Intelligent integration of heterogeneous data for answering analytics queries in multi-cloud environments

Genoveva Vargas-Solar[1], Chirine Ghedira-Guégan[2] and Nadia Bennani[3]

[1]CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221 Villeurbanne, France
[2]Univ Lyon, Université Jean Moulin Lyon 3, LIRIS, UMR5205, iaelyon School of Management, France
[3]Univ Lyon, INSA Lyon, CNRS, UCBL, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69621 Villeurbanne, France

Abstract
This position paper discusses the design of an approach to enable trusted data integration in a multi-cloud environment in the presence of heterogeneous and large data sources. This approach is based on mechanisms to compute trust in data and its providers by applying statistical and probabilistic methods on its provenance. The result is a solution expressing analytics queries as services coordination that can be enacted on multi-cloud settings. This paper describes the associated challenges and possible ways of addressing them. The approach and challenges are based on concrete requirements stemming from a medical scenario related to understanding, modelling, and predicting patients' conditions associated with sleep apnoea.

Keywords
Intelligent data integration, data analytics queries, multi-cloud, data services, data driven e-health

1. Introduction

The digitalization of companies implies the explosion of data to be integrated and analyzed to answer different types of questions for retrieving and analyzing data. Over the past decade, service-oriented environments have made it easier for many users to access data through data services. A data service is a software entity accessible through APIs that describe the methods that provide data either on-demand or continuously.

Companies willing to make data-driven decisions currently use services and process large amounts of data and resources. However, these companies have to deal with the efficiency and cost of processing linked to this avalanche of complex multi-source data deployed on several clouds for economic reasons. In this context, several economic actors are led to cross-reference this voluminous data through data integration, guided by queries that specify the data required by an application, a user, or a community of users.

For example, in the e-health context, there can be services providing information about physiological metrics of people, apnoea events, during sleeping hours, glucose measures by day, dietary and training sessions information. Through these data services, applications can be, for example, access training programs performed by the patient Alice? The number of apnoea intervals of my patients during the last 10 days? The number of apnoea intervals/night when the patient Bob slept 5H and had a 200 glucose level 2H before going to sleep?

These queries can be answered by composing on demand and stream data services. Beyond the methods described by the APIs, the services are tagged with QoS measures such as data freshness, response time, execution price, among others. As a result, the data services composition includes a combination of QoS preferences that can be interpreted as constraints.

The notion of service level agreement (SLA) can define the agreement between the users and the set of services used. From this contract, the user's preferences and requirements emerge. For example, a user can:

1. privilege the data freshness of the data used to answer her/his query, like 1month old data;
2. the availability of the service;
3. the latency of data delivery constraint to milliseconds.

In this context, the challenges and questions that drive the our work are:

- What models, mechanisms, algorithms are suitable for data integration when qualitative and quantitative criteria guide it?
- What are the criteria, constraints & requirements on the data that guide queries evaluation and data integration across multiple providers?
- When we invoke machine intelligence, the question is what intelligence mechanisms to consider for making integration intelligent?
- Finally, does the cloud bring specific challenges when making data integration intelligent?

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
genoveva.vargas-solar@cnrs.fr (G. Vargas-Solar);
chirine.ghedira-guegan@univ-lyon3.fr (C. Ghedira-Guégan);
nadia.bennani@insa-lyon.fr (N. Bennani)
http://www.vargas-solar.com (G. Vargas-Solar)
0000-0001-9545-1821 (G. Vargas-Solar)

This position paper discusses the interest of treating analytics queries as trustworthy service-based coordinations and considering them as first-class citizens. We introduce an approach to answer analytics queries on medical data by building trustworthy data integration. The approach considers that data is provided by services deployed on different cloud providers. Therefore, the paper highlights the challenges to running service-based analytics queries in a multi-cloud environment. The approach uses mechanisms to infer trust levels in data and providers by applying statistical and probabilistic methods.

Accordingly, the remainder of the paper is organized as follows. Section 2 discusses related work regarding data distribution on services and service-based querying approaches. Section 3 describes the solution proposed for implementing data analytics on apnea conditions. Section 4 illustrates use cases addressed through data centred strategies that use machine learning and data analytics algorithms. Section 5 concludes the paper and discusses future work.

2. Related Work

The classical view of data integration has been widely addressed in databases through data model equivalence and transformation, schema integration, and query rewriting algorithms. The main feature of these approaches is the prior knowledge of the data sources when performing the integration process. The emergence of data services has revolutionised the problem of data integration because services provide data whose sources and format are not known. Data integration with data provider services starts with a query expressing the data requirements and searches for services that can provide them. Also, in the approaches [1, 2, 3], the services export their APIs (Application Programming Interfaces) and data according to a pivotal model that can be used to integrate the results. Thus, the problem of data integration becomes a rewriting problem where queries are rewritten using matching and service composition mechanisms [4]. The problem is more challenging because of the diversity of data requesting and consuming devices and the absence of meta-data (i.e. veracity, freshness, etc.) specifying the associated conditions of data provision and use (user preferences). Services guarantees are established through service level agreements (SLAs) between producers and consumers. These agreements have the advantage of guiding the integration process by pruning the data but significantly increase the complexity of the process, especially in the presence of a large quantity (of the order of several thousand) of data sources.

Data integration [5] is driven by queries that specify the data required by an application, a user or a community of users. The classical view of data integration, where data sources are known in advance, has been widely addressed in databases: data model equivalence and transformation, schema integration, and query rewriting algorithms. In data integration in the presence of services that act as data providers, the starting point is a query that expresses needs in terms of the data required and must search for services that can meet these needs. In the approaches [1, 2, 3], the services export their APIs (Application Programming Interface) and data in a pivotal model that can be used to integrate results. Thus, the data integration problem becomes a query rewriting problem using matching and service composition mechanisms [4].

With the evolution of technology, queries are issued from multiple devices with different constraints and results are consumed under other conditions (energy consumption, network bandwidth consumption, economic cost, privacy, trust and criticality). Data producers do not export the properties of their data and the conditions under which it is delivered. Consumers express the expected quality of data [6] and the conditions under which it will be consumed, such as its veracity, freshness, etc. These qualitative aspects necessary for data consumption and the conditions under which queries are to be evaluated are specified through contracts (SLAs) between data producers and data consumers and user profiles informing about data usage preferences. These specifications have the advantage of guiding and pruning data in the integration process. Still, they add significant complexity to an already complex process, especially in the case of queries using a large number of data sources. Indeed, evaluating a query (i.e. rewriting a query) becomes a combinatorial problem whose complexity increases with the expression of quality requirements which are non-orthogonal constraints.

Heuristics and "best-effort" approaches have already been proposed in the fields of databases [4] and services (SOA) [7]. Given the heterogeneous context of multi-cloud, multi-device and multi-objective Internet of Things (IoT), the primary challenge is to deliver responses to user requests in a reasonable time and at an acceptable cost, given the complexity of the rewriting process. Furthermore, guaranteeing the quality of integrated data requires considering the source of the data and the level of trust in the services that provide it. The idea is to capitalise on the execution history of requests by proposing an intelligent process that can learn from previous integration experiences. The difficulty would then lie in the diversity of the execution contexts of each request (deployed services, expressed needs, critical situation, required level of confidence, response time, etc.).

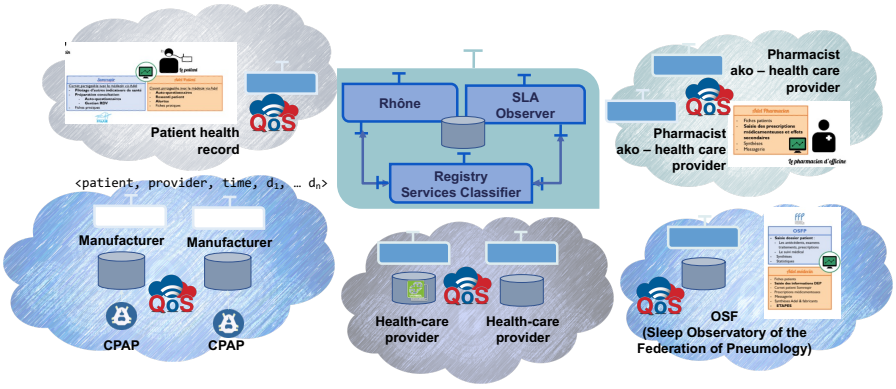

Figure 1: ADEL servitization on the cloud.

3. Trusted Apnoea Data Integration for Answering Analytics Queries

The approach proposed in this paper has been proposed in the context of the project SUMMIT. As a technology transfer exercise, the project has a partner, the startup Datamedcare, that promotes the platform ADEL that integrates different actors that intervene in treating a condition called sleep apnoea.

Neumologists treat the condition, and part of the treatment includes using devices called CPAPs intended to be used every day during people's sleep. The device is technically calibrated according to patients' physical and physiological characteristics. Doctors receive records that include the use of the device and other information from the patient and decide how to adjust the treatment if things are not going well.

The platform ADEL integrates the information and provides a global view of the protocols implemented to follow patients. The integration is loosely coupled, preventing doctors from automatically analysing conditions by correlating data. Besides, given the medical context, it is essential to ensure the respect of SLA so that doctors, patients, and Care providers can be confident about the quality of the data and the data services used to exchange them among these actors. The data provision strategy of ADEL does not technically consider this SLA notion.

3.1. Servitization of the ADEL platform

First of all, behind the scenes, we worked on a servitization phase of the current setting of the ADEL platform (see Figure 1). The objective of this phase was to ensure the independence of the actors and the type of data and infrastructure they use to produce and manage this data.

Servitization was important because we included a QoS dimension considering that the function of data providers and consumers is not enough to be able to expose SLAs.

We defined five services groups hosted in secure private or public cloud providers (Azure, Google Cloud and AWS) to ensure different quality guarantees. The groups consist of services related to the following data: (1) Observations collected by CPAPs (the masks used by patients to treat their apnoea). (2) Questionnaires used by neumologists to follow patients. (3) Health care providers' data regarding patients' subscriptions. (4) Pharmacies data about treatments for apnoea patients. (5) Data collected by the association of neumologists studying apnoea.

In this context, analytics queries issued by actors can specify SLA contracts regarding, for example, the availability of the services managing data, their response time, and the quality of the data like freshness, frequency of upload, etc.

In this setting, we assumed that analytics queries with their SLA specifications are rewritten into services compositions enacted to answer them [6]. Query rewriting from our point of view implies:

- Determining which are the data services to be used?
- Figuring out to which extent potential services and compositions fulfil SLA?
- Studying which deployment strategy is adapted for enacting a target service composition?

Given the complexity and non-triviality of the apnoea case, the questions and requirements can be summarized in two categories:

Apnoea questions to answer. In the context of the Apnoea application to which we transfer our SLA guided intelligent data integration through queries we have a kind of baseline questions provided by Apnoea specialists. These are essentially analytics questions willing to

understand the conditions and evolution of the disease, particularly concerning the CPAP's use.

Questions include an analysis of the way CPAPs are used. The idea is to determine the extent to which patients adapt to a given CPAP model and how this adaptation determines whether they will be assiduous. Assiduity can result in a positive evolution of their condition.

Other questions concern the analytics by the companies that have to know how they perform interacting with patients to adopt their product.

Query Rewriting as a Service Considering our objectives, we have proposed a rewriting service consisting of three main components (see the centre in Figure 1)):

- Rhone a services composition module guided by SLA requirements
- An SLA observer that monitors the services continuously and evaluates QoS metrics
- A registry which is a services classifier that tags and ranks services with a quality index used by Rhone to choose the services to be composed given a query and its SLA specification.

3.2. Selecting trustworthy services

As aforementioned, we target a trustable integration; therefore, we describe the SLA aspects and how we integrated them into services compositions to answer questions. Recall that we need to choose the services that will provide data, ensuring SLA expectations for answering a question. For addressing SLA and given that we work for medical applications, we considered a definition of Trust that includes service performance and data quality, knowing that there is no or few access to meta-data what we call black-box services.

In this context we addressed three questions:

- P1. What is the appropriate model for describing individual data services trust using service performance and data quality factors?
- P2. How to collect the necessary information for this trust evaluation model?
- P3. How to define data quality metrics using the collected information?

To this end, we propose a data quality observability protocol [8], defining data timeliness metrics for data services as a black box. The overall objective is enabling data consumers to select the most reliable data service according to their needs providing a trust-sorted list of data services. Thus, the system is composed of three main modules:

- performance measuring module, which collects and measures performance metrics;
- data quality evaluation module which implements our observability protocol;
- data service trust measuring module which collects both data quality and performance measurements and computes data services trust scores. This meta-data is not included in SLA's models. In our future work, we will propose the extension of SLA's for including such measurements.

4. Use Cases

SUMMIT (http://summit.imag.fr) is a technology transfer project funded by the Auvergne-Rhone-Alpes region that addresses multi-clouds, intelligent data integration, service level agreement and focuses on the context of multi-device environments in the medical context.

We have the following services available implemented during the servitization process of the ADEL platform (see Figure 1)): (i) the Healthcare providers with information about the patients' insurances; (ii) a service dealing with information about patients' medical procedures; (iii) CPAPs data services managing the observations about apnoea episodes during sleep collected when patients use their devices; and (iv) computing services that provide analytics functions deployed in different cloud providers.

We implemented experiments as service-based queries to coordinate the ADEL platform's data services with data processing operators to answer analytics questions regarding the apnoea condition. Experiments focus on (1) classifying patients according to their CPAP frequency of use (i.e., compliance). The objective is to observe the evolution of their physiological metrics as they use their CPAP. (2) Other experiments address the study of metrics regarding the apnoea condition and the use of CPAPs in time seeking behaviour patterns. (3) Finally, experiments are devoted to predicting the evolution of patients' condition according to the evolution of their physiological status and the use of their treatment.

4.1. Services based analytics queries

The first aspect to consider is translating questions expressed in natural language by topic experts into queries. We have used reference questions of the SUMMIT project partners and translated them into service coordinations that implement queries. In this project, we do not address the problem of processing natural language specifications. Still, we assume that we have a query language like the one proposed in our previous work [9, 10].

Classification queries. Q-1: *Are there any key characteristics that differentiate patients whose compliance is less than 2 hours/night, between 2 and 4 hours/night and more than 4 hours/night on average?*

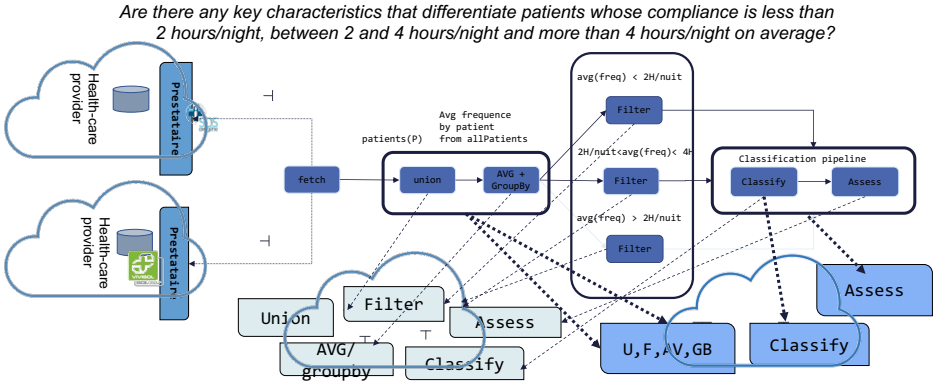

Figure 2: Classification query.

A general services composition schema to answer this question can be the following (see Figure 2):

1. Fetch data from health care providers.
2. Union, avg and group by the use of the device by patient.
3. Filter the right duration and then classify patients according to the average duration of use and assess the classification quality.

The whole data preparation and engineering step can be implemented by composing services or a composition ready to use. Similarly, the filtering process involves 3 filtering tasks that can be executed in parallel by one or different computing services.

Modelling and correlation queries. We expressed queries that included joining and filtering data provided by different services as data preparation tasks. Then the resulting data sets are used as input to identify the role of several attributes in a level of apnoea condition and discover their role in this value.

- Q-2: *Correlation between Epworth*[1] *and the apnoea-hypopnea index (AHI) at diagnosis.* Figure 3 shows the corresponding service coordination that implements this question. Data are fetched in parallel from two providers, managing the apnoea data collected from the CPAPs and data from the patients' records. The coordination in the figure is generic, but there can be other possibilities. For example, having two parallel sequences of "fetch, filter and project operations" for each service provider, joining both results and finally applying a model that can estimate correlation.

[1]The Epworth Sleepiness Scale (ESS) is a scale intended to measure daytime sleepiness that is measured by use of a very short questionnaire. This can be helpful in diagnosing sleep disorders

Other questions that we implemented concern aggregation queries like Q-3: where physiological metrics collected by CPAPs are analysed to observe their evolution along time (see below). Filters can be applied to select data items for specific intervals. Data preparation can be done first to see the data distribution along time and include assiduity factors.

- Q-3: *Evolution of pressures, leaks, AHI over time under treatment.*
- Q-4: *How does compliance changes over time since initiation of treatment. Which factors influence compliance (including type of mask and brand of CPAP).*

The first part of Q-4: is implemented under the same principle of the Q-3 . Then other operations can be applied to address the second part and determine the factors that can influence compliance. To answer the second part of Q-4: a service of type health care provider must be used to fetch the history of the CPAPs versions and brands that every patient has tested.

The first tasks of all the service coordinations expressing queries (see Figures 2 and 3) start by fetching and preparing data by selecting, filtering (e.g., items with(out) specific values) or projecting data and then applying operations like the union. These operations can be executed in parallel or sequentially become the input operators (tasks) that can infer/discover/model or compute aggregations.

Prediction queries. *On the basis of clinical data, including history and self-administered questionnaires (OSFP), is it possible to predict the severity of the condition with AHI<15; 15 <AHI<30; AHI>30?* Having data providers with labelled data collections can support prediction queries. The tasks can include analysing the properties of the attributes/variables of the data (e.g., linear

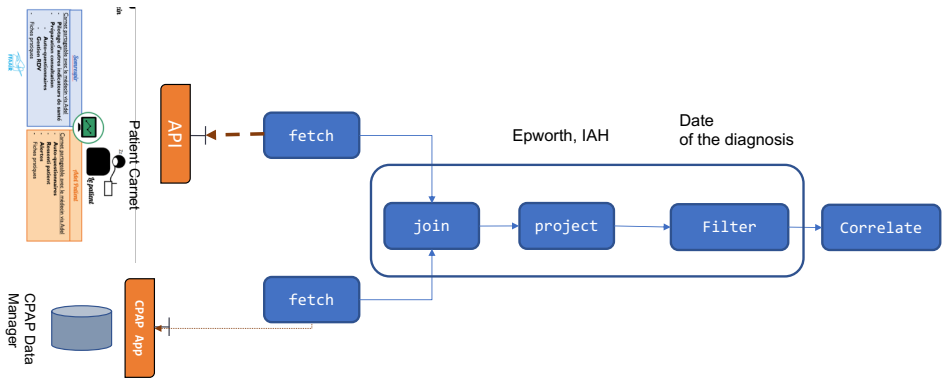

Figure 3: Correlating Epworth and AHI for understanding the context of diagnosis.

dependency of attributes with the target variable). It can also include tasks to engineering the data fetched from providers. Then it can consist of the application of prediction models and other assessment tasks.

4.2. Open Challenges: discussion and position

Having worked on the design and implementation of service coordinations that can implement the queries, the next step to address is the cloud or environment in which queries are executed. Besides, several services can be used and available for a given task or a given type of data. There is a decision making problem to address to **choose the services** that will execute the tasks of a query. Our project can deploy services in different clouds and guarantee various services, including trust guarantees. Thus, this concerns a rewriting process that can **generate a solution space** for a given query rather than one coordination.

In our current work, we have proposed data quality observability protocol and its associated service TUTOR [8, 11]. The overall objective is to select the most reliable services according to given needs and provide a trust-sorted list of services.

The open challenge is to include the query rewriting process within an **optimisation process** that can use a cost model to rank or propose the coordinations that can potentially provide expected results in the best conditions possible. These conditions can include trust aspects associated with the services participating in a services coordination.

Another critical challenge is to reason about deploying a services coordination that implements a query. The coordination process can be executed in a distributed setting. Should the coordination run on different clouds or in only one cloud? Therefore, trust aspects can concern the cloud providers hosting the execution of a query. The decision making associated with the deployment is an open issue that we are currently addressing.

5. Conclusions and Future Work

This paper introduced open challenges and possible directions for integrating data for answering analytics queries on multi-cloud environments. The problems discussed are inspired by a concrete use case related to the analysis of medical data to understand, model and predict the condition of sleep apnoea. Our current work concerns the stabilisation, profiling, testing, and scaling of Rhone [6], an algorithm for SLA guided data services composition. We are also consolidating services monitoring for computing services trust indices based on technical metrics and data quality metrics [8].

We observe the following main perspectives: (1) Evolving towards trust-based data services recommendation (multi-cloud). (2) Addressing more database challenges like capitalising on case observations to design deployment patterns for data services compositions. (3) Monitoring for collecting knowledge. (4) Proposing enactment and optimisation strategies.

6. Acknowledgement

This work is funded by the project SUMMIT, pack ambition program of the region Auvergne Rhône Alpes - P089 - 0718-184-ARA, https://summit.imag.fr.

References

[1] C. Ba, U. Costa, M. Halfeld-Ferrari, R. Ferre, M. A. Musicante, V. Peralta, S. Robert, Preference-driven refinement of service compositions, in: Proceedings of CLOSER 2014 International Conference on Cloud Computing and Services Science, 2014.

[2] M. Barhamgi, D. Benslimane, B. Medjahed, A query rewriting approach for web service composition, IEEE Transactions on Services Computing 3 (2010) 206–222.

[3] M. Lenzerini, Data integration: A theoretical perspective, in: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2002, pp. 233–246.

[4] U. S. Costa, M. H. Ferrari, M. A. Musicante, S. Robert, Automatic refinement of service compositions, in: International Conference on Web Engineering, Springer, 2013, pp. 400–407.

[5] R. Pottinger, A. Halevy, Minicon: A scalable algorithm for answering queries using views, The VLDB Journal 10 (2001) 182–198.

[6] D. A. Carvalho, P. A. S. Neto, C. Ghedira-Guegan, N. Bennani, G. Vargas-Solar, Rhone: A quality-based query rewriting algorithm for data integration, in: East European Conference on Advances in Databases and Information Systems, Springer, 2016, pp. 80–87.

[7] K. Benouaret, D. Benslimane, A. Hadjali, M. Barhamgi, Fudocs: A web service composition system based on fuzzy dominance for preference query answering, Proceedings of the VLDB Endowment 4 (2011) 1430–1433.

[8] S. Romdhani, G. Vargas-Solar, N. Bennani, C. G. Guegan, Qos-based trust evaluation for data services as a black box, in: C. K. Chang, E. Daminai, J. Fan, P. Ghodous, M. Maximilien, Z. Wang, R. Ward, J. Zhang (Eds.), 2021 IEEE International Conference on Web Services, ICWS 2021, Chicago, IL, USA, September 5-10, 2021, IEEE, 2021, pp. 476–481. URL: https://doi.org/10.1109/ICWS53863.2021.00067. doi:10.1109/ICWS53863.2021.00067.

[9] V. Cuevas-Vicenttín, G. Vargas-Solar, C. Collet, N. Ibrahim, C. Bobineau, Coordinating services for accessing and processing data in dynamic environments, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, 2010, pp. 309–325.

[10] V. Cuevas-Vicenttín, G. Vargas-Solar, C. Collet, Evaluating hybrid queries through service coordination in hypatia, in: Proceedings of the 15th International Conference on Extending Database Technology, 2012, pp. 602–605.

[11] S. Romdhani, N. Bennani, C. G. Guegan, G. Vargas-Solar, Trusted data integration in service environments: A systematic mapping, in: S. Yangui, I. B. Rodriguez, K. Drira, Z. Tari (Eds.), Service-Oriented Computing - 17th International Conference, ICSOC 2019, Toulouse, France, October 28-31, 2019, Proceedings, volume 11895 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 237–242. URL: https://doi.org/10.1007/978-3-030-33702-5_18. doi:10.1007/978-3-030-33702-5_18.