Data & Knowledge Engineering 111 (2017) 114–136

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

A Fine-Grained Distribution Approach for ETL Processes in Big Data Environments

Mahfoud Bala[a,*], Omar Boussaid[b], Zaia Alimazighi[c]

[a] Department of informatics, Saad Dahleb University, Blida 1, Blida, Algeria
[b] University of Lyon 2, Lyon, France
[c] Department of informatics, USTHB, Algiers, Algeria

ARTICLE INFO

Keywords:
Data Warehousing
ETL
Parallel and Distributed Processing
Big Data
MapReduce

ABSTRACT

Among the so-called "4Vs" (volume, velocity, variety, and veracity) that characterize the complexity of Big Data, this paper focuses on the issue of "*Volume*" in order to ensure good performance for Extracting-Transforming-Loading (ETL) processes. In this study, we propose a new fine-grained parallelization/distribution approach for populating the Data Warehouse (DW). Unlike prior approaches that distribute the ETL only at coarse-grained level of processing, our approach provides different ways of parallelization/distribution both at process, functionality and elementary functions levels. In our approach, an ETL process is described in terms of its core functionalities which can run on a cluster of computers according to the MapReduce (MR) paradigm. The novel approach allows thereby the distribution of the ETL process at three levels: the "process" level for coarse-grained distribution and the "functionality" and "elementary functions" levels for fine-grained distribution. Our performance analysis reveals that employing 25 to 38 parallel tasks enables the novel approach to speed up the ETL process by up to 33% with the improvement rate being linear.

1. Introduction

The widespread use of novel technologies has lead to the generation of huge volumes of data. For instance, MapReduce (MR) jobs run continuously on Google's clusters, processing more than 20 petabytes of data per day [1] and over 10 petabytes of data are generated monthly by Facebook [2]. Indeed, according to [3], the New York Stock Exchange generates about one terabyte of new trade data per day and the Internet Archive stores around 2 petabytes of data, and is growing at a rate of 20 terabytes per month. Such data becomes so large and so it will be difficult to continue using traditional processing tools. Big data is characterized by the four "Vs" [4] where "*Volume*" relates to the large amount of data that goes beyond the usual units, "*Velocity*" to the speed with which this data is generated and is processed, "*Variety*" to the diversity of formats and structures, and "*Veracity*" to data accuracy and reliability.

This work aims to provide solutions to the problems posed by Big Data in the context of decision-support systems (DSS). We are particularly interested in the Extracting-Transforming-Loading (ETL) functionalities facing large data which is often processed for cleansing, filtering, standardization and conforming purposes. The ETL process is a collection of software components dedicated to the refreshment of data warehouses (DWs). Nowadays, the Internet and Web 2.0 are generating data at growing rates, and as a result the conventional DSS, and ETL in particular, should be revisited in order to deal with the complexity of Big Data. To achieve this, the

[*] Corresponding author at: Department of informatics, Saad Dahleb University, Blida 1, Algeria.
E-mail addresses: mbala@univ-blida.dz (M. Bala), omar.boussaid@univ-lyon2.fr (O. Boussaid), zalimazighi@usthb.dz (Z. Alimazighi).

http://dx.doi.org/10.1016/j.datak.2017.08.003
Received 21 June 2016; Received in revised form 19 July 2017; Accepted 18 August 2017
Available online 26 August 2017

ETL, being one of the key components of the DSS, should be adapted to the Big Data environment. Data distribution and parallel/distributed processing offer promising solutions in this context. The traditional ETL system that usually operates on a single machine (ETL server) cannot deal with very large volume of data (at the terabytes and petabytes scale). On the other hand, the ETL should be revisited following the emergence of new paradigms such as *Cloud Computing* [5], *MapReduce* (MR) [6], and *NoSQL* (Not Only SQL) data models [7]. The parallelization/distribution of the data processing on clusters is a promising solution to address the complexity of Big Data. The solution offered by the DSS community, in this context, is to distribute the ETL process on multiple cluster nodes [8–10]. Each instance of the ETL process handles a source data partition in a parallel manner to improve the performance of the ETL [8–10]. However, the existing solutions have been defined only at a "process" level which tend to be coarse-grained in nature. Moreover, they have not considered the "ETL functionalities"/"elementary functions" which are often fine-grained and allow the deep understanding of the ETL complexity and consequently improve significantly the ETL process.

This paper extends our research work introduced by [11] that suggests a new fine-grained parallel/distributed ETL approach for Big Data, consisting of a set of MR-based ETL functionalities. We first present a fine-grained description of an ETL process in terms of its core functionalities and elementary functions; these are then parallelized/distributed according to the MR paradigm. Our approach allows thereby the parallelization/distribution of the ETL at three levels: notably "functionality" and "elementary function" levels as well as the "process" level. Our experimental results will reveal that our approach improves the ETL performance for handling Big Data. A number of approaches have been proposed by the DSS community in the ETL field [12–15,8,16,9,10]. We propose a classification of the existing research studies based on the parallelization criteria. We have developed a prototype and conducted some experiments in order to validate our approach. The rest of this paper is structured as follows. Section 2 presents the fundamental concepts related to this research study. Section 3 presents a state-of-the-art in the ETL field followed by a classification of ETL approaches proposed in the literature according to the parallelization criteria. Section 4 describes our novel approach which is illustrated with ETL functionalities examples in Sections 5 and 6. Section 7 describes our prototypical implementation and the conducted experiments. Section 8 concludes this work and presents some suggestions for future research.

2. Key concepts

The main purpose of this section is to briefly introduce the key concepts and terms related to our research work. The reader is referred to [17,18,6] for more details.

Decision Support System: From various operational data sources, a decision-support system (DSS) produces and stores in a central repository, called data warehouse (DW), valuable information that synthesizes the organization's activities. The DSS uses the DW to provide strategic information for decision-makers.

ETL process: The data integration phase of a DSS is based on an Extracting-Transforming-Loading (ETL) process dedicated to extracting (E) relevant source data that will be transformed (T) for cleansing, standardization and conforming purposes and then loaded (L) in the DW that provides relevant information for analysis and decision-making applications.

Fig. 1 depicts the ETL environment that consists of the data store layer (*Data Sources*, *Data Staging Area (DSA)*, *Data Warehouse (DW)*) and the processing layer (*Extract*, *Transform*, *Load*). The Figure shows also the interaction between the ETL tasks and the three data storage areas.

Parallelization vs Distribution of ETL processes: We focus in this study on two aspects in order to speed-up the ETL process: (i) "*parallelization*" and (ii) "*distribution*". The "*distribution*" consists of assigning data partitions or processing units (processes, instances of process, functions or instances of functions, etc.) to multiple nodes of the cluster. As for "*parallelization*", the latter allows running simultaneously multiple independent processing units in order to speed-up the whole process. While the two concepts "*Parallelization*" and "*Distribution*" are different, they are closely related. Indeed, the "*distribution*" of data partitions and processing units on clusters allows efficient "*parallel*" processing of data. Moreover, the "*distribution*" enables the processing system to insure load balancing and fault-tolerance.

Big Data: Big Data is characterized mainly by: (i) the huge data sets with excessive size (*Volume*), (ii) the various sources and

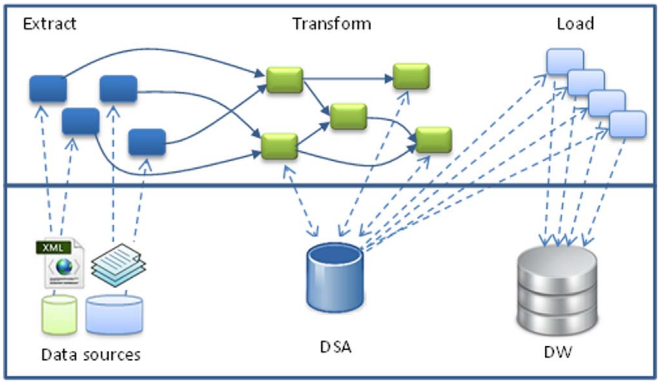

Fig. 1. The ETL process and its environment.