Neurocomputing 174 (2016) 439–453

Contents lists available at ScienceDirect

Neurocomputing

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

An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication

Awais Ahmad, Anand Paul*, M. Mazhar Rathore

The School of Computer Science and Engineering, Kyungpook National University, Daegu 702-701, Republic of Korea

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
24 March 2015
Accepted 14 April 2015
Available online 1 September 2015

Keywords:
M2M
Big Data
Divide-and-conquer
Data fusion domain

ABSTRACT

Machine-to-Machine (M2M) communication relies on the physical objects (e.g., satellites, sensors, and so forth) interconnected with each other, creating mesh of machines producing massive volume of data about large geographical area (e.g., living and non-living environment). Thus, the M2M is an ideal example of Big Data. On the contrary, the M2M platforms that handle Big Data might perform poorly or not according to the goals of their operator (in term of cost, database utilization, data quality, processing and computational efficiency, analysis and feature extraction applications). Therefore, to address the aforementioned needs, we propose a new effective, memory and processing efficient system architecture for Big Data in M2M, which, unlike other previous proposals, does not require whole set of data to be processed (including raw data sets), and to be kept in the main memory. Our designed system architecture exploits divide-and-conquer approach and data block-wise vertical representation of the database follows a particular petitionary strategy, which formalizes the problem of feature extraction applications. The architecture goes from physical objects to the processing servers, where Big Data set is first transformed into a several data blocks that can be quickly processed, then it classifies and reorganizes these data blocks from the same source. In addition, the data blocks are aggregated in a sequential manner based on a machine ID, and equally partitions the data using fusion algorithm. Finally, the results are stored in a server that helps the users in making decision. The feasibility and efficiency of the proposed system architecture are implemented on Hadoop single node setup on UBUNTU 14.04 LTS core™*i*5 machine with 3.2 GHz processor and 4 GB memory. The results show that the proposed system architecture efficiently extract various features (such as River) from the massive volume of satellite data.

1. Introduction

Big Data is one of the central and influential research challenges for the 2020 horizon. The archetype relies on the acquisition and aggregation of the massive volume of data to support innovation in the upcoming years. The data is considered as big when it meets the requirements of "*four V's*", such as *Volume, Variety, Velocity, and Value*. The groundwork of Big Data exploitation is to empower the existence data sets to extract new information, helps in enrichment of business values chains. According to the IDC group, the quantity of world data will be 44 times bigger in the next few years (such as 0.8–35 zettabytes). Therefore, in this context, the machine-to-machine archetype relies on the world of interconnected object [1], which can be used for acquisition, aggregation and analyzing the data depending on context. In these days', vehicles, smart phones, buildings, satellites, sensors, etc. collects various information about physical environment (i.e., large geographic area), where majority of them are generating zettabytes of the sensed data. For instance, the Gartner group predicts that up to 26 billion of the machines will be connected to the internet by 2020. Furthermore, Intechno Consulting estimates that up to 180 billion Euros will be generated worldwide. The aforementioned estimations are the examples of the Big Data aggregation and analysis as it can be dealt with the large Volume in large Variety, aggregated data with a high Velocity to define application with added-Value [2].

The coupling between M2M and the Big Data communities is strong [3–5]. As such, there is no widespread approach that supports data acquisition, data aggregation, and data analysis from numerous objects (such as, satellites, sensors, and so forth), and their exploitations. Based on the aforementioned needs, recent research efforts are focused on the data acquisition from the data generation tiers [6], the aggregation tiers [7], or the exploitation [8], and lastly, the data analysis tiers. Such approaches are somehow challenging task in

* Corresponding author.
E-mail addresses: awais.ahmad@live.com (A. Ahmad), paul.editor@gmail.com (A. Paul), rathoremazhar@gmail.com (M.M. Rathore).

http://dx.doi.org/10.1016/j.neucom.2015.04.109
0925-2312/© 2015 Elsevier B.V. All rights reserved.

M2M communication than locating, identifying, understanding and citing data [9]. Having a large scale data, all of this has to happen in a mechanized manner since it requires diverse data structure as well as semantics to be articulated in forms of computer readable format. However, by analyzing simple data having one data set, an intelligently designed database is required. Apparently, there might be alternative ways to store all of the same information. In such conditions, the existing designs might have an advantage over others for individual process and possible drawbacks for some other purposes.

In the current scenarios of Big Data analytics, various platforms have been provided by relational database vendors, which are used for data aggregation and data analyzes. Such platforms are either software only or they just provide analytical services, which runs in a third party hosted environment. Furthermore, these platforms are meant for new technologies, which are used to analyze massive volume of data, such as, web traffic (e.g., social media) and global positioning system (GPS) data. Now a days, various analytical platforms are available on the market that could be used for specific applications (i.e., each of these platforms is designed for a particular goal).

The incredible growth in the data also posing new challenges, such as, how to aggregate massive volume of data? How to store such data in a limited amount of memory allocated for the particular task? Moreover, how to process and analyze these data when there is no such intelligent algorithm is available? Moreover, large-scale data cannot be tackled by standard reduction techniques since their runtime becomes impractical. Several other approaches have been developed that helps in enabling data reduction techniques, which deal with this problem. In the case of Prototype Reduction (PR), the data level partitioning is based on a distributed partitioning model that sustains the class distribution. Such type of reduction splits the original data into various subsets that could be individually addressed. Afterward, it combines each partially reduced set into a global solution. Furthermore, torrents of event data are required to be distributed over various databases, and large process mining problems need to be distributed over a network of computer. Several other approaches could be found in the literature [2,4,25,27]. However, the generic divide-and-conquer approach based on fusion technique could be the optimal solution for the said challenges.

Hence, in a nutshell, the following two main problems appear when we increase the data set size during analysis.

- The existing Big Data architectures are not capable of processing and analyzing a large amount of data, i.e., the data that is generated by the various remote sensory satellites.
- Continuous feature extraction, such as rivers or highways detection from remote sensory Big Data is a challenging issue. Such scheme requires efficient algorithms in handling large scale earth observatory datasets on a limited timescale.

Therefore, in this work, we propose a system architecture designed for analyzing Big Data in M2M using a divide-and-conquer mechanism that welcomes real-time and offline data. The proposed system architecture handles the drawbacks mentioned above. To do so, various machine are used for Earth Observatory System (i.e., satellites and sensors), which are used to collect data, and directly transmits the data to the ground base station. The ground base station pre-process the raw data in which they extract useful information from the raw data in order to create data blocks (called granules). Afterward, these data blocks are transmitted to Data Aggregation Unit (DAgU), where it aggregates all the data blocks, and arrange them in a sequential manner based on their unique machine ID. Furthermore, fusion algorithm is employed, which is used for partitioning the data blocks. The exploitation of the fusion algorithm is used to disseminate equal size data in each divide-and-conquer servers. Such technique not only helps in enhancing the efficiency of the divide-and-conquer servers but also helps the system in fast data processing. Finally, the partitioned data blocks are sent to the Decision Making Unit (DMU), which are used for analysis as well storing of the results. The Decision Making server can utilize those results depending on the requirement of the user.

The proposed architecture welcomes both real-time as well as offline data (e.g., GPRS, xDSL, or WAN). The contribution of the proposed system architecture is summarized as follows.

1. The data aggregation technique concatenates all the data being generated by various machines in larger block.
2. The larger data blocks are arranged in a sequential manner based on the machine ID. Afterward, these data block are partitioned into smaller data blocks (granules).
3. These granules are then sent to D&CPU, in which each granule is sent to one server for final processing.
4. The result storage device helps the user to get their desired results at any times, which can be used for future comparison, if needed.

The proposed divide-and-conquer data analytical architecture for Big Data in M2M has several advantages, such as, at data acquisition stage, the data is concatenated to form a Big Data block that helps the system to combine the same data type, the fusion domain helps in enhancing the efficiency of D&CPU by dividing the data into smaller data blocks. Each block is then sent to a single server for further processing, which helps in increasing the processing efficiency, and finally, users can use the desired results for comparison purpose.

The remainder of this paper is organized as follows. In Section 2, we give a detailed survey of the Big Data. In Section 3, we briefly explained the requirements for the Big Data using divide-and-conquer approach for M2M. In Section 4, we describe the proposed system architecture for divide-and-conquer approach in M2M. In Section 5, we present a detailed analytical and simulation results using Hadoop. Finally, Section 6 offers a conclusion and future work of the paper.

2. Related work

Big Data and its analysis are at the verge of modern science and business, where author highlights the identity of number of sources on Big Data such as online transactions, emails, audios, videos, search queries, health records, social networking interactions, images, click-streams, logs, posts, search queries, health records, social networking interactions, mobile phones and applications, scientific equipment, and sensors [3]. The proposed model is using conventional database tools. The challenge is to capture, form, store, manage, share, analyze and visualize the Big Data. In addition, the characteristics of Big Data, such as a variety, volume, and velocity, the three main characteristics of Big Data are elaborated briefly in Section 1.

The concept of Big Data is stimulating a broad range of curiosity in the industrial sector [11]. The report provides concrete examples of Big Data generated by sensors. For instance, manufacturing companies use various machines (e.g., sensors) which are embedded in monitoring usage patterns, predicting maintenance problems and enhancing the product quality in their machinery equipment. On analyzing data streams generated by the embedded machines, allows manufacturers to improve their products in their machinery. A massive volume of data is generated by numerous machines deployed in the supply lines of utility providers, which are constantly monitoring the production quality,