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 Context[1]

The data integration is a well-known and widely studied problem in the database domain. It consists in merging data from different data sources and granting a unified view [8]. 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 [6, 7]. Data processing and analytics are costly tasks that can benefit from the cloud elastic resources provision, coupled with programming paradigms like Map-Reduce. [5] 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 [4].

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 [3]; (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) [2]; 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 [9].

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

[1] You can find the references at https://www.dropbox.com/s/dyg08a622ucv6xl/references.pdf?dl=0

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 [2].

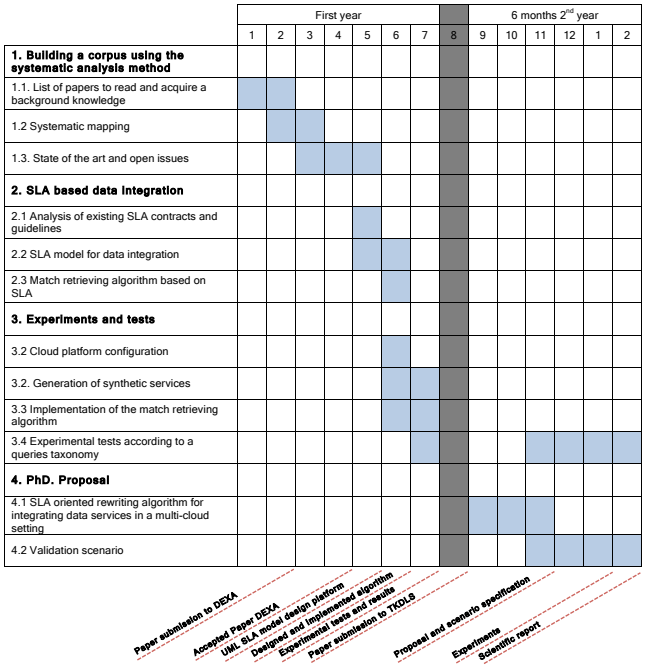
Publications and thematic schools.

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.

[2] 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.