International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 5 Issue 7, July-2016

Big Data Resource Discovery Considering Semantics in Grid Environment

Imen Ketata[1], Gilles Zurfluh[2]

[1]Gabes University, Tunisia
[2]Toulouse 1 University, France

ABSTRACT

Nowadays, everybody talks about the famous phenomenon called 'Big Data'. No one can escape this term particularly when we talk about large-scale distributed databases, i.e., data grid environment. Resource discovery (data source discovery) is an important step in the management, integration and querying of big data. The addressing protocol adopted for this discovery should respect not only the properties of hardware and software environment (evolution, dynamicity, scalability...) but also big data's properties (volume, variety, velocity...). We propose in this paper an addressing system for big data resource discovery taking into account semantics in grid environment.

Keywords: Big Data, Grid, Maintenance, Resource Discovery.

1. INTRODUCTION

Currently, one on the biggest problems of Business Intelligence (BI) is the hard access to the large amounts of data (big data) stored on the Web. So, these data, with huge volume, are only partially exploited for decision making aid.
According to the 3V principle characterizing big data, the volume is not the only constraint posed by masses of data. There is also variety, that is to say, the mix of data types: graphics, video, documents. Adding to that the velocity, implying that data are undergoing treatment in real time (aggregation, trigger alerts, etc.) before being stored [6] and [3].

In addition, grid (grid computing) is an environment where software and hardware resources are distributed over a network; it provides incremental power of processing and / or storage. Even that, grid [6] and cloud [4] concepts are very close, grid does not offer an "anonymous" environment. Indeed, it allows users to select the servers on which they want to work. This environment provides a suitable solution for storing large amounts of data such as big data [7] [15] [3] [11]. Compared to a classical system, a grid has the following advantages: 1) sharing resources by group of users [6] [11], 2) lower costs thanks to choosing adapted servers to an issue, 3) scalability [6] [11], 4) bottlenecks and fault resistance [15].

Today, exploit and process data sources in large-scale and dynamic environments like grid [11] is a real challenge due to the following three criteria: 1) large number of data sources continuously in evolution (big data) 2) dynamicity of the environment (the joining / leaving of grid nodes) and 3) data source heterogeneity. The resource discovery process (which is also called discovery data sources process) is an important step for the query processing in big data grid environment. The transparency of grid facing its users is a barrier against its integration and large-scale use. Therefore, data and meta-data placement system must be established [11].

In this work we focus on business intelligence area, dealing with resource discovery for the decision support aid. We supposed that all sources are considered as a big data and are stored in a data grid. To integrate these sources, we proposed a data resource discovery mechanism (addressing protocol) suitable for NoSQL databases.

The paper is organized as follows. The following section studies the state of the art (related work). In section 3, we show the architecture of big data resource discovery mechanism for NoSQL databases in grid environment. Adding to that, we formalize related addressing protocol. Finally, in last section we conclude.

2. RELATED WORK

Resource discovery get the attention of many researchers. The first works were based on key word search, quickly proved inefficient due to the centralization which is not scalable. The latter has given birth to decentralized solutions based on more scalable systems, such as peer-to-peer systems. These systems are classified according to peer architecture: unstructured, structured and hybrid. However, all resource discovery methods employed do not take into account the semantic aspect. In this context, and to have more relevant discovery result, various works have included semantics. Indeed, three approaches have been defined [9]: 1) using name correspondences in schemas, 2) employing a global ontology and 3) applying several domain ontologies. For the first approach, establish necessarily connections and update them continuously, remains a complex and costly task [2]. For the second approach, designing a global ontology (known as global schema) reveals much ambiguity [1]. Therefore, the third approach, using several domain ontologies, is the most viable one compared to the others.

Whereas, methods applied in the latter approach still impose mapping topology: two by two, mapping table and Super Peer. This requires a strong hypothesis to the administrator which has to respect and follow this topology. Hence, the scope of [10] work, proposing a method that can be adapted to any topology type. [10] presents a resource discovery method taking into account not only the semantic heterogeneity of data sources but also peer dynamicity to query execution. However, to illustrate the discovery mechanism established, the user query used is an SQL query related to relational models. These models are distinguished by their highly structured static schema. However, today, in big data area and with grid environment, imposing such strong hypothesis is a blocking constraint [11]. For these reasons, we deal, in this work, with resource discovery in grid environment allowing the use of NoSQL (Not Only SQL) queries.

3. PAPER BEFORE STYLING BIG DATA RESOURCE DISCOVERY PROCESS WITH NOSQL DATABASES IN GRID ENVIRONMENT

A. NoSQL Systems

Big data requires both of the following properties: 1) consistency and large data volume, 2) velocity and data dynamicity and 3) variety and data format heterogeneity [14] [13]. Whereas, the variety can not be always verified face to a highly structured and static schema (such as in SQL databases). Indeed, research works are oriented towards NoSQL databases (Not Only SQL). NoSQL-based systems are more expressive with their flexible data structure. Thus, they can better support the data variety. In this context, we introduce this work. It extended [10] work. We propose a method of resource discovery, taking into account semantics in grid environment, based on NoSQL databases to query execution.

B. Architecture

Given the high number of domains in a grid, to manage all data from a centralized DBMS is hard to conceive even impossible. Therefore, we adopt the following solution of decomposition in VO (Virtual Organization). Each domain of the grid corresponds to a VO. Each VO is associated to a DBMS. In this way, each DBMS and its appropriate VO's elements are managed independently. As we reported earlier, the databases available now are not necessarily relational databases. Consequently, our DBMS must support NoSQL databases (Not Only SQL), eg., Hbase [8], MongoDB [12], Cassandra [5]... Thanks to the completeness of its platform, we adopt the Hbase system [8]. Hence, data grid can be considered as a network of several Hbase systems where each one is associated to an OVi and each OVi is associated to a domain i. This distribution allows taking into account the principle of locality and also autonomy of each VO.

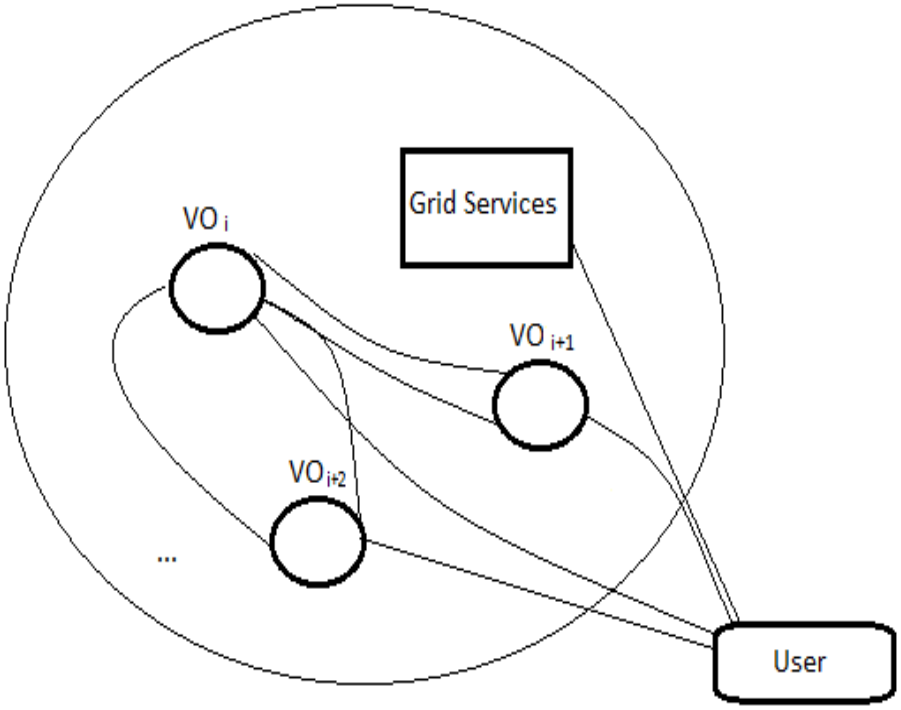

Figure 1: Big Data Resource Discovery Multi-VO in Grid Environment

C. Addressing Protocol

The addressing protocol used between peers (also called nodes) must be defined. We have two types of addressing protocols: (1) the addressing protocol intra-VO (research into the same VO) and (2) the inter-VO addressing protocol (discovery between VO). For intra-VO discovery, within the same VO, the Hbase system platform is able to establish addressing system thanks to MapReduce system functionalities. Thus, resource discovery in a single VO is then a classic discovery with MapReduce system functionalities. Now, it remains to define inter-VO discovery. In other terms, we should determine how virtual organizations communicate between each other. To guarantee completeness of the resource discovery and mainly its results, graph topology of nodes and VO must be connected [10]. In other words, query data resource discovery must be propagated to all other VO through existing connections (domain ontology mappings) [10]. Indeed, we employ an addressing system allowing permanent access from one VO to another.

After describing how the VO and their associated Hbase systems are distributed, we apply the RDMTS method (Resource Discovery Method Taking into account Semantics) for data resources discovery mechanism and also its maintenance system [10].

D. Formalisation

Let S(OVi) be a set of virtual organizations. For $i \neq j$, a VOi is connected to a VOj through domain ontology's mapping relationships established between several domains in grid [10]. Suppose that |S(VOI)| the number of neighbors VOi. Also, assume that all our VO and related mapping relationships form a graph noted: G(S, A), with S the set of vertices presenting all VO and A the edges (mapping relationships). There is an arc Aij if and only if there is a mapping between ontologies associated to OVi and OVj, respectively. As we reported previously, to ensure the full research results of the resource discovery mechanism, the graph must be connected. So, we suppose that the graph G is connected. There is a path $Pij \in A$ from VOi to VOj. Every peer Pk in a VOi must be able to initiate, at any time, a resource discovery process on VOj ($i \neq j$) because of the instability of grid environment.

```
//APk: Access point from one VO towards another.
//Path: Resource discovery process path.
//Lookup(C, VOAPk, Pi, Path): Discover the concept C in the VO via the //APk node.
//TTL: Time-To-Live (limit of the propagation range of a message).

Metadata <-- Lookup(C, VOAPk, Pi, Path);
//Intra-domain-ontology search.
TTL <-- TTL – 1;
If(TTL != 0) then
For each APk ∈ APS
Metadata <-- Metadata U Lookup(Translate(C, VOj, VOAPk),
VOAPk, Pi, Path U VOj);
//Inter-domain-ontology search.
If(not Empty(Metadata))
then Return(Metadata, Pi, Path U VOj);
```

Result of the research is sent to the first query sender peer (node). This result includes metadata describing discovered resource. After that, we keep query path established along the whole resource discovery process for its translations between VO. In fact, we use this path to translate user's query concepts between domain ontologies.

4. CONCLUSION

In this work we introduce a resource discovery method for big data sources in a large scale environment, as data grid. This method takes into account semantic aspect. It allows the discovery of all data sources despite the big number of data available through these sources, semantic heterogeneity, node dynamicity and scalability of the environment. An Hbase system is associated with each domain whose nodes constitute a virtual organization (VO). Our method represents an addressing protocol based on a combined technique contributing to two types of addressing systems (intra-VO and inter-VO). This latter allows permanent access between VO without any topology restrictions. Such method allows an efficient resource discovery process and maintenance system face to the specific and ambiguous properties of such an environment.

REFERENCES

[1]. Raddad Al King, Abdelkader Hameurlain, Franck Morvan, "Metadata Lookup for Distributed Query Optimization in P2P Environment", in: International Conference on Parallel and Distributed Computing Systems (PDCS 2007), Las Vegas, Nevada, 24/09/07-26/09/07, International Society for Computers and their Applications (ISCA), p. 36-43, 2007.

[2]. R. Alking, A. Hameurlain and F. Morvan, "Ontology-Based Data Source Localization in a Structured Peer-to-Peer Environment". IDEAS, Coimbra, Portugal, 2008.

[3]. Lawal Muhammad Aminu, "Implementing Big Data Management on Grid Computing Environment" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 9, Page No. 8455-8459 September 2014.

[4]. Muhammad Adnan; Department of CS, UET, Lahore, Pakistan; Muhammad Afzal; Muhammad Aslam; Roohi Jan; A. M. Martinez-Enriquez, "Minimizing big data problems using cloud computing based on Hadoop architecture", 2014 11th Annual High Capacity Optical Networks and Emerging/Enabling Technologies (Photonics for Energy), p. 99 – 103, ISSN: 1949-4092, IEEE, Print ISBN: 978-1-4799-6939-5 INSPEC Accession Number: 14903865location Charlotte, NC, 15-17 Dec. 2014.

[5]. http://cassandra.apache.org/.

[6]. C.Chandhini, Megana L.P, "Grid Computing-A Next Level Challenge with Big Data", International Journal of Scientific & Engineering Research Volume 4, Issue3, 1 ISSN 2229-5518 IJSER © 2013 http://www.ijser.org, March-2013.

[7]. Dan Garlasu, Core Technology Oracle Romania Bucharest, Romania, Virginia Sandulescu; Ionela Halcu; Giorgian Neculoiu; Oana Grigoriu; Mariana Marinescu; Viorel Marinescu, "A big data implementation based on Grid computing", in Roedunet International Conference (RoEduNet 2013 11th), ISSN: 2068-1038, p. 1-4, Print ISBN: 978-1-4673-6114-9 INSPEC Accession Number: 13500804 Conference Location: Sinaia DOI: 10.1109/RoEduNet.2013.6511732 Publisher: IEEE 17-19 Janvier 2013.

[8]. http://hbase.apache.org/index.html.

[9]. Imen Ketata, Riad Mokadem, Franck Morvan, "Biomedical Resource Discovery considering Semantic Heterogeneity in Data Grid Environments", in: International Conference on Integrated Computing Technology (InTech 2011), Sao Carlos-Brazil, Mai-Juin 2011.
[10]. Imen Ketata, Riad Mokadem, Franck Morvan, "Resource Discovery Considering Semantic Properties in Data Grid Environments" (regular paper), in: International Conference on Data Management in Grid and P2P Systems (GLOBE 2011), Toulouse, 01/09/2011-02/09/2011, Springer, LNCS 6864, p. 61-72, Septembre 2011.
[11]. Ajay Kumar and Seema Bawa, International, "Distributed and Big Data Storage Management in Grid Computing", Journal of Grid Computing & Applications (IJGCA) Vol.3, No.2, June 2012.
[12]. https://www.mongodb.com/.
[13]. Philip Chen C. L. et Zhang C. Y. (2014), "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences", available online 21 January 2014, ISSN 0020-0255, http://dx.doi.org/10.1016/ j.ins. 2014.01.015.
[14]. Big data analytics, TDWI Best Practices Report, Fourth Quarter.
[15]. Yuvraj S. Sase , Pratik A.Yadav, "Big Data Implementation Using Hadoop and Grid Computing", International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 4, April 2014 Two days National Conference – VISHWATECH 2014.