



REVUE ALGERIENNE DES SCIENCES

Section -A-

Sciences de la Nature et de la Vie

Sciences techniques

<http://univ-eltarf.dz/fr/>



Semantic Big Data integration: A survey

Hassiba Laifa^{a*}, Raoudha Khcherif^a, Henda Hajjami BEN GHEZALA^a

^aNational School for Computer Sciences (ENSI), University of Manouba, Manouba, Tunisia, 2010.

Informations	Abstract
Key words : Big Data Data integration ETL Ontology *Correspondence : hassiba.laifa@ensi-uma.tn	The emergence of Big Data has increased the amount of daily produced data. These data are in various formats namely: email, audio, video and structured/unstructured text data, etc. In fact, traditional databases and software techniques cannot handle such huge and heterogeneous information. Therefore, new data management systems should be adopted to integrate this distributed data and to abstract this heterogeneity. In this context Semantic Web (SW) technology, especially ontologies were widely used in the data integration system to eliminate semantic data heterogeneity. Mining the state of the art, two main techniques exist, physical and virtual data integration. In this context, several research works around the aforementioned methods were proposed. In this paper, we studied the most relevant research works in the semantic Big Data integration area. In fact, we performed a comparative study and its discussion of the different approaches proposed to achieve both physical and virtual techniques.

1. Introduction

Nowadays, a huge amount of data comes from several sources. Indeed, these data can be disparate, on different formats, from various sources and without uniformity; it means they can be structured or unstructured. It is generally characterized by the 3V's: a large volume of data to be processed, a wide variety of information from several sources and velocity, which is the frequency, or speed of creating, collecting and sharing this information. Thus, Big Data requires a specific infrastructure for their storage, processing and analysis. Companies are struggling on how to integrate, query and analyze such data.

Data integration is one of the oldest research areas in the field of databases. This area deals with the problem of providing a unified view about heterogeneous and distributed data to the user by combining data residing in different sources [1]. Besides, applications that need to query multiple data sources face the challenge of integrating heterogeneous and autonomous data. Therefore, data integration is an important phase in large enterprises having a multitude of data sources [2]. Thus, offering uniform access to a set of autonomous and heterogeneous data sources is the goal of data integration systems [3].

Big data integration differs from the traditional data integration in terms of (1) the number of data sources, (2) dynamicity of data sources, (3) extreme heterogeneity of data sources and (4)

the wide difference in quality between data sources, with significant differences in the coverage, accuracy and timeliness of the provided data [4]. In fact, the task of data integration has become more difficult and more complex and the semantics of the data can be lost if it arrives with huge volume and high velocity. In this context, ontologies were widely used in semantic data integration systems [5][6]. According to [7], an ontology is a formal, explicit specification of a shared conceptualization for a domain. They have been developed in Artificial Intelligence to facilitate knowledge sharing and reuse [8]. They were employed in data integration systems to solve syntactic and semantic heterogeneity problems by providing a semantic model of data sets to be integrated.

In the literature, several works and approaches addressed the issue of semantic Big Data integration. Those approaches can be classified into physical (or materialized) integration, where the data is physically copied to dedicated storage (such as Data Warehouse), and virtual integration in which data remains in the original sources and is accessible during the execution of the query on an intermediate interface.

In this paper, we summarized different semantic data integration approaches. Then, we compare the aforementioned approaches and discuss the comparative study. This manuscript is organized as follows: The first section regroups a state of

the art of semantic data integration approaches. The second section contains a comparative study of works presented. The comparative studies are discussed in the third section. We end the work with a conclusion in the fourth section.

2. State of the art

In physical data integration, data is extracted from different sources and loaded into common storage after performing a set of transformation and cleaning operations to make them uniform.

The data warehouse, which is the main approach of physical data integration defined by Inmon, W.H in 1992 as a "subject-oriented, integrated, time-varying, non-volatile method of data collection, is used primarily in organizational decision making". Also known as data warehouse population, data warehousing is done through the ETL (Extract, Transform, and Load) process. The ETL process is composed of 3 phases. The first phase consists in extracting data from different sources. The second step is the transformation phase involving the conversion of data into the target schemas [9] by applying a series of cleaning operations including eliminating duplicates, filling empty fields, converting data types, etc. The final phase (loading) is to store the transformed data in the data warehouse.

In this data integration type, the response time to the user's requests is short because the query will look for a result from a single data source represented by the common storage (DW). Therefore, the physical data integration approaches suffer from the freshness of data as enough fresh data cannot be found only after the updating phase.

Virtual integration (also known as mediator [10]) is defined as an approach applied to provide an intermediate tool between users or applications, on one hand, and a set of autonomous, heterogeneous, distributed and scalable information sources, on the other hand. This tool offers a transparent access service to sources through a unique interface and a single query language" [11]. In the virtual integration, data remains in the sources and the query is executed at the intermediary interface. The query executed in the intermediate is converted into the formats of the local data sources. Then, results of sub-query retrieved data need to be combined to answer the query.

In the virtual integration, the query is slow because it will be divided into sub-requests and sent to the wrappers of the individual sources to be executed on local models. Then, the intermediate interface receives the responses of the wrappers, combines them into one answer and sends it to the user. This process is time-

consuming. However, in the virtual integration approaches, data is always fresh and updated because it is extracted from the sources directly.

We presented in the next section several works for Big Data integration approach, in particularly semantics ones.

- In [12], authors have proposed a data integration framework where the target schema was represented as a semantic web ontology and the sources corresponded to NoSQL databases. Currently, they are implementing their data integration solution on two popular NoSQL databases: MongoDB, as a document database, and Cassandra as a column family store. Their contributions are: (1) Analyzing a set of schema-less NoSQL databases to generate local ontologies; (2) Generating a global ontology based on the discovery of correspondences between the local ontologies; and finally (3) proposing a query translation solution from SPARQL to query languages of the sources.

- In [13], the authors proposed a semantic ETL framework that uses semantic technologies to integrate and publish data from multiple sources. Those technologies were also applied to connect, link and load data into a data warehouse. The proposed framework includes: (1) creating semantic data model via ontologies to provide a basis for integration and understanding knowledge from multiple sources; (2) forming integrated semantic data using Resource Description Framework (RDF) as the graph data model and (3) extracting useful knowledge and information from the combined web of data using SPARQL as the semantic query language. This approach uses ontologies to provide a common vocabulary for the integrated data and generated semantic data as part of the transformation phase of ETL.

- Handling the semantic heterogeneity for Big Data application in industrial automation using Semantic Web technologies is the main topic of [14]. Authors, in this study, dealt with structural heterogeneity (or syntactic heterogeneity) by taking into consideration various data sources to ensure quick and valuable decisions as well as semantic heterogeneity by creating a shared ontology, which guaranteed the transformation of the data sources into the same "language". The developed approach includes the following steps: (1) Construction of upper ontology describing all data sources, (2) Transformation of data according to this ontology and (3) Analysis of data with the help of Big Data paradigm.

- In [15], researchers described a (Python-based) programmable Semantic ETL framework (SETL) to process and integrate data semantically. This approach uses the Semantic Web (SW) standards and tools to produce a

semantic data warehouse based on RDF triples. SETL involves traditional relational data, semantically annotated data (RDF data) in the analytical process, and provides a method to produce semantic data (in RDF triples format) from the source data. The main phases are the following: The first step consists in defining an ontology that describes the relevant data and the target data warehouse applying the Ontology Definition module. After mapping between data sources and the target ontology using the Define Mapping module, the Extraction module extracts data from multiple heterogeneous data sources. The third phase is to transform the source data. In this step, the Traditional Transformation module cleanses and formats the extracted data and the Semantic Transformation module converts the data into RDF triples according to the target ontology. The final phase consists of loading the data into a triple store and/or publishing the data on the Web as Linked Data.

- Authors of [16] have presented a semantic mediation architecture based on ontology for reducing the semantic problems, and on the summary for reducing the quantity of data with minimum execution time. A domain ontology on alimentation risks field is built to use as well as a global schema of the mediator. Their architecture is split into three separate layers: Mediator layer, Wrapper layer and data source layer. (1) The mediator layer represents the unified interface for querying heterogeneous data sources by using the global ontology. (2) A wrapper layer is a middle layer between the mediator layer and data source layer. The initial query is written by SPARQL language of this ontology and the wrapper will translate this query to sources query language and extraction of answers. (3) The source layer contains external data sources. These local sources are voluminous and it associated by its own ontology represented the knowledge in an independent with other local ontologies.

- Paper [17] introduce SemLinker, an ontology-based data integration system. SemLinker is part of a metadata management framework for personal data lake (PDL). It adopts an automatic approach that only operates on the schema metadata level without involving the physical transformation of data during integration. Thus, it preserves the data in their native formats and structures while, at the same time, allowing the data to be easily analyzed and queried by casual users. The SemLinker architecture consists of three layers: (1) a global schema layer G, which is modelled as a global ontology and is described using web ontology language (OWL). (2) a local schemata layer which is described using the resource description framework (RDF) and is stored in

the schema's repository S., and (3) the relationships between these layers were SemLinker is responsible for automatically mapping the local schema Si of the data source i to a semantically corresponding concept in the global ontology G.

- In [18], an ontology-based Big Data integration approach, exploiting a NoSQL database (namely MongoDB), was depicted. It is based on three main steps: (1) by wrapping data sources to MongoDB databases, the content of each data source was converted into a MongoDB database; (2) mapping MongoDB databases into ontology modules where each MongoDB database was mapped into an OWL ontology module by means of transformation rules; and (3) Creating a global ontology module by merging the modules obtained in the previous step. They employed two tools: M2Onto (MongoDB To Ontology), implementing transformation rules from MongoDB to ontology modules, and MOOM (Modular Owl Ontology Merging) allowing to merge the ontology modules resulting from the previous step to get a global one.

3. Comparative study

We analyzed the above approaches using the four following criteria: (1) Approach type to identify the type of data integration approach (physical or virtual), (2) Data sources (depicting the formats of the data sources), (3) Target schema was taken into account to describe the output of the process and (4) Ontology construction automaticity degree grouped the approaches to build ontology as Manual (M), semi-manual (SM) or automatic (A).

Work	Approach type	Data sources	Target schema	Ontology construction automaticity		
				M	SM	A
[12]	Virtual	NoSql DB (MongoDB and Cassandra)	Ontology		+	
[13]	Physical	Big data	DW	+		
[14]	Physical	Big data	RDF shared storage		+	
[15]	Physical	Big data	DW		+	
[16]	Virtual	External data sources	Ontology		+	
[17]	Physical	Big data	key-value tuple store			+
[18]	Virtual	Databases wrapped to MongoDB	Ontology			+

Tab.1.Comparative study

4. Discussion

From the exhaustive study of both physical and virtual integration approaches, we mentioned that the main challenge faced is the data heterogeneity. Indeed, heterogeneous data requires more effective solutions to perform the integration process. To overcome this problem, Semantic Web (SW) technologies were used. In this context, ontologies were widely utilized in data integration system, especially in the transformation stage in physical approaches, to create a semantic data model and generate semantic linked data to be stored in the common storage, also as a querying model in virtual approaches.

One of the challenges of [13] is the engineering of ontology, which requires the intervention of a human expert and it is often a time-consuming process. Otherwise, there are systems that need real-time data processing. The manual ontology construction in [14] is a complicated and time-consuming process.

Work in [12] does not support NoSQL stores corresponding to a graph model, is dedicated only for both the MongoDB and Cassandra NOSQL databases. In addition, [14, 18] cannot support all NoSQL system.

These works tried to solve the problem of semantic heterogeneity without taking into account the volume and velocity of the data. The semantics of the data can be lost if it arrives with huge volume and high velocity.

As perspectives, there is a need to consider the solving of automaticity problem to limit manual intervention. It is also necessary to incorporate another category of NOSQL stores: graph databases.

5. Conclusion

In this paper, we analyzed a set of works that dealt with big data integration. These works were divided into 2 types according to their integration process. The first type includes physical approaches where data from different sources are copied to common storage after a set of transformation operations to make them uniform. In the other hand, in virtual integration type, data remain in the origin's sources. To interrogate them, the user's query will be executed in an intermediate interface. To solve semantic problem in data integration system ontologies were widely utilized, especially in the transformation stage in physical approaches, to create a semantic data model and generate semantic linked data to be stored in the common storage, also as a querying model in virtual approaches.

Reference:

- [1] Lenzerini, M.: Data integration: A theoretical perspective. In: Proceedings of the twenty first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 233{246. ACM (2002)
- [2] Halevy, A., Rajaraman, A., Ordille, J.: Data integration: the teenage years. In: Proceedings of the 32nd international conference on Very large databases. pp. 9{16. VLDB Endowment (2006)
- [3] Doan, A., Halevy, A., Ives, Z.: Principles of data integration. Elsevier (2012)
- [4] Dong, X.L., Srivastava, D.: Big data integration. In: 2013 IEEE 29th international conference on data engineering (ICDE). pp. 1245{1248. IEEE (2013)
- [5] Doan, A., & Halevy, A. Y. (2005). Semantic integration research in the database community: A brief survey. *AI magazine*, 26(1), 83-83.
- [6] Noy, N. F. (2004). Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record*, 33(4), 65-70.
- [7] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2), 199-221.
- [8] Fensel, D. (2001). Ontologies. In *Ontologies* (pp. 11-18). Springer, Berlin, Heidelberg.
- [9] Vassiliadis, P., Simitsis, A., Baikousi, E.: A taxonomy of etl activities. In: Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP. pp. 25{32. ACM (2009)
- [10] Wiederhold, G.: Mediators in the architecture of future information systems. *Computer* 25(3), 38{49 (1992)
- [11] Zellou, A.: Contribution a la reecriture lav dans le contexte de wassit, vers un framework d'integration de ressources. These de Doctorat. Rabat. Maroc. Avril (2008)
- [12] Cure, O., Lamolle, M., Duc, C.L.: Ontology based data integration over document and column family oriented nosql. *arXiv preprint arXiv:1307.2603* (2013)
- [13] Bansal, S.K.: Towards a semantic extract-transform-load (etl) framework for big data integration. In: 2014 IEEE International Congress on Big Data. pp. 522{529. IEEE (2014)
- [14] Jirkovsky, V., Obitko, M.: Semantic heterogeneity reduction for big data in industrial automation. In: ITAT (2014)
- [15] Deb Nath, R.P., Hose, K., Pedersen, T.B.: Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In: Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP. pp. 15{24. ACM (2015)
- [16] Aggoune, A., Bouramoul, A., & Kholliadi, M. K. (2016, March). Big data integration: A semantic mediation architecture using summary. In 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 21-25). IEEE.
- [17] Alrehamy, H., & Walker, C. (2018). SemLinker: automating big data integration for casual users. *Journal of Big Data*, 5(1), 1-26.
- [18] Abbes, H., Gargouri, F.: Mongodb-based modular ontology building for big data integration. *Journal on Data Semantics* 7(1), 1{27 (2018)