Chapter 3

# Thesis Advancement

Contents



## 3.1 Context and motivation

The emergence of multi-cloud environments opens new challenges to data processing. Current data integration implies consuming data by matching and composing different data services, and integrating the results while respecting data consumers quality requirements. These requirements are exposed in service level agreement (SLA) contracts established between data consumers and cloud providers. The SLA defines what a data consumer can expect as system behavior, but also the properties of the data such as its provenance, veracity, freshness, whether the consumer accepts to pay for data, and how much is he/she ready to pay for the resources necessary for integrating his/her expected result.

Several researches have introduced algorithms for data integration in service-oriented architectures. They have designed methods for matching, selecting and composing services according to some data requirements and constraints. These works share a performance problem while combining and producing compositions. To tackle it, authors have proposed heuristics for computing and identifying the best services and compositions based on quality aspects. However, current works focus on performance capabilities of services neglecting data properties and the new constraints imposed by the multi-cloud context.

In this sense, the objective of this work is to address data integration in a multi-cloud context. The originality of our approach consists in guiding the entire data integration solution taking into consideration (i) data consumer requirement statements with respect to performance capabilities of data services (for instance, availability and response time) and to data collection properties (provenance, freshness,

veracity, data type, among others); (ii) infrastructure properties (reliability, computing, storage and memory capacity, and cost) imposed by the multi-cloud context; and (iii) SLA contracts exported by different data services and cloud providers.

## 3.2 Problem statement

The multi-cloud architecture brings new challenges to data integration and data processing applications. Instead of taking into consideration the user query and his/her requirements with respect to services' performance capabilities (such as percentage of availability and response time) in isolation, the integration process must consider in addition the new constraints imposed by context:

- User requirements concerns not only data services' performance capabilities but also quality requirements of the data which is provided (such as freshness, cost, provenance, data type, veracity among others).

- Data provision is constrained to the available computing resources agreed between data services and cloud providers on service level agreement (SLA) contracts.

- The data integration process requires a high level of computing resources. The huge amount of data and data services on multi-cloud settings increases even more the complexity of the solutions.

In this sense, current data integration solutions introduces a multi-dimensional matching problem that should take into account:

- Expressing data consumer requirements with respect to performance capabilities of data services and to the quality of the expected data.

- Matching and selecting data services according to the data consumer expectations (concerning performance and data quality requirements) with respect to data services quality measures defined in SLA contracts.

- Matching and selecting data services which have available resources according to different SLAs that they have agreed with different cloud providers.

- Delivering results with respect to the data consumer requirements and expectations depending on the context which he/she consumes the data.

Thus, the intention of this thesis project is to address data integration on multi-cloud environments. The aim is propose a data integration approach adapted to the multi-cloud context in which data is delivered according to data consumers expectations by profiting from previous integration results instead of launching the expensive data integration process from the first step.