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

Cloud architectures have been widely used for hosting different types of resources and applications. Current data integration solutions imply consuming data from different data services, integrating the results, and delivering them to the consumer in a homogeneous way. Nowadays, these services could be deployed in different clouds (multi-cloud) profiting from the elasticity, distributed processing and *pay-per-use* model imposed by this scenario. However, the multi-cloud context adds another complexity to the integration process, as it imposes and introduces new constraints and characteristics associated to data consumer, data producers, the cloud infrastructure and the data itself. Such constraints and characteristics are related to data security, data quality, processing and memory limits, storage, budget, among others.

Thus, given a consumer query, the integration process in this context should match and take into consideration all the constraints and characteristics of each entity (data consumer, data producers, the infrastructure and the data) while delivering the results to the consumer. To achieve this, we believe that service level agreements (SLA) can be used to unify all constraints and requirements information to serve as a meaning for the integration. However, the current SLA models are not able to cover all the integration aspects and the new issues in this process. Due to this reason, the first challenge is to propose a new kind of SLA (called Integration SLA) that matches the user's integration preferences (including constraints and characteristics) with the SLA's provided by cloud services, given a specific user cloud subscription. In this context, the matching process can lead to (i) an exhaustive search in the chain of SLAs; (ii) deal with SLA incompatibilities; and (iii) deal with heterogeneous SLA specifications (different schemata, different measures semantics and granularities).

Another 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 his/her requirements (for example, concerning quality, freshness, context and resources consumption). The integrated SLA guides services selection and filtering considering all the constraints and characteristics of each actor. In addition, it helps to control the amount of data to be retrieve and processed according to the user's consumption rights.

The objective of this PhD is to propose data integration strategies adapted to the economic model of the cloud taking into account all the constraints and characteristics imposed by the multi-cloud context.

2. Synthesis of the Research Activities

During the second year, we have organized our research activities as follows:

1. **Development of our data integration approach**. We have been working on our data integration approach, which is briefly described as follows. Given a query and a set of user requirements associated to it, the query execution process is divided in three phases. The first phase creates a SLA for the user request. It consists in looking for a (stored) integration SLA for a similar request. If a similar SLA is found, it is reused and the request is forwarded to the query evaluation phase. Otherwise, a new SLA to the integration (called integration SLA) is produced. The query is expressed as a service composition with associated user preferences and constraints. 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.
2. **Extension of the query-rewriting algorithm**. In collaboration with our colleagues in Brazil (authors in [1][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. The algorithm extension proposes two original aspects: (i) the user can express his/her quality preferences and associated them

[1] Ba, C., Costa, U., H. Ferrari, M., Ferre, R., A. Musicante, M., Peralta, V., Robert, S.: **Preference-driven refinement of service compositions**. In: Int. Conf. on Cloud Computing and Services Science. Proceedings of CLOSER 2014.

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*.

3. **SLA Model**. We have been working on the SLA model to data integration. The model considers different entities (such as data producers, data consumers, the infrastructure and the data). Each concept has its constraints and characteristics: (i) data producers describe the type of data they produce, production rate and time, location, cost, and they have an associated SLA defining access policies and resources limits over the infrastructure; (ii) data consumers have could subscriptions defining access policies, resources limits, but they also describe their requirements in terms of data quality, budget and privacy aspects; (iii) the infrastructure is characterized in terms of processing limit, storage, memory and access policies; and (iv) the data itself can be tagged considering its type, security issues, and quality aspects (such as provenance, freshness, degree of rawness, and veracity). Each concept and its constraints and characteristics are taken into account during the integration process. In the next step, we are going to analyze how the constraints and characteristics can be represented on SLA in order to be mapped and matched during the integration.
4. **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).
    - In addition, we are working on another paper to be submitted to ICSOC PhD Symposium (8 June).
5. **Presentations**. In order to have an external feedback, we have presented our work during the monthly group meeting in the SOC-Team research group from Lyon 1.

3. Perspectives

Currently, SLA incompatibilities are not taken in account. We are working on this issue, enriching our model and approach. We have also to work and study heuristics to be applied in our algorithm in order to reduce the composition search space making in this sense the integration more efficient. Moreover, it is necessary to analyze how the query execution plan should be parallelized to let the execution more efficient in a multi-cloud environment. Finally, evaluate and validate the entire quality-based data integration approach in a multi-cloud configuration. The calendar below describes our intended activities.