REMOTE ONLINE BIG DATA ANALYTICS ARCHITECTURE

Nagalakshmi D R[1], Suma R[2], Bhavya N Javagal[3]

[1]M. Tech (Student), T. John Institute of Technology
nagalakshmidr333@gmail.com
[2]Assistant Professor, T. John Institute of Technology
sumar@tjohngroup.com
[3]Assistant Professor, T. John Institute of Technology
bhavya@tjohngroup.com

Abstract

There is a massive amount of real-time data produced from remote sense analytics which is generally referred as "Big Data". There is an increased change in the Big Data and its analysis majorly in the research field and applications such as processing, mining, querying and distributing. There is a great need to collect the real-time data and extract the useful information which is later used for computational purposes. This paper describes Big Data Analytics architecture which is used in data analysis. The architecture which is designed consists of three modules, such as data collection module, data filtration module and finally analysis module. It has the capability of equally distributing the load, filtering and parallel data execution. Hadoop is used for processing the remotely collected data. Hadoop is used to implement the algorithms using Map Reduce.

Keywords: *Big Data, Map Reduce, Hadoop*

---***---

I. INTRODUCTION

Organizations use various analytical techniques to manage large datasets, it can be both structured and unstructured types of data. The data is obtained from various sources such as social media, sensor networks, public profiles, government data holdings, company databases.

Although data mining is one of the technique to maintain the large amount of data but varied data types can be handled efficiently using Big Data. This results in increasing processing speed. Big Data analytics is the field which deal with data which are massive in size and velocity. Big Data analytics comprises of distributed file system framework like Hadoop.

Big Data is used in various researches for processing, storing and extraction of useful data. Big Data is used in corporations to understand more about their workforce, to increase productivity and business processes, hence there is a demand for the technique which makes data collection and analysis easier and simpler.

II. MOTIVATION FOR REMOTE ONLINE BIG DATA ANALYTICS

A. Challenges of Big Data

Big Data can retrieve useful information, but it has its own challenges. Many Big Data applications are not intended and leads to unpredictable output.

Hence there is a need to design architecture which handles remote data and data which is stored for offline analysis [2]. The data from different applications have some common individualities such as: 1. Extracting the useful information from the data and makes it easy to analyze. 2. Scalability issue, is the major problem when it comes to managing the large datasets. 3. Managing the inter-machine communication and handling failure of machine. 4. The large datasets lead to the more expanded database and hence difficult to manage the remoter sensing application [4].

B. Backgroud Reviews

In research field large amount of data is obtained from the sensors, hence there is a need to collect this data. This task is carried out in data collection module. The data which is not required is discarded on the basis of magnitude. At the same time the required and useful information is retained without losing it.

In some applications we might require entire document or sometimes a part of the document. There is a necessity to maintain the metadata which gives detailed information about the data collection and its analysis [5]. Since there are different sources for each dataset, it's difficult to manage the metadata.

This paper is organized with a brief introduction in part I and followed by challenges of Big Data and Background work in part II and part III is enriched with our implementation strategies and methodologies. Part IV describes the system design, followed by conclusion on results obtained in section IV.

III. METHODOLOGIES

Remote online Big Data architecture is proposed to analyze the Big Data in an efficient way. The architecture has been classified into three tasks.

Primarily the data is acquired and cleaned under Hadoop environment [6]. The preprocessing step involves data pre-processing with data aligning and refining. The alignment of data is also achieved with filling an unfilled attributes. This is considered to be our primary processing step. In future steps, the data is analyzed and processed under filtering algorithm, the Map Reduce operation is also performed in this step. The data collected and proceed is now stored offline under local regional servers.

Decision making and Analysis is considered "Eq. (1)" to be most difficult and standout step performed in our proposed system and thus fetches generic data sets for acquiring data analysis and controlling. The system also helps the users to achieve image redundancy optimization from regular and trivial storage systems [7]. The system also contributes in this step to identify a distinct difference in processing a time gap from regular JAVA and Hadoop clustering.

$$E_i = \sum_{i=0}^{n} ED\,(I_i) \qquad (1)$$

The major contribution of this system application is to achieve reliable results on selecting Hadoop clustering environment v/s java environment in terms of data processing and efficiency matching. The application also makes use of high data processing record and thus converts the given input data sets into gray images and extracts edges for achieving a secure data redundancy.

IV. SYSTEM DESIGN

Generally the server acquires the data from the satellite and thus provides a backup at servers connected directly to satellite. These servers are unauthorized and highly preserved from public accessing [8]. Apart from this, the servers also store the images in reliably higher ratio of memory and space. On demand from national or regional servers, the data is optimized and forwarded. These set of data is considered as internet data. In our proposed work we have considered a demanded earth Arial pictorial data.

The system consists of a planned server accessing of data images from satellite and is fetched on request as shown in Fig. 3.1. The data acquired is stored in local servers and thus the request is processed in these servers for fetching the data. The remote servers contain data greater than regular size thus processing time and storage is considerably high and thus we propose the system to process such complicated images in a simpler manner under Hadoop Cluster.

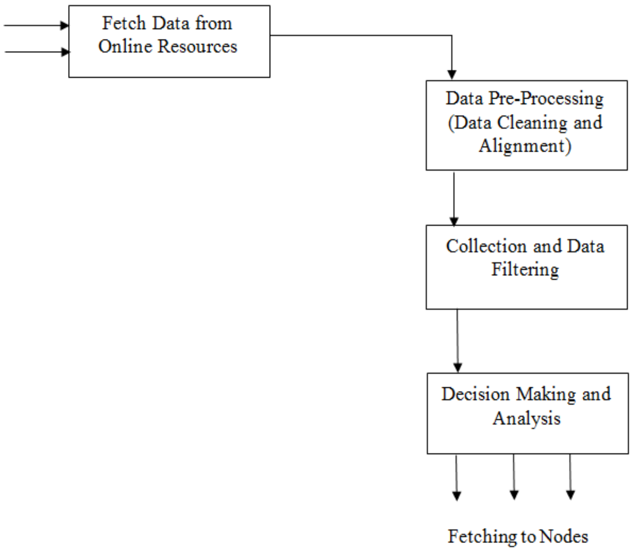
Fetch Data from Online Resources
Data Pre-Processing (Data Cleaning and Alignment)
Collection and Data Filtering
Decision Making and Analysis
Fetching to Nodes

Fig. 4.1 Overview on System Design

Primarily the data is acquired and cleaned under Hadoop environment. The pre-processing step involves data pre-processing with data aligning and refining. The alignment of data is also achieved with filling an unfilled attributes [9]. This is considered to be our primary processing step. In future steps, the data is analyzed and processed under filtering algorithm, the Map Reduce operation is also performed in this step. The data collected and proceed is now stored offline under local regional servers.

These main servers are connected with inland servers and also proceed with connection with other primary servers. On request, the data is flowed from main server to inland servers via internet and thus we achieve the input data sets [10]. Future, the samples are connected and stored into the regional servers and databases. The acquired images for Hadoop node is fetched from regional servers in our proposed system. Apart from storage, the data preprocessing and filtering is performed in this stage. Data cleaning is requires as acquired data is accompanied with other relevant and irrelevant attribute set.

V. RESULTS

Map Reduce based data reduction and monitoring under a scalable and most efficient time of processing. The data acquired here is considered as a primary asset for processing and achieving image redundancy. The system also aims to focus on time consummation during processing huge data sets in Hadoop environment and Core Java environment.

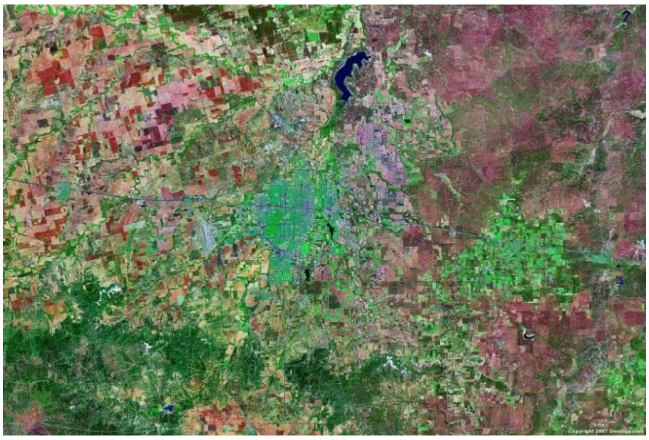

Fig. 5.1 Input Sample Image

In Fig. 5.1 , we have showcased a sample of input image under Hadoop processing environment.

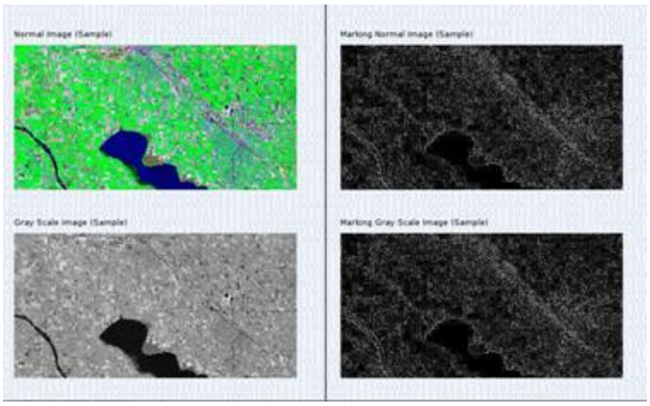
Normal Image (Sample)
Marking Normal Image (Sample)
Gray Scale Image (Sample)
Marking Gray Scale Image (Sample)

Fig. 5.2 Output Sample Image

The proposed system has successfully achieved predominant results on processing time and efficiency with respect to processing time in Hadoop Cluster and Core Java.

Apparently, we have also observed a high redundancy optimization of earth images with an overall input processed with datasets as shown in Fig. 3. The system is successfully performed a processing of 60MB data under an interval of 22,000 millisecond under Java and less than 2500 millisecond in Hadoop cluster. The system also showcases a new edge on understanding a Hadoop MapReduce operation for redundancy.

VI. CONCLUSION

This proposed system has successfully achieved predominant results on processing time and efficiency with respect to processing time in Hadoop Cluster and Core Java. Apparently, we have also observed a high redundancy optimization of earth images with an overall input processed with datasets. The system is successfully performed a processing of 60MB data under an interval of 11,000 sec under Java and less than 2500 sec in Hadoop cluster. The system also showcases a new edge on understanding a Hadoop MapReduce operation for redundancy. Enhancement in terms of direct server dataset processing can be proposed in the upcoming version of this system.

REFERENCES

[1] D. Agrawal, S. Das, and A. E. Abbadi, “Big Data and cloud computing: Current state and future opportunities,” in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp. 530–533.

[2] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Mad skills: New analysis practices for Big Data,” PVLDB, vol. 2, no. 2, pp. 1481–1492, 2009.

[3] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[4] H. Herodotou et al., “Starfish: A self-tuning system for Big Data analytics,” in Proc. 5th Int. Conf. Innovative Data Syst. Res. (CIDR), 2011, pp. 261–272.

[5] K. Michael and K. W. Miller, “Big Data: New opportunities and new challenges [guest editors’ introduction],” IEEE Comput., vol. 46, no. 6, pp. 22–24, Jun. 2013.

[6] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

[7] R. D. Schneider, Hadoop for Dummies Special Edition. Hoboken, NJ, USA: Wiley, 2012.X. Yi, F. Liu, J. Liu, and H. Jin, “Building a network highway for Big Data: Architecture and challenges,” IEEE Netw., vol. 28, no. 4, pp. 5–13, Jul. /Aug. 2014.

[8] R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd Ed. New York, NY, USA: Academic Press, 1997.

[9] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ, USA: Wiley, 2003.

[10] C.-I. Chang, Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Norwell, MA, USA: Kluwer, 2003.