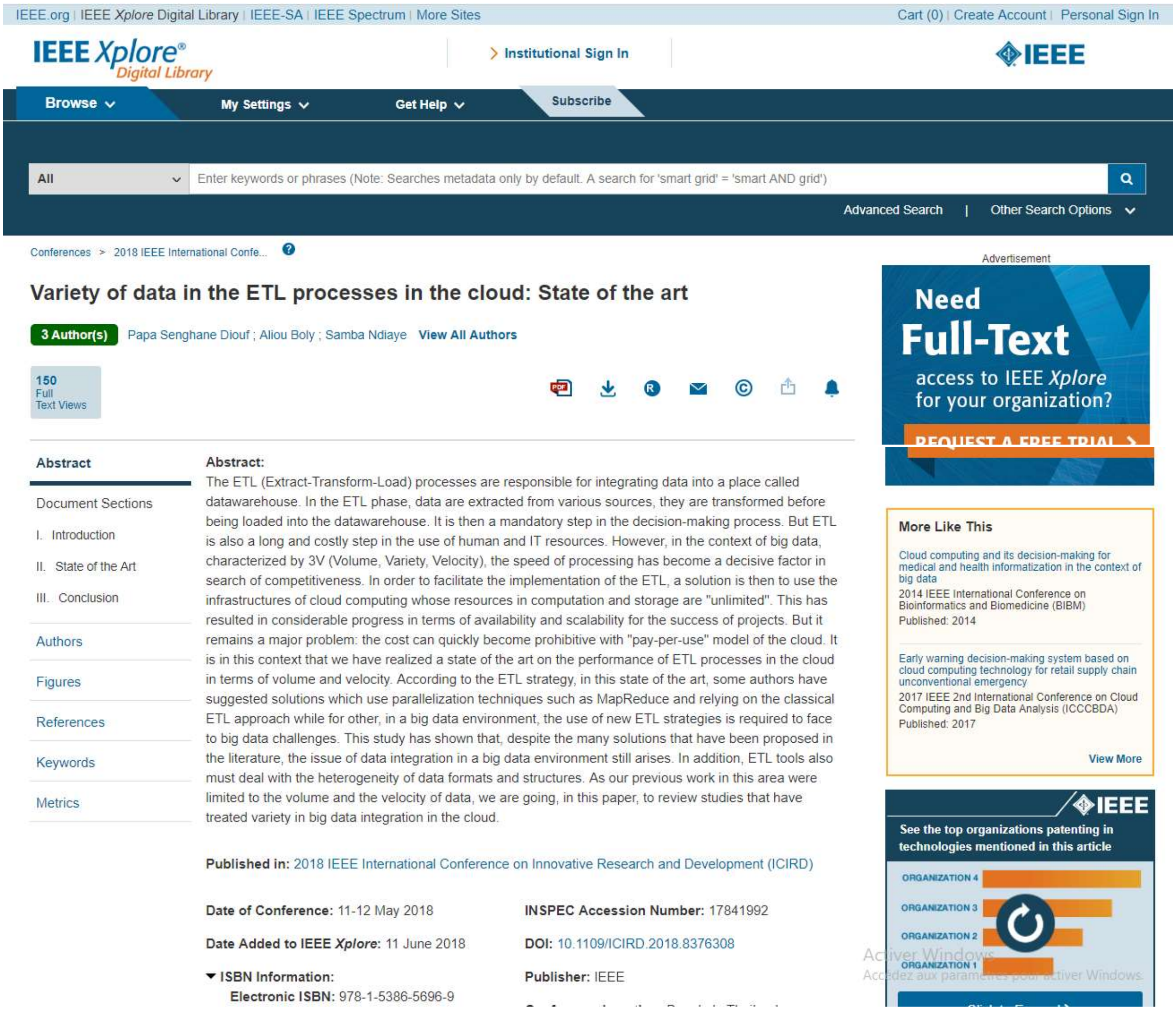
Publication pour la présente candidature LAFPT

Titre	Variety of data in the ETL processes in the cloud: state of the art
Auteurs	Papa Senghane Diouf ; Aliou Boly ; **Samba Ndiaye**
Référence	2018 IEEE International Conference on Innovative Research and Development (ICIRD)
Editeur	IEEE
Pages	95-99
Année	2018
DOI	10.1109/ICIRD.2018.8376308
URL	https://ieeexplore.ieee.org/document/8376308
Index	https://www.scopus.com/authid/detail.uri?authorId=6701604512
ISBN	978-1-5386-5696-9
Encadreur	Oui
Extraire d'une thèse	Non



IEEE.org | IEEE *Xplore* Digital Library | IEEE-SA | IEEE Spectrum | More Sites   Cart (0) | Create Account | Personal Sign In

IEEE *Xplore*® Digital Library   > Institutional Sign In   IEEE

Browse   My Settings   Get Help   Subscribe

All   Enter keywords or phrases (Note: Searches metadata only by default. A search for 'smart grid' = 'smart AND grid')

Advanced Search | Other Search Options

Conferences > 2018 IEEE International Confe...

Variety of data in the ETL processes in the cloud: State of the art

3 Author(s) Papa Senghane Diouf ; Aliou Boly ; Samba Ndiaye View All Authors

150 Full Text Views

Abstract

Document Sections

I. Introduction

II. State of the Art

III. Conclusion

Authors

Figures

References

Keywords

Metrics

Abstract:
The ETL (Extract-Transform-Load) processes are responsible for integrating data into a place called datawarehouse. In the ETL phase, data are extracted from various sources, they are transformed before being loaded into the datawarehouse. It is then a mandatory step in the decision-making process. But ETL is also a long and costly step in the use of human and IT resources. However, in the context of big data, characterized by 3V (Volume, Variety, Velocity), the speed of processing has become a decisive factor in search of competitiveness. In order to facilitate the implementation of the ETL, a solution is then to use the infrastructures of cloud computing whose resources in computation and storage are "unlimited". This has resulted in considerable progress in terms of availability and scalability for the success of projects. But it remains a major problem: the cost can quickly become prohibitive with "pay-per-use" model of the cloud. It is in this context that we have realized a state of the art on the performance of ETL processes in the cloud in terms of volume and velocity. According to the ETL strategy, in this state of the art, some authors have suggested solutions which use parallelization techniques such as MapReduce and relying on the classical ETL approach while for other, in a big data environment, the use of new ETL strategies is required to face to big data challenges. This study has shown that, despite the many solutions that have been proposed in the literature, the issue of data integration in a big data environment still arises. In addition, ETL tools also must deal with the heterogeneity of data formats and structures. As our previous work in this area were limited to the volume and the velocity of data, we are going, in this paper, to review studies that have treated variety in big data integration in the cloud.

Published in: 2018 IEEE International Conference on Innovative Research and Development (ICIRD)

Date of Conference: 11-12 May 2018   **INSPEC Accession Number:** 17841992

Date Added to IEEE *Xplore*: 11 June 2018   **DOI:** 10.1109/ICIRD.2018.8376308

▾ **ISBN Information:**   **Publisher:** IEEE

Electronic ISBN: 978-1-5386-5696-9   **Conference Location:** Bangkok, Thailand

More Like This

Cloud computing and its decision-making for medical and health informatization in the context of big data
2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
Published: 2014

Early warning decision-making system based on cloud computing technology for retail supply chain unconventional emergency
2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)
Published: 2017

View More

2018 IEEE International Conference on Innovative Research and Development (ICIRD)
11-12 May 2018,Bangkok Thailand

Variety of data in the ETL processes in the cloud: state of the art

Papa Senghane Diouf
Department of mathematics and computer science
Cheikh Anta Diop University
Dakar, Senegal
papasenghane.diouf@ucad.edu.sn

Aliou Boly
Department of mathematics and computer science
Cheikh Anta Diop University
Dakar, Senegal
aliou.boly@ucad.edu.sn

Samba Ndiaye
Department of mathematics and computer science
Cheikh Anta Diop University
Dakar, Senegal
samba.ndiaye@ucad.edu.sn

***Abstract*—The ETL (Extract-Transform-Load) processes are responsible for integrating data into a place called datawarehouse. In the ETL phase, data are extracted from various sources, they are transformed before being loaded into the datawarehouse. It is then a mandatory step in the decision-making process. But ETL is also a long and costly step in the use of human and IT resources. However, in the context of big data, characterized by 3V (Volume, Variety, Velocity), the speed of processing has become a decisive factor in search of competitiveness. In order to facilitate the implementation of the ETL, a solution is then to use the infrastructures of cloud computing whose resources in computation and storage are "unlimited". This has resulted in considerable progress in terms of availability and scalability for the success of projects. But it remains a major problem: the cost can quickly become prohibitive with "pay-per-use" model of the cloud. It is in this context that we have realized a state of the art on the performance of ETL processes in the cloud in terms of volume and velocity. According to the ETL strategy, in this state of the art, some authors have suggested solutions which use parallelization techniques such as MapReduce and relying on the classical ETL approach while for other, in a big data environment, the use of new ETL strategies is required to face to big data challenges. This study has shown that, despite the many solutions that have been proposed in the literature, the issue of data integration in a big data environment still arises. In addition, ETL tools also must deal with the heterogeneity of data formats and structures. As our previous work in this area were limited to the volume and the velocity of data, we are going, in this paper, to review studies that have treated variety in big data integration in the cloud.**

Keywords*—*etl, big data, cloud, cost, variety

I. INTRODUCTION

In order to improve the decision-making process, Business Intelligence (BI) tools are used. They allow the production of relevant information and knowledge within organizations. That is why their use has now become widespread. A fundamental step in the process of implementing a decision-making system is the design and implementation of a datawarehouse. In this process, data from various sources are first and for most extracted, then cleaned and eventually standardized before being stored into the datawarehouse. This very important step in the decision-making process is called ETL (Extract - Transform - Load). It is the biggest task of the datawarehousing process. According to Kimball [1], ETL easily consumes 70 percent of the resources needed for implementation and maintenance of a typical datawarehouse. Nowadays, organizations produce large amounts of data in a wide variety of formats at a rapid rate [2] [3]: this is the era of big data. All this makes the ETL step difficult, time consuming and costly. The latter must adapt to the context of big data because the traditional approaches prove to be inadequate or even impossible to implement [4] [5] [6]. This involves using cloud features such as availability and resource elasticity [7] [8], usage billing, data remoteness, and so on. Based on these observations, first, we have realized a state of the art about the performance of ETL processes in the cloud in terms of volume and velocity [9]. In this paper we note that some authors have proposed ETL solutions built on the classical ETL approach and which use parallelization techniques for improving the performance of the ETL processes. For others, a new ETL approach as ELT (Extract-Load-Transform) or TEL (Transform – Extract - Load) is needed for dealing with data in a big data context. This study [9] has shown that, despite the many solutions proposed in the literature, the issue of big data integration is still topical. As a result of this study, in this paper, we have reviewed the works that have addressed the issue of the variety in the big data integration. Indeed, in the big data environment, data come from heterogeneous sources and they have diverse formats and structures (structured, semi-structured and unstructured). These properties of data are referred to as the variety [10]. ETL processes have then to deal with these kinds of data. Addressing big data integration is then a challenge. It requires a large computational infrastructure to ensure successful data processing [6]. As cloud provides for dynamic resource scaling, this makes it a natural fit for big data applications [11]. In the literature, there are some proposals that have been made for solving the issue of the variety in big data integration. In the following section we present these studies.

II. STATE OF THE ART

Solutions are proposed to solve the problem of variety in the context of big data integration. Starting from the fact that their deployment in the cloud is without incident, we have distinguished mainly the following works:

A. Research works done by Skoutas et al. [12]:

In [12], authors have proposed an ontology-based approach for facilitating the conceptual design of ETL processes. A graph representation is used as a conceptual model for the datastores. This makes possible, according to authors [12], to deal with and to support structured and semi-structured data. This proposed approach [12] rely on the use of semantic web technologies to annotate the data sources and the datawarehouse. The OWL-DL language is then chosen for

978-1-5386-5283-1/18/$31.00 ©2018 IEEE
978-1-5386-5696-9/18/$31.00 ©2018 IEEE

describing the ontology. This has solved, the question of the mapping between data sources and the datawarehouse. However, this proposal is only a solution for the variety of data sources (heterogeneity). It only takes into account structured and semi-structured data. Thus, the unstructured data are not taken into consideration in this study. In summary, the issue of variety has not been fully addressed by Skoutas et al. [12].

B. Research works done by Salmen et al [13]:

Salmen et al. [13] have designed a strategy based on the Semantic Enhancement (SE) for data integration. Using a framework called Data Representation and Integration Framework (DRIF), the proposed solution has permitted to deal the full spectrum of data sources, types, models and modalities. Indeed, a simple data representation scheme is used to encapsulate every piece of data from heterogeneous sources into a unified representation. In this integration approach, [13], DRIF, is applied on the large heterogenous data. The result obtained is called dataspace. This dataspace is organized as follows:

- Segment 0: it is the level where are the heterogeneous data sources;
- Segment 1: this segment manages unstructured data;
- Segment 2: it stores the structured data coming from Segment 0;
- Segment 3: data models and ontologies are defined at this level.

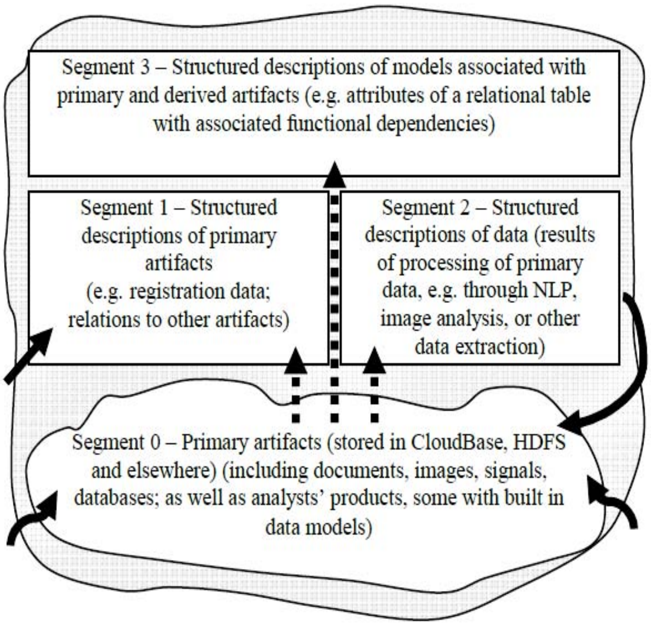

Segment 3 – Structured descriptions of models associated with primary and derived artifacts (e.g. attributes of a relational table with associated functional dependencies)
Segment 1 – Structured descriptions of primary artifacts (e.g. registration data; relations to other artifacts)
Segment 2 – Structured descriptions of data (results of processing of primary data, e.g. through NLP, image analysis, or other data extraction)
Segment 0 – Primary artifacts (stored in CloudBase, HDFS and elsewhere) (including documents, images, signals, databases; as well as analysts' products, some with built in data models)

Fig 1. Organization of the dataspace [13]

For authors, the dataspace can integrate hundreds of millions of unstructured documents and large quantities of images, signals data, and other structured and unstructured data [13]. Once the dataspace is building, the Semantic Enhancement (SE) strategy is applied on it. Then, SE add a semantic layer to this dataspace. The proposals made in [13] allows to meet the challenge of integrating with various formats and structures. But, this solution use a unified representation of data which leads to the loss of information on the data and reduces the accuracy of the model. Moreover the proposed model is static, the appearance of data with new structures requires the reconstruction of the dataspace. This can be very tight with the velocity of data.

C. Research works done by Boury-Brisset [14]:

For dealing with data that have various format and structures, Boury-Brisset [14] has proposed to rely on semantic and big data technologies. In this study, author has suggested the design and implementation of a framework for scalable Multi-Intelligence Data Integration Services (MIDIS). Through this framework, the goal of Boury-Brisset [14] is then to facilitate the integration of heterogeneous unstructured and structured data [14]. The Fig. 2 illustrates the architecture of MIDIS and the data flow from data collected from various sources, their integration, to intelligence analysis [14].

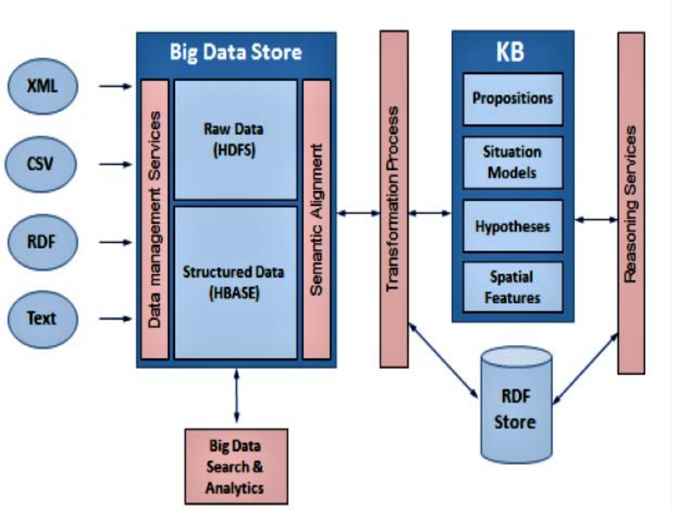


Fig 2. Multi-Intelligence Data Integration Services [14]

Although this proposal is very interesting, it does not cover all aspects of the variety. Indeed, the model doesn't consider very common formats in a big data environment such as image and video formats. It doesn't effectively represent the semantic relation in the big data from heterogeneous sources [14].

D. Research works done by Bansal et al. [15] [16]:

Bansal et al. [15] have found that in the big data era, it is difficult for traditional ETL systems to process data which come from heterogenous sources and having various formats and structures. To ensure this issue, they have proposed a semantic ETL framework that uses semantic technologies to integrate heterogeneous data. To do this, as shown in Fig. 3, a semantic data model is designed through ontologies. This data model offers a basis for integration and understanding knowledge from multiple sources [15]. Also, an integrated semantic data is created using Resource Description Framework (RDF) as the graph data model and the extraction of useful knowledge and information from the combined web is done via a semantic query language called SPARQL. This solution has been implemented on a few public data sets with information on vehicles, household, transportation and fuel economy. However, experimental results obtained with this approach were not presented by authors. In this study, we therefore do not have a precise idea on the impact of the use of this approach on the performance of ETL processes.

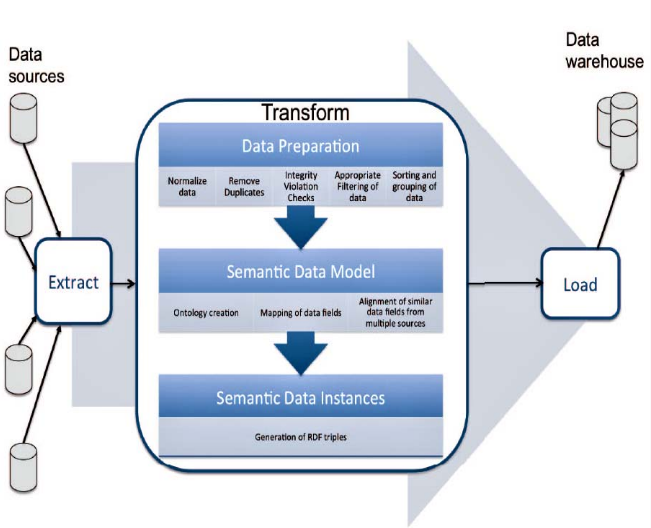


Fig 3. Overview of Semantic ETL Process [15].

E. *Research works done by Souissi et al. [17]:*

In their studies, Souissi et al. [17] were interested in the variety aspect of data in big data integration. They have proposed an ETL tool called GENUS which is able to deal with the issue of the variety of data in a big data environment. As shown in Fig. 4, their solution extracts data from different document types (text, image, video), transform and load them into a datawarehouse. The problem of the variety of data is treated at the transformation stage of the ETL process. The transformation phase is then divided into two parts: data cleaning and extracting main concepts. In data cleaning, data are cleaned to prevent the errors resulting from the extraction phase. After cleaning data, main concepts and metadata are extracted. These latter will be stored in XML files and loaded into datawarehouse. Although this approach allows in some cases to answer the question of the variety of data in the ETL processes, it is too limited to text, image, video types. Moreover, in this proposal the data integration approach used is ETL. This latter is not suitable for large-scale data integration and is considered as outdated. Transformations are made before the loading of data into the datawarehouse. This create a bottleneck. Considering the number of treatments to be done on data, the performances can deteriorate even if parallelization techniques are used. [9]

F. *Research works done by Rani et al. [18]:*

The difficulty of integrating heterogeneous sources is an obstacle. Indeed, unstructured data makes the data integration more complex due to the lack of a well-defined schema. Relying on these facts, in their research works, Rani et al. [18] have proposed MOUNT, a multi-level annotation and integration framework. In this approach, authors [18] have suggested two levels of annotation: the coarse-grained and the fine-grained. In the coarse-grained, Yago ontology and SEeds SEarch (SESE) are used for categorizing the domain information of the heterogeneous data sources. The fine-grained annotation handles the unified data integration of structured and unstructured data.

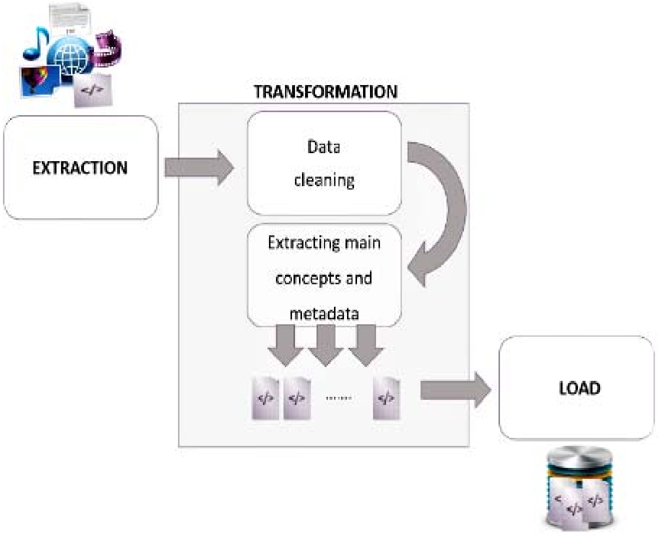


Fig 4. The proposed ETL tool [17].

Based on these levels of annotation, this proposal has focused on applying the semantic enrichment on the big data sources [18]. This solution has permitted to treat with structured and unstructured data and it allows to face to the challenge of the variety of data. The Fig. 6 illustrate the MOUNT methodology. Experimental results have estimated the accuracy of MOUNT at 94%. Despite this good result, the loss of information related to the use of this model can have negative consequences for the decisions made. Also, the accuracy of this model needs to be improved for a quality datawarehouse

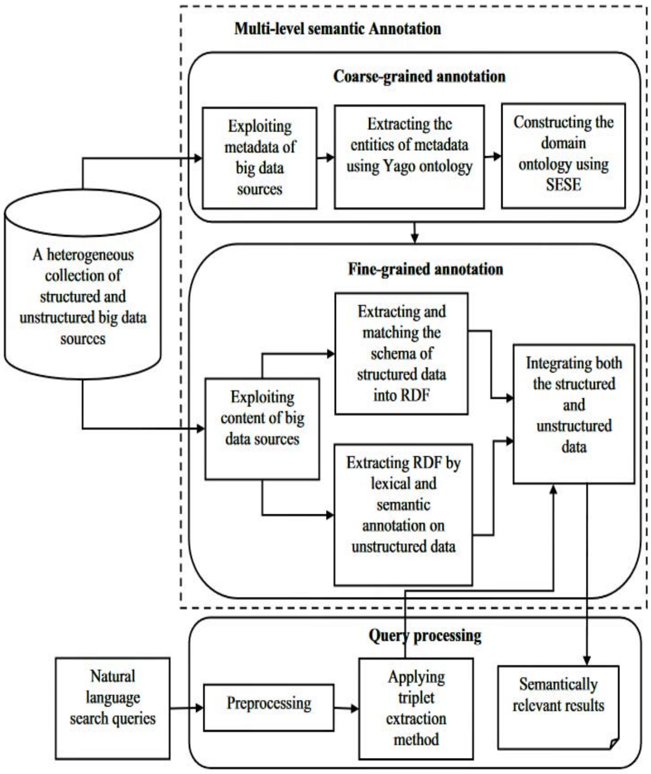

Multi-level semantic Annotation
Coarse-grained annotation
Exploiting metadata of big data sources
Extracting the entities of metadata using Yago ontology
Constructing the domain ontology using SESE
A heterogeneous collection of structured and unstructured big data sources
Fine-grained annotation
Extracting and matching the schema of structured data into RDF
Exploiting content of big data sources
Extracting RDF by lexical and semantic annotation on unstructured data
Integrating both the structured and unstructured data
Query processing
Natural language search queries
Preprocessing
Applying triplet extraction method
Semantically relevant results

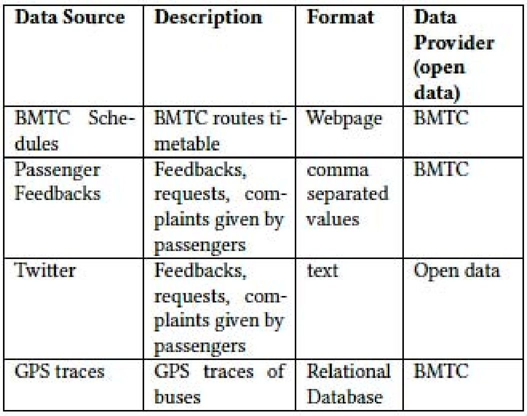
Fig 5. The MOUNT methodology [18]

G. *Research works done by Jaybal et al. [19]:*

In this study [19], authors have made a data analytics framework, called *HDSAnalytics*, for heterogeneous data sources. Indeed, to solve the question of the variety of data, the widely used solution for data integration is often to integrate them into a unified data source. This leads in loss of information due to semantic, syntactic, and schematic differences that appear between data sources [19] Faced with this problem, Jaybal et al. [19] have proposed *HDSAnalytics* for using heterogeneous data sources "As-Is" without integrating into a single data source. In [19], data sources can be used by individually accessing each of these data sources through the user interface of their proposed framework. The latter includes the following components:

- **User interface (UI)** where the user selects some concepts;
- **Analytics Engine (AE)** invokes analytics model based on selected concepts;
- **Data Source Interface (DSI)** includes different data interface(DI) routines implemented to retrieve data in the format requested by the analytic model;
- **Heterogeneous data sources (DS).**

The *HDSAnalytics* solution has been applied to data from Bangalore Metropolitan Transport Corporation (BMTC).



Data Source	Description	Format	Data Provider (open data)
BMTC Schedules	BMTC routes timetable	Webpage	BMTC
Passenger Feedbacks	Feedbacks, requests, complaints given by passengers	comma separated values	BMTC
Twitter	Feedbacks, requests, complaints given by passengers	text	Open data
GPS traces	GPS traces of buses	Relational Database	BMTC

Fig 6. Data Sources for Bangalore City Bus Fleet Analysis [19]

This tool is flexible and allows to process data with new structures. However, the accuracy of this tool has not been compared to that of solutions which integrate data into a unified data source

III. Conclusion

The ETL phase is an essential step in the datawarehousing process. This is by far the longest and most expensive phase in an implementation project of decision-making system. The adaptation of this ETL phase to the big data / cloud context, characterized by large volumes of heterogeneous and remote data, has therefore become a necessity. Nowadays, the cloud has become a very promising solution for handling data at large scale. It is in this context that we are interested in the problem of the big data integration in the cloud. Firstly, we have made a state of the art on the performance of ETL processes in the cloud in terms of volume and velocity. In this state, we have found that several approaches are proposed to address the issue of big data integration. According to the ETL strategy, some authors have suggested solutions which rely on the classical ETL approach while for other, in a big data environment, the use of new ETL strategy is needed to face to big data challenges. Therefore, despite these interesting solutions a problem remains: the variety of data is not considered. To solve this issue, some proposals have been made in the literature. These latter rely mainly on the use of semantic technologies for meeting this challenge of big data. However, we have noted that the proposed solutions haven't consider the dynamicity or even the elasticity of data in the cloud. Also, the issue of the cost of resource used has not been deeply studied in these different approaches. And, since the cloud billing model is in use, exploiting the elasticity of resources would reduce the cost. The implementation of ETL solution that take all these aspects of the cloud and big data into account is a promising prospect. Based on these remarks, we propose, in future, to work on the design and implementation of an ETL tool, that exploit the advantages of cloud computing but also consider the three (3) main characteristics of big data (Volume, Variety and Velocity).

References

[1] Ralph Kimball and Joe Caserta. The data warehouse etl toolkit: practical techniques for extracting. Cleaning, Conforming, and Delivering Data, page 528, 2004.

[2] Anureet Kaur. Big data : A review of challenges, tools and techniques. International journal of scientific research in science, engineering and technology, 2(2):1090–1093, 2016.

[3] Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao. Big data analytics= machine learning+ cloud computing. arXiv preprint arXiv:1601.03115, 2016.

[4] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on,pages 42–47. IEEE, 2013.

[5] Xiufeng Liu, Nadeem Iftikhar, and Xike Xie. Survey of real-time processing systems for big data. In Proceedings of the 18th International Database Engineering & Applications Symposium, pages 356–361. ACM, 2014.

[6] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of big data on cloud computing: Review and open research issues. Information Systems, 47:98–115, 2015.

[7] Barrie Sosinsky. Cloud computing bible, volume 762. John Wiley & Sons, 2010.

[8] Jakobus S Van der Walt. Business intelligence in the cloud. South African Journal of Information Management, 12(1):1–15, 2010.

[9] Papa Senghane Diouf, Aliou Boly, and Samba Ndiaye. Performance of the etl processes in terms of volume and velocity in the cloud: State of the art. In Engineering Technologies and Applied Sciences (ICETAS), 2017 4th IEEE International Conference on, pages 1–5. IEEE, 2017.

[10] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2):157–164, 2013.

[11] Svetoslav Zhelev and Anna Rozeva. Big data processing in the cloudchallenges and platforms. In AIP Conference Proceedings, volume 1910, page 060013. AIP Publishing, 2017.

[12] Dimitrios Skoutas and Alkis Simitsis. Ontology-based conceptual design of etl processes for both structured and semi-structured data. International Journal on Semantic Web and Information Systems (IJSWIS), 3(4):1–24, 2007.

[13] David Salmen, Tatiana Malyuta, Alan Hansen, Shaun Cronen, and Barry Smith. Integration of intelligence data through semantic enhancement. 2011.

[14] Anne-Claire Boury-Brisset. Managing semantic big data for intelligence. In STIDS, pages 41–47, 2013.

[15] Srividya K Bansal. Towards a semantic extract-transform-load (etl) framework for big data integration. In Big Data (BigData Congress), 2014 IEEE International Congress on, pages 522–529. IEEE, 2014.

[16] Srividya K Bansal and Sebastian Kagemann. Integrating big data: A semantic extract-transform-load framework. Computer, 48(3):42–50, 2015.

[17] S. Souissi and M. BenAyed. Genus: An etl tool treating the big data variety. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pages 1–8, Nov 2016.

[18] P Shobha Rani, RM Suresh, and R Sethukarasi. Multi-level semantic annotation and unified data integration using semantic web ontology in big data processing. Cluster Computing, pages 1–13, 2017.

[19] Yogalakshmi Jaybal, Chandrashekar Ramanathan, and S Rajagopalan. Hdsanalytics: a data analytics framework for heterogeneous data sources. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pages 11–19. ACM, 2018.