International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Survey Paper on Big Data Analytics in Real Time Satellite Data

Pattanshetty Shashikala G.[1], Dr. Kini S. N.[2]

[1]Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar Pune-28
Savitribai Phule Pune University, Pune, India

[2]Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar Pune-28
Savitribai Phule Pune University, Pune, India

Abstract: ***Digital World needs to deal with tremendous data having any format like text, auido, video, images, etc genertaed at high speed and it is vast in volume. This data is real time data named as Big Data. Analytics of Big Data requires deep analysis to deal with real time Big Data. Big Data comes from various sources and it is generated at high speed. To achieve performance of system, distributed processing is the best approach. This paper is a survey paper which focuses on distributed image processing. But this is not the normal image processing because it uses a Real Time Satellite Data(Big Data).It focuses on various features of various fields and tries to provide survey of Recent Advancements can be made to improve the perfomance. It includes mainly Big Data, Big Data Analytics, Distributed Image Processing and Real Time Satellite Application and Data.***

Keywords: Big Data, Big Data Analytics, Distributed Image Processing, Real Time Satellite Application and Data

1. Introduction

This is paper entitles Survey Paper on Big Data Analytics in Real Time Satellite Data. It puts focus on Big Data and its Analytics. As we know, todays world is digital one. Very large amount of data is generated at every seconds resulting into large data sets. Using traditional analytics techniques show their inadequateness to deal with real time data. Big Data Analytics overcomes the challenges of traditional analysis. It also focuses on distributed processing so workload will be distributed to increase the performance of the system. In this paper, the survey of recent advancements is made to perform distributed Image processing where images are from Real Time Satellite Data or we can say Big Data. So focus is on:

- Big Data
- Big Data Analytics
- Real Time Satellite Data Sets
- Distributed Image Processing

Normal image processing techniques can be used as base to process the real time satellite data with more advancements to achieve parallel processing. System should be defined in such a way so that it will meet all the objectives including greater performance, low storage cost, accurate analysis.

A. Big Data

Big Data is so large and complex data sets where data is generated continuously at high speed. The growth of data sets is increasing exponentially as the data are generating and gathered from numerous information-sensing mobile devices, remote sensing, software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks [1][2].Characteristics of Big Data are given as follows[3][4]:

- Volume: It specifies the quantity of data which is gathered and stored. It focuses on observing and track.
- Velocity: Big data is a real time data generated continuously at high speed. Gathering and processing of data done at Real time.
- Variety: It specifies the kind of data and its nature. It can be in any format, structured, unstructured, images, text, audio, video, etc.

Big data provides great potential to deal with Business Intelligence problems. But the knowledge discovery from Big Data is the more challenging task. Traditional analytics techniques are insufficient to deal with it. So we need Big Data Analytics Techniques to enhance the overall performance of the system and to meet all objectives of Big Data.

B. Big Data Analytics

As stated above, traditional data analytics techniques are not sufficient. Big Data analytics provides more utilized way to process data at real time. Big Data collection and analysis is challenge. The following four types of Big Data Analytics need to be considered [5] :

- Predictive: It works on past data patterns and provides possible outcomes for specific given situation. It focuses on what may happen in future by using historical data.
- Prescriptive: Prescriptive analysis tells us what actions to be performed for specific condition. It provides recommendations.
- Diagnostic: Diagnostic analysis uses past data to draw some conclusions which specifies what happened and why. This is also known as root cause analysis.
- Descriptive: It is the simplest form of analysis similar to the data mining.It works in real time and tells what is happening.

C. Real Time Satellite Data

Here we are considering remote sensing as data source. Remote Sensing can be used in vast application do-main in which outcomes of RS are utilized. Satellite Data is

generated at high speed and its is tremendous in volume. We do not have control on the incoming data. We need to deal with storing of data also. Remote sensing promotes the expansion of earth observatory system as cost-effective parallel data acquisition system to satisfy specific computational require-ments [6].Remote sensing has advantages as given below[7] :

It provides Monitoring of change detection in proper way.

It is efficient and cost effective as it covers more area. It provides solution over a collection of isolated data.

It provides information updating quickly.

Figure shows images taken from the satellite.

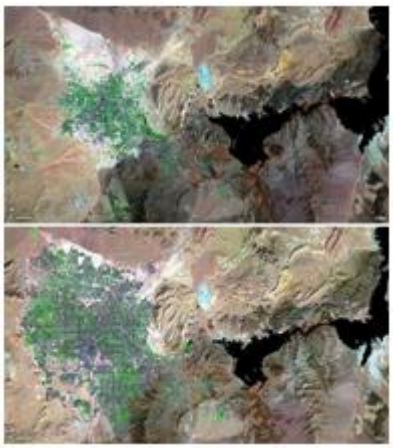

Figure 1: Las Vegas urban sprawl,1984-2011. Credit: NASA/USGS

D. Distributed Image Processing

Distribute image processing aims to process real time satellite data in such way so that the performance of system can be enhanced. We can use Apache Hadoop to distribute the workload and perform processing in parallel resulting into greater system performance at low cost and achieve storage efficiency. Basically image processing includes collection of data, pre-processing, feature extraction, pattern recognition and interpretation. These steps lead to knowledge discovery. As we are working with real time satellite data (Big data),these steps require to be more descriptive to provide more accuracy in analysis of data. Distributed processing allows advanced communication and computational cost and also provides scalability[8].

2. Literature Review

We are focusing on Big Data and Big Data Analytics. Big Data Analytics provides Capability of collecting scattered data. It helps to understand user behaviour. It is a real time processing which provides monitoring infrastructure for operators. A remote sensing application is a software application which processes remote sensing data[9].Data can be collected from satellite. RS Satellite can be divide into two types, i.e. geostationary and near-polar[7]. Geostationary satellites are boosted into a high geosynchronous orbit at approximately 35 900 kms above the equator[7].Near-polar orbiting satellites orbit the earth[7]. We can made advancements in basic image processing techniques such as Corner detection, Focal plane, segmentation etc to allow distributed image processing. Distributed processing allows images to be processed and transmitted efficiently without considering the global knowledge of visual information [8].

3. Related Work

This section focus on literature survey including Big Data, Big Data Analytics Tools and Distributed Image Processing.

Big Data and Cloud Computing: Current State and Future Opportunities

Divyakant Agrawal Sudipto Das Amr El Abbadi Department of Computer Science University of California, USA
Topic: Big Data and Cloud Computing Description :

It focuses on Cloud Computing and Big Data.Scalable Database management supports heavy applications and ad-hoc analytics and decision support

Advantages:
Scalability, elasticity, fault-tolerance, self-manageability, and ability to run on commodity hardware

Disadvantages:
To provide feasibility of system to make effective use of available resources and minimize operation cost

MAD Skills: New Analysis Practices for Big Data Jeffrey Cohen Greenplum Brian Dolan Fox Audience Network Mark Dunlap Evergreen Technologies Joseph M. Hellerstein U.C. Berkeley Caleb Welton Greenplum

Topic: Big Data Analytics Description:
Big Data Collection and analysis is a challenge. It provides Magnetic ,Agile and deep aanalysis.

Advantages:
Quick Import and Frequent iterations

Disadvantages:
Requires standardizing a vocabulary for objects like vectors, matrices, functions and functionals

Starfish: A Selftuning System for Big Data Analytics Herodotos Herodotou, Harold Lim, Gang Luo, Nedyalko Borisov, Liang Dong, Fatma Bilgen Cetin, Shivnath Babu Department of Computer Science Duke University

Topic: Big Data Analytics

Description:
It uses MAD skills to express the features and overcomes challenges of MAD. Starfish is a MADDER and self-tuning system for analytics on big data.

Advantages:
Data-life cycle awareness, Elasticity, and Robustness, minimizes cost and improves performance.

Disadvantages:
Leading us to different design choices

MapReduce Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat jeff@google.com, sanjay@google.com Google, Inc.
Topic: Map-Reduce Description:

MapReduce is a programming model and an associated implementation for processing and generating large data sets

Advantages:
Parallelize and distribute computations and to make such computations fault-tolerant, redundant execution can be used to reduce the impact of slow machines, and to handle machine failures and data loss.

Disadvantages:
Information-centric network function virtualization over 5G mobile wireless networks

C. Liang, F. R. Yu, and X. Zhang
Topic: Mobile Wireless Network

Description:
Integrating wireless network virtualization with ICN techniques can significantly improve the end-to-end network performance

Advantages:
Virtual resource allocation and in-network caching strategy as an optimization problem, which maximizes the utility function of mobile virtual network operations

Disadvantages:
Future work is in progress to consider admission control in the proposed architecture

Wireless Communications in the Era of Big Data Suzhi Bi, Rui Zhang, Zhi Ding, and Shuguang Cui
Topic: Wireless Communication Description:

The challenges and opportunities in the design of scalable wireless systems to embrace such a bigdata era. It introduces methods to capitalize from the vast data traffic, for building a bigdata-aware wireless network with better wireless service quality and new mobile applications

Advantages:
Allows us to effectively manage and in fact take advantage of wireless bigdata traffic

Disadvantages:
Future Investigation: Mobile data security and privacy, Distributed network traffic control

Research Article Change Detection in Synthetic Aper-ture Radar Images Based on Fuzzy Active Contour Models and Genetic Algorithms

Jiao Shi,JiajiWu, Anand Paul, Licheng Jiao, andMaoguo Gong

Topic: Image Processing Description :It focuses on change detection for synthetic radar images. It uses fuzzy active counter model and genetic algorithm. Advantages :Robust analysis of difference image.

Disadvantages: It requires proper selection of fuzzy co-efficient.

Analytical study of parallel and distributed image processing Harshad B. Prajapati Information Technology Department Dharmsinh Desai University Nadiad-387001, Gujarat, IN-DIA,Dr. Sanjay K. Vij Director, Sardar Vallabhbhai Patel Institute of Technology Sardar Vallabhbhai Patel Institute of Technology Vasad-388306, Gujarat, INDIA

Topic: Distributed Image Processing

Description:
Focuses on feature "Using more than one computation resource for solving certain time consuming problems"

Advantages: It focuses on image processing at low ,intermediate and high level

Disadvantages: Requires robust and advanced algorithms to implement.

Machine Learning in Image Processing
Olivier Lezoray, Christophe Charrier, Hubert Cardot,and Sebastien Lef ‚evre

Topic: Image Processing and Machine Learning

Description
Machine learning in Image Processing puts great impact on analytics.

Advantages
Better understanding of image results.

Disadvantages:
More complex to develope algorithms for the machine learning.

4. System Overview

This section provides an example of an architecture pro-posed previously named as Real time Big Data architecture for remote sensing application [6]. Following figure gives architecture:

This architecture encompasses three basic units to provide real time analytics of remote sensing Big Data.These units are listed below [6]:

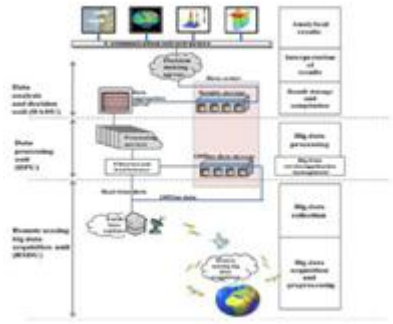

Figure 2: Remote Sensing Big Data Architecture[6]

Remote sensing data acquisition unit (RSDU):
This unit collects the data from various soucres around the globe and performs pre-processing on data.

Data Processing Unit (DPU):
This unit provides filteration and load balancing to maximize the performance of the system to meet parallelism.

Data Analysis and Decision Unit (DADU):
This unit performs aggregation and compilation of data.It stores the compiled result on Results Storage Server and then interprets the data to provide decision and accurate analytics to applications.

This architecture provides only descriptive analytics of remote sensing Big Data. If we extend this architecture by providing distributed image processing, then it can be used for the predictive analytics which will specify what might happen in future.

A. Distributed Image Processing
Basically image processing includes preprocessing, classification, feature generation, feature selection and extraction, pattern matching and recognition and it provides knowledge discovery that can be used by various applications such as biometric application, forecasting applications etc. In normal image processing ,all work goes sequentially which results into time consuming process and degrades the performance of the system. As we stated, we are dealing with real time remote sensing big data,so traditional image processing is not sufficient. Real time Big data analytics aims to process vast amount of data in such way so that the storage cost will be low and performance will be high. This can be implemented only when all task goes in parallel. An image processing is decomposed in subtasks and each task will be assigned to node in clusters. This will achieve the working of node in parallel resulting into high performance. We need to focus on incorporation of semantic model and behavior recognition [8].

5. System Analysis

There is no recent technology that can provide Big data Analytics of distributed image processing. Following figure 3 shows block diagram of the sequential processing of image. As shown in figure 3, Image data go through various phases.

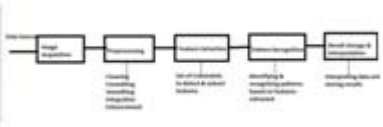

Figure 3: Basic block diagram of Image Processing

It shows the simples form of image processing. Each image is given as input and we assume that some primary preprocessing has been done. We can implement secondary preprocessing which may include formatting of image, smoothing, enhancement etc. Next phase includes feature generation, detection, selection and extraction. For the feature generation, the set constraints or rule needs to be defined. Defining set of constraints allows to extract features from data sets which are to be processed. It results into patterns and these patterns can be used for future use. Recognized patterns are stored in the results and get interpreted. These results are made available to various application for further analytics and knowledge discovery.

This scenario focuses on sequential processing of image which results into time consuming process and might halt the performance. This system works well with limited data sets. But we need to work with Big data in real time processing so there is need for distributed image processing. Basic idea is to perform segmentation of images into blocks and distribute over the clusters of node for processing. So the task will go in parallel increasing the performance. This can be implemented by using Big Data Implementation tool i.e Apache Hadoop.

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of large-scale datasets. There are two components of Apache Hadoop including HDFS and Map-Reduce.

HDFS: Hadoop Distributed File System
HDFS is a virtual file system which allows partitioning and replication. It focuses on distributed storage.

Map-Reduce
It is a programming model used for processing and generating large data sets. Map function uses key/value pair and generates intermediate key/value pairs.A reduce function merges all intermediate values associated with the same intermediate key[10].

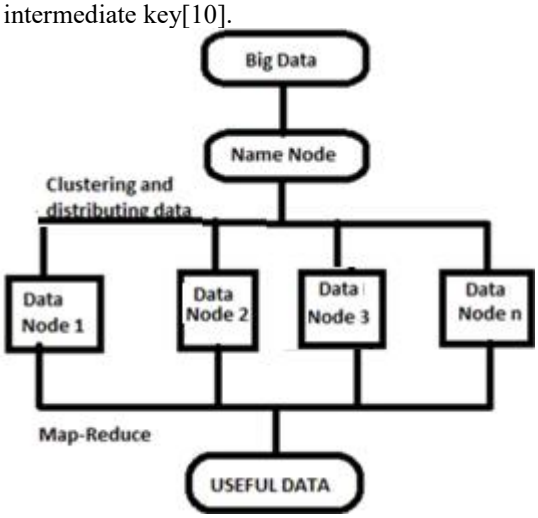

Figure 3: Basic block diagram of Hadoop Framework

Figure 4 shows the distributed processing in Hadoop Framework. This framework can be used for distributed image processing of large data sets to achieve accuracy of big data analytics and increase the performance of system. It is not the easy task, it has so many challenges to overcome n improve the working of process.

6. Conclusion

In this paper, we focused on mainly the analytics of big data. It is difficult to work with large and complex data sets. Traditional analytics shows their insufficiency to deal with large amount data where the data size varies from TB to PB. While working with real time remote sensing big data, we need to provide advancements in distributed image processing. Distributed image processing can be used in various applications. These applications are based on predictive analytics which want to know what might happen in future. This type of analytics is used in forecasting applications, tsunami prediction, earthquake prediction. Real time processing of remote sensing big data can be used in various fields like Health care systems, agriculture all types of mining etc. Distributed image processing using real time satellite data has so many challenges. There is no recent technology that can be used to implement it. We need to provide more advanced system to deal with as the collaboration and communication in distributed system is challenging.

7. Acknowledgment

This is to acknowledge and thank all the individuals who played defining role in shaping this paper. I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Dr. S. N. Kini. and Prof. A. S. Devare for giving valuable guidance, inspiration and encouragement to embark this paper. I would like to thank all respected authors of various papers, blogs, conferences which I have referred in my paper. Their innovative investigation in different fields and topics helped me to write this paper.

References

[1] Hellerstein, Joe (9 November 2008). ”Parallel Programming in the Age of Big Data”. Gigaom Blog.
[2] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.
[3] Hilbert, Martin. ”Big Data for Development: A Review of Promises and Challenges. Development Policy Review.” martinhilbert.net. Retrieved 2015-10-07.
[4] DTSC 7-3: What is Big Data?. 12 August 2015 via YouTube.
[5] http://www.smartshifttech.com/big-data-and-analytics-which-type-analytics-does-your-business-need
[6] Muhammad Mazhar Ullah Rathore, Anand Paul, Bo-Wei Chen, Bormin Huang, and Wen Ji, Real-Time Big Data Analytical Architecture for Remote Sensing Application, IEEE journal of selected topics in applied earth observations and remote sensing, 2015.
[7] http://www.fao.org/docrep/003/t0446e/T0446E04.htm
[8] Gary Chan, Pascal Frossard, and Anthony Vetro ”Distributed Image Processing”,IEEE Xplore Guest Editors
[9] https://en.wikipedia.org/wiki/Remotesensingapplication
[10] Jeffrey Dean and Sanjay Ghemawat jeff@google.com, sanjay@google.com Google, Inc. ”MapReduce: Simplified Data