Thesis Advancement Report 2016-2017 (Third 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

User-guided data integration implies consuming data by matching and composing different data services and integrating the results while respecting data consumers quality requirements. These requirements could be defined in service level agreement (SLA) contracts established between data consumers and data providers. The SLA defines what a data consumer can expect as system behavior, but also the properties of the data such as its provenance, veracity, freshness, whether the consumer accepts to pay for data, and how much is he/she ready to pay for the resources necessary for integrating his/her expected result.

Several authors have introduced algorithms for data integration in service-oriented architectures (such as [Barhamgi 2010, Costa 2013, Benouaret 2011, Ba 2014]). They have focused on the query-rewriting problem, which in this context includes matching, selecting and composing services according to some data requirements and constraints. The problem of composing services has been proven to be an NP-hard problem, which implies a performance problem while searching for service combinations. To tackle it, heuristics for computing and identifying the best services and compositions based on quality aspects have been proposed (such as [Cardoso 2004, Berbner 2006, Menascé 2008, Sasikaladevi 2014]). However, current work focus on services' properties neglecting data properties and the new constraints imposed by the service-oriented context (for example, a given data service could be out of resources according to what he agreed on SLA with his service provider).

The service-oriented context brings challenges to query rewriting in data integration solutions. Instead of taking into consideration only the user query and his/her requirements with respect to services' properties (such as percentage of availability and response time), the integration process must consider in addition the new constraints imposed by the context:

- User requirements may concern not only data services' properties but also quality requirements of the data, which is being provided (such as freshness, cost, provenance, data type, veracity among others).
- Data provision is constrained to the available computing resources agreed between data services and service providers on service level agreement (SLA) contracts. For example, a data service could have agreed to perform a limited quantity of requests per day.
- The data integration process requires a high level of computing resources while searching for services and producing compositions. The huge amount of data and data services in the service context increases even more the complexity of the solutions.

Concerning the aforementioned, current data integration solutions introduce a multi-dimensional matching problem that should take into account:

- Matching and selecting data services according to the data consumer requirements (concerning service and data's properties) with respect (i) to data services quality measures defined in SLA contracts; and (ii) to data services' available resources according to different SLAs that they have agreed with different service providers.
- Delivering results with respect to the data consumer requirements depending on the context which he/she consumes the data (for example, using a mobile phone)

Thus, this thesis addresses data integration on service-oriented environments. The aim is to propose a data integration approach targeting (multi-)cloud context in which data is delivered according to data consumers' expectations and financial constraints. The multi-cloud introduces a new vision to the problem considering that data services could deploy services in different cloud providers, under different quality conditions (with respect to performance and data properties) which are agreed in several SLA contracts under different pricing conditions.

Furthermore, the explosion in the number of services increases the complexity of service selection and combination, and the decision of which rewriting will be done according to the different quality aspects and cost. To achieve this, this thesis aims to fulfill these three main contributions:

- A new query rewriting algorithm which takes into account user requirements regarding data properties and service properties according to the deployment context (see appendix B);
- Enhancing the algorithm by proposing an heuristic-based approach to avoid the combinatorial problem while searching and composing services (Future work); and
- Reducing the rewriting frequency by building a framework that reuses previous data integration based on a query history (work in progress).

2. Synthesis of the Research Activities

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

Data integration meta-model and SLA schemas: to describe our context and to illustrate the approach, we have designed a meta-model for data integration. Current SLA schemas are focused on service properties aspects such as availability and response time. Thus, to better fit on the data integration requirements, we have proposed cloud SLA that is an agreement between a data service and a cloud provider, and a service SLA which is a new kind of agreement defined by data services exposing the properties of the data they provide. The results of this part of our work have been accepted in the ICSOC PhD Symposium 2016.

A method for service and composition selection: Current data integration implies dealing with a huge number of data services. Consequently, query rewriting activities become more expensive in terms of computing resources and generation time. In this scenario, it is mandatory to develop a method to identify the best services to produce the best compositions that can achieve the user satisfaction. Thus, we have started working in an heuristic to rank data services and compositions based on SLA measures concerning service properties (percentage of availability, response time, throughput and others) and data properties (data type, freshness, veracity, provenance and others).

Definition and formalization of a taxonomy of queries: We have defined and formalized a set of possible relations between queries, which differ in terms of abstract services, service properties and data properties.

A method for reusing queries: It is well known that query rewriting is expensive. Thus, based on the proposed query taxonomy, we have designed and formalized a reusability approach, which allows reusing data services and compositions from previous integration in order to profit from them.

Query history data model and implementation: The reusability approach is based on a history of previous integrated queries. While defining the query taxonomy and reusability issues, we have identified the key information that should be part of the history allowing a satisfactory reusability process. A query history data model, which includes queries, abstract services, data services and compositions, was designed using the collected information. Further, the Rhone algorithm was adapted to be in accordance with the model.

3. Perspectives

Our perspective is to tackle what is remaining in the contributions while writing the final thesis document. Our intended calendar presented below:

Activities:	**Feb - April**			**May - July**			**Aug - Dec**				
Formalization of the query taxonomy and reusability functions	🟦	🟦	🟦								
Implementing reusability functions and including in the *Rhone*		🟦	🟦								
Building a proof of concept to the approach		🟦	🟦	🟦	🟧	🟧	🟧				
Writing a scientific paper			🟦			🟧	🟧				
Optimizing the first version of the approach				🟦	🟧						
Heuristics for optimizing the service selection and composition					🟧	🟧					
Writing the final thesis document				🟧	🟧	🟧	🟧	🟧	🟧		
Thesis Defense											🟧