Thesis Advancement Report 2014-2015 (First 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. Objectifs et l'Originalité du sujet de thèse (2 pages maxi)

Data integration is a widely studied issue in the database domain. It consists in merging data from different databases and providing a unified view of this data to the user [16]. Commonly, data integration is referred in the literature as a problem of answering queries using views. Many authors have reported their algorithms for this purpose [14]. Levy *et al.* proposed the *bucket algorithm*. It builds a *bucket* for each subgoal in the query. A *bucket* includes all views that can mapped to a given subgoal. Then, the algorithm generates combinations of the different buckets produced, and checks wheter each one is a rewriting of the query. Duschka and Genesereth introduced the *inverse-rules algorithm* [9]. It produces a set of inverse rules (one for each subgoal in the view) from local views to the global view. Rewritings are obtained by unfolding the query in terms of the inverse rules produced. Pottinger and Halevy presented the *MiniCon algorithm* in [20]. It identifies a set of views that contains query subgoals. Once a views is selected, the algorithm creates mapping from the query subgoal to the view subgoal if the mapping of variables are possible. These mappings are named MiniCon description (MCD). Finally, a rewriting is a combination of MCDs covering all query subgoals and satisfying the relations between predicates. In general, algorithms in this domain share the same performance problem while combining views to produce a rewriting. Depending on the size of the query and the amount of available views, these algorithms requires a lot of computing resources and time to process the rewriting and integration. New architectures like the cloud open challenges to data integration.

Data integration can be seen in the cloud computing as a service composition problem. Selecting services and producing service compositions is computationally costly. Moreover, executing compositions can lead to retrieve and process data collections that can require an important amount of memory, storage and computing resources. Data integration solutions on the service-oriented domain deal with query rewriting problems. In our previous work [8], we have identified trends and open issues regarding the use of SLA in data integration solutions on multi-cloud environments. Barhamgi *et al.* proposed a query rewriting approach which processes queries on data provider services [5]. The query and data services are modeled as RDF views. A rewriting answer is a service composition in which the set of data service graphs fully satisfy the query graph. Benouaret *et al.* introduced a service composition framework to answer preference queries [6]. In that approach, two algorithms based on [5] are presented to rank the best rewritings based on previously computed scores. Ba *et al.* presented an algorithm based on *MiniCon* that produces and order rewritings according to user preferences [4]. The user preference concept is a score used to rank the order in which services are selected. As in the database domain, these approaches requires an important amount of resources to process the rewriting and integration. It is important to review the existing data integration studies to be adapted to the cloud model.

The cloud computing paradigm provides computing resources (as services) in an on-demand and scalable manner to cloud consumers. This on-demand provision and the pricing model imposed by the cloud changes the way data integration solutions should be tackled. Instead of designing processes and algorithms taking into account the limits on resources availability, the cloud sets focus on the economic cost allowing resource consumption and producing results while delivering data under subscription oriented cost models.

When a service is billed to a consumer, the provider and customer agree on price, quality guarantees and penalties associated to its violation in service level agreement (SLA) contracts. Several researches have reported their studies on SLA in different domains [2]. In the cloud context, Rak *et al.* proposed an approach to specify security requirement and to associate them to cloud services [21]. Mavrogeorgi *et al.* introduced a SLA management framework that allows the creation and enforcement of customized SLAs [18]. Leitner *et al.* presented a approach to monitor and predict SLA violations before they impacted the provider's SLA [15]. In general, proposals regarding SLAs in the deployment of services focus on two aspects: (*i*) approaches focusing on the life cycle of the SLA mainly interested in the contract negotiation phase between the cloud and the service consumer; and (*ii*) works monitoring contracts and cloud resources in order to avoid SLA violations, and consequently penalties due to its violation. In this sense, to the best of our knowledge, we have not identified any other approach that proposes the use of SLA associated to a data integration solution in a multi-cloud environment.

Computing resources are delivered as services by the cloud. Services are billed and agreed between service providers and service customers under service level agreement (SLA) contracts. Both sides must agree together on quality conditions and penalties under which the service is delivered.

SLA proposals for cloud computing could be divided in two groups: (i) works developing tools and methods to help on SLA negotiation and enforcement phase [21, 18]; and (ii) approaches that monitor contracts and cloud resources in order to detect and avoid SLA violations [15, 17]. To our knowledge, SLA publications have not been yet integrated to data integration in a multi-cloud environment.

Considering the aforementioned, 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 process 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 (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 labelled 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. The project development is occurring naturally and well. In fact, this happens due to several aspects: (i) the close relationship and the easy access to my directors; (ii) the good frequency of meetings. Normally, we have at least one meeting for a week (that can be web conferences) in order to evaluate the status of the work; (iii) the research center infrastructure and team; and (iv) the monthly group meetings where we can discuss about the development of the projects with other colleagues. In addition, there are meetings in another laboratory, LIRIS. These moments are important because we can have an external view and comments about our research.

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.

The main contributions of the area are: (i) providing a global representation of heterogeneous data by defining a schema (e.g., global and local as view approaches); (ii) tagging data with meta-data or by associating them to knowledge (e.g. semantic Web approaches); and (iii) architectures used for integrating data (i.e. distributed databases, multi-databases, federated databases, etc). The emergence of cloud computing and service oriented computing opens new challenges to data integration. The possibility of an unlimited access to resources changes the problems associated to data processing. Cloud-based data management using data sharing enables the collaboration of different entities to perform design tasks [12, 13]. Data processing and analytics are costly tasks that can benefit from the cloud elastic resources provision, coupled with programming paradigms like Map-Reduce. [11] proposes SODIM that works on a pool of collaborative services and can process a large number of databases represented as web services.

In the cloud scenario resources are not necessarily located in the same cloud. One cloud cannot be expected to provide the necessary resources to fulfill application requirements. With growing needs and requirements, applications use different cloud providers for externalizing different data processing and management resources adding more challenges to data integration, considering the large amount and diversity of data, user quality and security requirements of the integration, and cloud heterogeneity in expressing and enforcing the corresponding clauses [10].

In cloud computing, a common way of defining requirements and obligations between the *provider* and *customer* is through service level agreement (SLA). SLAs have been adopted in the cloud, focussing (i) on the lifecycle of a security SLA on hybrid clouds [7]; (ii) on SLA models for addressing management

capabilities as a service, Pcloud services, performed through agreed and negotiated in contracts (elasticity, high availability, scalability and on demand provisioning) [3]; and (iii) on functional and non-functional requirements of the different cloud delivery models [1]. Summarizing, SLA contributions focus on: (i) the SLA negotiation phase; and (ii) resources monitoring and allocation to detect and avoid SLA violations. We identified one single approach regarding data integration in a grid environment guided by SLA [19].

2 Problem Statement

The problem addressed in our work is how can a user **efficiently** obtain results for her queries, **meeting her QoS requirements**, respecting all her subscribed contracts with the involved cloud provider(s), and without neglecting services contracts? Particularly, for queries that call several services deployed on different clouds.

Hypothesis: We assume that data are provided as services that export APIs with methods to retrieve and process data. Data integration is done (i) on a (multi)-cloud service oriented environment; (ii) under new conditions with respect to the type of data sources, the environment where it is performed and the preferences of data consumers and the SLA. We assume also that (iii) SLA measures can be monitored and negotiated in all cloud providers; and that (iv) cloud services and data services are listed in a registry.

We believe that data integration on multi-cloud environments can take advantage by integrating SLA on its solutions. To the best of our knowledge, we have not identified any other proposal adopting the use of SLAs combined with a data integration approach on a (multi)-cloud context.

3 Objectives

The objective is to propose a data integration solution in a multi-cloud environment guided by user preferences and SLA exported by different clouds. This new approach brings different challenges and open issues: (1) Identify and classify quality measures linked to data quality, data security and to cloud resources; (2) Propose an unified formalism to represent them; (3) Propose and implement a mechanism that ensures SLA within the data integration process performed on a multi-cloud and cope this with application requirements; and (4) Design a new matching-retrieving algorithm to perform the integration process, selecting the best service composition according to the user requirements and the SLAs.

4 Synthesis and Perspectives of the Research Activities

During the fist year we have organized our research activities in three groups (see below). These activities were organized and discussed in meetings with advisors and with individual work.

Problem statement and state of the art. The objective has been to acquire background knowledge on data integration, cloud and SLA building a corpus with publications and reading selected papers. Therefore, we applied the systematic mapping methodology consisting in retrieving papers from scientific databases, filtering them according to inclusion and exclusion criteria and research interests expressed in research questions. A classification schema was proposed, consisting in facets and dimensions. The abstracts of the final papers collection were read to classify each paper within the scheme. We identified the trends and open issues in our research topic and proposed the general lines of an original data integration solution according to current trends in the area.

Results: We built a collection of 114 papers and analytics results. We proposed a data integration classification scheme that serves as initial entry for building a state of the art.

Experimentation. We are currently configuring an experimentation platform on the cloud using Open Stack to implement and evaluate our match-retrieving algorithm [1].

Publications and thematic schools.

D. A. S. Carvalho, P. A. Souza Neto, G. Vargas-Solar, N. Bennani, C. Ghedira, Can Data Integration Quality be Enhanced on Multi-cloud using SLA?, Short paper, *In Proceedings of the 26th International Conference on Database and Expert Systems Applications*, LNCS, Valencia, Spain, 2015 (to appear)

Additionally, in April, I attended to the *1st French Brazilian School on Smart cities and Big Data* at the University of Grenoble Alpes (http://fr-br-school.imag.fr).

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

[1] You can check the detailed list of activities in https://www.dropbox.com/s/2cf6gncumzrjacd/sla-matching-experiment.docx?dl=0

Bibliography

[1] Mohammed Alhamad, Tharam Dillon, and Elizabeth Chang. Conceptual SLA framework for cloud computing. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 606–610. IEEE, April 2010.

[2] Mohammed Alhamad, Tharam S. Dillon, and Elizabeth Chang. A survey on sla and performance measurement in cloud computing. In *OTM Conferences (2)*, volume 7045 of *Lecture Notes in Computer Science*, pages 469–477. Springer, 2011.

[3] Ines Ayadi, Noemie Simoni, and Tatiana Aubonnet. SLA Approach for "Cloud as a Service". In *2013 IEEE Sixth International Conference on Cloud Computing*, pages 966–967. IEEE, June 2013.

[4] 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.

[5] M. Barhamgi, D. Benslimane, and B. Medjahed. A query rewriting approach for web service composition. *Services Computing, IEEE Transactions on*, 3(3):206–222, July 2010.

[6] Karim Benouaret, Djamal Benslimane, Allel Hadjali, and Mahmoud Barhamgi. FuDoCS: A Web Service Composition System Based on Fuzzy Dominance for Preference Query Answering, September 2011. VLDB - 37th International Conference on Very Large Data Bases - Demo Paper.

[7] Karin Bernsmed, Martin Gilje Jaatun, Per Hakon Meland, and Astrid Undheim. Security slas for federated cloud services. *2012 Seventh International Conference on Availability, Reliability and Security*, 0:202–209, 2011.

[8] Daniel A. S. Carvalho, Plácido A. Souza Neto, Genoveva Vargas-Solar, Nadia Bennani, and Chirine Ghedira. *Database and Expert Systems Applications: 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part II*, chapter Can Data Integration Quality Be Enhanced on Multi-cloud Using SLA?, pages 145–152. Springer International Publishing, 2015.

[9] Oliver M. Duschka and Michael R. Genesereth. Answering recursive queries using views. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '97, pages 109–116, New York, NY, USA, 1997. ACM.

[10] Schahram Dustdar, Reinhard Pichler, Vadim Savenkov, and Hong-Linh Truong. Quality-aware service-oriented data integration: Requirements, state of the art and open challenges. *SIGMOD Rec.*, 41(1):11–19, April 2012.

[11] Ghada ElSheikh, Mustafa Y. ElNainay, Saleh ElShehaby, and Mohamed S. Abougabal. SODIM: Service Oriented Data Integration based on MapReduce. *Alexandria Engineering Journal*, 52(3):313–318, September 2013.

[12] Hector Gonzalez, Alon Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, and Warren Shen. Google fusion tables: Data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, pages 175–180, New York, NY, USA, 2010. ACM.

[13] Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. Google fusion tables: Web-centered data management and collaboration. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1061–1066, New York, NY, USA, 2010. ACM.

[14] Alon Y. Halevy. Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294, December 2001.

[15] P. Leitner, A. Michlmayr, F. Rosenberg, and S. Dustdar. Monitoring, prediction and prevention of sla violations in composite services. In *Web Services (ICWS), 2010 IEEE International Conference on*, pages 369–376, July 2010.

[16] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA, 2002. ACM.

[17] A. Maarouf, M. El Hamlaoui, A. Marzouk, and A. Haqiq. Combining multi-agent systems and mde approach for monitoring sla violations in the cloud computing. In *Cloud Technologies and Applications (CloudTech), 2015 International Conference on*, pages 1–6, 2015.

[18] N. Mavrogeorgi, V. Alexandrou, S. Gogouvitis, A. Voulodimos, D. Kiriazis, T. Varvarigou, and E. K. Kolodner. Customized slas in cloud environments. In *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on*, pages 262–269, Oct 2013.

[19] Tiezheng Nie, Guangqi Wang, Derong Shen, Meifang Li, and Ge Yu. Sla-based data integration on database grids. In *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, volume 2, pages 613–618, July 2007.

[20] Rachel Pottinger and Alon Halevy. Minicon: A scalable algorithm for answering queries using views. *The VLDB Journal*, 10(2-3):182–198, September 2001.

[21] M. Rak, N. Suri, J. Luna, D. Petcu, V. Casola, and U. Villano. Security as a service using an sla-based approach via specs. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, volume 2, pages 1–6, 2013.