IN PARTNERSHIP WITH:
CNRS

Université de Montpellier

Activity Report 2017

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:

A1. - Architectures, systems and networks
A1.1. - Architectures
A1.1.6. - Cloud
A1.1.7. - Peer to peer
A1.3. - Distributed Systems
A3.1. - Data
A3.3. - Data and knowledge analysis
A3.5. - Social networks
A3.5.2. - Recommendation systems
A4. - Security and privacy
A4.8. - Privacy-enhancing technologies
A5.4.3. - Content retrieval
A5.7. - Audio modeling and processing

Other Research Topics and Application Domains:

B1. - Life sciences
B1.1. - Biology
B1.1.9. - Bioinformatics
B6. - IT and telecom
B6.5. - Information systems

1. Personnel

Research Scientists

Patrick Valduriez [Team leader, Inria, Senior Researcher, HDR]
Reza Akbarinia [Inria, Researcher]
Alexis Joly [Inria, Researcher]
Antoine Liutkus [Inria, Researcher, from Sep 2017]
Florent Masseglia [Inria, Researcher, HDR]
Didier Parigot [Inria, Researcher, HDR]
Dennis Shasha [NYU, Inria, Advanced Research Position, from Apr 2017 until Jun 2017]

Faculty Members

Esther Pacitti [Univ Montpellier II (sciences et techniques du Languedoc), Associate Professor, HDR]
Michel Riveill [Univ de Nice - Sophia Antipolis, Professor, from Aug 2017]

Post-Doctoral Fellow

Ji Liu [Inria, until Nov 2017]

PhD Students

Paule Bondiombouy [OGES Congo, until Aug 2017]
Christophe Botella [INRA]
Lea El Beze [Univ de Nice - Sophia Antipolis]

Mathieu Fontaine [Inria, from Sep 2017]
Gaetan Heidsieck [Inria, from Nov 2017]
Titouan Lorieul [Univ de Montpellier]
Sakina Mahboubi [Univ de Montpellier]
Khadidja Meguelati [Univ de Montpellier]
Djamel Edine Yagoubi [Univ de Montpellier]

Technical staff

Antoine Affouard [Inria]
Benjamin Billet [Inria, until Sep 2017]
Boyan Kolev [Inria, granted by FP7 ClouddbAppliance project]
Oleksandra Levchenko [Inria]
Valentin Leveau [Inria, granted by Agropolis Fondation]
Sen Wang [Inria, until Sep 2017]

Visiting Scientists

Teresa Branch-Smith [Inria, until Feb 2017]
Vitor Sousa Silva [UFRJ, from Oct 2017]
Mehdi Zitouni [Univ de Tunis]

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). 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 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

5.1.1. Awards

BEST PAPER AWARD:

[32]

R. Souza, V. Silva, P. Miranda, A. Lima, P. Valduriez, M. Mattoso. *Spark Scalability Analysis in a Scientific Workflow*, in "SBBD 2017: 32th Brazilian Symposium on Databases", Uberlandia, Brazil, October 2017, pp. 1-6, Best paper award, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620161

6. New Software and Platforms

6.1. LogMagnet

Keywords: Data mining - Data stream

Functional Description: 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.

- Participants: Florent Masseglia and Julien Diener
- Contact: Florent Masseglia
- URL: https://team.inria.fr/zenith/?s=LogMagnet

6.2. Pl@ntNet - Mobile - Androïd

Keywords: Bioinformatics - Biology

Functional Description: This is the Android front-end of the Pl@ntNet platform, publicly available on Google play: https://play.google.com/store/apps/details?id=org.plantnet&hl=fr The main feature of the app is to identify plant species from photographs, through a visual recognition software making use of deep learning technologies. The number of species and the number of images used by the application evolve with the contributions of the Pl@ntNet community.

- Participant: Julien Champ
- Partners: INRA - CIRAD - IRD
- Contact: Alexis Joly
- URL: https://play.google.com/store/apps/details?id=org.plantnet&hl=en

6.3. Pl@ntNet - Mobile -IOS

Keywords: Bioinformatics - Biology

Functional Description: This is the iOS front-end of the Pl@ntNet platform, publicly available on Apple store: https://itunes.apple.com/fr/app/plantnet/id600547573?mt=8 The main feature of the app is to identify plant species from photographs, through a visual recognition software making use of deep learning technologies. The number of species and the number of images used by the application evolve with the contributions of the Pl@ntNet community.

- Participant: Hervé Goëau
- Partners: INRA - CIRAD - IRD
- Contact: Alexis Joly
- URL: https://itunes.apple.com/fr/app/plantnet/id600547573?mt=8

6.4. Pl@ntNet - Web - Angular

Keywords: Bioinformatics - Biology

FUNCTIONAL DESCRIPTION: This is the web front-end of the Pl@ntNet platform, publicly available at: http://identify.plantnet-project.org/ The main feature of the app is to identify plant species from photographs, through a visual recognition software making use of deep learning technologies. The number of species and the number of images used by the application evolve with the contributions of the Pl@ntNet community.

- Participant: Alexis Joly
- Partners: INRA - CIRAD - IRD
- Contact: Alexis Joly
- URL: https://identify.plantnet-project.org/

6.5. Pl@ntNet - DataStore

KEYWORDS: Bioinformatics - Biology
FUNCTIONAL DESCRIPTION: Datastore of the Pl@ntNet platform dedicated to the management of botanical data (observations + taxonomy) based on Apache CouchDB, Node.JS, Angular.JS, Apache Lucene.

- Participant: Hervé Goëau
- Partners: INRA - CIRAD - IRD
- Contact: Alexis Joly
- URL: https://plantnet.org/

6.6. Pl@ntNet - API

KEYWORDS: Bioinformatics - Biology
FUNCTIONAL DESCRIPTION: REST API of the Pl@ntNet platform. It provides services for data access, authentication, logging, contribution management, etc. It is mainly based on Node.JS + CouchDB.

- Authors: Samuel Dufour Kowalski, Alexis Joly, Pierre Bonnet and Antoine Affouard
- Partners: INRA - CIRAD - IRD
- Contact: Alexis Joly
- URL: https://plantnet.org/

6.7. Snoop

FUNCTIONAL DESCRIPTION: Snoop is a C++ framework dedicated to large-scale content-based image retrieval. Its main features are (i) the extraction and efficient indexing of visual features (hand-crafted or learned through deep learning), (ii) the search of similar images through approximate k-nearest neighbors and (iii), the supervised recognition of trained visual concepts. The framework can be used either as a set of C++ libraries or as a set of web services through a RESTFUL API.

- Participants: Alexis Joly, Jean-Christophe Lombardo and Olivier Buisson
- Partner: INA (Institut National de l'Audiovisuel)
- Contact: Alexis Joly

6.8. PlantRT

KEYWORDS: Bioinformatics - Biology
FUNCTIONAL DESCRIPTION: PlantRT is a distributed gossip-based platform for content sharing enabling plants observation keywords search and GPS position based recommendation. It combines advantages from centralized and P2P systems.

- Participants: Alexis Joly, Esther Pacitti, Julien Champ, Maximilien Servajean and Miguel Liroz-Gistau
- Contact: Maximilien Servajean

6.9. SciFloware

Scientific Workflow Middleware
KEYWORDS: Bioinformatics - Distributed Data Management
FUNCTIONAL DESCRIPTION: SciFloware is 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.

- Participants: Didier Parigot and Dimitri Dupuis
- Contact: Didier Parigot
- URL: http://www-sop.inria.fr/members/Didier.Parigot/pmwiki/Scifloware

6.10. CloudMdsQL Compiler

FUNCTIONAL DESCRIPTION: CloudMdsQL (Cloud Multidatastore Query Language) is a functional SQL-like language, capable of querying multiple cloud data stores (SQL, NoSQL, HDFS, etc.). The compiler parses a CloudMdsQL query and generates an optimized query execution plan to be processed by a query operator engine.

- Authors: Boyan Kolev and Patrick Valduriez
- Contact: Patrick Valduriez

6.11. Triton Server

End-to-end Graph Mapper
KEYWORD: Web Application
FUNCTIONAL DESCRIPTION: A server for managing graph data and applications for mobile social networks. The server is built on top of the OrientDB graph database system and a distributed middleware. It provides an End-to-end Graph Mapper (EGM) for modeling the whole application as (i) a set of graphs representing the business data, the in-memory data structure maintained by the application and the user interface (tree of graphical components), and (ii) a set of standardized mapping operators that maps these graphs with each other.

- Participants: Didier Parigot, Patrick Valduriez and Benjamin Billet
- Contact: Didier Parigot
- Publication: End-to-end Graph Mapper

6.12. Hadoop_g5k

FUNCTIONAL DESCRIPTION: Apache Hadoop provides an open-source framework for reliable, scalable, parallel computing. It can be deployed and used in large-scale platforms such as Grid 5000. However, its configuration and management is very difficult, specially under the dynamic nature of clusters. Therefore, we built Hadoop_g5k (Hadoop easy deployment in clusters), a tool that makes it easier to manage Hadoop clusters and prepare reproducible experiments. Hadoop_g5k offers a set of scripts to be used in command-line interfaces and a Python interface. It is actually used by Grid5000 users, and helps them saving much time when doing their experiments with MapReduce.

- Participants: Miguel Liroz-Gistau, Patrick Valduriez and Reza Akbarinia
- Contact: Patrick Valduriez
- URL: https://www.grid5000.fr/mediawiki/index.php/Hadoop_On_Execo

7. New Results

7.1. Data Management

7.1.1. Top-k Query Processing Over Encrypted Data in Clouds

Participants: Sakina Mahboubi, Reza Akbarinia, Patrick Valduriez.

Cloud data outsourcing provides users and companies with powerful capabilities to store and process their data in third-party data centers. However, the privacy of the outsourced data is not guaranteed by the cloud providers. One solution for protecting the user data against security attacks is to encrypt the data before being sent to the cloud servers. Then, the main problem is to evaluate user queries over the encrypted data.

In [41], we address the problem of top-k query processing over encrypted data, and propose an efficient approach called BuckTop. Our approach uses the bucketization technique to manage the encrypted data in the remote server. It includes a top-k query processing algorithm that works on the encrypted data of the buckets, and returns a set that contains the encrypted top-k results. It also has a filtering algorithm that efficiently eliminates the false positives in the server side. We implemented BuckTop, and compared its response time for processing top-k queries over encrypted data with that of the TA algorithm over original (plaintext) data. Our results show excellent performance gains. They show that the response time of BuckTop over encrypted data is close to TA over plaintext data.

7.1.2. End-to-end Graph Mapper

Participants: Benjamin Billet, Didier Parigot, Patrick Valduriez.

The growth of linked data in web and mobile applications motivates software developers to model their business data as graphs, enabling them to leverage the capabilities of various graph databases. Going one step further, we introduce an End-to-end Graph Mapper (EGM) [22], for modeling the whole application as (i) a set of graphs representing the business data, the in-memory data structure maintained by the application and the user interface (tree of graphical components), and (ii) a set of standardized mapping operators that maps these graphs with each other. As a benefit, the application becomes a complex live query over multiple graph databases, making the development process simpler and safer, thanks to the automation of repetitive development tasks. This work has been done in collaboration with Beepeers (http://www.beepeers.com), a startup that develops and markets social network mobile applications for small communities in the context of the Triton I-lab.

7.1.3. Management of Simulation Data

Participants: Vitor Silva, Patrick Valduriez.

In complex simulations, users must track quantities of interest (residuals, errors estimates, etc.) to control as much execution as possible. However, this tracking is typically done only after the simulation ends. We are designing techniques to extract, index and relate strategic simulation data for online queries while simulation is running. We consider coupling these techniques with largely adopted libraries such as libMesh (for numerical solvers) and ParaView (for visualization), so that queries on quantities of interest are enhanced by visualization and provenance data. Interactive data analysis support is planned for post simulation and runtime as in-situ and in-transit, taking advantage of memory access at runtime.

In [21], we propose a solution (architecture and algorithms) to combine the advantages of a dataflow-aware scientific workflow management system (SWfMS) and the raw data file analysis techniques to allow for queries on raw data file elements that are related, but reside in separate files. Armful is the name of the architecture and its main components are a raw data extractor, a provenance gatherer and a query processing interface, which are all dataflow aware. We show ARMFUL instantiated with the Chiron SWfMS.

In [31], we instantiate Armful without the SWfMS, plugging the components directly in the simulation code of highly optimized parallel applications. With support of sophisticated online data analysis, scientists get a detailed view of the execution, providing insights to determine when and how to tune parameters.

We also started investigating the combination of in-transit analysis and visualization, with the development of SAVIME (Scientific Analysis and Visualization In-Memory). The system adopts a multi-dimensional data model TARS (Typed Array Schema) [29] that enables the representation of simulation output data, the topology mesh and simulation metadata. Data produced by the simulation is ingested into the system without any transformation as a Typed Array (TAR). We intend SAVIME to implement an algebra on TARs that enables simulation output analysis and direct production of visualization output.

7.2. Scientific Workflows

7.2.1. *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. Recently, some Scientific Workflow Management Systems (SWfMSs) with provenance support (e.g. Chiron) have been deployed in the cloud. However, they typically use a single cloud site.

In [16], we consider a multisite cloud, where the data and computing resources are distributed at different sites (possibly in different regions). Based on a multisite architecture of SWfMS, i.e. multisite Chiron, and its provenance model, we propose a multisite task scheduling algorithm that considers the time to generate provenance data. We performed an extensive experimental evaluation of our algorithm using Microsoft Azure multisite cloud and two real-life scientific workflows (Buzz and Montage). The results show that our scheduling algorithm is up to 49.6% better than baseline algorithms in terms of total execution time.

In [28], 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.

7.2.2. *Parallel Execution of Scientific Workflows in Spark*

Participants: Reza Akbarinia, Esther Pacitti.

The success of using workflows for modeling large-scale scientific applications has fostered the research on parallel execution of scientific workflows in shared-nothing clusters, in which large volumes of scientific data may be stored and processed in parallel using ordinary machines. However, most of the current scientific workflow management systems do not handle the memory and data locality appropriately. Apache Spark deals with these issues by chaining activities that should be executed in a specific node, among other optimizations such as the in-memory storage of intermediate data in RDDs (Resilient Distributed Datasets). However, to take advantage of the RDDs, Spark requires existing workflows to be described using its own API, which forces the activities to be reimplemented in Python, Java, Scala or R, and this demands a big effort from the workflow programmers.

In [24], we propose a parallel scientific workflow engine called TARDIS, whose objective is to run existing workflows inside a Spark cluster, using RDDs and smart caching, in a completely transparent way for the user, i.e., without needing to reimplement the workflows in the Spark API. We evaluated our system through experiments and compared its performance with Swift/K. The results show that TARDIS performs better (up to 138% improvement) than Swift/K for parallel scientific workflow execution.

In [32], we evaluate a parameter sweep workflow also in the Oil and Gas domain, this time using Spark to understand its scalability when having to execute legacy black-box code with a DISC system. The source code of the dataflow implementation for Spark is available on github ((github.com/hpcdb/RFA-Spark).

7.3. Data Analytics

7.3.1. Massively Distributed Indexing of Time Series

Participants: Djamel Edine Yagoubi, Reza Akbarinia, Florent Masseglia.

Indexing is crucial for many data mining tasks that rely on efficient and effective similarity query processing. Consequently, indexing large volumes of time series, along with high performance similarity query processing, have became topics of high interest. For many applications across diverse domains though, the amount of data to be processed might be intractable for a single machine, making existing centralized indexing solutions inefficient.

In [42], we propose a parallel indexing solution that scales to billions of time series, and a parallel query processing strategy that, given a batch of queries, efficiently exploits the index. Our experiments, on both synthetic and real world data, illustrate that our index creation algorithm works on 1 billion time series in less than 2 hours, while the state of the art centralized algorithms need more than 5 days. Also, our distributed querying algorithm is able to efficiently process millions of queries over collections of billions of time series, thanks to an effective load balancing mechanism.

In [43], we propose RadiusSketch, a sketch/random projection-based approach that scales nearly linearly in parallel environments, and provides high quality answers. We illustrate the performance of our approach on real and synthetic datasets of up to 1 Terabytes and 500 million time series. The sketch method, as we have implemented, is superior in both quality and response time compared with the state of the art centralized algorithm. In a parallel environment with 32 processors, on both real and synthetic data, our parallel approach improves by a factor of up to 100 in index time construction and up to 15 in query answering time. Finally, our data structure makes use of idle computing time to improve the recall and precision yet further.

7.3.2. 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 [18], 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.

7.3.3. Closed Itemset Mining in Massively Distributed Environments

Participants: Mehdi Zitouni, Reza Akbarinia, Florent Masseglia.

Data analytics in general, and data mining primitives in particular, are a major source of bottlenecks in the operation of information systems. This is mainly due to their high complexity and intensive call to IO operations, particularly in massively distributed environments. Moreover, an important application of data analytics is to discover key insights from the running traces of information system in order to improve their engineering. Mining closed frequent itemsets (CFI) is one of these data mining techniques, associated with great challenges. It allows discovering itemsets with better efficiency and result compactness.

However, discovering such itemsets in massively distributed data poses a number of issues that are not addressed by traditional methods. One solution for dealing with such characteristics is to take advantage of parallel frameworks, e.g. MapReduce. In [33], [44], we address the problem of distributed CFI mining by introducing a new parallel algorithm, called DCIM, which uses a prime number based approach. A key feature of DCIM is the deep combination of data mining properties with the principles of massive data distribution. We carried out exhaustive experiments over real world datasets to illustrate the efficiency of DCIM for large real world datasets with up to 53 million documents.

7.3.4. Optimal Data Placement for Fast Parallel Mining of Frequent Itemsets

Participants: Saber Salah, Reza Akbarinia, Florent Masseglia.

Frequent itemset mining presents one of the fundamental building blocks in data mining. However, despite the crucial recent advances that have been made in data mining literature, few of both standard and improved solutions scale. This is particularly the case when (i) the quantity of data tends to be very large or (ii) the minimum support is very low.

In [19], we address the problem of parallel frequent itemset mining (PFIM) in very large databases, and study the impact and effectiveness of using specific data placement strategies in a massively distributed environment. By offering a clever data placement and an optimal organization of the extraction algorithms, we show that the arrangement of both the data and the different processes can make the global job either completely inoperative or very effective. In this setting, we propose two different highly scalable, PFIM algorithms, namely P2S (Parallel-2-Steps) and PATD (Parallel Absolute Top Down). P2S algorithm allows discovering itemsets from large databases in two simple, yet efficient parallel jobs, while PATD renders the mining process of very large databases more simple and compact. Its mining process is made up of only one parallel job, which dramatically reduces the mining runtime, the communication cost and the energy power consumption overhead in a distributed computational platform. Our different proposed approaches have been extensively evaluated on massive real-world data sets. The experimental results confirm the effectiveness and scalability of our proposals by the important scale-up obtained with very low minimum supports compared to other alternatives.

7.4. Data Search

7.4.1. Adversarial Autoencoders For Novelty Detection

Participants: Valentin Leveau, Alexis Joly.

In this work [40], we addressed the problem of novelty detection, i.e recognizing at test time if a data item comes from the training data distribution or not. We focus on Adversarial autoencoders (AAE) that have the advantage to explicitly control the distribution of the known data in the feature space. We show that when they are trained in a (semi-)supervised way, they provide consistent novelty detection improvements compared to a classical autoencoder. We further improve their performance by introducing an explicit rejection class in the prior distribution coupled with random input images to the autoencoder.

7.4.2. Going deeper in the automated identification of Herbarium specimens

Participants: Alexis Joly, Herve Goeau.

Hundreds of herbarium collections have accumulated a valuable heritage and knowledge of plants over several centuries. Recent initiatives started ambitious preservation plans to digitize this information and make it available to botanists and the general public through web portals. However, thousands of sheets are still unidentified at the species level while numerous sheets should be reviewed and updated following more recent taxonomic knowledge. These annotations and revisions require an unrealistic amount of work for botanists to carry out in a reasonable time. Computer vision and machine learning approaches applied to herbarium sheets are promising but are still not well studied compared to automated species identification from leaf scans or pictures of plants in the field. In this work [14], we proposed to study and evaluate the accuracy with which herbarium images can be potentially exploited for species identification with deep learning technology. In addition, we proposed to study if the combination of herbarium sheets with photos of plants in the field is relevant in terms of accuracy, and finally, we explore if herbarium images from one region that has one specific flora can be used to do transfer learning to another region with other species; for example, on a region under-represented in terms of collected data. This is, to our knowledge, the first study that uses deep learning to analyze a big dataset with thousands of species from herbaria. Results show the potential of Deep Learning on herbarium species identification, particularly by training and testing across different datasets from different herbaria. This could potentially lead to the creation of a semi, or even fully automated system to help taxonomists and experts with their annotation, classification, and revision works.

7.4.3. Crowdsourcing Thousands of Specialized Labels: a Bayesian active training approach

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

The use of crowdsourced and more generally user-generated annotations became the de facto methodology for building training data in a variety of data indexing and search tasks. When the labels correspond to well known or easy-to-learn concepts, it is straightforward to train the annotators by giving a few examples with known answers. Neither is true when there are thousands of complex domain specific labels. In this work, we focused on the particular case of crowdsourcing domain-specific annotations that usually require hard expert knowledge (such as plant species names, architectural styles, medical diagnostic tags, etc.). We considered that common knowledge is not sufficient to perform the task but any people can be taught to recognize a small subset of domain-specific concepts. In such a context, it is best to take advantage of the various capabilities of each annotator through teaching (annotators can enhance their knowledge), assignment (annotators can be focused on tasks they have the knowledge to complete) and inference (different annotator propositions can be aggregated to enhance labeling quality). In this work [20], we proposed a set of data-driven algorithms to (i) train image annotators on how to disambiguate among automatically generated candidate labels, (ii) evaluate the quality of annotators' label suggestions 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 measured the benefits of these algorithms by a live user experiment related to image-based plant identification involving around 1,000 people (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.

7.4.4. Evaluation of Content-Based Biodiversity Identification techniques

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

We ran a new edition of the LifeCLEF evaluation campaign [26] with the involvement of 15 research teams working on content-based biodiversity identification worldwide. The main novelties of the 2017 edition of LifeCLEF compared to the previous years were the following:

- **Scalability**: To fully reach its objective, an evaluation campaign such as LifeCLEF requires a long term research effort so as to (i) encourage non incremental contributions, (ii) measure consistent performance gaps and (iii), progressively scale up the problem. Therefore, the number of species was increased considerably between the 2016 and 2017 editions. The plant task, in particular, made a big jump with 10,000 species instead of 1,000 species in the training set. This makes it one of the largest image classification benchmark. Besides, the data set of the bird task was increased by 50% up to 1,500 species which makes it the largest audio classification benchmark as well.
- **Noisy vs. clean data**: The focus of the plant task this year was to study the impact of training identification systems on noisy Web data rather then clean data [35]. Collecting clean data massively is actually prohibitive in terms of human cost whereas noisy Web data can be collected at a very cheap cost. Therefore, we built two large-scale datasets illustrating the same 10K species: one with clean labels coming from the Web platform Encyclopedia Of Life, and one with a high degree of noise - domain noise as well as category noise - crawled from the Web without any filtering. The main conclusion of our evaluation was that convolutional neural networks (CNN) appear to be amazingly effective in the presence of noise in the training set. All networks trained solely on the noisy dataset did outperform the same models trained on the trusted data. Even at a constant number of training

iterations (i.e. at a constant number of images passed to the network), it was more profitable to use the noisy training data. This means that diversity in the training data is a key factor to improve the generalization ability of deep learning. The noise itself seems to act as a regularization of the model. Beyond technical aspects, this conclusion is of high importance in botany and biodiversity informatics in general. Data quality and data validation issues are of crucial importance in these fields and our conclusion is somehow disruptive.

- **Time-coded soundscapes**: As the soundscapes data appeared to be very challenging in 2016 (with an accuracy below 15%), we introduced in 2017 new soundscape recordings containing time-coded bird species annotations thanks to the involvement of expert ornithologists. In total, 4,5 hours of audio recordings were collected and annotated manually with more than 2000 identified segments. The main outcome of our evaluation [36], was that the best performing system on that data was based on a purely image-based convolutional neural network architecture (Inception V4) applied to a standard time-frequency representation. This shows the convergence of the best performing methods whatever the targeted domain.
- **New organisms and identification scenarios**: The SeaCLEF task was extended with novel scenarios involving new organisms, i.e (i) salmons detection for the monitoring of water turbine, and (ii), marine animal species recognition using weakly-labeled images and relevance ranking.

7.4.5. Pl@ntNet Business Venture proposal

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

The ACM Multimedia conference (rank A) introduced in 2017 a new "Business Venture Track" soliciting business venture proposals that combine multimedia technology. The aim is to bridge the gap between academia and industry on multimedia research, innovation and application. The track was open for submissions by all multimedia researchers and entrepreneurs. In this context, we have been working on a business venture proposal around the Pl@ntNet project that has been accepted for publication [25]. Our business proposal is to allow enterprises or organizations to set up their own private collaborative workflow within Pl@ntNet information system. The main added value is to allow them to work on their own business object (e.g. plant disease diagnostic, deficiency measurements, railway lines maintenance, etc.) and with their own community of contributors and end-users (employees, sales representatives, clients, observers network, etc.). This business idea answers to a growing demand in agriculture and environmental economics. Actors in these domains acknowledge that machine learning techniques are mature enough but the lack of training data and efficient tools to collect them remains a major problem. A collaborative platform like Pl@ntNet extended with the technical innovations presented in this paper is the ideal tool to bridge this gap. It will initiate a powerful positive feedback loop boosting the production of training data while improving the work of the employees.

8. Bilateral Contracts and Grants with Industry

8.1. Microsoft ZcloudFlow (2013-2017)

Participants: 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 (2014-2017)

Participants: Benjamin Billet, Didier Parigot.

Triton is a common Inria lab (i-lab) between Zenith and Beepeers (http://www.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.

8.3. SAFRAN (2018)

Participants: Reza Akbarinia, Florent Masseglia.

SAFRAN and Inria are involved in the DESIR frame-agreement (Florent Masseglia is the scientific contact on "Data Analytics and System Monitoring" topic. In this context, SAFRAN dedicates 80K€ for a joint study of one year on time series indexing. The specific time series to be exploited are those of engine benchmarking with novel characteristics for the team (multiscale and multidimensional).

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 6 year project (2012-2018) 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. Projet 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.

9.2.1.2. Institut de Convergence Agriculture numérique #DigitAg, (2017-2023), 275Keuro.

Participants: Esther Pacitti, Florent Masseglia, Patrick Valduriez.

#DigitAg brings together in a partnership of seventeen actors (public research and teaching organizations, transfer actors and companies) with the objective of accelerating and supporting the development of agriculture companies in France and in southern countries based on new tools, services and uses. Based in Montpellier with an office in Toulouse and Rennes and led by Irstea, #DigitAg's ambition is to become a world reference for digital agriculture. In this project, Zenith is involved in the analysis of big data from agronomy.

9.2.2. Others

9.2.2.1. 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. H2020 Projects

9.3.1.1. 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 industral 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.

9.3.1.2. 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: (i) 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; (ii) 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; (iii) 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. Inria International Partners

9.4.1.1. 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)

9.4.2. Inria Associate Teams Not Involved in an Inria International Lab

9.4.2.1. SciDISC

Title: Scientific data analysis using Data-Intensive Scalable Computing

Inria principal investigator:Patrick Valduriez

International Partner:

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

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

Universidade Federal Fluminense, Niteroi (Brazil), Daniel Oliveira

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

Start year: 2017

See also: https://team.inria.fr/zenith/scidisc/

Data-intensive science requires the integration of two fairly different paradigms: high-performance computing (HPC) and data-intensive scalable computing (DISC). Spurred by the growing need to analyze big scientific data, the convergence between HPC and DISC has been a recent topic of interest. This project will address the grand challenge of scientific data analysis using DISC (SciDISC), by developing architectures and methods to combine simulation and data analysis. The expected results of the project are: new data analysis methods for SciDISC systems; the integration of these methods as software libraries in popular DISC systems, such as Apache Spark; and extensive validation on real scientific applications, by working with our scientific partners such as INRA and IRD in France and Petrobras and the National Research Institute (INCT) on e-medicine (MACC) in Brazil.

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.

9.4.3.1. International Initiatives

BD-FARM

Title: Big Data Management and Analytics for Agriculture and Farming

International Partner (Institution - Laboratory - Researcher):

Chubu University - International Digital Earth Applied Science Research Center (IDEAS), Kiyoshi Honda

Duration: 2016 - 2017

Start year: 2016

See also: https://team.inria.fr/zenith/bdfarm-2016-2018-stic-asia/

World population is still growing and people are living longer and older. World demand for food rises sharply and current growth rates in agriculture are clearly not sufficient. But extreme flood, drought, typhoon etc, caused by climate change, give severe damages on traditional agriculture. Today, an urgent and deep redesign of agriculture is crucial in order to increase production and to reduce environmental impact. In this context, collecting, managing and analyzing dedicated, large, complex, and various datasets (Big Data) will allow improving the understanding of complex mechanisms behind adaptive, yield and crop improvement. Moreover, sustainability will require detailed studies such as the relationships between genotype, phenotype and environment. In other words, data science and ICT for agriculture must help improving production. Moreover, it has to be done while getting properly adapted to soil, climatic and agronomic constraints as well as taking into account the genetic specificities of plants.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Several international scientists visited the team and gave seminars

- Tamer Özsu (University of Waterloo, Canada): “Approaches to RDF Data Management and SPARQL Query Processing” on March 9.
- Dennis Shasha (NYU) “Reducing Errors by Refusing to Guess (Occasionally)” on June 1.
- Fabio Porto (LNCC, Brazil): “Database System Support of Simulation Data” on January 27 and “Simulation Data Management” on June 1.
- Marta Mattoso (UFRJ, Brazil): “Human-in-the-loop to Fine-tune Data in Real Time ” on December 14.

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 2017 international workshop [1] dedicated to content-based biodiversity identification techniques, Dunlin, sept. 2017
- Alexis Joly was in the organizing committee of the Floris'tic national workshop held in Toulouse, nov. 2017 (http://floristic.org/journeefloristic/)

Conference program committees :

- International Conference on Very Large Data Bases (VLDB), 2017: R. Akbarinia, F. Masseglia
- International Workshop on Big Data Management in Cloud Systems, 2017: R. Akbarinia
- International Conference on XLDB, 2017: P. Valduriez
- 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
- ACM Multimedia conference (ACMMM), 2017: A. Joly
- IEEE International Conference on Image Processing (ICIP), 2017: A. Joly
- ACM International Conference on Multimedia Retrieval (ICMR), 2017: A. Joly
- International Conference and Labs of the Evaluation Forum (CLEF), 2017: A. Joly
- IEEE Int. Conf. on Data Mining, 2017: F. Masseglia
- ACM Symposium on Applied Computing 2017: F. Masseglia
- IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA), 2017: F. Masseglia
- Int. Symposium on Information Management and Big Data (SimBig), 2017: F. Masseglia
- Int. Symposium on Methodologies for Intelligent Systems (ISMIS), 2017: F. Masseglia

[1] http://www.imageclef.org/lifeclef/2017

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
- 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
- Transactions on Parallel and Distributed Systems (TPDS): F. Masseglia
- Data Mining and Knowledge Discovery (DMKD): F. Masseglia
- International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS): F. Masseglia
- Transactions on Large-Scale Data and Knowledge-Centered Systems (TLKDS): F. Masseglia

Other activities (national):

- P. Valduriez is the scientific manager for the Latin America zone at Inria's Direction of Foreign Relationships (DPEI). In 2017, he has been elected President of the Scientific Committee of the BDA conference for 4 years.
- P. Valduriez gave an invited talk "Data science and science data" on Dec. 1, at the first seminar of Axis 4 "Information systems, data storage and transfer" of the #Digitag Institut de Convergence on numerical agriculture, in Montpellier.
- Alexis Joly gave invited talks in Conference of the biology society of Montpellier 2017 (60 participants), Transfer-LR seminar 2017 on machine learning (100 participants), CIRAD BIOS department seminar 2017 (100 participants).
- Alexis Joly served as a reviewer for the French International Cooperation program (STIC AmSud)
- F. Masseglia participated to a Panel for Inria's 50th anniversary in Grenoble on "digital litteracy", November 2017.
- F. Masseglia gave invited talks: on "Big Data Analytics" at Univ. of Marseille, January 27 ; on "Massively Distributed Time Series Analytics" for LIRMM's 25th anniversary ; on "Computer Science for Everyone" at "Université du Tiers Temps" (UTT), March 2017.
- F. Masseglia served as a reviewer for international programs (STIC AmSud, ECOS SUD).
- F. Masseglia is scientific referent for Inria on the frame agreement with SAFRAN about "System Monitoring and Data Analytics".
- F. Masseglia is "Chargé de mission pour la médiation scientifique Inria" and heads Inria's national network of colleagues involved in science popularization.
- E. Pacitti is responsible for new comers at Polytech' Montpellier's Direction of Foreign Relationships.

Other activities (international):

- E. Pacitti gave an invited talk at CEFET, Rio de Janeiro on nov 27 on “uncertainty analysis of big simulation data”.
- P. Valduriez gave several invited talks: “The CloudMdsQL Multistore System” on may 2 at the Data Systems Group Seminar Series, University of Waterloo, Canada, and on may 4 at McGill University, Montréal, Canada; “Data Science: opportunities and risks” on july 26 at the Colloquium of PESC/COPPE, UFRJ, Rio de Janeiro; “An Overview of Polystores” on oct 11 at the XLDB conference in Clermont-Ferrand; “Data science and science data” on nov 27 at CEFET, Rio de Janeiro.

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, 10h, level M2, Faculty of Science, UM

Florent Masseglia:

Science Popularization: 2 Ph.D students, from 2 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

10.2.2. Supervision

- PhD in progress: Gaetan Heidsieck Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping, started Oct 2017, Univ. Montpellier, Esther Pacitti, Christophe Pradal, François Tardieu
- PhD in progress: Christophe Botella, Large-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)
- PhD in progress: Vitor Silva, Supporting Human-in-the-Loop in Large-scale Workflows, started 2014, UFRJ, Brazil Advisors: Marta Mattoso (UFRJ), Daniel Oliveira (UFF), Patrick Valduriez
- PhD in progress: Renan Souza, Massively Distributed Clustering, started 2015, UFRJ, Brazil Advisors: Marta Mattoso (UFRJ), Daniel Oliveira (UFF), Patrick Valduriez

10.2.3. Juries

Members of the team participated to the following PhD committees:

- A. Joly: Sanaa Chafik (ENSEM Casablanca)
- E. Pacitti: Vincent Leroy (HDR, Univ. Grenoble), Douglas Ericson Marcelino de Oliveira (UFF, Rio de Janeiro), Pauline Folz (Univ. Nantes, reviewer), Uras Tos (Univ. Toulouse, reviewer), Carlyna Bondlombouy (Univ. Montpellier).
- P. Valduriez: Patrick KamnangWanko (Univ. Bordeaux), Julien Pilourdault (Univ. Grenoble, reviewer), Yvan Brondino (Univ. Madrid, reviewer)
- F. Masseglia: Andres Moreno (Univ. Nice-Sophia Antipolis), Kenza Kellou-Menouer (Univ. Versailles St-Quentin), Raef Mousheimish (Univ. Versailles St-Quentin, reviewer)

Members of the team participated to the following hiring committees:

- A. Joly: associate professor position, Univ. Toulon
- P. Valduriez: associate professor position, Univ. Nantes

10.3. 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). Florent Masseglia is co-author of the first course and scientific referent of the other courses.

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. The extension of this book for the French "Collège" has been released this year with new activities and new scientific content.

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, 14 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.

F. Masseglia is member of the pedagogic committee of "Edu'up", a project from France-IOI on learning code and computational thinking.

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 used in a large-scale program in the Czech republic (with an objective of 40 classrooms in the end). 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.

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 co-organized the regional Code-Week events with the local network of media-library ("réseau des médiathèques de Montpellier Méditerranée Métropole").
- F. Masseglia is member of the project selection committee for “La fête de la science” in Montpellier.
- A. Joly designed/animated several stands and public demos in events: “La fête de la science” (Montpellier, 1 day demo), 50 ans de l'Inria (Paris, 2-days demo), Expomobile Homonumericus (Itinerant expo in France).

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. *Crowdsourcing Thousands of Specialized Labels: A Bayesian Active Training Approach*, in "IEEE Transactions on Multimedia", June 2017, vol. 19, n^{o} 6, pp. 1376 - 1391 [*DOI :* 10.1109/TMM.2017.2653763], https://hal.archives-ouvertes.fr/hal-01629149

[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

Articles in International Peer-Reviewed Journals

[13] J. CAMATA, V. SILVA, P. VALDURIEZ, M. MATTOSO, A. L. G. A. COUTINHO. *In situ visualization and data analysis for turbidity currents simulation*, in "Computers & Geosciences", January 2018, vol. 110, pp. 23-31 [*DOI :* 10.1016/J.CAGEO.2017.09.013], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620127

[14] J. CARRANZA-ROJAS, H. GOEAU, P. BONNET, E. MATA-MONTERO, A. JOLY. *Going deeper in the automated identification of Herbarium specimens*, in "BMC Evolutionary Biology", December 2017, vol. 17, n^{o} 1, 181 p. [*DOI :* 10.1186/S12862-017-1014-Z], https://hal.inria.fr/hal-01580070

[15] S. COHEN-BOULAKIA, K. BELHAJJAME, O. COLLIN, J. CHOPARD, C. FROIDEVAUX, A. GAIGNARD, K. HINSEN, P. LARMANDE, Y. LE BRAS, F. LEMOINE, F. MAREUIL, H. MÉNAGER, C. PRADAL, C. BLANCHET. *Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities*, in "Future Generation Computer Systems", 2017 [*DOI :* 10.1016/J.FUTURE.2017.01.012], https://hal.archives-ouvertes.fr/hal-01516082

[16] J. LIU, E. PACITTI, P. VALDURIEZ, M. MATTOSO. *Scientific Workflow Scheduling with Provenance Data in a Multisite Cloud*, in "Transactions on Large-Scale Data- and Knowledge-Centered Systems", 2017, vol. 33, pp. 80-112 [*DOI :* 10.1109/IPDPS.2007.370305], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620224

[17] 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", February 2017, vol. 67, pp. 341–353 [*DOI :* 10.1016/j.future.2016.06.002], https://hal.inria.fr/hal-01336655

[18] S. Salah, R. Akbarinia, F. Masseglia. *A Highly Scalable Parallel Algorithm for Maximally Informative k-Itemset Mining*, in "Knowledge and Information Systems (KAIS)", January 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01288571

[19] S. Salah, R. Akbarinia, F. Masseglia. *Data placement in massively distributed environments for fast parallel mining of frequent itemsets*, in "Knowledge and Information Systems (KAIS)", 2017, vol. 53, n^{o} 1, pp. 207-237 [*DOI :* 10.1007/s10115-017-1041-5], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620383

[20] M. Servajean, A. Joly, D. Shasha, J. Champ, E. Pacitti. *Crowdsourcing Thousands of Specialized Labels: A Bayesian Active Training Approach*, in "IEEE Transactions on Multimedia", June 2017, vol. 19, n^{o} 6, pp. 1376 - 1391 [*DOI :* 10.1109/TMM.2017.2653763], https://hal.archives-ouvertes.fr/hal-01629149

[21] V. J. Silva, J. J. Leite, J. J. Camata, D. De Oliveira, A. L. G. A. Coutinho, P. Valduriez, M. J. Mattoso. *Raw data queries during data-intensive parallel workflow execution*, in "Future Generation Computer Systems", January 2017, vol. 75, pp. 402-422 [*DOI :* 10.1016/j.future.2017.01.016], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01445219

International Conferences with Proceedings

[22] B. Billet, M. Jurret, D. Parigot, P. Valduriez. *End-to-end Graph Mapper*, in "BDA: Conférence sur la Gestion de Données — Principes, Technologies et Applications", Nancy, France, November 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620239

[23] J. Carranza-Rojas, A. Joly, P. Bonnet, H. Goëau, E. Mata-Montero. *Automated Herbarium Specimen Identification using Deep Learning*, in "TDWG 2017 - Annual Conference on Biodiversity Information Standards", Ottawa, Canada, October 2017 [*DOI :* 10.3897/tdwgproceedings.1.20302], https://hal.archives-ouvertes.fr/hal-01629142

[24] D. Gaspar, F. Porto, R. Akbarinia, E. Pacitti. *TARDIS: Optimal Execution of Scientific Workflows in Apache Spark*, in "DaWaK 2017: Data Warehousing and Knowledge Discovery", Lyon, France, LNCS, August 2017, n^{o} 10440, pp. 74-87 [*DOI :* 10.1007/978-3-319-64283-3_6], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620060

[25] A. Joly, P. Bonnet, A. Affouard, J.-C. Lombardo, H. Goëau. *Pl@ntNet -My Business*, in "ACM Multimedia 2017", Mountain View, United States, October 2017, pp. 1-11, https://hal.inria.fr/hal-01638263

[26] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, J.-C. Lombardo, R. Planque, S. Palazzo, H. Müller. *LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges*, in "CLEF: Cross-Language Evaluation Forum for European Languages", Dublin, Ireland, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (editors), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, September 2017, vol. LNCS, n^{o} 10456, pp. 255-274 [*DOI :* 10.1007/978-3-319-65813-1_24], https://hal.archives-ouvertes.fr/hal-01629191

[27] A. Khatibi, F. Porto, J. G. Rittmeyer, E. Ogasawara, P. Valduriez, D. Shasha. *Pre-processing and Indexing techniques for Constellation Queries in Big Data*, in "DaWaK 2017: 19th International Conference on Big Data Analytics and Knowledge Discovery", Lyon, France, Big Data Analytics and Knowledge Discovery, Springer, August 2017, n^{o} 10253, pp. 74-87, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620398

[28] J. Liu, L. Pineda-Morales, E. Pacitti, A. Costan, P. Valduriez, G. Antoniu, M. Mattoso. *Efficient Scheduling of Scientific Workflows using Hot Metadata in a Multisite Cloud*, in "BDA: Conférence sur la Gestion de Données — Principes, Technologies et Applications", Nancy, France, November 2017, 13 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620231

[29] H. Lustosa, N. Lemus, F. Porto, P. Valduriez. *TARS: An Array Model with Rich Semantics for Multidimensional Data*, in "ER FORUM 2017: Conceptual Modeling : Research In Progress", Valencia, Spain, November 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620376

[30] O. Rodriguez, C. Colomier, C. Rivière, R. Akbarinia, F. Ulliana. *Querying Key-Value Stores Under Simple Semantic Constraints : Rewriting and Parallelization*, in "BDA: Conférence sur la Gestion de Données — Principes, Technologies et Applications "", Nancy, France, November 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620207

[31] R. Souza, V. Silva, J. Camata, A. L. G. A. Coutinho, P. Valduriez, M. Mattoso. *Tracking of Online Parameter Fine-tuning in Scientific Workflows*, in "Workflows in Support of Large-Scale Science (WORKS), in conjunction with ACM/IEEE Supercomputing", Denver, United States, November 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620974

[32] *Best Paper*
R. Souza, V. Silva, P. Miranda, A. Lima, P. Valduriez, M. Mattoso. *Spark Scalability Analysis in a Scientific Workflow*, in "SBBD 2017: 32th Brazilian Symposium on Databases", Uberlandia, Brazil, October 2017, pp. 1-6, Best paper award, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620161.

[33] M. Zitouni, R. Akbarinia, S. Ben Yahia, F. Masseglia. *Massively Distributed Environments and Closed Itemset Mining: The DCIM Approach*, in "CAiSE: Advanced Information Systems Engineering", Essen, Germany, June 2017, vol. LNCS, n^{o} 10253, pp. 231-246 [*DOI :* 10.1007/978-3-319-59536-8_15], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620238

Conferences without Proceedings

[34] A. Affouard, H. Goëau, P. Bonnet, J.-C. Lombardo, A. Joly. *Pl@ntNet app in the era of deep learning*, in "ICLR 2017 - Workshop Track - 5th International Conference on Learning Representations", Toulon, France, April 2017, pp. 1-6, https://hal.archives-ouvertes.fr/hal-01629195

[35] H. Goeau, P. Bonnet, A. Joly. *Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017)*, in "CLEF 2017 - Conference and Labs of the Evaluation Forum", Dublin, Ireland, September 2017, pp. 1-13, https://hal.archives-ouvertes.fr/hal-01629183

[36] H. Goeau, H. Glotin, W.-P. Vellinga, R. Planqué, A. Joly. *LifeCLEF Bird Identification Task 2017*, in "CLEF 2017 - Conference and Labs of the Evaluation Forum", Dublin, Ireland, September 2017, pp. 1-9, https://hal.archives-ouvertes.fr/hal-01629175

Scientific Books (or Scientific Book chapters)

[37] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Deep neural network based multichannel audio source separation*, in "Audio Source Separation", Springer, 2017, forthcoming, https://hal.inria.fr/hal-01633858

Books or Proceedings Editing

[38] L. BELLATRECHE, P. VALDURIEZ, T. MORZY (editors). *Advances in Databases and Information Systems*, Elsevier, October 2017, vol. 70 [*DOI :* 10.1016/J.IS.2017.08.003], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01630719

[39] A. HAMEURLAIN, J. KÜNG, R. WAGNER, R. AKBARINIA, E. PACITTI (editors). *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXIII*, Springer, Berlin, Heidelberg, 2017, vol. LNCS, n^{o} 10430 [*DOI :* 10.1007/978-3-662-55696-2], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01624805

Research Reports

[40] V. LEVEAU, A. JOLY. *Adversarial autoencoders for novelty detection*, Inria - Sophia Antipolis, February 2017, https://hal.inria.fr/hal-01636617

[41] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Top-k Query Processing Over Outsourced Encrypted Data*, Inria Sophia Antipolis - Méditerranée, April 2017, n^{o} RR-9053, 24 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01502142

Scientific Popularization

[42] D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, T. PALPANAS. *DPiSAX: Massively Distributed Partitioned iSAX*, in "ICDM 2017: IEEE International Conference on Data Mining", New Orleans, United States, November 2017, pp. 1-6, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620125

[43] D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, D. SHASHA. *RadiusSketch: Massively Distributed Indexing of Time Series*, in "DSAA 2017: IEEE International Conference on Data Science and Advanced Analytics", Tokyo, Japan, October 2017, pp. 1-10, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620154

[44] M. ZITOUNI, R. AKBARINIA, S. BEN YAHIA, F. MASSEGLIA. *Massively Distributed Environments and Closed Itemset Mining: The DCIM Approach*, in "BDA 2017: 33ème Conférence sur la Gestion de Données — Principes, Technologies et Applications", Nancy, France, November 2017, vol. 4, pp. 1-15 [*DOI :* 10.1145/1837934.1837995], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620354

Other Publications

[45] N. KERIVEN, A. DELEFORGE, A. LIUTKUS. *Blind Source Separation Using Mixtures of Alpha-Stable Distributions*, November 2017, working paper or preprint, https://hal.inria.fr/hal-01633215

[46] B. KOLEV, O. LEVCHENKO, F. MASSEGLIA, R. AKBARINIA, E. PACITTI, P. VALDURIEZ. *Highly Scalable Real-Time Analytics with CloudDBAppliance*, October 2017, XLDB: Extremely Large Databases Conference, Poster, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01632355