Towards Quality Guided Data Integration on Multi-cloud Settings

Daniel A.S. Carvalho(✉)

Université Jean Moulin Lyon 3, Centre de Recherche Magellan,
IAE, Lyon, France
daniel.carvalho@univ-lyon3.fr

Abstract. This PhD project addresses data integration considering data quality (freshness, provenance, cost, availability) properties in a multi-cloud context. In fact, in a multi-cloud context, data is made available through a huge offer of services deployed on different clouds with heterogeneous quality of service features. By users who thank to their contracts with the clouds expressed by traditional SLA according to their rights. Consequently, data integration in this context needs to take into account these new constraints. The aim of our work is to revisit previously proposed data integration solutions in order to adapt them to the multi-cloud context. Our solution consists in defining over the clouds a layer that provides a reasoning on the best services combination that meets services and user constraints and willings, the best way to deploy the integration process. This layer should let further data integration easier thank to the definition of a new kind of SLA called *Integration SLA*. This paper gives a model-oriented vision of our proposal.

Keywords: Data integration · Query rewriting algorithm · Cloud computing · SLA

1 Introduction

Our work addresses data integration considering data quality properties (freshness, provenance, cost, availability) and service level agreements (SLA). Existing approaches - guided by heterogeneous data structures and formats, semantics and integrity constraints - have already tackled quality issues. Furthermore, this work explicitly considers infrastructure properties (reliability, computing, storage and memory capacity, and cost) imposed by the multi-cloud context and data providers quality to guide the integration process. In this new context, existing solutions are not sufficient as they need an infrastructure over the clouds that (*i*) allow services to express quality aspects; (*ii*) integration solutions to take into account the huge service offer and the multi-cloud paradigm constraints/advantages; and (*ii*) an intelligent entity to decide which services to select, where and

D.A.S. Carvalho—(Supervised by Chirine Ghedira-Guegan, Genoveva Vargas-Solar and Nadia Bennani, with inputs from Plácido A. Souza Neto).

K. Drira et al. (Eds.): ICSOC 2016 Workshops, LNCS 10380, pp. 139–144, 2017.
https://doi.org/10.1007/978-3-319-68136-8_14

in which conditions further integration demands could be treated using the past integration experience. The objective is to customize data providers (services) look up and the data integration considering different data consumers requirements and expectations depending on the context in which they consume data (e.g., mobile devices with few physical capacities, critical decision making). Our work relies on two assumptions: (1) the data integration process is totally or partially externalized on different clouds that provide necessary resources under different conditions (SLA); (2) data can be retrieved from several data providers (i.e., services) with different quality properties.

Let us suppose that during Brazilian Olympic games in 2016, Lucas wants to know two days in advance the weather forecast near his location to make decisions about the events he wants to attend. According to the weather, OGApp is an application that proposes possible matches in different stadiums with available seats (sunny seats or not, in the middle or in the sides, and on the side of a specific team). Lucas has several preferences regarding privacy (i.e. he wants his personal data to be anonymous), time, schedule, budget, cost (e.g., using free data services or not). Several data provision and computing services can be composed by OGApp to integrate data that can help Lucas to make his decision. Furthermore, since Lucas often looks for data in his mobile devices he is subscribed to several clouds to externalize "costly" processes (e.g., storage of retrieved data, correlation and aggregation of data coming from different providers, data transmission on 3G). OGApp will rely on the clouds to perform the integration process for Lucas respecting his preferences and the conditions of his subscriptions in the clouds. Suppose now that later Geraldine asks for the same result as Lucas but her constraint is to obtain the results with a minimum cost. Using Lucas' previous integration plan, the OGApp could be able to answer partially her query. Consequently, the same integration plan could be replayed. Thus, the data integration process becomes a combinatorial problem where a query result is a data collection integrated by composing different data providers and data processing (cloud) services that fulfill quality constraints and SLAs specified by a data consumer. Given a user query, the integration process deals with different matching problems: (i) matching the *query* and *data provider services* - the data provider services should be able to produce a (complete or partial) result for the query; (ii) matching the *user preferences* and the *quality guarantees* provided by the data provider (iii) matching the *user preferences* and *user's type of subscriptions* - the user may have several subscriptions with different clouds that should influence the way to choose the services according to the cloud resources offered thank to user subscription; and (iv) the *data provider services* and *their type of subscriptions* - the data provider services also have subscription with the clouds, and this imposes to adapt the way the service is delivered according to the resources to which it has access.

We assume the quality conditions that the user can expect from a service are defined in service level agreements (SLA). In our context, we need to identify which SLAs measures apply to the data integration process and how they should

be taken into consideration for providing a final result that fulfills data consumers requirements.

This PhD project proposes an approach for data integration guided by quality and SLAs partially or totally performed over a multi-cloud settings. The originality of our approach consists in guiding and personalizing the entire data integration process - while selecting, filtering and composing cloud services, and delivering the results - taking into account (*a*) user preferences statements; (*b*) SLAs exported by different cloud providers; and (*c*) QoS measures associated to data collections (for instance, trust, privacy, economic cost).

The reminder of this paper is organized as follows. Section 2 discusses the related works. Section 3 gives an overview of our SLA-based data integration approach. Section 4 describes the research plan, and Sect. 5 concludes the paper.

2 Related Works

Related works rely on four topics: (*i*) data integration and data quality in the database domain; (*ii*) data integration approaches in the cloud and in service-oriented contexts; (*iii*) query rewriting approaches; and (*iv*) service level agreements for cloud computing.

Data integration has been widely discussed in the database domain. [10] discussed theoretical aspects in data integration including modeling applications, query evaluation, dealing with inconsistencies and reasoning queries. Moreover, [9] reviewed several query rewriting approaches. [3] surveyed data quality aspects in data integration systems. [11] presented a data quality broker that allows to submit queries with associated quality requirements over a global schema and to provide results according to them.

[6,8] performed data integration in service-oriented contexts, particularly considering data services. However, they consider computing resources consumption versus performance for guiding the data integration process. [12] proposed an inter-cloud data integration system considering privacy requirements and the cost for protecting and processing data. [11–13] tackled quality aspects of the integration, but do not consider crucial aspects such as data consumers and data providers requirements and constraints, the associated infrastructures and the data quality itself.

As traditional databases theory, data integration on cloud and service-oriented context deals with query rewriting issues. Existing works like [1,2,4,7] have refered it as a service composition problem. Given a query, the objective is to lookup and compose data services that can contribute to produce a result. In general, these works must address performance issues, because they use algorithms that can become expensive according to the complexity of the query and on the number of available services. Although [1,4] have considered preferences and scores to produce rewritings, the multi-cloud context introduces new requirements and constraints to the integration process. Currently, the approaches are not sufficient to cover the new challenges. Thus, they should be revisited and adapted in order to make the integration efficient in this new environment.

Research contributions related to SLA in cloud computing concern (i) SLA management; (ii) inclusion of security requirements on SLAs; (iii) SLA negotiation; (iv) SLA matching; and (v) monitoring and allocation of cloud resources to detect and avoid SLA violations. We strongly believe that SLAs can be used to explicitely introduce the notion of quality in the current data integration solutions. In this sense, the use of SLAs to guide the entire data integration in a multi-cloud context seems original and promising for providing new perspectives to the data integration problem.

3 An SLA-Based Data Integration Approach: A Model Oriented Vision

To explain our solution, in this section we present a metamodel that depicts implied entities and their relationship. Then a meta-process will introduce the functionalities of the proposed solution.

According to our metamodel (Fig. 1), the *Multi-Cloud* is viewed as a set of *Cloud Infrastructures*. *Data producers* and *Data consumers* subscribe to *Cloud infrastructures*. Their subscription credentials are illustrated thanks to a *SLA* (*Consumer SLA* or *Producer SLA*) defining what the *Cloud infrastructure* offers to them through their subscription. *Data* are provided and consumed by *Data Producers* and *Data Consumers*, respectively. The *integration SLA* is a new type of SLA we introduce to reflect a multi-cloud contract between the user and the implied services according to the constraints imposed by the environment.

The data integration meta-process (see Fig. 2) implies the entities presented in the metamodel. It consists of three macro-steps. *First*, *query management* activities to process the user query and preferences; *second*, *SLA management* activities to enforce the SLA associated to the involved services, search and reuse previous *integration SLAs*, and create a new one for the current request; *third*, *query rewriting* activities [5] to search and filter *data producers*, to generate and execute the integration plan, and to compute results.

The activities defined in our meta-process bring the following challenges to our research:

1. SLA design. The issue is what are the important information that should be inserted in the integration SLA to facilitate further integration? How these information should be collected and stored during the integration to help next integrations.
2. Integration reuse. How to exploit cleverly the past integration processes?
3. Rewriting process. How to optimize it to make the execution time viable? Retrieving, integrating and delivering are tasks that requires a large amount of resources and processing time. Thus, it is necessary to study a efficient manner to make efficient the overall execution.

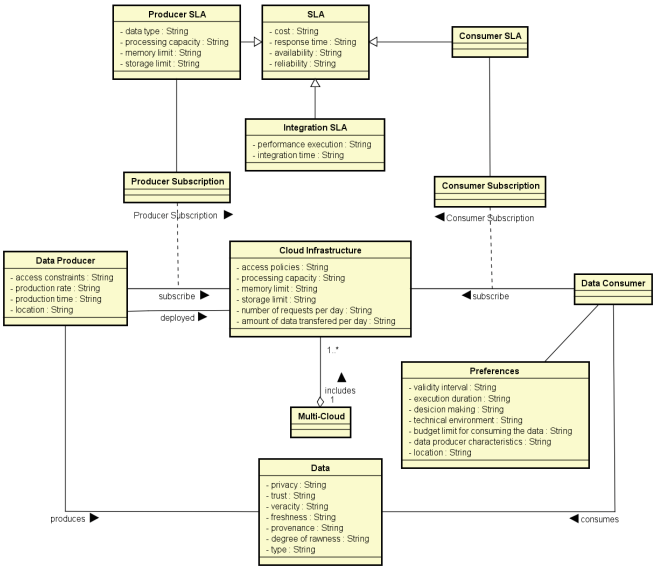

Fig. 1. Data integration metamodel

Fig. 2. Data integration meta-process

4 Research Plan

We started proposing a query rewriting algorithm called *Rhone*. It serves as proof of concept for the feasibility of our data integration process guided by cloud constraints and user preferences [5]. The first results are promising: *Rhone* reduces the rewriting number and processing time while considering user preferences and services' quality aspects extracted from SLAs to guide the service selection and rewriting. Furthermore, the integration quality is enhanced, and the integration total cost is reduced. For the time being, quality enhancement is assessed through benchmarks and use cases deployed on an experimental multi-cloud environment.

We are currently working on an SLA model for the integration process to express the constraints and the quality feature of a previous data integration. Other important research aspects are how to make efficient the rewriting process by reducing the composition search space.

5 Conclusions

This paper introduced a new data integration solution, adapted to the multi-cloud context. The solution is described thank to a metamodel describing the implied entities and a meta-process presenting the activities and the corresponding challenges.

References

1. Ba, C., Costa, U., Halfeld-Ferrari, M., Ferre, R., Musicante, M.A., Peralta, V., Robert, S.: Preference-driven refinement of service compositions. In: Proceedings of CLOSER 2014 International Conference on Cloud Computing and Services Science (2014)
2. Barhamgi, M., Benslimane, D., Medjahed, B.: A query rewriting approach for web service composition. IEEE Trans. Serv. Comput. Serv. Comput. (2010)
3. Batini, C., Scannapieco, M.: Data Quality Issues in Data Integration Systems, pp. 133–160. Springer, Heidelberg (2006). doi:10.1007/3-540-33173-5_6
4. Benouaret, K., Benslimane, D., Hadjali, A., Barhamgi, M.: FuDoCS: a web service composition system based on fuzzy dominance for preference query answering. In: 37th International Conference on Very Large Data Bases (VLDB 2011) (2011)
5. Carvalho, D.A.S., Souza Neto, P.A., Ghedira-Guegan, C., Bennani, N., Vargas-Solar, G.: *Rhone*: a quality-based query rewriting algorithm for data integration. In: Ivanović, M., Thalheim, B., Catania, B., Schewe, K.-D., Kirikova, M., Šaloun, P., Dahanayake, A., Cerquitelli, T., Baralis, E., Michiardi, P. (eds.) ADBIS 2016. CCIS, vol. 637, pp. 80–87. Springer, Cham (2016). doi:10.1007/978-3-319-44066-8_9
6. Correndo, G., Salvadores, M., Millard, I., Glaser, H., Shadbolt, N.: SPARQL query rewriting for implementing data integration over linked data. In: Proceedings of the 1st International Workshop on Data Semantics - DataSem 2010. ACM, New York (2010)
7. Costa, U.S., Ferrari, M.H., Musicante, M.A., Robert, S.: Automatic refinement of service compositions. In: Daniel, F., Dolog, P., Li, Q. (eds.) ICWE 2013. LNCS, vol. 7977, pp. 400–407. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39200-9_33
8. ElSheikh, G., ElNainay, M.Y., ElShehaby, S., Abougabal, M.S.: SODIM: service oriented data integration based on mapreduce. Alexandria Eng. J. (2013)
9. Halevy, A.Y.: Answering queries using views: a survey. VLDB J. **10**(4), 270–294 (2001)
10. Lenzerini, M.: Data integration: A theoretical perspective. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. PODS 2002, pp. 233–246. ACM, New York (2002)
11. Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., Baldoni, R.: The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. Inf. Syst. **29**(7), 551–582 (2004)
12. Tian, Y., Song, B., Park, J., Huh, E.-N.: Inter-cloud data integration system considering privacy and cost. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS, vol. 6421, pp. 195–204. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16693-8_22
13. Yau, S.S., Yin, Y.: A privacy preserving repository for data integration across data sharing services. IEEE T. Services Computing 1 (2008)