Thesis Advancement Report 2014-2015 (First Year)

Thesis title: Trusted-SLA Guided on Multi-cloud Environments
PhD. student: Daniel Aguiar da Silva Carvalho
Supervisor: Chirine Ghedira-Guegan
Co-supervisors: Nadia Benani and Genoveva Vargas-Solar

1 Context

The data integration is a well-known and widely studied problem in the database area. It consists in merging data from different data sources and granting a unified view of the data [8]. In this area, the contributions can be divided in: (i) providing a global integrated representation of different data collections. This can be done either by defining a schema (e.g., global and local as view approaches), by tagging data with meta-data or by associating them to knowledge (e.g. semantic Web approaches); and (ii) The integration and the deployment architectures used for integrating data (i.e. Cloud, federated databases, etc). Cloud computing opens new challenges to data integration. The possibility of an unlimited access to resources that arises changes the way to process data. In this context, some data integration approaches have been proposed. [7] proposed a cloud-based data management and integration system. It enables data sharing, integration and collaboration between multiple users according to some design foundations (such as integration in the Web, incentives for sharing and facilitate collaboration). [6] described in detail the system architecture, integration process, query processing proposed by [7]. [5] combined data integration, service oriented architecture and distributed processing. The Service Oriented Data Integration based on MapReduce System (SODIM) works on a pool of collaborative services and can process a large number of databases represented as web services.

In the cloud scenario, one cloud cannot be expected to provide the necessary resources to fulfill application requirements. Therefore, applications have started to address different cloud providers for externalizing different data processing and management resources. This new (multi)-cloud configuration add more challenges to data integration, considering the large amount and diversity of data, and quality and security aspects of the integration. Data privacy is the most popular aspect in this context. [11] proposed a privacy-preserving repository in order to integrate data. Based on users' integration requirements, the repository supports the retrieval and integration of data across different services. [10] introduced an inter-cloud data integration system that considers a trade-off between users' privacy requirements and the cost for protecting and processing data. To be synthesized, other quality aspects of data integration services have been highlighted in [4].

In cloud computing, a common way of defining requirements and obligations between the *cloud provider* and *cloud customer* is through service level agreement (SLA). SLAs have been widely adopted in the cloud context. [3] presented a approach for security service level agreements on hybrid clouds focussing on the lifecycle of a security SLA, considering some security mechanisms (i.e. secure resource pooling, secure elasticity, access control, audit, verification and compliance, and incident management and response). [2] introduced a generic SLA model that includes management capabilities as a service which are agreed and negotiated in contracts. These management capabilities (elasticity, high availability, scalability and on demand provisioning) are performed by managing services called Pcloud services. The idea is to help the cloud customer to choose the appropriated providers that fits his requirements. [1] designs SLA based on functional and non-functional requirements of the different cloud delivery models.

Summarizing, the contributions can be divided in two groups: (i) approaches focussing on the SLA negotiation phase; and (ii) approaches for monitoring and allocation of resources to detect and avoid SLA violations. Among them, we identified one single approach regarding data integration in a grid environment guided by SLA [9].

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.

2 Problem Statement

We assume that data integration is done on a (multi)-cloud service oriented environment. We also consider that data integration is done 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. SLA measures can be monitored in all cloud providers and negotiated. Data are provided as services that export APIs to retrieve data and processing methods. We suppose that cloud services and data services are listed in a registry.

Let us show an example from the domain of energy management to illustrate our problem. For instance, we assume we are interested in queries expressed in an SQL-like language associated to a set of QoS preferences expressing the requirements of the user 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 problem here is how can the user efficiently obtain results for her queries, meeting her QoS requirements, respecting 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.

3 Objectives

The general objective of our work is to propose a data integration solution in a multi-cloud environment guided by user preferences and service level agreements (SLA) exported by different clouds. This new approach brings different challenges and open issues.

- In order to enhance data integration by integrating SLA we have to identify and classify quality measures associated to data and to cloud resources;
- Propose and implement a mechanism that ensures SLA within the data integration process which is performed on different clouds and cope this with application requirements; and
- 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

The research activities are organized in three groups: *problem statement and state of the art*, *setting an experiment platform*, and *publications and thematic schools*.

Problem statement and state of the art. During the first year, we have been working on the state of the art. The idea is to be aware of all types of publications close related to the thesis proposal. To reach this, we proceeded with a literature analysis using a systematic mapping methodology.

Briefly, the methodology consists in retrieving papers from scientific databases using the same search string. These papers are filtered according to an inclusion and exclusion criteria that should be defined based on the research interests. The papers will be classified in different categories (called facets) and for each facet in a specific dimension. The facets and dimensions are defined based on the authors knowledge and interests. Taking the final papers collection, the abstracts should be read in order to classify each paper into the dimensions for each facet. This methodology allowed us to identify trends and open issues regarding our research topic and proposing an approach that fills some gaps and proposes an original data integration solution according to current trends in the area.

As result to this work, we have produced a data integration classification scheme and retrieved a collection of 114 that builds the state of the art to the thesis.

Setting an experiment platform. In parallel with the state of the art analysis, we have been working on configuring a cloud environment to evaluate our approach [1].

Publications and thematic schools. Based on the publications extracted from the mapping process methodology, we have written an article that was **accepted** to the 26th International Conference on Database and Expert Systems applications (DEXA 2015). Additionally, in April, I attended to the *1st French Brazilian School on Smart cities and Big Data* at the University of Grenoble Alpes.

The figure below presents the intended calendar.

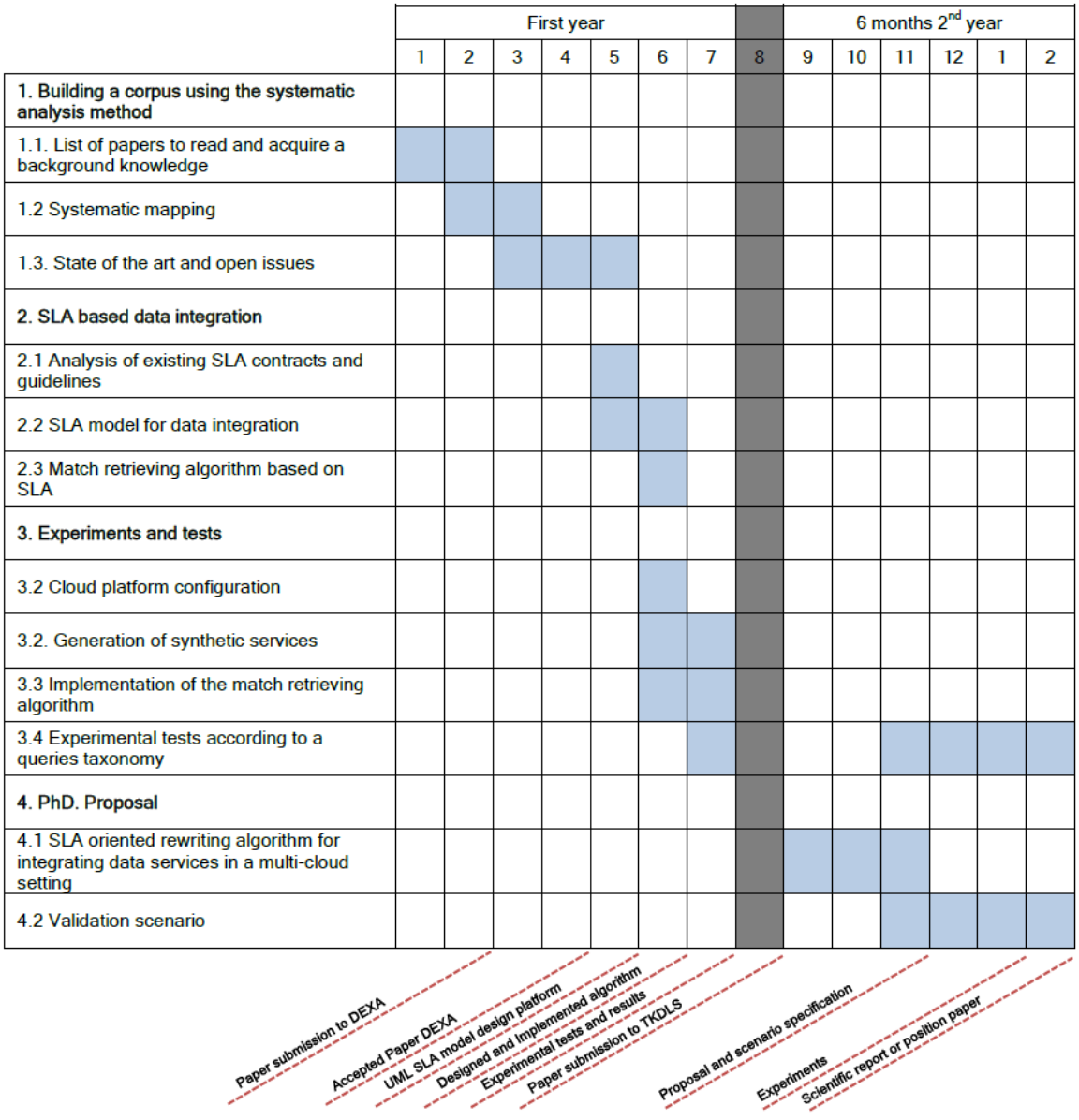

Figure 1: Calendar

[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] 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.

[10] Yuan Tian, Biao Song, Jimupimg Park, and Eui-Nam Huh. Inter-cloud data integration system considering privacy and cost. In Jeng-Shyang Pan, Shyi-Ming Chen, and NgocThanh Nguyen, editors, *Computational Collective Intelligence. Technologies and Applications*, volume 6421 of *Lecture Notes in Computer Science*, pages 195–204. Springer Berlin Heidelberg, 2010.

[11] Stephen S. Yau and Yin Yin. A privacy preserving repository for data integration across data sharing services. *IEEE T. Services Computing*, 1(3):130–140, 2008.