A Research on Data Lakes and their Integration Challenges

Davide Piantella

Politecnico di Milano – Dipartimento di Elettronica, Informazione e Bioingegneria
Via G. Ponzio 34/5, 20133 Milano, Italy

Abstract
With the advent of IoT and big data, we observed a huge variety of types of data (e.g. semi-structured data, conversational data, sensor data, photos, and videos) and sources (e.g. social networks, open data, webpages, and sensors). Data integration addresses the problem of reconciling data from different sources, with inconsistent schemata and formats, and possibly conflicting values. In this paper, I describe my PhD research topic: the enhancement of data integration, discovering new techniques capable of handling the peculiar characteristics of big data, and the study of novel frameworks and logical architectures to support the integration process.

Keywords
Data integration, Big data, Data lakes

1. Introduction

In this paper, I will describe my PhD research, the contributions we developed until now, and the research topics we plan to analyze in my last year as a PhD student. The paper is organized as follows: Section 1 contains an introduction to the research area and an overview of the related challenges, Section 2 describes our contributions, and Section 3 outlines possible future works.

1.1. Big data integration

The data integration process has the goal of aligning different data sources to provide uniform access to data, possibly addressing sources with different database schemata, different data formats, semantic and representation ambiguity, and data inconsistency [1]. Nowadays, the extensive use of user-generated content, along with the Internet Of Things and the digital transformation of industries, has made available a huge mass of data. Since the value of data explodes when it can be analyzed after having been linked and fused with other data, addressing the *big data integration* challenge is critical to realize the promises of the big data phenomenon [2].

The initial focus of data integration was on structured (typically table-based) data, and it had traditionally three main phases: the first phase is *schema alignment*, with the purpose of aligning different database schemata and understanding which attributes have the same semantics; the

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy
davide.piantella@polimi.it (D. Piantella)
0000-0003-1542-0326 (D. Piantella)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3194/paper73.pdf

second phase is *entity linkage*, in which a given set of records is partitioned in such a way that each partition corresponds to a distinct real-world entity; finally, through *data fusion* we resolve all possible conflicts that can arise when different sources provide different values for the same attribute of an entity. Sources can provide conflicting values for many reasons (e.g. errors, outdated information, mistyping, etc.) and different possible reconciliation methods can be exploited (e.g. keep all values, choose a random one to store, compute the average, keep the most recent, etc.), each of which may be suitable or improper depending on the specific domain and usage context [3].

Nowadays, the needs and characteristics of big data, with many other data kinds and formats such as texts, social media data, images, and videos, have introduced new challenges such as heterogeneity and data quality. The heterogeneity of formats is reflected in how data is structured: we can identify three main data categories: structured data (e.g. relational databases), semi-structured data (e.g. XML documents), and unstructured data (e.g. natural language texts and images). In this context, with also the increasing usage of NoSQL databases that are capable of handling document-based data and graphs [4], schema alignment is very hard or even impossible, since a schema description may not be present at all.

With semi-structured or unstructured data, especially if we have to integrate sources that use different data formats, also the entity linkage problem becomes really difficult to solve since we often cannot leverage any information from database schemata. There are techniques (such as [5]) able to reduce the computational complexity of the entity linkage phase, exploiting statistics computed directly from the data to better describe data sources. Moreover, format heterogeneity introduces the need for *data extraction*, to be performed before the three classical steps of data integration [6].

In general, the problems encountered during these steps are not completely solved, and are often aggravated by the big data context: we can find much more incomplete, dirty, and outdated information than before, therefore new data integration and data cleaning techniques have to be developed, with the purpose of reducing noise and errors in the values provided by sources.

1.2. Data lakes

An emerging trend is to use data lakes to collect a huge amount of data and documents, exploiting the usage of metadata and modern storage techniques. Data lakes are schema-less, thus they can be used to store raw data potentially in every format (e.g. relational data, texts, logs, images, external APIs, streaming data, etc.), without any preprocessing.

This is made possible by a complex functional architecture (shown in Figure 1), composed of several components which are responsible, among others, for data cleaning, data discovery, data integration, and data catalog processes [7]. Many approaches can be exploited to ensure proper integration of data and concepts, often leveraging and refining ontologies [8] to create a mapping between concepts that are expressed in different datasets ingested by the data lake.

All the components of a data lake usually leverage machine learning techniques and profiling tools to extract metadata useful for describing data and creating connections among datasets. The stored data is continuously analyzed, in order to discover new information leveraging the novel data that is loaded into the data lake, thus refining the inherent latent knowledge.

The data lake paradigm was theorized back in 2011 and, even if there are some implementations [9, 10], given its peculiarities and complex architecture there are many research challenges still to be solved [11].

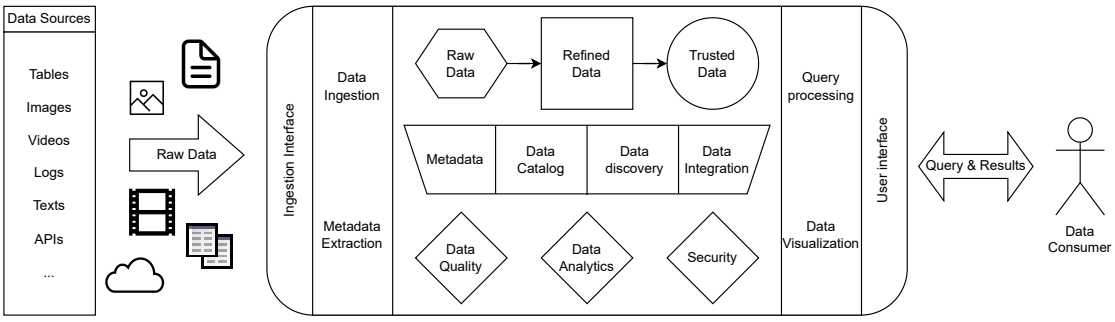

Figure 1: Common architecture of a data lake.

2. Contributions

In this Section we will present our contributions to open challenges regarding the topics described in Section 1. More specifically, Section 2.1 describes the *multi-truth data fusion* problem and presents our approach to solve it, Section 2.2 shows how we exploited *ontology reasoning* to analyze unstructured data. Both these approaches can be used as part of the data integration and data discovery components of a data lake.

We are also contributing to two research projects: (i) INAIL BRiC RECKOn, trying to reduce the number of work accidents by analyzing historical reports and real-time data; (ii) HBD Health Big Data[1], developing an infrastructure to easily share research data among different medical institutes. The contributions presented, when possible, leverage these research projects as case studies.

2.1. Multi-truth data fusion

With the abundance of data sources, the data fusion step must necessarily be carried out automatically, without any human support. In this context is born the problem of *multi-truth data fusion*, related to the cardinality of the values composing the truth: examples being the cases of work accidents, where more than one worker might be involved, or of data coming from multiple sensors at different sampling rates, or of patients having more than one pathology. The main difficulty of the multi-truth context is that, if two data sources provide different values for the same data object, we cannot conclude that they necessarily oppose each other as if we were in a single-truth context. This is particularly challenging if neither the true values nor their cardinality is known a priori, and therefore we have no clue on how the values provided by the sources are interrelated.

[1]https://www.alleanzacontroilcancro.it/progetti/health-big-data/

To tackle this interesting problem, we developed a novel domain-aware Bayesian algorithm for data fusion, designed for the multi-truth case, which models the trustworthiness of sources by taking into account a new definition of their *authority*. Our algorithm also provides a value-reconciliation step, to group together the values that have been recognized as variants representing the same real-world entity. Our approach is described in a paper submitted to an international journal, currently under reviewing process.

In the context of the RECKOn project, we have also defined a framework handling the integration of context-aware real-time sensor data, historical data, and external services, to determine if the current state is a possible dangerous working situation. We have published this framework in [12].

2.2. Ontology reasoning

We can also exploit ontologies to resolve the heterogeneity of data formats, semantics, and representations, within the data integration process. In the RECKOn project, we have applied modern techniques of ontology reasoning and concept extraction to better analyze and compare the official government reports of past work injuries, which are available only in Italian natural language texts. We published in [13] a new pipeline for the extraction of medical concepts from Italian texts, leveraging NLP operations and UMLS as reference ontology.

This methodology is also applicable to the analysis of Electronic Health Records (EHR), which is a comprehensive, cross-institutional, and longitudinal collection of healthcare data, trying to group the entire clinical life of a patient [14]. EHR can be categorized into structured (e.g. personal information, diagnosis codes, laboratory results, etc.) and unstructured (e.g. clinical notes, discharge summaries, etc.), the latter being the most complete and thus complex to inspect. Using our pipeline, we can automatically analyze and compare unstructured EHR, extracting valuable knowledge that can be exploited, for example, in patient-modeling and clinical decision support systems.

3. Future works

Many works [15, 16, 17] show that a key element to maximize the potentialities of the data lake paradigm is the capability to properly handle *metadata*. We are currently exploring new techniques to extract metadata from heterogeneous sources, to automatically and efficiently discover relations among data, and build data catalogs.

Another interesting challenge is the *continuous integration* and analysis needed in the data lake paradigm: when new input data become available, it could serve as a connection among data that were previously unlinked. As an example, consider having in a data lake a dataset regarding the air pollution level of all the major cities of Italy. We can assume that this dataset is simply a table with a few attributes: city, date, and amount of PM10. We now provide the data lake with another dataset regarding the quality of public water, which has the following attributes: city, province, region, and quality. As a result, the data lake should be able to automatically link the common data at different levels, discovering newly available information such as the average amount of PM10 for each province and region, and a possible correlation between air pollution and public water quality. Moreover, it should leverage the hierarchical structure of

Italian regions, provinces, and cities in any other stored dataset. The concepts of continuous and real-time data integration have already been studied in the context of data warehouses [18], but the paradigm shift introduced by data lakes brings this challenge to a completely different level, still to be explored.

Acknowledgement

I wish to acknowledge my advisor Prof. Tanca and my fellow colleagues at TISLab for their massive and experienced support.

References

[1] 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.
[2] X. L. Dong, D. Srivastava, Big data integration, Synthesis Lectures on Data Management 7 (2015) 1–198.
[3] A. Doan, A. Halevy, Z. Ives, Principles of data integration, Elsevier, 2012.
[4] K. Sahatqija, J. Ajdari, X. Zenuni, B. Raufi, F. Ismaili, Comparison between relational and nosql databases, in: 2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0216–0221.
[5] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. Jagadish, Scaling entity resolution: A loosely schema-aware approach, Information Systems 83 (2019) 145–165.
[6] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, W. Zhang, From data fusion to knowledge fusion, arXiv preprint arXiv:1503.00302 (2015).
[7] N. Miloslavskaya, A. Tolstoy, Big data, fast data and data lake concepts, Procedia Computer Science 88 (2016) 300–305.
[8] S. Bergamaschi, S. Castano, M. Vincini, Semantic integration of semistructured and structured data sources, ACM Sigmod Record 28 (1999) 54–59.
[9] R. Hai, S. Geisler, C. Quix, Constance: An intelligent data lake system, in: Proceedings of the 2016 international conference on management of data, 2016, pp. 2097–2100.
[10] A. Beheshti, B. Benatallah, R. Nouri, V. M. Chhieng, H. Xiong, X. Zhao, Coredb: a data lake service, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 2451–2454.
[11] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, P. C. Arocena, Data lake management: challenges and opportunities, Proceedings of the VLDB Endowment 12 (2019) 1986–1989.
[12] P. Agnello, S. M. Ansaldi, E. Lenzi, A. Mongelluzzo, D. Piantella, M. Roveri, F. A. Schreiber, A. Scutti, M. Shekari, L. Tanca, Reckon: a real-world, context-aware knowledge-based lab, in: 29th Italian Symposium on Advanced Database Systems, SEBD 2021, volume 2994, CEUR-WS, 2021, pp. 466–473.
[13] P. Agnello, S. M. Ansaldi, F. Azzalini, G. Gangemi, D. Piantella, E. Rabosio, L. Tanca, Extraction of medical concepts from italian natural language descriptions, in: 29th Italian

Symposium on Advanced Database Systems, SEBD 2021, volume 2994, CEUR-WS, 2021, pp. 275–282.

[14] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE journal of biomedical and health informatics 22 (2017) 1589–1604.

[15] P. N. Sawadogo, E. Scholly, C. Favre, E. Ferey, S. Loudcher, J. Darmont, Metadata systems for data lakes: models and features, in: European conference on advances in databases and information systems, Springer, 2019, pp. 440–451.

[16] F. Ravat, Y. Zhao, Metadata management for data lakes, in: European Conference on Advances in Databases and Information Systems, Springer, 2019, pp. 37–44.

[17] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, B. Mitschang, Modeling metadata in data lakes—a generic model, Data & Knowledge Engineering 136 (2021) 101931.

[18] R. M. Bruckner, B. List, J. Schiefer, Striving towards near real-time data integration for data warehouses, in: International Conference on Data Warehousing and Knowledge Discovery, Springer, 2002, pp. 317–326.