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. This paper identifies trends and open issues regarding the use of SLA in data integration solutions on multi-cloud environments. Therefore it presents results of a Systematic Mapping [8] that analyzes the way SLA, data integration and multi-cloud environments are correlated in existing works. The main result is a classification scheme consisting of facets and dimensions namely (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). The proposed classification scheme is used to organize a collection of representative papers and discuss the numerical analysis about research trends in the domain.

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.

Integrating and processing heterogeneous huge data collections (i.e., Big Data) calls for efficient methods for correlating, associating, and filtering them according to their "structural" characteristics (due to data variety) and their quality (veracity), e.g., trust, freshness, provenance, partial or total consistency.

Existing data integration techniques must 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. For example, 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 shared [11]. Naturally, a collaboration between cloud providers becomes necessary [4] but this should be done in a user-friendly way, with some degree of transparency.

In this context, the main contribution of our work is 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 SLA. The classification scheme results from applying the methodology defined in [8] called *systematic mapping*. It consists of dimensions 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 to be used as dimensions 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. Section 3 gives a quantitative analysis of our study and identifies open issues in the field. Section 6 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 [8] 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 identifies 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 Searching and screening 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. As a result of the filtering process proposed by the systematic mapping methodology [8] 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 to identify frequent terms. We used these terms for proposing a classification scheme consisting of three facets that group dimensions. The following lines define the facets and dimensions of the classification scheme we propose.

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 approaches used for describing the databases content in order to integrate them. Data integration can be done by using *meta-data, schema*, and *knowledge*.

[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

Data Quality: This facet groups the dimensions representing data quality measures. 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 facets *data quality* and *SLA*. 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.

3 Quantitative Analysis

This section discusses the quantitative analysis 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. The number of publications has increased as cloud infrastructures have become more popular and accessible. It seems 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 schema-less philosophy [9].

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

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?

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

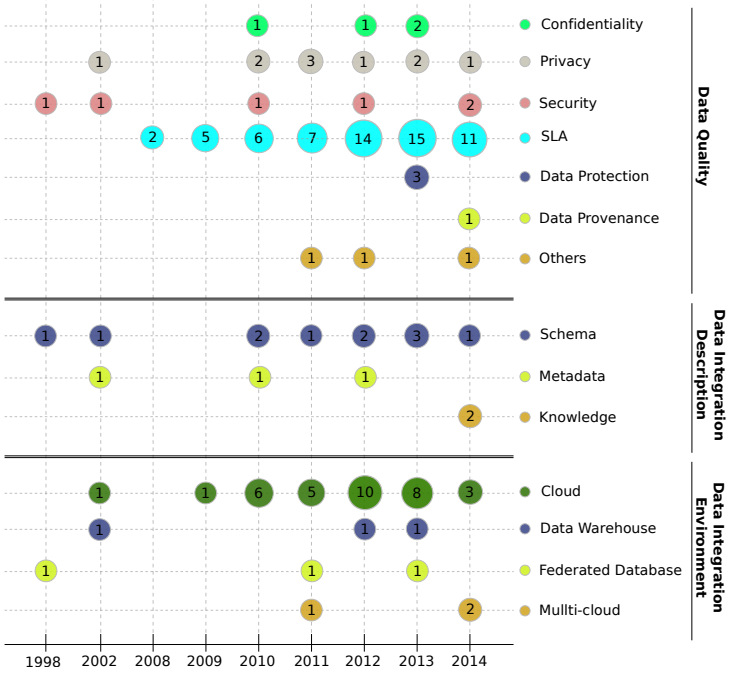

Fig. 1: Publications Per Year

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?

We answered RQ3 by combining the facet *data quality* with the facets *data integration environment* and *data integration description* (Figure 4). 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. The 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 [7] but that needs to be specialized for data integration aspects. Therefore, it is important to iden-

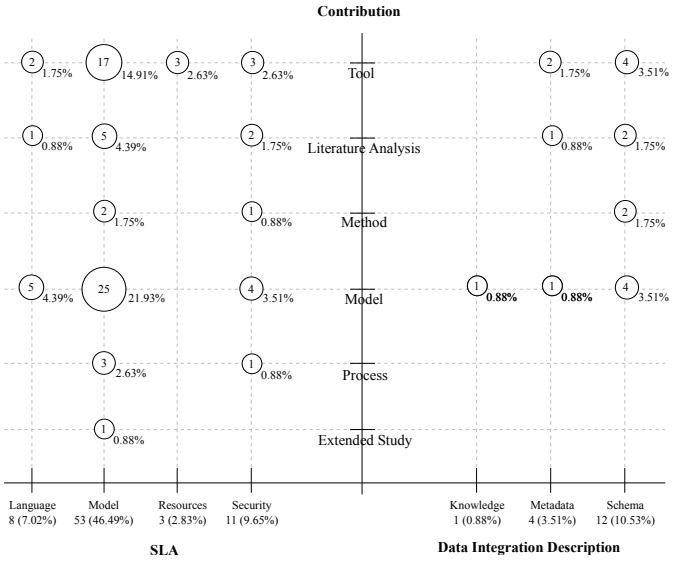

Fig. 2: Facets Contribution, SLA and Data Integration Description

tify the measures that characterize the quality of data and the quality measures associated to different phases of data integration. These phases include selecting data services, retrieving data, integrating and correlating them and building a query result that can be eventually stored and that must be delivered. The data integration 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. These resources 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 including SLAs.

4 Enhancing Data Integration on Multi-Cloud environments with SLA

In order to illustrate our vision, let us consider an example from the domain of energy management. 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 labeled as green?* We consider a simplified SLA cloud contract inspired in the cheapest contract provided by Azure: *cost of $0,05 cents per call, 8 GB of I/O volume/month, free data transfer cost within the same region, 1 GB of storage.* Suppose that the user is ready to pay a maximum of *$5 as total query cost*; she requests

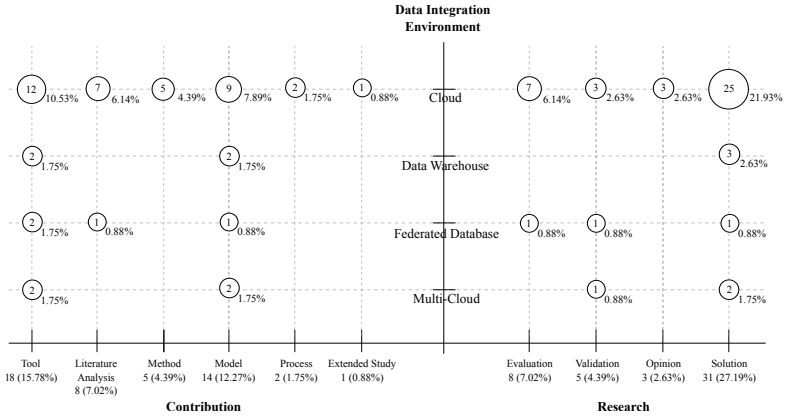

Fig. 3: Facets Data Integration Environment, Contribution and Research

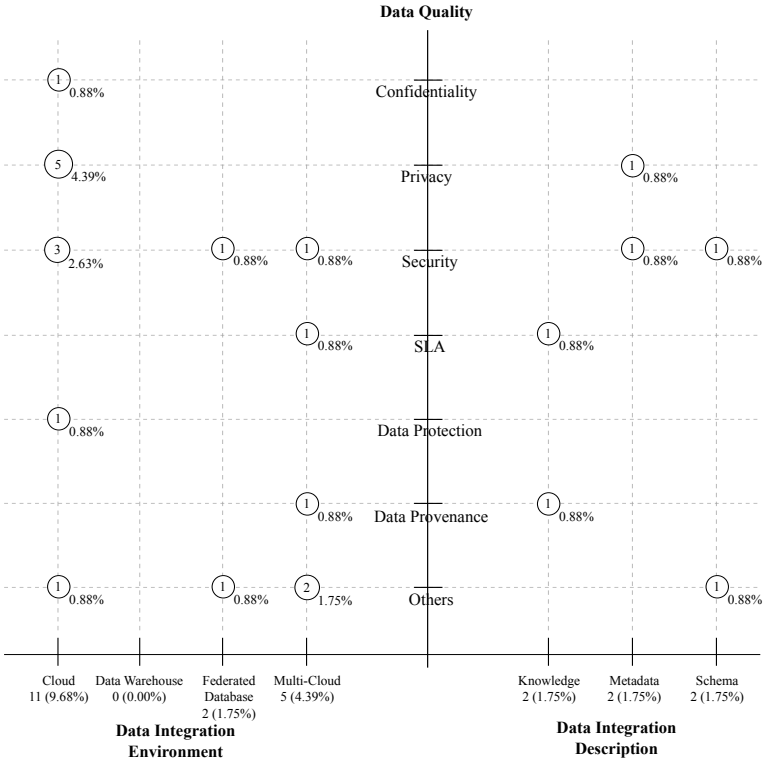

Fig. 4: Facets Data Quality, Data Integration Environment and Data Integration Description

that only *green* energy providers should be listed (provenance), with at least *85%* of precision of provided data, even if they are not fresh; she requires an availability rate of at least 90% and a response time of *0,01 s*. 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?

According to our classification scheme that resulted from our systematic mapping, we propose a new vision of data integration. This vision includes the description of the context in which data integration is done in modern environments. It also identifies the phases of the data integration process with their associated problems and challenges when they must include SLAs and QoS preferences expressed by data consumers.

4.1 Data integration context

We assume that data integration is done on a (multi)-cloud service oriented environment shown in Figure 5. We 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.

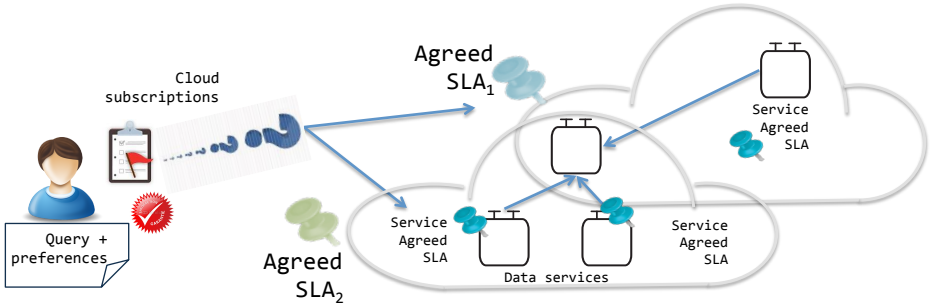

Fig. 5: New data integration context

There are data providers that are services possibly deployed in clouds and available through their API or in a REST architecture. We assume that each service exports an agreed SLA that specifies the economic cost per call, the maximum number of calls that can be done per day, the availability of the service, the average response time when a method is called, the reliability, the privacy of the produced data (whether they can be stored or not), the precision of their responses, freshness and provenance of the produced data.

Cloud providers define also their SLA contracts expressing subscription contracts that specify, the cost per request (cost/request), the volume of data that can be exchanged per month (I/0 volume/month), the cost of transferring data or applications within the same data centre or between data centres (datatransferCost/region), and storage space (storageSpace). For example some cloud providers

enable the customer to choose the zone to install PaaS services and deploy applications (e.g. zone 1 is Europe). If the customer wishes to deploy services in zone 1 but store data in zone 2 the transfer cost will change.

Some of these measures (cost/call, maxCall/day) are static and explicitly specified by the service provider. In contrast, the other measures should be computed by monitoring the conversations between the service and the applications that contact it.

In our vision a query expressed in an SQL-like language is associated to a set of QoS preferences expressing the requirements of the user. For example, the economic cost she is ready to pay for executing the query, the provenance of the data, the reputation of data services and the expected time response. The answer of such a query is the result of integrating data from different services according to a series of phases described in the following section.

4.2 SLA guided data integration

Given a query, its associated QoS preferences, cloud providers and services that can potentially be data providers (see Figure 5), SLA guided data integration can consist in four steps.

Generating a derived SLA The key and original aspect of our proposed data integration and provision process is to define a vertical mapping of user QoS preferences and agreed SLAs. This leads to a *derived SLA* that guides the evaluation of a query.

A query has associated preferences expressed as macroscopic constraints (i.e. user preferences statement): execution time, pay / no pay, data reliability, provenance, freshness, privacy, partial/full results, delivery mode. These constraints are coupled with the profile of the user which is in general stated in her cloud subscription (amount of assigned storage space, number of requests, I/O transferred Mega bytes, etc.).

Given agreed SLA's and a user preferences statement the challenge is to compute a *derived SLA* that maps SLA measures and preferences attributes. The derived SLA is defined as a set of measures that correspond to the user preferences computed as a function of different static, computed and hybrid measures. The *derived SLA* will guide the way the query will be evaluated, and the way results will be computed and delivered. Therefore, we propose to classify SLA measures to represent the relationship between fine grained measures used by agreed SLAs and coarse grained measures used in user preferences statements. It is also necessary to specify how to compute coarse grained measures with fine grained ones. For example, data precision will be computed as a function of availability, freshness and provenance exported by data services.

Filtering data services The derived SLA is used for filtering possible data services that can be used for answering the query. This is done using a set of matching algorithms based on graph structures and RDF specifications. This step may lead either to the rejection of integration in case of total incompatibility, or to a

negotiation between SLA which will lead to the proposal for a negotiated SLA integration and thus the need for an adaptive setting.

Query rewriting Given a set of data services that can potentially provide data for integrating the query result, we compute possible data service compositions that give partial or exhaustive results according to the derived SLA and the agreed SLA of each data service. The objective is to generate a number k of service compositions, combining as much as possible the services available such that the constraints of the derived SLA are verified.

Integrating a query result The service compositions are executed in one or several clouds where the user has a subscription. The execution cost of service compositions must fulfill the derived SLA (that expresses user requirements). In this phase we generate an execution plan considering the derived SLA and the subscription of the user to one or several clouds. We consider for example the economic cost determined by the data to be transferred, the number of external calls to services, data storage and results delivery costs and we decide how to use clouds resources for executing the composition. A first approach for performing this phase has been addressed in [5].

The first phase is an open issue for dealing with SLAs and particularly for adding quality dimensions to data integration. The problem is complex because SLA describe different elements participating in the data integration process: data services, cloud services at the different levels of the architecture (i.e., IaaS, PaaS, SaaS), data consumers subscriptions to cloud providers. The SLAs contain measures related to the way services are provided but also related to the data they provide. All these aspects must be considered for matching resources (i.e., services) with data consumers preferences. As shown in the following section and in our study SLA models and languages have been proposed. In contrast efficient preferences and SLAs matching algorithms need to be proposed to compute derived SLAs. Concerning the other data integration phases, they have been partially addressed by existing works, where some quality dimensions are considered (e.g., data privacy). In our vision there are open issues to be addressed in order to have solutions that consider SLA in order to enhance data integration in multi-cloud environments.

5 Related Works

Existing works addressing data integration can be grouped according to two different lines of research that correspond to the facets of the classification scheme that we propose: *(i)* data integration and services; and *(ii)* service level agreements and data integration.

5.1 Data integration and services

As shown in our classification scheme data integration description is a major topic. Existing works address knowledge oriented approaches for addressing the

problem. For example, [2] proposes a query rewriting method for achieving RDF data integration using SPARQL. The principle of the approach is to rewrite the RDF graph pattern of the query using data manipulation functions in order to: (i) solve the entity co-reference problem which can lead to ineffective data integration; and (ii) exploit ontology alignments with a particular interest in data manipulation. [3] introduces the Service Oriented Data Integration based on MapReduce System (SODIM) which combines data integration, service oriented architecture and distributed processing. SODIM works on a pool of collaborative services and can process a large number of databases represented as web services. The novelty of these approaches is that they perform data integration in service oriented contexts, particularly considering data services. They also take into consideration the requirement of computing resources for integrating data. Thus, they exploit parallel settings for implementation costly data integration processes.

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 shared. [11] focusses on data privacy based on 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] proposes an inter-cloud data integration system that considers a trade-off between users' privacy requirements and the cost for protecting and processing data. According to the users' privacy requirements, the query plan in the cloud repository creates the users' query. This query is subdivided into sub-queries that can be executed in service providers or on a cloud repository. Each option has its own privacy and processing costs. Thus the query plan executor decides the best location to execute the sub-query to meet privacy and cost constraints. As said before, the most popular "quality" property addressed in clouds when dealing with data is privacy. The majority of works addressing data integration in the cloud tackle security issues. We believe that other SLA measures need to be integrated in the data integration solutions if we want to provide solutions that cope to the characteristics of the cloud and the expectations of data consumers.

5.2 Service level agreement and data integration

Service level agreement (SLA) contracts have been widely adopted in the context of Cloud computing. Research contributions mainly concern (i) SLA negotiation phase (step in which the contracts are established between customers and providers) and (ii) monitoring and allocation of cloud resources to detect and avoid SLA violations.

[6] proposes a data integration model guided by SLAs in a Grid environment. Their architecture is subdivided into four parts: (i) a *SLA-based Resource Description Model* that describes the database resources; (ii) a *SLA-based Query Model* that normalizes the different queries based on the SLA information; (iii) an *SLA-based Matching Algorithm* selects the databases and finally (iv) a *SLA-based Evaluation Model* to obtain the final query solution. Considering our previous

work [1], to the best of our knowledge, we have not identified more proposals concerning the use of SLAs combined with a data integration approach in a multi-cloud context.

6 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 statements. 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 (fresh, precise, partial). 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 that includes steps that imply greedy computations. It is important to combine and revisit well-known solutions adapted to these contexts. We are currently developing the strategies and algorithms of our vision applied to energy consumption applications and also to elections and political campaign data integration in order to guide decision making on campaign strategies.

References

1. N. Bennani, C. Ghedira-Guegan, M.A. Musicante, and G. Vargas-Solar. Sla-guided data integration on cloud environments. In *2014 IEEE 7th International Conference on Cloud Computing (CLOUD)*, pages 934–935, June 2014.
2. Gianluca Correndo, Manuel Salvadores, Ian Millard, Hugh Glaser, and Nigel Shadbolt. SPARQL query rewriting for implementing data integration over linked data. In *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*, page 1, New York, New York, USA, March 2010. ACM Press.
3. 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.
4. 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.
5. Carlos-Manuel Lopez-Enriquez, Victor Cuevas-Vicenttin, Genoveva Vargas-Solar, Christine Collet, and Jose-Luis Zechinelli-Martini. A planning-based service composition approach for data-centric workflows. In *First International Workshop on Knowledge Aware Service Oriented Applications, KASA 2014, Paris, France, November 3, 2014. Proceedings*, 2014.

6. 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.
7. 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.
8. 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.
9. Pramod J Sadalage and Martin Fowler. *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education, 2012.
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 Ngoc Thanh Nguyen, editors, *ICCCI (1)*, volume 6421 of *Lecture Notes in Computer Science*, pages 195–204. Springer, 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.