Can Data Integration Quality be Enhanced on Multi-cloud using SLA?

Daniel A. S. Carvalho[1], Plácido A. Souza Neto[3], Genoveva Vargas-Solar[4], Nadia Bennani[2], Chirine Ghedira[1]

[1] Université Jean Moulin, Lyon 3 MAGELLAN, IAE – France
daniel.carvalho@univ-lyon3.fr, chirine.ghedira-guegan@univ-lyon3.fr
[2] CNRS INSA-Lyon, LIRIS, UMR5205 – France
nadia.bennani@insa-lyon.fr
[3] Instituto Federal do Rio Grande do Norte, Natal – Brazil
placido.neto@ifrn.edu.br
[4] CNRS, LIG-LAFMIA, Saint Martin d'Hères – France
genoveva.vargas@imag.fr

Abstract. The aim of this paper is to identify trends and open issues regarding the use of SLA associated with data integration solutions on multi-cloud environments. To reach the target, we performed a Systematic Mapping [3] concerning the aforementioned topics in order to analyze the way in which they are correlated. To do so, after retrieving scientific productions on the subject, a classification of the results has been done according to several facets. In this paper, we have considered a subset of them: (i) data integration environment (cloud; data warehouse; federated database; multi-cloud); (ii) data integration description (knowledge; metadata; schema); and (iii) data quality (confidentiality; privacy; security; SLA; data protection; data provenance; others). We combine the facets then expressed and analyzed the results as bubble charts.

Keywords: Systematic Mapping, Service Level Agreement, Data Integration, Multi-cloud Environment.

1 Introduction

The emergence of new architectures like the cloud opens new opportunities for data integration. The possibility of having unlimited access to cloud resources and the "pay as U go" model make it possible to change the hypothesis for processing big data collections. Instead of designing processes and algorithms taking into consideration limitations on resources availability, the cloud sets the focus on the economic cost implied when using resources and producing results by parallelizing their use while delivering data under subscription oriented cost models.

Integrating and processing heterogeneous huge data collections (i.e., Big Data) calls for efficient methods for correlating, associating, filtering them taking into consideration their "structural" characteristics (due to data variety) but

also their quality (veracity), e.g., trust, freshness, provenance, partial or total consistency. Existing data integration techniques need to be revisited considering weakly curated and modeled data sets provided by different services under different quality conditions. Data integration can be done according to (i) quality of service (QoS) requirements expressed by their consumers and (ii) Service Level Agreements (SLA) exported by the cloud providers that host huge data collections and deliver resources for executing the associated management processes. Yet, it is not an easy task to completely enforce SLAs particularly because consumers use several cloud providers to store, integrate and process the data they require under the specific conditions they expect. A major concern when integrating data from different sources (services) is privacy that can be associated to the conditions in which integrated data collections are built and share [5]. Naturally, a collaboration between cloud providers becomes necessary [1] but this should be also done in a user-friendly way, with some degree of transparency. On the other hand, the definition of a SLA extension seems to be a reasonable way to better guide data integration on a (multi-)cloud environment.

In this context, the contribution of our work is proposes a classification scheme of existing works fully or partially addressing the problem of integrating data in multi-cloud environments taking into consideration an extended form of Service Level Agreement.

The classification scheme results from applying the methodology defined in [3] called *systematic mapping* for defining a classification of a field. A classification consists of categories clustered into facets in which publications (i.e., papers) are aggregated according to frequencies (i.e., number of published papers). According to the methodology, the study consists in five interdependent steps including (i) the definition of a research scope by defining research questions; (ii) retrieving candidate papers by querying different scientific databases (e.g. IEEE, Citeseer, DBLP); (iii) selecting relevant papers that can be used for answering the research questions by defining inclusion and exclusion criteria; (iv) defining a classification scheme by analyzing the abstracts of the selected papers to identify the terms that will be used as categories for classifying the papers; (v) producing a systematic mapping by sorting papers according to the classification scheme.

The remainder of this paper is organized as follows. Section 2 describes our study of data integration perspectives and the evolution of the research works that address some aspects of the problem. It gives a quantitative analysis of our study and identifies open issues in the field. Section 3 concludes the paper and discusses future work with reference to the stated problem.

2 Data integration challenges: classification scheme

The aim of our bibliographic study using the systematic mapping methodology [3] is to (i) categorize and quantify the key contributions and the evolution of the research done on *SLA-guided data integration in a multi-cloud environment* and (ii) discover open issues and limitations of existing works. Our study is guided by three research questions:

RQ1: *Which are the SLA measures that have been mostly applied in the cloud?* This question will help to identify the type of properties used for characterizing and evaluating the services provided by different clouds.

RQ2: *How have published papers on data integration evolved towards cloud topics?* This question is devoted to identify the way data integration problems addressed in the literature started to include issues introduced by the cloud.

RQ3: *In which way and in which context has data integration been linked to Quality of Service (QoS) measures in the literature?* The objective of this question is to understand which QoS measures have been used for evaluating data integration and to determine the conditions in which specific measures are particularly used.

2.1 Search and screening of papers

According to our research questions and our expertise in data integration we chose a set of keywords to define a complex query to be used for retrieving papers from four target publication databases: IEEE [5], ACM [6], Science Direct [7] and CiteSeerX [8]. We used the following conjunctive and disjunctive general query which was completed with associated terms from a thesaurus and rewritten according to the expression rules of advanced queries in each database:

("Service level agreement" AND ("Data integration" OR "Database integration") AND ("Cloud" OR "Multi-cloud "))

We retrieved a total of 1832 publications. According to the systematic mapping methodology, the initial collection was cleaned and filtered according to inclusion and exclusion criteria applied as filters when analyzing the titles and abstracts of the papers. In general, we only kept publications written in English, addressing SLA models and languages, quality measures, and/or (multi)-cloud topics related to data integration. As a result of the filtering process we excluded 1718 publications. The number of papers included for building the final collection were 114 publications [9].

2.2 Defining classification facets

We analyzed the titles and abstracts of the papers derived in the previous phase using information retrieval techniques in order to identify the frequent relevant terms. We used these terms for building a classification scheme consisting of three facets that group frequent relevant terms. We added two facets for classifying the

[5] http://ieeexplore.ieee.org/
[6] http://dl.acm.org/
[7] http://www.sciencedirect.com/
[8] http://citeseerx.ist.psu.edu/
[9] List of references available in: https://github.com/danielboni/DEXA-2015-Can-Data-Integration-Quality-be-Enhanced-on-Multi-cloud-using-SLA.git

type of papers (e.g., position, survey, etc.) and the type of contributions (e.g., model, system, language, etc). According to the systematic mapping methodology, the relevant frequent terms in papers can be considered as dimensions that represent subcategories within the facets that group them. The following lines define the facets and dimensions of the classification scheme that we propose for studying SLA guided data integration in multi-cloud environments.

Data Integration Environment: This facet groups the dimensions that characterize the architectures used for delivering data integration services (*data warehouse* and *federated database*) and architectures used for deploying these services (*cloud* and *multi-cloud*).

Data Integration Description: This facet groups the dimensions describing the type of data used for describing the databases content in order to integrate them. Data integration can be done by using *meta-data, schema*, and *knowledge*.

Data Quality: This facet groups the dimensions that represent the parameters that can be used for measuring data quality. Measures can be related directly to data for instance *confidentiality, privacy, security, protection and provenance* and to the conditions in which data is integrated and delivered (i.e., dimension *SLA*).

The original vision of our classification scheme is that of adding the notion of *quality* to data integration represented by the facet *data quality* that groups the measures used for determining data quality that can be integrated in SLAs; and the facet *SLA* characterizing the way SLAs are expressed and associated to data integration (e.g., model, language). With these facets our classification scheme shows the aspects that must be considered when addressing data integration in the cloud taking into account (i) the quality of data, (ii) the systems that integrate data and (iii) the quality warranties that a data consumer can expect expressed in SLAs.

2.3 Quantitative Analysis

This section discusses the quantitative analysis that we propose as a result of applying the Systematic Mapping methodology. Quantitative results are aggregated and presented in bubble charts that combine different facets. In order to observe the evolution of the publication trends we defined a time screen between the years 1998 and 2014 (see Figure 1). SLA has emerged when Cloud issues started to be addressed around 2009 and publication has increased as cloud infrastructures have become more popular and accessible. It seems indeed that data integration is an open issue when it is combined with SLA and cloud trends. Less recent papers seem to be devoted to the way data is described under schemata or knowledge representation strategies, this could be due to the fact that these strategies are consolidated today and to the emergence of NoSQL approaches with their schemaless philosophy [4].

We combined facets for answering the research questions proposed for guiding our study. The following lines discuss the answers.

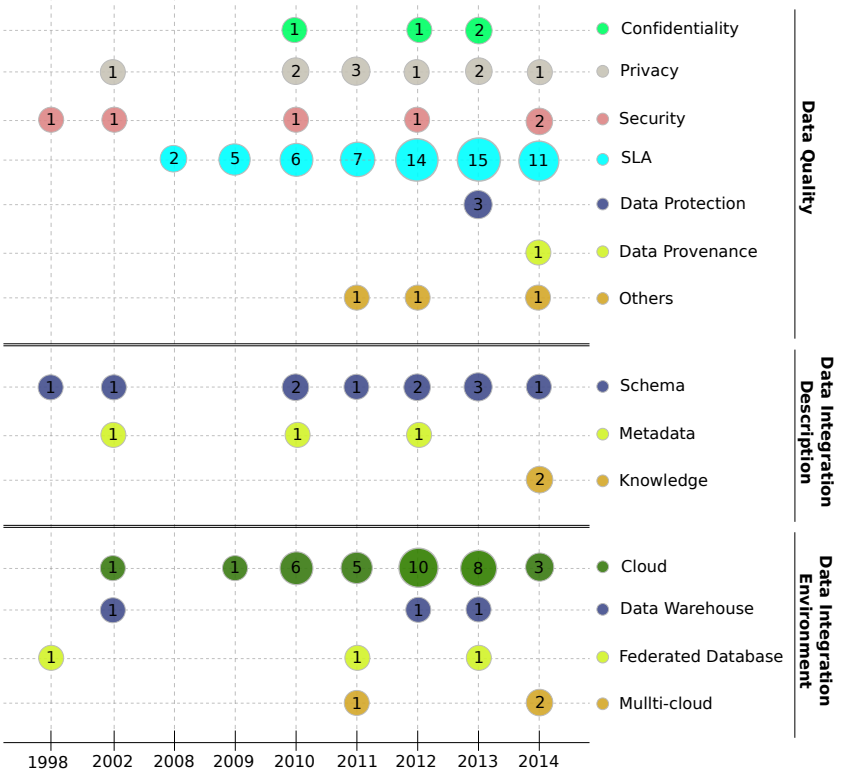

Fig. 1: Publications Per Year

RQ1: Which are the SLA measures that have been mostly applied in the cloud?

The facets SLA expression, data integration description and contribution give elements for determining which SLA measures have been applied to the cloud (Figure 2). The resulting bubble chart shows that most contributions propose SLA models and that *privacy* and *security* (11 papers - 9.65%) are the most popular measures considered by SLA models for the cloud. These measures concern the network, information, data protection and confidentiality in the cloud. Most contributions propose SLA models (53 papers - 46.49%) but some languages (8 papers - 7.02%) have also emerged. *Data provenance* is also a measure that emerges but only in papers dealing with multi-cloud environments. Data integration is merely addressed by using schemata (12 papers - 10.53%) and meta-data (4 papers - 3.51%) particularly through models (34 papers - 29.82%) and tools (25 papers - 21.93%). Still, some works propose surveys (8 papers - 7.02%).

RQ2: How have published papers on data integration evolved towards cloud topics?

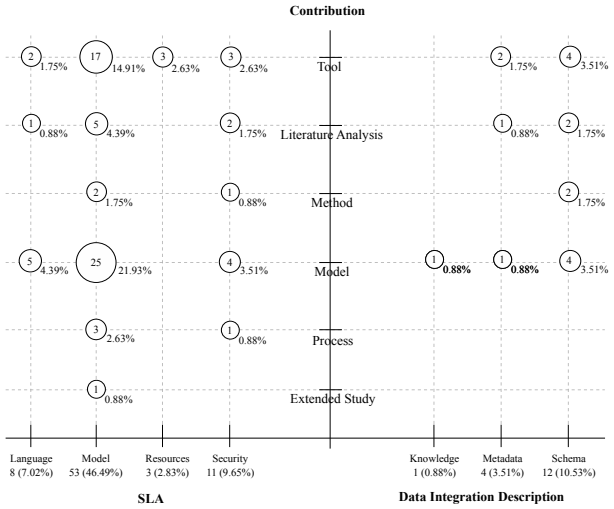

Fig. 2: Facets Contribution, SLA and Data Integration Description

Combining the facets data integration environment, contribution and research it is possible to observe the evolution of publications on data integration towards the cloud (Figure 3). *Data warehouse* environments are the most common architecture. This can be explained by the increase of scientific and industrial applications needing to build integrated data sets for performing analysis and decision making tasks. The proposals are delivered as *models* (14 papers - 12.27%) and *tools* (18 papers - 15.78%) used for facilitating data integration, mostly done in the *cloud*. The most popular deployment environment of recent papers is the *cloud*. Given the importance and crucial need of data integration most papers present concrete solutions as algorithms, methods and systems (31 papers - 27.19%).

RQ3: In which way and in which context has data integration been linked to QoS measures in the literature?

Combining the facet data quality with the facets data integration environment and data integration description (Figure 4) we answered RQ3. Particularly data integration and QoS measures are associated within environments like cloud (9.68%) and multi-cloud (4.39%).

According to our quantitative analysis we observe that QoS has started to be considered for integrating data, yet the type of measures and properties are diverse. They address measures related to the conditions in which data are accessed like security and privacy. Cloud is becoming a popular environment to perform data integration in which security issues are most frequently addressed. We identify a promising research area concerning the need of studying SLA which is currently addressed for the cloud as a whole [2] but that needs to be specialized for data integration aspects. Therefore, it is important to identify the

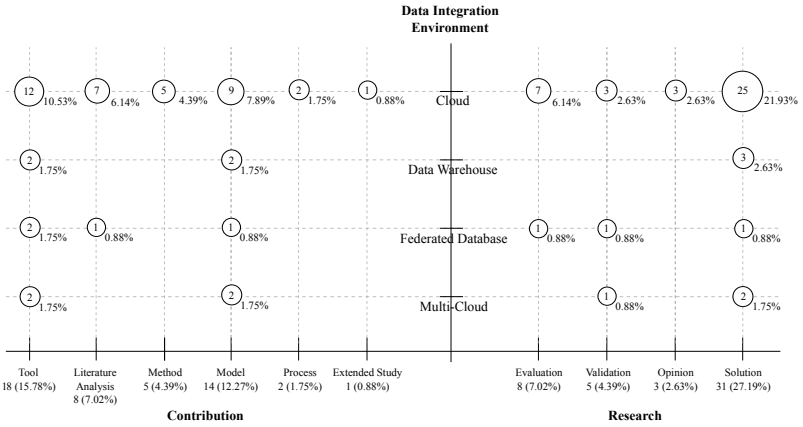

Fig. 3: Facets Data Integration Environment, Contribution and Research

measures that characterize the quality of data but also the quality measures associated to different phases of data integration: selecting data services, retrieving data, integrating and correlating them and building a query result that can be eventually stored and that must be delivered. These phases are implemented by greedy algorithms and generate intermediate data that can be stored for further use. Therefore they consume storage, computing, processing and communication resources that have an associated economic cost but that must ensure some QoS guarantees to data consumers. This problem seems to be open in the domain, and we believe that it must be part of a new vision of data integration. We believe that it is possible to add and enhance the quality of data integration by considering data integration based SLAs.

3 Conclusion and final remarks

This paper introduces the challenge of integrating data from distributed data services deployed on different cloud providers guided by service level agreements (SLA) and user preferences statement. The data integration problem is stated as a continuos data provision problem that has associated SLAs and that uses techniques for ensuring different qualities of delivered data (freshness, precision, completeness). The problem statement was derived from a classification scheme that resulted from a study of existing publications identified by applying the systematic mapping method. Our contribution is the definition of a classification scheme that shows the aspects that characterize a modern vision of data integration done in multi-cloud environments and that can be enhanced by including SLAs in its process.

Current big data settings impose to consider SLA and different data delivery models. We believe that given the volume and the complexity of query evaluation

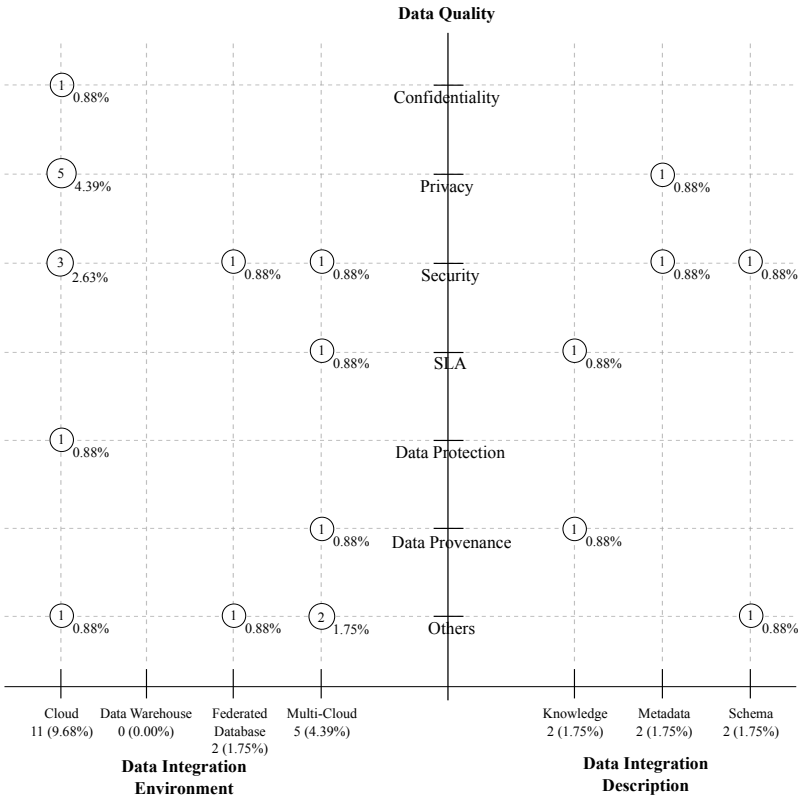

Fig. 4: Facets Data Quality, Data Integration Environment and Data Integration Description

that includes steps that imply greedy computations. It is important to combine and revisit well-known solutions adapted to these contexts.

From the results of our systematic analysis, (i) we identified 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; and (ii) we are also currently developing the strategies to better define a SLA extension and data consumers preferences description for guiding data integration in multi-cloud environments.

References

1. Mohamad Hamze, Nader Mbarek, and Olivier Togni. Self-establishing a Service Level Agreement within autonomic cloud networking environment. In *2014 IEEE Network Operations and Management Symposium (NOMS)*, pages 1–4. IEEE, May 2014.

2. Carlos Pedrinaci, Jorge Cardoso, and Torsten Leidig. Linked USDL: A vocabulary for web-scale service trading. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 68–82, 2014.
3. Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swinton, UK, UK, 2008. British Computer Society.
4. Pramod J Sadalage and Martin Fowler. *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education, 2012.
5. 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.