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

Context

We need one or two sentences describing the main results of the state of the art naming, schema integration, semantic mapping among data collections, data integration in the Web, distributed query evaluation and optimization, architectures like distributed db, multi-database, federated db, data warehouses.,

The data integration is a well-known and widely ~~discussed~~ *studied* problem in the database area. consists in merging data from different data sources and granting a unified view of the data [8]. ~~The~~ cloud computing opens new challenges to data integration. The possibility of an unlimited access to resources that arises ~~with the cloud model~~ *it is possible to process and integrate* 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 ~~with~~ *in* the ~~web~~ *Web*, ~~easy of use~~, incentives for sharing and facilitate collaboration), ~~the authors described the system and some examples of applications that can take advantages from~~ it. [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, it is ~~a hard task to one single cloud, deliver the resources necessaries to~~ *one cloud cannot be expected to provide the necessary resources to fulfill application requirements* ~~fit customer requirements. To avoid this, cloud providers began to share their computing resources.~~ *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. ~~The~~ data privacy is the most ~~tackled~~ *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. In general, 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) ~~contracts~~. SLAs have been widely adopted in the cloud context. [3] presented a approach for security service level agreements on hybrid clouds. ~~The author focus~~ *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 ~~management~~ *managing* services called Pcloud services ~~that are defined in order to achieve application requirements.~~ The idea is to help *the* cloud customer to choose the appropriated providers that fits his requirements. [1] ~~focused on the~~ *designs* ~~designing step of the~~ SLA. ~~Based~~ *based* on functional and non-functional requirements of the different cloud delivery models, ~~a conceptual framework for cloud computing is proposed. The objective is to enhance trust and reliability between the parts in the negotiation phase.~~

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

on multicloud environments

We believe that data integration can take advantages ~~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.

Problem Statement   *Problem needs further definition. Example OK as long as U use it to explain your 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.

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 question 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.

These are not hypothesis. This are the key aspects of our approach. You can present it as the first step towards an approach for SLA based data integration on multi-cloud environments. Concerning hypothesis: SLA measures can be monitored in all cloud providers and negotiated. Data are provided as services that export provided data retrieval and prcessing methods in APIs. We suppose that cloud services and data services are listed in a registry.

~~Hypothesis~~

We adopted the following hypothesis for defining the methodology of work:

- Data integration can take advantages ~~by~~ integrating SLA on its solutions on a (multi)-cloud context;

- ~~The key and original aspect of data integration and provision process considering SLA in a (multi)-cloud context can be defining a vertical mapping of user QoS preferences and agreed SLAs;~~

- ~~One possible solution can be defining a *derived SLA* that guides the evaluation of queries to be executed on the multi-cloud environment;~~

- ~~Have a framework for an integrated solution can possible add quality to our proposal.~~

Objectives

The general objective of our work is to propose

~~Based on the aforementioned works and our hypothesis, we are interested in proposing~~ 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. ~~So, the objective of my PhD thesis are:~~

In order to enhance data integration by integrating SLA we have to identify and classify quality measures associated to data and to cloud resources

- ~~We believe that we can enhance data integration quality by integrating SLA on its solutions. So, To reach this, it is necessary to identify which quality measures associated to data (that the data consumers what to query) and to the multi-cloud context should be present in SLA;~~

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.

- ~~Considering the nature of the multi-cloud environment where we dealing with private and public clouds, it is necessary to proposed a mechanism to grant quality and security aspects of the integration to the data consumer;~~

- ~~It is necessary~~ design a new matching-retrieving algorithms ~~in order~~ to perform the integration process, selecting the best service composition according to the user requirements and the SLAs;

- ~~We will develop an algorithm(s) and a framework to reach our target.~~

Synthesis and Perspectives of the Research Activities

During the first year of PhD, 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.

This work performed, we thus have written an article that was **accepted** to the 26th International Conference on Database and Expert Systems applications (DEXA 2015). We selected 114 papers based on our inclusion and exclusion criteria. This final data collection builds the state of the art to the thesis and, in the next step, we will perform an in-depth analysis of theses papers in order to (i) analyze what have been made and (ii) propose my approach.

Based on the publications extracted from the mapping process methodology, we will proceed the analysis of the current state of the art in order to formalize our proposal. The analysis will be the basis to our model proposal. As a natural result, we will write a paper describing our approach and a survey. In parallel, we will carry on the first steps of implementation of the proposed approach. The figre presents the intended calendar.

Organize activities into three groups indicating the results:
Say that you have regular meetings with advisors and
describe your methodology
1. Problem statement and state of the art
Here a brief description using the above text

RESULTS
Publications collection and associated analysis
Data integration classification scheme
2. Setting an experiment platform
Here a brief description making reference to the feuille de
route on onedrive
RESULTS
- Specification of the experiment: definition of synthetic
services, service directory
- SLA model
3. Publication and thematic schools
Description

- DEXA paper
- Attendance to thematic school
- Lectures you attended during the year

	First year								6 months 2nd year					
	1	2	3	4	5	6	7	8	9	10	11	12	1	2
1. Building a corpus using the systematic analysis method														
1.1. List of papers to read and acquire a background knowledge	▓	▓												
1.2 Systematic mapping		▓	▓											
1.3. State of the art and open issues			▓	▓	▓									
2. SLA based data integration														
2.1 Analysis of existing SLA contracts and guidelines					▓									
2.2 SLA model for data integration					▓	▓								
2.3 Match retrieving algorithm based on SLA						▓								
3. Experiments and tests														
3.2 Cloud platform configuration						▓								
3.2. Generation of synthetic services						▓	▓							
3.3 Implementation of the match retrieving algorithm						▓	▓							
3.4 Experimental tests according to a queries taxonomy							▓				▓	▓	▓	▓
4. PhD. Proposal														
4.1 SLA oriented rewriting algorithm for integrating data services in a multi-cloud setting									▓	▓	▓			
4.2 Validation scenario											▓	▓	▓	▓

Accepted paper DEXA 2015

UML SLA model design platform

Designed and Implemented algorithm

Experimental tests and Results

Paper submission to TKDLS

Proposal and scenario specification
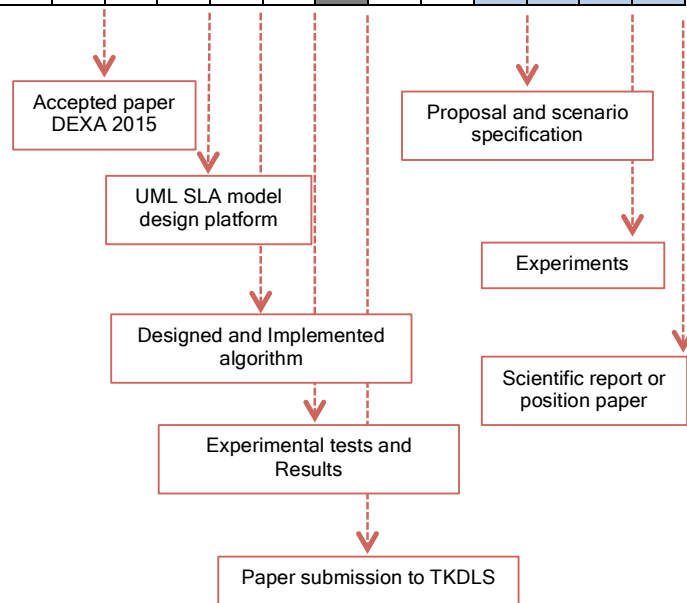
Experiments

Scientific report or position paper

Figure 1: Calendar

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.