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

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 and how much is the user ready to pay for the resources necessary for integrating her expected result. Moreover, data provision can be done by services according to different data quality measures. Such measures describe the conditions in which a service can provide or process data. These measures can be expressed in a service level agreement (SLA). An SLA states, what the user can expect from a service or system behavior. For example, whether it implements an authentication process, if it respects data consumer's privacy and the quality of the data the service can deliver, like freshness, veracity, reputation and other non-functional conditions like the business model that controls data delivery.

Data provision and data processing services may need a considerable amount of storage, memory and computing capacity that can be provided by cloud architectures. Their SLA includes the measures about the cloud services that they require to execute their requests. The cloud, itself exports a general SLA that specifies the conditions in which users can access the services (infrastructure, platform and software) deployed in it. A user willing to use the cloud services establishes a contract with the cloud provider guided by an economic model that defines the services she can access, the conditions in which they can be accessed (duplication, geographical location) and their associated cost. Different cloud providers have different possible contracts to establish with users (i.e., platinum, silver, gold, ivory users). Thus, for a given requirement, a user could decide which cloud services (from one or several cloud providers) to use for retrieving, processing and integrating data according to the type of contracts she can establish with them.

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).

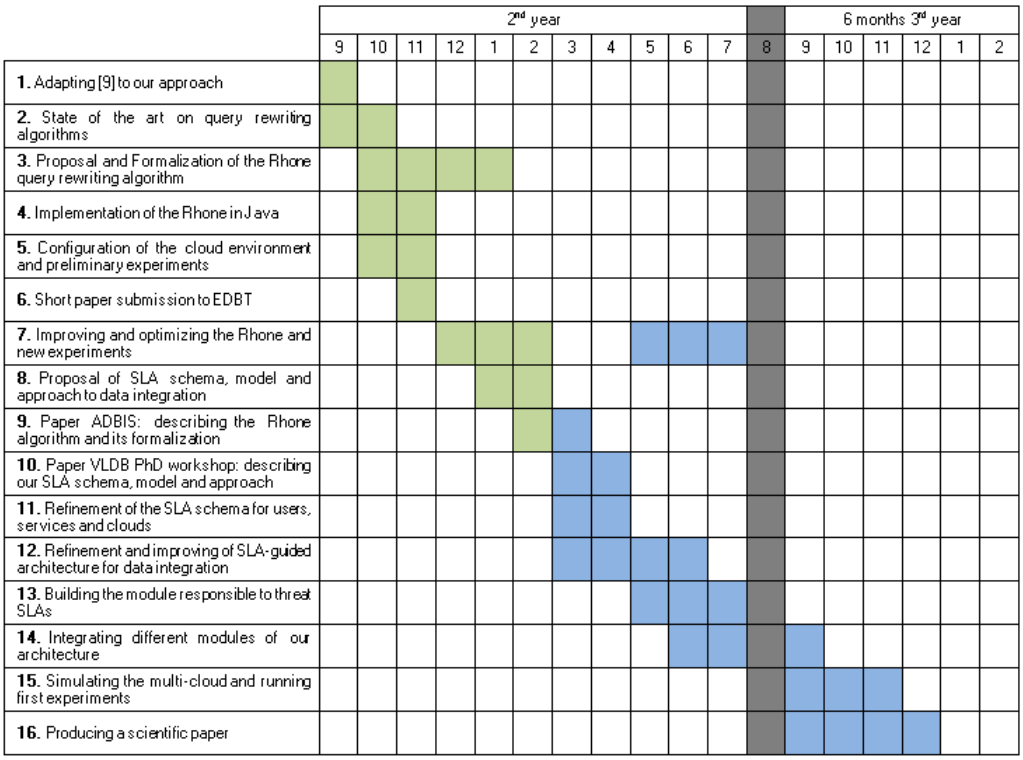
The **second challenge** is to guide data integration taking into consideration the integrated SLA. Here, the data integration process includes (i) looking up services that can be used as data providers, and for services required to process retrieved data and build an integrated result; (ii) performing 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.

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*.

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.
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).



The figure below presents the perspectives described as activities in the following calendar.

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.