Thesis Advancement Report 2015-2016 (Second Year)

Thesis title: Trusted-SLA Guided Data Integration on Multi-cloud Environments
PhD. student: Daniel Aguiar da Silva Carvalho
Supervisor: Chirine Ghedira-Guegan **Co-supervisors:** Nadia Bennani and Genoveva Vargas-Solar

1 Context[1]

Current data integration implies consuming data from different data services and integrating the results while meeting users' quality requirements. For example, whether the user accepts to pay for data, its provenance, veracity and freshness. The data itself can be delivered by services according to different data quality measures describing the conditions in which a service can provide or process data. These measures can be expressed in a service level agreement (SLA) stating what the user can expect from a service or system behavior.

Data processing may need a considerable amount of storage, memory and computing capacity that can be provided by cloud architectures. Authors have presented their integration approaches and systems in cloud or service-oriented contexts [2, 3, 4, 5]. [2, 3] have considered the requirement of computing resources for integrating data focusing on performance aspects. [4, 5] tackled quality aspects of the integration such as privacy and cost. Although these works have considered data integration, particularly by using data services, we believe that there are other crucial aspects that should be studied regarding the requirements and constraints of data consumers, data producers, the associated infrastructure, and the data itself:

- Data consumers have cloud subscription which has an associated SLA describing characteristics and constraints such as data quality, current location, validity interval, budget, and resources limit over the cloud infrastructure.
- Data producers describe the type of data they produce, the production rate, production time, location, cost, service availability, response time, for instance. In addition, they also have cloud subscription defining access policies and resources limit over the infrastructure.
- The infrastructure (the cloud) are characterized considering access policies, processing (VMs scale up and down), storage and memory limit.
- The data itself can be tagged considering security issues (privacy and trust), quality (veracity, freshness, provenance, degree of rawness) and the type of data.

We believe that the current SLA models can be extended to cover these aspects. Thus, given a consumer query and his/her integration requirements, the integration process deals with a multi-dimensional matching problem taking into account the different contracts established between consumers and producers with their associated cloud infrastructure.

Thus, the **first challenge** is to compute what we call an integrated SLA that matches the user's integration preferences (including quality constraints and data requirements) with the SLA's provided by cloud services, given a specific user cloud subscription. In this context, matching the user' integration preferences with services can lead to an exhaustive search in the chain of SLAs and to deal with SLA incompatibilities. Furthermore, the matching deals also with heterogeneous SLA specifications (different schemata, different measures semantics and granularities).

The **second challenge** is to guide data integration taking into consideration the integrated SLA. Here, the data integration process includes (i) looking up services that can be used as data providers, and for services required to process retrieved data and build an integrated result; (ii) performing data retrieval, processing and integration and (iii) deliver results to the user considering her preferences (quality requirements, context and resources consumption). The integrated SLA guides services selection and filtering; it can help to control the amounts of data to retrieve and process according to consumption rights depending on the user subscription to the participating cloud providers and how to deliver data considering the user's context.

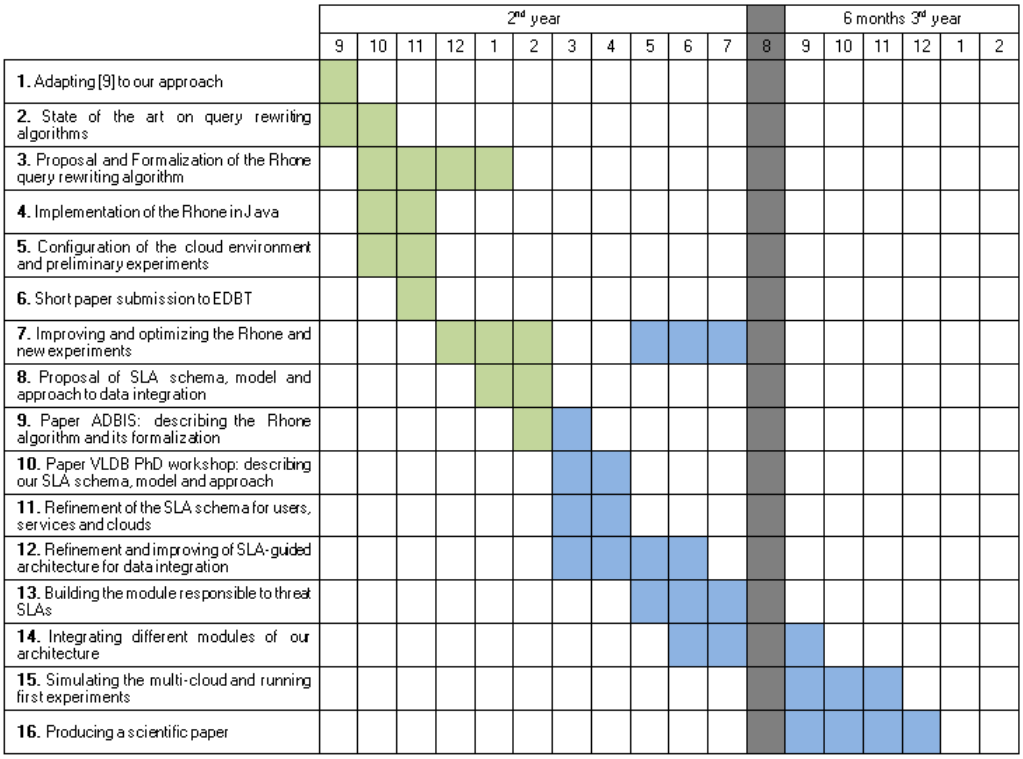
[1] The complete list of references can be found at: https://www.dropbox.com/s/98mith5xtn1dnb0/references-doctoral-school.pdf?dl=0

This thesis project intends to address data integration in a multi-cloud hybrid context. The originality of our approach consists in guiding the entire data integration solution taking into account (i) user preferences statements; (ii) SLA contracts exported by different cloud providers; and (iii) several QoS measures associated to data collections properties (for instance, trust, privacy, economic cost). The objective is to propose data integration strategies adapted to the vision of the economic model of the cloud. 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.

2 Synthesis and Perspectives of the Research Activities

During the second year, we have organized our research activities as follows:
Development of our data integration approach. Given a query and a set of user preferences associated to it, the query execution process is divided in three phases. The first phase is the *SLA derivation* in which a SLA for the user request is created. It consists in looking for a (stored, integrated) SLA derived for a similar request. If a similar SLA is found, the request is forwarded to the query evaluation phase. Otherwise, a new SLA to the integration (called integrated SLA) is produced. The query is expressed as a service composition with associated user preferences. In the second phase, service composition, the query is rewritten in terms of different services considering the user preferences and the SLAs of each service involved in the composition. The rewriting result is stored for further uses. Finally, in the query evaluation phase, the query is optimized in terms of user preferences and SLAs concerning the consumed resources and the economic cost of the query. Once optimized, the query processed in the execution engine. In addition, we are assuming a SLA management module and monitoring system responsible to verify if the SLA contracts are being respected.
Extension of the query rewriting algorithm. In collaboration with our colleagues in Brazil (authors in [1]), we have worked on an adapted version of [1] to our data integration solution extending their data structure to map services to the query, and adding the concepts of user preferences to the query and quality measures to the services. In addition, we have developed and formalized the *Rhone* service-based query rewriting algorithm guided by service level agreements (SLA). Our work proposes two original aspects: (i) the user can express her quality preferences and associated them to his query; and (ii) service's quality aspects defined in SLAs guide the service selection and the whole rewriting process taking into consideration that services and rewritings should meet the user requirements, and the different cases of incompatibilities of SLAs, uncompleted SLA and the integration SLA. Preliminary experiments were produced to evaluate the *Rhone*.
Publications. D. A. S. Carvalho, P. A. Souza Neto, G. Vargas-Solar, N. Bennani, C. Ghedira, Rhone: a quality-based query rewriting algorithm for data integration, Short paper, *20th East-European Conference on Advances in Databases and Information Systems*, ADBIS 2016 (to appear).

Currently, we have been working on the SLA model to data integration, on the schema for user SLA, cloud SLA and integration SLA, and on a scenario description to illustrate our approach. Our perspectives and intended calendar considering the project are described in the figure:
https://www.dropbox.com/s/risn3fo2ostgrbk/calendar-perspectives-second-year.PNG?dl=0.


2nd year
6 months 3rd year
9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2
1. Adapting [9] to our approach
2. State of the art on query rewriting algorithms
3. Proposal and Formalization of the Rhone query rewriting algorithm
4. Implementation of the Rhone in Java
5. Configuration of the cloud environment and preliminary experiments
6. Short paper submission to EDBT
7. Improving and optimizing the Rhone and new experiments
8. Proposal of SLA schema, model and approach to data integration
9. Paper ADBIS: describing the Rhone algorithm and its formalization
10. Paper VLDB PhD workshop: describing our SLA schema, model and approach
11. Refinement of the SLA schema for users, services and clouds
12. Refinement and improving of SLA-guided architecture for data integration
13. Building the module responsible to threat SLAs
14. Integrating different modules of our architecture
15. Simulating the multi-cloud and running first experiments
16. Producing a scientific paper

Bibliography

[1] Cheikh Ba, Umberto Costa, Mirian H. Ferrari, Rémy Ferre, Martin A. Musicante, Veronika Peralta, and Sophie Robert. Preference-driven refinement of service compositions. In *Int. Conf. on Cloud Computing and Services Science, 2014*, Proceedings of CLOSER 2014, 2014.

[2] Gianluca Correndo, Manuel Salvadores, Ian Millard, Hugh Glaser, and Nigel Shadbolt. SPARQL query rewriting for implementing data integration over linked data. In *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*, New York, New York, USA, 2010. ACM Press.

[3] Ghada ElSheikh, Mustafa Y. ElNainay, Saleh ElShehaby, and Mohamed S. Abougabal. SODIM: Service Oriented Data Integration based on MapReduce. *Alexandria Engineering Journal*, 2013.

[4] Yuan Tian, Biao Song, Jimupimg Park, and Eui nam Huh. Inter-cloud data integration system considering privacy and cost. In *ICCCI*, volume 6421 of *Lecture Notes in Computer Science*, pages 195–204. Springer, 2010.

[5] Stephen S. Yau and Yin Yin. A privacy preserving repository for data integration across data sharing services. *IEEE T. Services Computing*, 1(3):130–140, 2008.