IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Processing of Real Time Big Data for Machine Learning

Namrata Gawande[1], Ramdas Gawande[2]

Computer Engineering, Pimri Chinchwad College of Engineering, Pune, India[1]

Information Technology, Pimri Chinchwad College of Engineering, Pune, India[2]

Abstract: Assets of real time digital world daily generate massive volume of real-time data (mainly referred to the term "Big Data"), where insight information has a potential significance if collected and aggregated effectively. In today's era, there is a great deal added to real-time remote sensing Big Data than it seems at first, and extracting the useful information in an efficient manner leads a system toward a major The computational challenges, such as to analyse, aggregate, and store, where data are remotely collected. Keeping in view the above mentioned factors, there is a need for designing a system architecture that welcomes both real time, as well as offline data processing. Therefore, in this paper, we propose real-time Big Data analytical architecture for processing such data environment.

Keywords: Big Data, Hadoop, HDFS, Cloud Computing, data analysis.

I. INTRODUCTION

Recently, a great deal of interest in the field of Big Data and its analysis has risen, mainly driven from extensive number of research challenges strappingly related to confide applications, such as modelling, processing, querying, mining, and distributing large-scale repositories. The term "Big Data" classifies specific kinds of data sets comprising formless data, which well in data layer of technical computing. Applications and the Web the data stored in the underlying layer of all these technical computing application scenarios have some precise individuality in common, such as 1. Large scale: data, which refers to the size and the data warehouse. 2. Scalability issues: which refer to the application's likely to be running on large scale (e.g., Big Data). 3. Sustain extraction transformation loading (ETL) method: from low, raw data to well thought-out data up to certain extent. 4. Development of uncomplicated interpretable: analytical over Big Data warehouses with a view to deliver an intelligent and momentous knowledge for them. Big Data are usually generated by online transaction, video/audio, email, and number of clicks, logs, posts, social network data, scientific data, remote access sensory, mobile phones, and their applications.

These data are accumulated in databases that grow extraordinarily and become complicated to confine, form, store, manage, share, process, analyze, and visualise via typical database software tools. Advancement in Big Data sensing and computer technology revolutionizes the way remote data collected, processed, analyzed, and managed. These platforms come in various shapes from software only to analytical services that run in third-party hosted environment. In remote access networks, where the data source such as sensors can produce an overwhelming amount of raw data. We refer it to the first step, i.e., data acquisition, in which much of the data are of no interest that can be filtered or compressed by orders of magnitude. With a view to using such filters, they do not discard useful information. For instance, in consideration of new reports, is it adequate to keep that information that is mentioned with the company name? Alternatively, is it necessary that we may need the entire report, or simply a small piece around the mentioned name? The second challenge is by default generation of accurate metadata that describe the composition of data and the way it was collected and analyzed.

In this project, we referred the high speed continuous stream of data or high volume offline data to Big Data which is leading us to a new world of challenges. Such consequences of transformation of remotely sensed data to the scientific understanding are a critical task. Hence the rate at which volume of the remote access data is increasing, a number of individual users as well as organizations are now demanding an efficient mechanism to collect, process, and analyze, and store these data and its resources.

II. LITURTURE SURVEY

The increase in the data rates generated on the digital universe is escalating exponentially. With a view in employing current tools and technologies to analyse and store, a massive volume of data are not up to the mark [2], since they are unable to extract required sample data sets.

Therefore, we must design an architectural platform for analyzing both remote access real time and offline data. When a business enterprise can pull-out all the useful information obtainable in the Big Data rather than a sample of its data set, in that case, it has an influential benefit over the market competitors. Big Data analytics helps us to gain insight and make better decisions. Therefore, with the intentions of using Big Data,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

modifications in paradigms are at utmost. To support our motivations, we have described some areas where Big Data can play an important role.

In healthcare scenarios, medical practitioners gather massive volume of data about patients, medical history, medications, and other details. The above-mentioned data are accumulated in drug-manufacturing companies. The nature of these data is very complex, and sometimes the practitioners are unable to show a relationship with other information, which results in missing of important information. With a view in employing advance analytic techniques for organising and extracting useful information from Big Data results in personalized medication, the advance Big Data analytic techniques give insight into hereditary causes of the disease.

III. PROBLEM DEFINITION

Big Data analysis is somehow a challenging task than locating, identifying, understanding, and citing data [3]. 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, a mechanism is required of how to design a database. There might be alternative ways to store all of the same information. In such conditions, the mentioned design might have an advantage over others for certain process and possible drawbacks for some other purposes. In order to address these needs, various analytical platforms have been provided by relational databases vendors [4].

These platforms come in various shapes from software only to analytical services that run in third party hosted environment. In remote access networks, where the data source such as sensors can produce an overwhelming amount of raw data. We refer it to the first step, i.e., data acquisition, in which much of the data are of no interest that can be filtered or compressed by orders of magnitude. With a view to using such filters, they do not discard useful information. For instance, in consideration of new reports, is it adequate to keep that information that is mentioned with the company name? Alternatively, is it necessary that we may need the entire report, or simply a small piece around the mentioned name?

The second challenge is by default generation of accurate meta data that describe the composition of data and the way it was collected and analysed. Such kind of metadata is hard to analyse since we may need to know the source for each data in remote access.

Normally, the data collected from remote areas are not in a format ready for analysis. Therefore, the second step refers us to data extraction, which drags out the useful information from the underlying sources and delivers it in a structured formation suitable for analysis. For instance, the data set is reduced to single-class label to facilitate analysis, even though the first thing that we used to think about Big Data as always describing the fact. However, this is far away from reality; sometimes we have to deal with erroneous data too, or some of the data might be imprecise.

To address the aforementioned needs, this paper presents a real time Big Data analytical architecture, which is used to analyse real time, as well as offline data. At first, the data are remotely pre-processed, which is then readable by the machines. Afterward, this useful information is transmitted to the base system for further data processing. The Base System performs two types of processing, such as processing of real-time and offline data. In case of the off-line data, the data are transmitted to offline data-storage device. The incorporation of offline data-storage device helps in later usage of the data, whereas the real-time data is directly transmitted to the filtration and load balancer server, where filtration algorithm is employed, which extracts the useful information from the Big Data.

On the other hand, the load balancer balances the processing power by equal distribution of the real-time data to the servers. The filtration and load-balancing server not only filters and balances the load, but it is also used to enhance the system efficiency. Furthermore, the filtered data are then processed by the parallel servers and are sent to data aggregation unit (if required, they can store the processed data in the result storage device) for comparison purposes by the decision and analysing server.

The proposed architecture welcomes remote access data as well as direct access network data (e.g., GPRS, 3G, xDSL, or WAN). The proposed architecture and the algorithms are implemented in Hadoop using Map Reduce programming by applying real time data sensing.

IV. PROPOSED SOLUTION

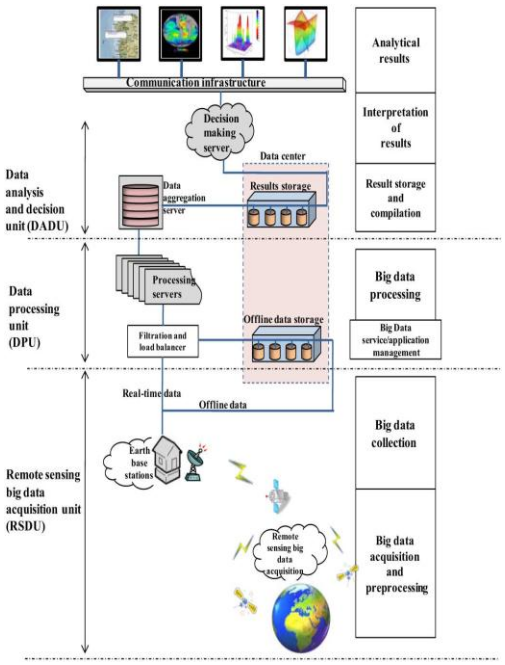

Fig.1: Remote sensing Big Data architecture [1].

DOI 10.17148/IJARCCE.2016.5583

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

We have divided real time Big Data processing architecture into three parts, i.e., 1) data acquisition unit 2) data processing unit and 3) data analysis and decision unit. The functionalities and working of the said parts are described as below.

A. Data Acquisition Unit

The need for parallel processing of the massive volume of data was required, which could efficiently analyze the Big Data. For that reason, the proposed unit is introduced in the real time Big Data processing architecture that gathers the data from various available data gathering unit around the globe. We assume that the data capturing unit can correct the erroneous data. For effective data analysis, the Base System pre-processes data under many situations to integrate the data from different sources, which not only decreases storage cost, but also improves analysis accuracy. Some relational data pre-processing techniques are data integration, data cleaning, and redundancy elimination.

The data must be corrected in different methods to remove distortions caused due to the motion of the platform. We divided the data processing procedure into two steps, such as real-time Big Data processing and off line Big Data processing. In the case of off line data processing, the Base System transmits the data to the data center for storage. This data is then used for future analyses. However, in real-time data processing, the data are directly transmitted to the filtration and load balancer server, since storing of incoming real-time data degrades the performance of real-time processing. shown in Fig. 1.

B. Data Processing Unit

In data processing unit, the filtration and load balancer server have two basic responsibilities, such as filtration of data and load balancing of processing power. Filtration identifies the useful data for analysis since it only allows useful information, whereas the rest of the data are blocked and are discarded.

Hence, it results in enhancing the performance of the whole proposed system. Apparently, the load-balancing part of the server provides the facility of dividing the whole filtered data into parts and assign them to various processing servers. The filtration and load-balancing algorithm varies from analysis to analysis; e.g., if there is only a need for analysis of sea wave and temperature data, the measurement of these described data is filtered out, and is segmented into parts.

Each processing server has its algorithm implementation for processing incoming segment of data from load balancer. Each processing server makes statistical calculations, any measurements, and performs other mathematical or logical tasks to generate intermediate results against each segment of data. Since these servers perform tasks independently and in parallel, the performance proposed system is dramatically enhanced, and the results against each segment are generated in real time. The results generated by each server are then sent to the aggregation server for compilation, organization, and storing for further processing.

C. Data Analysis and Decision Unit

This unit contains three major portions, such as aggregation and compilation server, results storage server(s), and decision making server. When the results are ready for compilation, the processing servers in data processing unit send the partial results to the aggregation and compilation server, since the aggregated results are not in organized and compiled form. Therefore, there is a need to aggregate the related results and organized them into a proper form for further processing and to store them.

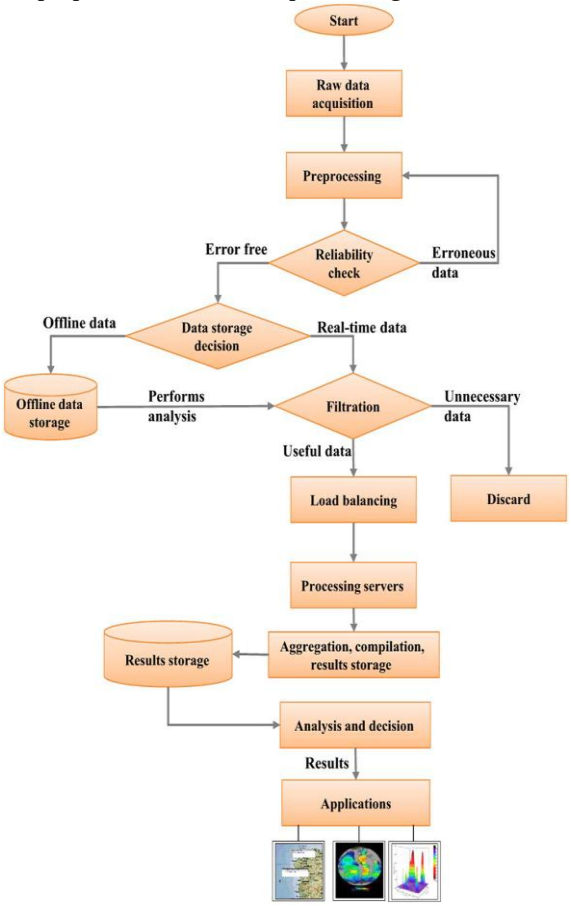

Fig. 2: Flowchart of the real time Big Data Processing architecture.

In the proposed architecture, aggregation and compilation server is supported by various algorithms that compile, organize, store, and transmit the results. Again, the algorithm varies from requirement to requirement and depends on the analysis needs. Aggregation server stores the compiled and organized results into the result's storage with the intention that any server can use it as it can process at any time. The aggregation server also sends the same copy of that result to the decision-making server to process that result for making decision. The decision-making server is supported by the decision algorithm, which inquire Aggregation server stores the compiled and organized results into the result's storage with the intention that any server can use it as it can process at any time. The aggregation server also sends the same copy of

DOI 10.17148/IJARCCE.2016.5583

that result to the decision-making server to process that result for making decision. The decision-making server is supported by the decision algorithm, which inquire different things from the result, and then make various decisions.

The decision algorithm must be strong and correct enough that efficiently produce results to discover hidden things and make decisions. The decision part of the architecture is significant since any small error in decision-making can degrade the efficiency of the whole analysis. The flowchart supporting the working of the proposed architecture is depicted in Fig. 2.

V. CONCLUSION

In this paper, we proposed architecture for real-time Big Data Processing. The proposed architecture efficiently processed and analyzed real-time and offline Big Data for decision-making. The architecture of real-time Big is generic (application independent) that is used for any type of real time Big Data processing. Furthermore, the capabilities of filtering, dividing, and parallel processing of only useful information are performed by discarding all other extra data. These processes make a better choice for real-time Big Data analysis. The technique proposed in this paper for each unit and subunits are used to analyze real time data sets, which helps in better understanding of data. The proposed architecture welcomes researchers and organizations for any type of real time Big Data analysis by developing algorithms for each level of the architecture depending on their analysis requirement.

REFERENCES

[1]. Real-Time Big Data Analytical Architecture for Remote Sensing Application Muhammad Mazhar Ullah Rathore, Anand Paul, Senior Member, IEEE, Awais Ahmad, Student Member, IEEE, Bo-Wei Chen, Member, IEEE, Bormin Huang, and Wen Ji, Member, IEEE
[2]. Wikibon Blog. (Oct. 14, 2014). [2310]. Big Data Statistics [Online]. Available: wikibon.org/blog/big-data-statistics/M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization," In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)
[3]. A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with Big Data," in Proc. 38th Int. Conf. Very Large Data Bases Endowment, Istanbul, Turkey, Aug. 27–31, 2012, vol. 5, no. 12, pp. 2032–2033.
[4]. P. Chandarana and M. Vijayalakshmi, "Big Data analytics frameworks," in Proc. Int. Conf. Circuits Syst.Commun. Inf. Technol. Appl. (CSCITA), 2014, pp. 430–434.