto Spampinatto), The Open University (Stefan Rüger)
- North Africa: Univ. of Tunis (Sadok Ben-Yahia)
- Australia: Australian National University (Peter Christen)
- Central America: Technologico de Costa-Rica (Erick Mata, former director of the US initiative Encyclopedia of Life)

9.4.3. Participation In other International Programs

We are involved in LifeCLEF lab, a self-organized research platform whose main mission is to promote research, innovation, and development of computer-assisted identification of living organisms. It was initiated by Alexis Joly in 2014 in collaboration with several European colleagues: Henning Müller (CH), Robert B Fisher (UK), Andreas Rauber (AU), Concetto Spampinato (IT), Hervé Glotin (FR). Each year, LifeCLEF releases large-scale experimental data covering tens of thousands of species (plants images, birds audio recordings and fish sub-marine videos). About 100-150 research groups register each year to get access to it and tens of them submit reports describing their conducted research (published in CEUR-WS proceedings). Results are then synthesized and further analyzed in joint research papers.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Marta Mattoso (UFRJ, Brazil) gave a seminar on "“Exploratory Analysis of Raw Data Files through Dataflows” in march.

Jose Mario Carranza Rojas (PhD student, Technologico de Costa-Rica) spent two days per week in the team in the context of a 4 months internship at the Montpellier research lab AMAP in the context of the Floris’Tic project).

10. Dissemination

10.1. Scientific Animation

Editorial board of scientific journals:

- VLDB Journal: P. Valduriez.
- Journal of Transactions on Large Scale Data and Knowledge Centered Systems: R. Akbarinia and E. Pacitti are guest editors of a special issue on data management in internet of things (IoT).
- Distributed and Parallel Databases, Kluwer Academic Publishers: E. Pacitti, 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): P. Valduriez.
- Multimedia Tools and Applications: A. Joly.

Organization of conferences and workshops:

- Alexis Joly was the chair of the LifeCLEF 2016 international workshop [1] dedicated to multimedia biodiversity data management, Evora, sept. 2016
- Alexis Joly was in the organizing committee of the Floris’tic national workshop held in Montpellier, nov. 2016 (http://floristic.org/8-nov-2016)

[1] http://www.imageclef.org/lifeclef/2016

Conference program committees :

- ACM SIGMOD Conf. 2016: R. Akbarinia, 2017: F. Masseglia
- IEEE Int. Conf. on Data Engineering (ICDE) 2016: R. Akbarinia, E. Pacitti, P. Valduriez (area chair)
- VLDB Joint Workshop on Big Data Open-Source Systems (BOSS) / Polyglot: P. Valduriez (co-chair)
- DataDiversityConvergence workshop, 6th International Conference on Cloud Computing and Services Science (CLOSER 2016): P. Valduriez (co-chair)
- Int. Conf. on Extending DataBase Technologies (EDBT), 2017: E. Pacitti
- 2nd Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems (BigD2M), 2016: E. Pacitti
- International Conference and Labs of the Evaluation Forum (CLEF), 2016: A. Joly
- ACM International Conference on Multimedia Retrieval (ICMR), 2016: A. Joly
- IEEE International Conference on Image Processing (ICIP), 2016: A. Joly
- ACM Multimedia conference (ACMMM), 2016: A. Joly
- Int. work.Multimedia Analysis and Retrieval for Multimodal Interactions (MARMI): A. Joly
- Int. Conf. on Scientific and Statistical Database Management (SSDBM), 2016: A. Joly
- European Conf. on Computer Vision (ECCV), 2016: A. Joly
- VLDB 2017: F. Masseglia
- IEEE Int. Conf. on Data Mining, 2016: F. Masseglia
- ACM Symposium on Applied Computing 2017: F. Masseglia
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD) 2017: F. Masseglia
- IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA), 2016: F. Masseglia
- Int. Symposium on Information Management and Big Data (SimBig), 2016: F. Masseglia
- Int. Conf. on Data Science, Technology and Applications (DATA), 2016: F. Masseglia

Reviewing in international journals :

- Distributed and Parallel Databases: R. Akbarinia
- ACM Transactions on Database Systems (TODS): A. Joly
- IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI): A. Joly
- IEEE Transactions on Knowledge and Data Engineering: R. Akbarinia
- Information Sciences: A. Joly
- Ecological Informatics: A. Joly
- Multimedia Tools and Applications Journal (MTAP): A. Joly
- Multimedia Systems: A. Joly
- Transactions on Information Forensics & Security: A. Joly
- International Journal of Computer Vision: A. Joly
- Transactions on Image Processing: A. Joly
- ACM Trans. on Database Systems: E. Pacitti
- Knowledge and Information Systems (KAIS): F. Masseglia
- IEEE Transactions on Knowledge and Data Engineering (TKDE): F. Masseglia

Other activities (national):

- Alexis Joly gave invited talks in national events: Inria scientific days (https://journees-scientifiques2016.inria.fr/), INRA-Inria days (https://journees.inra.fr/maths-info2016), Award ceremony of « La Recherche 2016 » (http://www.leprixlarecherche.com/).
- P. Valduriez is the scientific manager for the Latin America zone at Inria Direction des Relations Internationales (DRI), Member of the Scientific Committee at Agence Nationale de la Recherche (ANR) - Défi 7 Information and communication society and Member of the Scientific Committee of the BDA conference.
- F. Masseglia gave an invited talk at the CNRS national seminar on IST about "Publication Data Analytics" (Meudon, 10 November), and an invited talk to the DSI service of IRD on "Scientific Data Mining" (Montpellier, 18 September). Florent also participates in the Class'Code PIA project dedicated to teaching computational thinking for professionals of education (head of the working group on the definition of fundamental notions).
- P. Valduriez gave invited talks at: “La Science des données à l'IRIT”, Toulouse, april 2016, the Junior Conference on Data Science and Engineering, Paris Saclay, sept. 2016, the Orange Summer University “Innovate in IT”, sept. 2016.

Other activities (international):

- Alexis Joly was a member of the scientific advisory board of the EU REVEAL project (http://revealproject.eu). He was invited by the ATC innovation lab in Athens to participate in a workshop on social media verification (http://revealproject.eu/reveal-workshop-on-16-sept-2016).
- E. Pacitti gave an invited talk at the 6th workshop on big data and analytics (WOS6), co-organized by Technicolor and Inria (http://www.bretagne-networking.org/wos6) on Experiences on Data Management Techniques for Scientific Application in Rennes, nov. 2016, and a talk at Fundação Getulio Vargas on Experiences on Data Management Techniques for Big Data in Rio de Janeiro, dec. 2016.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Most permanent members of Zenith teach at the Licence and Master degree levels at UM2.

Reza Akbarinia:

Master Research: New approaches for data storage, 9h, level M2, Faculty of Science, UM

Florent Masseglia:

Science Popularization: 4 Ph.D students, from 3 different doctoral schools are having a 30h doctoral module under Florent Masseglia's supervision.

Esther Pacitti:

IG3: Database design, physical organization, 54h, level L3, Polytech'Montpellier, UM2

IG4: Networks, 42h, level M1, Polytech' Montpellier, UM2

IG4: Object-relational databases, 32h, level M1, Polytech' Montpellier, UM2

IG5: Distributed systems, virtualization, 27h, level M2, Polytech' Montpellier, UM2

Industry internship committee, 50h, level M2, Polytech' Montpellier

Patrick Valduriez:

Professional: Distributed Information Systems, Big Data Architectures, 75h, level M2, Capgemini Institut

Professional: XML, 10h, level M2, Orsys Formation

Alexis Joly:

Master Research: Large-scale Content-based Visual Information Retrieval, 3h, level M2, Faculty of Science, UM

10.2.2. Supervision

- PhD: Ji Liu, Multisite Management of Scientific Worflows in the Cloud, Univ. Montpellier, 3 Nov 2016, Advisors: Esther Pacitti and Patrick Valduriez
- PhD: Valentin Leveau, Spatially Consistent Nearest Neighbor Representations for Fine-Grained Classification, Univ. Montpellier, 9 Nov 2016, Advisor: Patrick Valduriez, co-advisor: Alexis Joly and Olivier Buisson
- PhD : Saber Salah, Optimizing a Cloud for Data Mining Primitives, Univ. Montpellier, 20 Apr 2016 Advisor: Florent Masseglia, co-advisor: Reza Akbarinia
- PhD in progress: Christophe Botella, Lareg-scale Species Distribution Modelling based on crowd-srouced image streams, started Oct 2016, Univ. Montpellier, Alexis Joly, François Munoz (IRD), Pascal Monestiez (INRA)
- PhD in progress: Titouan Lorieul, Pro-active crowdsourcing, started Oct 2016, Univ. Montpellier, Advisor: Alexis Joly
- PhD in progress: Mehdi Zitouni Closed Pattern Mining in a Massively Distributed Environment started sept. 2014, Univ. Tunis, Advisor: Florent Masseglia, co-advisor: Reza Akbarinia
- PhD in progress : Djamel-Edine Yagoubi, Indexing Time Series in a Massively Distributed Environment, started oct. 2014, Univ. Montpellier, Advisors: Florent Masseglia and Patrick Valduriez, co-advisor: Reza Akbarinia
- PhD in progress : Sakina Mahboubi, Privacy Preserving Query Processing in Clouds, started oct. 2015, Univ. Montpellier, Advisor: Patrick Valduriez, co-advisor: Reza Akbarinia
- PhD in progress: Khadidja Meguelati, Massively Distributed Clustering, started Oct 2016, Univ. Montpellier, Advisor: Florent Masseglia, co-advisor : Nadine Hilgert (INRA)

10.2.3. Juries

Members of the team participated to the following PhD committees:

- A. Joly: Zongyuan Ge (Queensland University of Technology), Amel Tuama Alhussainy (Univ. of Montpellier)
- E. Pacitti: Saliha Lallali (UVSQ), Nupur Mittal (Univ. Rennes 1)
- P. Valduriez: Damien Graux (Univ. Grenoble)
- F. Masseglia: Hadi Hashem (Telecom SudParis), Manuel Pozo (CNAM), Martin Kirchgessner (Univ. Grenoble)

10.3. Popularization

F. Masseglia is now “Chargé de mission pour la médiation scientifique Inria” and heads Inria’s national network of colleagues involved in science popularization.

Zenith has major contributions to science popularization, as follows.

10.3.1. Code Teaching for Kids

Teaching code is now officially in the school programs in France. Class’Code is a PIA project that aims at training the needed 300,000 teachers and professionals of education France. The project is a hybrid MOOC (both online courses and physical meetings).

Along with Class'Code, the association “La main à la pâte” has coordinated the writing of a school book on the teaching of computer science teaching, with Inria (Gilles Dowek, Pierre-Yves Oudeyer, Florent Masseglia and Didier Roy), France-IOI and the University of Lorraine. The book has been requested by and distributed to 15,000 readers in less than one month.

F. Masseglia is giving a doctoral training at different doctoral schools in Montpellier, in order to train facilitators for helping teachers and people of the education world to better understand the "computational thinking". So far, 12 people have been trained. He is also a member of the management board of "Les Petits Débrouillards" in Languedoc-Roussillon and the scientific responsible for school visits in the LIRMM laboratory.

10.3.2. Science Outreach

In the context of the Floris'tic project, A. Joly participates regularly to the set up of popularization, educational and citizen science actions in France (with schools, cities, parks, etc.). The softwares developed within the project (Pl@ntNet, Smart'Flore and ThePlantGame) are used in a growing number of formal educational programs and informal educational actions of individual teachers. For instance, Smart'Flore is used by the French National Education in a program for reducing early school leaving. Pl@ntNet app is used in the Reunion island in an educational action called Vegetal riddle organized by the Center for cooperation at school. It is also planned to be used in a large-scale program in the Czech republic that is being finalized (in 40 classrooms). An impact study of the Pl@ntNet application did show that 6% of the respondents use it for educational purposes in the context of their professional activity. The Inria movie “Pl@ntNet, the application that helps people identify plants” enjoyed about 350 thousand views on Youtube.

10.3.3. Events

Zenith participated to the following events:

- F. Masseglia co-organized and co-animated the Inria's stand at “La fête de la science”, Montpellier, held by Genopolys (a science village).
- F. Masseglia is member of the project selection committee for “La fête de la science” in Montpellier.
- M. Servajean and A. Joly animated a stand at “La fête de la science”, Montpellier, held by the LIRMM laboratory.
- A. Joly and J. Champ participated to the set-up of a Pl@ntNet demo within the French pavillon at the Universal Exposition hold in Milan (about 2M visitors on the French pavillon).
- As a member of the organizing committee of the Floris'tic project, A. Joly participated to several popularization and educational actions in collaboration with Tela Botanica NGO (cities, parks, schools, etc.).
- P. Valduriez published an article in Interstices on “the data in question” [58].

11. Bibliography

Major publications by the team in recent years

[1] T. Allard, G. Hébrail, F. Masseglia, E. Pacitti. *Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering*, in "34th International ACM Conference on Management of Data (ACM SIGMOD)", Melbourne, Australia, ACM SIGMOD, May 2015 [*DOI :* 10.1145/2723372.2749453], https://hal.inria.fr/hal-01136686

[2] A. Joly, P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, A. Affouard, J. Carré, J.-F. Molino, N. Boujemaa, D. Barthélémy. *A look inside the Pl@ntNet experience*, in "Multimedia Systems", 2015, 16 p. [*DOI :* 10.1007/s00530-015-0462-9], https://hal.inria.fr/hal-01182775

[3] A. JOLY, H. GOEAU, P. BONNET, V. BAKIC, J. BARBE, S. SELMI, I. YAHIAOUI, J. CARRÉ, E. MOUYSSET, J.-F. MOLINO, N. BOUJEMAA, D. BARTHÉLÉMY. *Interactive plant identification based on social image data*, in "Ecological Informatics", 2013 [*DOI :* 10.1016/J.ECOINF.2013.07.006], http://www.sciencedirect.com/science/article/pii/S157495411300071X

[4] B. KOLEV, P. VALDURIEZ, C. BONDIOMBOUY, R. JIMENEZ-PERIS, R. PAU, J. O. PEREIRA. *CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language*, in "Distributed and Parallel Databases", December 2016, vol. 34, n^{o} 4, pp. 463-503 [*DOI :* 10.1007/S10619-015-7185-Y], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01184016

[5] M. LIROZ-GISTAU, R. AKBARINIA, D. AGRAWAL, P. VALDURIEZ. *FP-Hadoop: Efficient Processing of Skewed MapReduce Jobs*, in "Information Systems", 2016, vol. 60, pp. 69-84 [*DOI :* 10.1016/J.IS.2016.03.008], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01377715

[6] J. LIU, E. PACITTI, P. VALDURIEZ, D. DE OLIVEIRA, M. MATTOSO. *Multi-Objective Scheduling of Scientific Workflows in Multisite Clouds*, in "Future Generation Computer Systems", 2016, vol. 63, pp. 76–95 [*DOI :* 10.1016/J.FUTURE.2016.04.014], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01342203

[7] H. LUSTOSA, F. PORTO, P. BLANCO, P. VALDURIEZ. *Database System Support of Simulation Data*, in "Proceedings of the VLDB Endowment (PVLDB)", September 2016, vol. 9, n^{o} 13, pp. 1329-1340, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01363738

[8] E. PACITTI, R. AKBARINIA, M. EL DICK. *P2P Techniques for Decentralized Applications*, Morgan & Claypool Publishers, 2012, 104 p. , http://hal.inria.fr/lirmm-00748635

[9] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Fast Parallel Mining of Maximally Informative k-Itemsets in Big Data*, in "IEEE International Conference on Data Mining (ICDM)", Atlantic city, United States, August 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01187275

[10] M. SERVAJEAN, R. AKBARINIA, E. PACITTI, S. AMER-YAHIA. *Profile Diversity for Query Processing using User Recommendations*, in "Information Systems", March 2015, vol. 48, pp. 44-63 [*DOI :* 10.1016/J.IS.2014.09.001], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01079523

[11] M. SERVAJEAN, A. JOLY, D. SHASHA, J. CHAMP, E. PACITTI. *ThePlantGame: Actively Training Human Annotators for Domain-specific Crowdsourcing*, in "ACM Multimedia 2016", Amsterdam, Netherlands, October 2016, https://hal.inria.fr/hal-01373769

[12] T. M. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, third edition*, Springer, 2011, 845 p. , http://hal.inria.fr/hal-00640392/en

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] V. LEVEAU. *Spatially Consistent Nearest Neighbor Representations for Fine-Grained Classification*, Université Montpellier, December 2016, https://hal.archives-ouvertes.fr/tel-01410137

[14] J. LIU. *Multisite Management of Scientific Workflows in the Cloud*, Université de Montpellier, November 2016, https://tel.archives-ouvertes.fr/tel-01400625

[15] S. SALAH. *Parallel Itemset Mining in Massively Distributed Environments*, Université de Montpellier, April 2016, https://hal-lirmm.ccsd.cnrs.fr/tel-01415587

Articles in International Peer-Reviewed Journals

[16] M. R. BOUADJENEK, H. HACID, M. BOUZEGHOUB. *Social Networks and Information Retrieval, How Are They Converging? A Survey, a Taxonomy and an Analysis of Social Information Retrieval Approaches and Platforms*, in "Information Systems", March 2016, vol. 56, pp. 1-18 [*DOI :* 10.1016/J.IS.2015.07.008], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01174843

[17] C. BONDIOMBOUY, B. KOLEV, O. LEVCHENKO, P. VALDURIEZ. *Multistore Big Data Integration with CloudMdsQL*, in "Transactions on Large-Scale Data- and Knowledge-Centered Systems", 2016, forthcoming, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01345712

[18] C. BONDIOMBOUY, P. VALDURIEZ. *Query processing in multistore systems: an overview*, in "International Journal of Cloud Computing", 2016, 38 p. , To appear, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01341158

[19] P. BONNET, A. JOLY, H. GOËAU, J. CHAMP, C. VIGNAU, J.-F. MOLINO, D. BARTHÉLÉMY, N. BOUJEMAA. *Plant identification: Man vs. Machine. LifeCLEF 2014 plant identification challenge*, in "Multimedia Tools and Applications", 2016, vol. 75, n^{o} 3, pp. 1647-1665 [*DOI :* 10.1007/S11042-015-2607-4], https://hal-sde.archives-ouvertes.fr/hal-01289798

[20] J. CHAMP, T. LORIEUL, P. BONNET, N. MAGHNAOUI, C. SERENO, T. DESSUP, J.-M. BOURSIQUOT, L. AUDEGUIN, T. LACOMBE, A. JOLY. *Categorizing plant images at the variety level: Did you say fine-grained?*, in "Pattern Recognition Letters", June 2016, forthcoming [*DOI :* 10.1016/J.PATREC.2016.05.022], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01348914

[21] B. KOLEV, P. VALDURIEZ, C. BONDIOMBOUY, R. JIMENEZ-PERIS, R. PAU, J. O. PEREIRA. *CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language*, in "Distributed and Parallel Databases", December 2016, vol. 34, n^{o} 4, pp. 463-503 [*DOI :* 10.1007/S10619-015-7185-Y], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01184016

[22] M. LIROZ-GISTAU, R. AKBARINIA, D. AGRAWAL, P. VALDURIEZ. *FP-Hadoop: Efficient Processing of Skewed MapReduce Jobs*, in "Information Systems", 2016, vol. 60, pp. 69-84 [*DOI :* 10.1016/J.IS.2016.03.008], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01377715

[23] J. LIU, E. PACITTI, P. VALDURIEZ, D. DE OLIVEIRA, M. MATTOSO. *Multi-Objective Scheduling of Scientific Workflows in Multisite Clouds*, in "Future Generation Computer Systems", 2016, vol. 63, pp. 76–95 [*DOI :* 10.1016/J.FUTURE.2016.04.014], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01342203

[24] H. LUSTOSA, F. PORTO, P. BLANCO, P. VALDURIEZ. *Database System Support of Simulation Data*, in "Proceedings of the VLDB Endowment (PVLDB)", September 2016, vol. 9, n^{o} 13, pp. 1329-1340, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01363738

[25] C. PRADAL, S. ARTZET, J. CHOPARD, D. DUPUIS, C. FOURNIER, M. MIELEWCZIK, V. NEGRE, P. NEVEU, D. PARIGOT, P. VALDURIEZ, S. COHEN-BOULAKIA. *InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid*, in "Future Generation Computer Systems", June 2016 [*DOI :* 10.1016/J.FUTURE.2016.06.002], https://hal.inria.fr/hal-01336655

[26] S. Salah, R. Akbarinia, F. Masseglia. *A Highly Scalable Parallel Algorithm for Maximally Informative k-Itemset Mining*, in "Knowledge and Information Systems (KAIS)", April 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01288571

[27] G. Sempéré, F. Philippe, A. DEREEPER, M. Ruiz, G. SARAH, P. Larmande. *Gigwa—Genotype investigator for genome- wide analyses*, in "GigaScience", May 2016 [*DOI :* 10.1186/s13742-016-0131-8], https://hal.archives-ouvertes.fr/hal-01411506

[28] V. Silva Souza, O. Daniel De, P. Valduriez, M. Mattoso. *Analyzing Related Raw Data Files through Dataflows*, in "Concurrency and Computation: Practice and Experience", 2016, vol. 28, n^{o} 8, pp. 2528-2545 [*DOI :* 10.1002/cpe.3616], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01181231

[29] J. Starlinger, S. Cohen-Boulakia, S. Khanna, S. Davidson, U. Leser. *Effective and Efficient Similarity Search in Scientific Workflow Repositories*, in "Future Generation Computer Systems", 2016, vol. 56, pp. 584-594 [*DOI :* 10.1016/j.future.2015.06.012], https://hal.archives-ouvertes.fr/hal-01170597

[30] G. Tagny Ngompe, A. Venkatesan, N. El Hassouni, M. Ruiz, P. Larmande. *AgroLD API : Une architecture orientée services pour l'extraction de connaissances dans la base de données liées AgroLD*, in "Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information", November 2016 [*DOI :* 10.3166/ISI.28.2-3.1-25], https://hal.archives-ouvertes.fr/hal-01411532

[31] S. Vrochidis, K. D. Karatzas, A. Karppinen, A. Joly. *Guest Editorial: Environmental Multimedia Retrieval*, in "Multimedia Tools and Applications", 2016, vol. 75, n^{o} 3, pp. 1557–1562 [*DOI :* 10.1007/s11042-016-3256-y], https://hal.archives-ouvertes.fr/hal-01373782

International Conferences with Proceedings

[32] T. Allard, G. Hébrail, F. Masseglia, E. Pacitti. *A New Privacy-Preserving Solution for Clustering Massively Distributed Personal Times-Series*, in "ICDE: International Conference on Data Engineering", Helsinki, Finland, May 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01270268

[33] C. Bondiombouy, B. Kolev, P. Valduriez, O. Levchenko. *Extending CloudMdsQL with MFR for Big Data Integration* *, in "BDA: Bases de Données Avancées", Poitiers, France, LIAS / ISAE-ENSMA, Poitiers, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01409104

[34] J. Champ, H. Goëau, A. Joly. *Floristic participation at LifeCLEF 2016 Plant Identification Task*, in "CLEF 2016 - Conference and Labs of the Evaluation forum", Évora, Portugal, September 2016, pp. 450–458, https://hal.archives-ouvertes.fr/hal-01373776

[35] H. Goëau, P. Bonnet, A. Joly. *Plant Identification in an Open-world (LifeCLEF 2016)*, in "CLEF 2016 - Conference and Labs of the Evaluation forum", Évora, Portugal, September 2016, pp. 428–439, https://hal.archives-ouvertes.fr/hal-01373780

[36] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, A. Joly. *LifeCLEF Bird Identification Task 2016: The arrival of Deep learning*, in "Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum", Evora, Portugal, September 2016, pp. 440–449, https://hal.archives-ouvertes.fr/hal-01373779

[37] A. JOLY, H. GOËAU, J. CHAMP, S. DUFOUR-KOWALSKI, H. MÜLLER, P. BONNET. *Crowdsourcing Biodiversity Monitoring: How Sharing your Photo Stream can Sustain our Planet*, in "ACM Multimedia 2016", Amsterdam, Netherlands, October 2016, https://hal.inria.fr/hal-01373762

[38] A. JOLY, H. GOËAU, H. GLOTIN, C. SPAMPINATO, P. BONNET, W.-P. VELLINGA, J. CHAMP, R. PLANQUÉ, S. PALAZZO, H. MÜLLER. *LifeCLEF 2016: Multimedia Life Species Identification Challenges*, in "CLEF 2016 - 7th International Conference of the CLEF Association", Evora, Portugal, N. FUHR, P. QUARESMA, T. GONÇALVES, B. LARSEN, K. BALOG, C. MACDONALD, L. CAPPELLATO, N. FERRO (editors), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, September 2016, pp. 286–310 [*DOI :* 10.1007/978-3-319-44564-9_26], https://hal.archives-ouvertes.fr/hal-01373781

[39] A. JOLY, J.-C. LOMBARDO, J. CHAMP, A. SALOMA. *Unsupervised Individual Whales Identification: Spot the Difference in the Ocean*, in "Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum", Evora, Portugal, September 2016, pp. 469–480, https://hal.archives-ouvertes.fr/hal-01373777

[40] C. JONQUET, A. TOULET, E. ARNAUD, S. AUBIN, E. D. YEUMO, V. EMONET, V. PESCE, P. LARMANDE. *AgroPortal: an open repository of ontologies and vocabularies for agriculture and nutrition data*, in "GODAN Summit", New York, NY, United States, B. SCHAAP (editor), September 2016, https://hal.archives-ouvertes.fr/hal-01398252

[41] B. KOLEV, C. BONDIOMBOUY, O. LEVCHENKO, P. VALDURIEZ, R. JIMENEZ-PÉRIS, R. PAU, J. PEREIRA. *Design and Implementation of the CloudMdsQL Multistore System*, in "CLOSER: Cloud Computing and Services Science", Roma, Italy, DataDiversityConvergence Workshop, April 2016, vol. 1, pp. 352-359 [*DOI :* 10.5220/0005923803520359], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01341172

[42] B. KOLEV, C. BONDIOMBOUY, P. VALDURIEZ, R. JIMÉNEZ-PERIS, R. PAU, J. PEREIRA. *The CloudMdsQL Multistore System*, in "SIGMOD", San Francisco, United States, June 2016, Verifier le Doi [*DOI :* 10.1145/2882903.2899400], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01288025

[43] B. KOLEV, R. PAU, O. LEVCHENKO, P. VALDURIEZ, R. JIMÉNEZ-PERIS, J. PEREIRA. *Benchmarking Polystores: the CloudMdsQL Experience*, in "IEEE BigData 2016: Workshop on Methods to Manage Heterogeneous Big Data and Polystore Databases", Washington D.C., United States, V. GADEPALLY (editor), IEEE Computing Society, December 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01415582

[44] V. LEVEAU, A. JOLY, O. BUISSON, P. VALDURIEZ. *Spatially Localized Visual Dictionary Learning*, in "ICMR '16 Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval", New York, United States, ACM, June 2016, pp. 367–370 [*DOI :* 10.1145/2911996.2912070], https://hal.archives-ouvertes.fr/hal-01373778

[45] J. LIU, E. PACITTI, P. VALDURIEZ, M. MATTOSO. *Scientific Workflow Scheduling with Provenance Support in Multisite Cloud*, in "VECPAR", Porto, Portugal, Faculty of Engineering of the University of Porto, Portugal, June 2016, 8 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01342190

[46] L. PINEDA-MORALES, J. LIU, A. COSTAN, E. PACITTI, G. ANTONIU, P. VALDURIEZ, M. MATTOSO. *Managing Hot Metadata for Scientific Workflows on Multisite Clouds*, in "BIGDATA 2016 - 2016 IEEE International Conference on Big Data", Washington, United States, December 2016, https://hal.inria.fr/hal-01395715

[47] M. SERVAJEAN, A. JOLY, D. SHASHA, J. CHAMP, E. PACITTI. *ThePlantGame: Actively Training Human Annotators for Domain-specific Crowdsourcing*, in "ACM Multimedia 2016", Amsterdam, Netherlands, October 2016, https://hal.inria.fr/hal-01373769

[48] R. SOUZA, V. SILVA, A. L. G. A. COUTINHO, P. VALDURIEZ, M. MATTOSO. *Online Input Data Reduction in Scientific Workflows*, in "WORKS: Workflows in Support of Large-scale Science", Salt Lake City, United States, ACM SIGHPC and IEEE, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01400538

National Conferences with Proceedings

[49] B. KOLEV, C. BONDIOMBOUY, P. VALDURIEZ, R. JIMÉNEZ-PERIS, R. SPAIN, J. PEREIRA. *Demonstration of the CloudMdsQL Multistore System*, in "BDA: Bases de Données Avancées", Poitiers, France, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01408621

[50] J. LIU, E. PACITTI, P. VALDURIEZ, D. DE OLIVEIRA, M. MATTOSO. *Scientific Workflow Execution with Multiple Objectives in Multisite Clouds*, in "BDA: Bases de Données Avancées", Poitiers, France, Principes, Technologies et Applications, LIAS / ISAE-ENSMA, Poitiers, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01409092

[51] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Privacy Preserving Query Processing in the Cloud*, in "BDA: Bases de Données Avancées", Poitiers, France, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01410395

[52] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Mining Maximally Informative k-Itemsets in Massively Distributed Environments*, in "BDA: Bases de Données Avancées", Poitiers, France, November 2016, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01411190

Conferences without Proceedings

[53] L. LE NGOC, A. TIREAU, A. VENKATESAN, P. NEVEU, P. LARMANDE. *Development of a knowledge system for Big Data: Case study to plant phenotyping data*, in "WIMS '16 Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics", Nimes, France, ACM, June 2016 [*DOI :* 10.1145/2912845.2912869], https://hal.archives-ouvertes.fr/hal-01411565

[54] A. VENKATESAN, N. EL HASSOUNI, F. PHILLIPE, C. POMMIER, H. QUESNEVILLE, M. RUIZ, P. LARMANDE. *Exposing French agronomic resources as Linked Open Data*, in "Ingenierie des Connaissances IC2016 - Workshop In Ovive", Montpellier, France, June 2016, https://hal.archives-ouvertes.fr/hal-01411759

[55] S. ZEVIO, N. EL HASSOUNI, M. RUIZ, P. LARMANDE. *AgroLD indexing tools with ontological annotations*, in "Semantic Web for Life Science - SWA4LS", Amsterdam, France, December 2016, https://hal.archives-ouvertes.fr/hal-01411713

Research Reports

[56] C. BONDIOMBOUY, P. VALDURIEZ. *Query Processing in Multistore Systems: an overview*, Inria Sophia Antipolis - Méditerranée, March 2016, n^{o} RR-8890, 38 p. , https://hal.inria.fr/hal-01289759

[57] S. COHEN-BOULAKIA, P. VALDURIEZ. *Bioinformatics big data processing*, Inria Sophia Antipolis ; LRI - CNRS, University Paris-Sud, May 2016, n^{o} RR-8915, 8 p. , https://hal.inria.fr/hal-01321033

Scientific Popularization

[58] S. Grumbach, P. Valduriez. *Les données en question*, in "Interstices", March 2016, https://hal.inria.fr/hal-01350453

Other Publications

[59] C. Jonquet, E. Dzalé-Yeumo, E. Arnaud, P. Larmande, A. Toulet, M.-A. Laporte. *AgroPortal : A Proposition for Ontology-Based Services in the Agronomic Domain*, Systems Biology and Ontologies, Poster session, January 2016, n^{o} P0343, PAG: Plant & Animal Genome, Poster, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01397465

[60] D. Robakowska Hyzorek, M. Mirouze, P. Larmande. *Integration and Visualization of Epigenome and Mobilome Data in Crops*, September 2016, Journees Ouvertes Biologie Informatique Mathematiques (JOBIM), Poster, https://hal.archives-ouvertes.fr/hal-01411668

[61] V. Silva, J. Camata, D. De Oliveira, A. L. G. A. Coutinho, P. Valduriez, M. Mattoso. *In Situ Data Steering on Sedimentation Simulation with Provenance Data*, November 2016, SC: High Performance Computing, Networking, Storage and Analysis, Poster, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01400532