Doctoral School Report 2014-2015

Daniel Aguiar da Silva Carvalho
Director: Chirine Ghedira-Guegan
Co-directors: Nadia Benani and Genoveva Vargas-Solar

Context

The emergence of new architectures like the cloud opens new challenges to data processing. The possibility of having unlimited access to cloud resources and the "pay as U go" model make it possible to change the hypothesis under which current technology and solutions address the processing of huge volumes of data. Instead of designing processes and algorithms taking into account the limits on resources availability, the cloud sets the focus on the economic cost implied of using resources and producing results by parallelizing the use of computing resources while delivering data under subscription oriented cost models.

The context imposes hence to consider SLA and different data delivery models. Indeed, we believe that given the volume and the complexity of query evaluation that includes steps that imply greedy computations, it is important to combine and revisit well-known data integration solutions and adapt them to this context. This can be done according to quality of service requirements expressed by the consumers and Service Level Agreement (SLA) contracts exported by the cloud providers that host data collections and deliver resources for executing the associated management processes.

This project intends to address data integration in a multi-cloud hybrid context guided by user preferences statements and SLA contracts exported by different cloud providers and by several QoS measures associated to data collections properties: trust, privacy, economic cost. To the best of our knowledge, we have not identified more proposals concerning the use of SLAs combined with a data integration approach on a multi-cloud context.

The objective is to propose data integration (lookup, aggregation, correlation) strategies adapted to the vision of the economic model of the cloud such as accepting partial results delivered on demand or under predefined subscription models that can affect the quality of the results; accepting specific data duplication that can respect user preferences but ensure data availability; accepting to launch a task that contributes to an integration on a first cloud whose SLA verifies a given requirement rather than a more powerful cloud but with less quality guarantees in the SLA.

In our work we consider an example from the domain of energy management. My directors are working on two national projects in this domain. So for instance, we assume we are interested in queries like: Give a list of energy providers that can provision 1000 KW-h, in the next 10 seconds, that are close to my city, with a cost of 0,50 Euro/KW-h and that are labeled as green? The question is how can the user efficiently obtain results for her queries such that they meet her QoS requirements, they respect her subscribed contracts with the involved cloud provider(s) and such that they do not neglect services contracts? Particularly, for queries that call several services deployed on different clouds. This work is part of an international collaboration with the DiMAP, Federal University of Rio Grande do Norte.

Synthesis of the research activities 2014-2015

During the first year of PhD project, we have been working on the state of the art. The idea is to be aware of all types of publications close related to the thesis proposal. To reach this, we proceeded with a literature analysis using a systematic mapping methodology.

Briefly, the methodology consists in retrieving papers from scientific databases using the same search string. These papers are filtered according to an inclusion and exclusion criteria that should be defined based on the research interests. The papers will be classified in different categories (called facets) and for each facet in a specific dimension. The facets and dimension are defined based on the authors' knowledge and interests. Taking the final papers collection, the abstracts should be read in order to classify each paper into the dimensions for each facet.

The final objective of this first work is to show the research trends of data integration as a result of the emergence of the cloud and the characteristics associated to Big Data that require resources in order to be processed. In other terms, this methodology allowed us to identify trends and open issues regarding our research topic and proposing an approach that fills some gaps and proposes an original data integration solution according to current trends in the area.

This work performed, we thus have written an article to be submitted to the 26th International Conference on Database and Expert Systems applications (DEXA 2015) in the beginning of March 2015. This work has been made in collaboration with the Informatics' Lab in Grenoble (co-supervisor Genoveva Vargas-Solar) and LIRIS Lab at INSA (co-supervisor Nadia Benani). The table below presents our results. We retrieved 1832 papers. 114 papers were selected based on our inclusion and exclusion criteria. This final data collection builds the state of the art to the thesis and, in the next step, we will perform an in-depth analysis of theses papers in order to (i) analyze what have been made and (ii) propose my approach.

Database	Amount	Included	Excluded
IEEE	658	56	602
ACM	649	31	618
Science Direct	106	6	100
CiteSeerX	419	21	398
Total	1832	114	1718

Perspectives for the next year 2015-2016

Based on the publications extracted from the mapping process methodology, we will proceed the analysis of the current state of the art in order to formalize our proposal. The analysis will be the basis to our model proposal. As a natural result, we will write a paper describing our approach and a survey. In parallel, we will carry on the first steps of implementation of the proposed approach. This will be a proof of concept that can be presented to partners in energy domain. The table below describes our intended calendar. The following activities are:

1. Paper submission: systematic mapping analysis.
2. Analysis of the current state of the art.
3. Writing the survey.
4. Approach proposal.
5. Writing the paper that describes our approach.
6. Carry on the first proof of concepts.

-	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan
1	●										
2	●	●	●	●	●						
3				●	●	●	●	●			
4							●	●	●		
5								●	●		
6									●	●	●