ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

HIPI Based Biometric probe image retrieval using Big Image data

Mohd Ahmed Abdul Mannan
Research Scholar, Department of Computer Engineering,
JJT University, Rajasthan, India.

Gulabchand K. Gupta
Seva Sadan College of Arts, Science & Commerce
Ulhasnagar, Dist. Thane, Maharashtra, India

Abstract: Everyday Terabyte of image data is generated through the social media which can generate large Biometric data. Storing and processing Big biometric Image data is a big problem. This paper proposes a model and method using hadoop image processing interface to store the large Biometric images using Hipi image Bundel to form a Hadoop cluster, and shows how it can be process using HIPI combined with OpenCv library. Proposed model uses the hadoop distributed file system for storing these images, and uses already available image processing functions from OpenCV by integrating it with the HIPI. Paper uses the cloudera manager to manage the Hadoop cluster.

Keywords: HIPI; Cloludera; HDFS; Hipi Image Bundle; Cull;OpenCV;

I. INTRODUCTION

Since the advent of social media generation of image data is increasing day by day and now it's growing exponentially. A number of tools and techniques have been continuously developed to handle the massive growth of data. RDBMS is one of the traditional database developed by Oracle and Microsoft SQL server is capable of handling a very large database but these database have limitation when the size of the data becomes in peta byte they require more CPU and hardware to handle this. Another problem with the traditional databases are they can handle only structured data where from social media and from images, audio and video high velocity of unstructured data are generated so the traditional database is totally incapable of handling such a large data. In this paper a model and method have been proposed which can be used to handle BIG data and Big Image data. This model is mainly consist of three major parts Firstly it will have a biometric image capturing system which will be used to capture the biometric images as there are number of types of biometric images like finger print, face iris etc. We will discuss only finger print in this paper but architecture proposed in the paper will support all the Biometric images. This capturing system takes the fingerprint of the person and converts it into feature vector this feature vector later on used for storing and matching against the database. Secondly the paper proposes a HIPI based architecture to store the BIG data of biometric feature vectors on Hadoop image processing interface. Thirdly a methodology is proposed which can be used to prepare Hadoop cluster on Ubuntu machine and Mac system, once Hadoop cluster is ready using the HIPI probe image can be retrieved using proposed methodology.

II. RELATED WORK

Seyyed Mojtaba et. al. (2014), This paper introduces a way to search the image using MapReduce on open source Hadoop framework for manipulating huge volume of data[1].

Luigi Mascolo et. al. (2015), Spatial agencies have to manage astronomically immense archives of Earth observation(EO) and require solution to make the data available to the student. This paper presented implementation of space partitioning algorithm on content predicated EO images for probing profoundly and immensely colossal databases of data. K-nearest neighbor proximate neighbor search taking immensely colossal duration, utilizing index structure solve such quandary. So the authors introduces a scalable indexing procedure, implemented on top of "big data" cluster computing framework. Experiments are carried out on publicly available fullpolarimetric products from NASA/JPL UAVSAR sensor archives. Additionally Scalability tests have been conducted by utilizing cloud based virtualized commodity machines, managed by a cluster analytics framework, Apache Spark. In particular performances of indexing procedure ameliorates and amends the algorithm performance[2].

Qinghua Luet. al.(2015) Author presented a conceptual framework CF4BDA to implement Big data analytics (BDA) application in the cloud. This framework analyze the subsisting work of BDA in two perspective i.e. life cycle of BDA application and the object involved in the context of BDA application in the cloud, this conceptual framework would significantly facilitate the identification of research gaps and opportunities in BDA. It withal enables engendering and evolving strategic approaches to guide implementing BDA applications in the Cloud[3].

Sai bharath S. et. al (2015), Cloud Systems are prone to assail so there is desideratum of cyber forensic mechanism for cloud. In this paper author proposes a cloud forensic clustering model across multiple virtual machine instances. Forensic clustering in cloud is performed utilizing the subsisting susceptibilities as its features on virtual machine (VM) disk images through which a cluster of kenned susceptibilities cloud be identified. Additionally investigators can perform their analysis on a minimal set of evidences predicated on their input criteria. It provides correlation between drives enabling the investigator to do multi cross drive analysis[4].

Baaskar Hari et. al. (2015), Notifying the performance measures in an authentic Cloud environment for different applications and accommodation models under different conditions is prodigiously arduous. In this paper a simulation implement is presented which make the possibility to anticipate the performance measures and cost estimates which would be occurred in case if there is an application to be deployed in cloud. The above simulation techniques enable the Cloud customers to: (i) Test their services in rehabilitatable and tractable environment free of cost; and (ii) tune the traffic afore deploying on authentic Clouds. Such studies could avail providers to optimize their cost to access the resource and target to ameliorate their profits. If these simulation platforms

do not subsist, Cloud customers and are coerced to depend either on theoretical or imprecise valuations. The tribulation-and-error approaches lead to inefficient accommodation performance and revenue generation[5].

Zhou Jingyu et. al. (2015), The paper proposes a hybrid approach for engendering proofs of cloud search results. To achieve this, search indices is model as sets, search operations is modeled as set intersections, which can be verified more frugally with collision-resistant RSA accumulators. Computing sizably voluminous set witnesses as proofs is computationally extravagant and results in long delays for returning proofs to clients. To address this quandary, Author have performed a number of optimizations. First, they utilize an interval-predicated witness to reduce the time for computing witnesses of sizably voluminous sets and unknown search keywords. Second, to further reduce the proof size, they introduce a hybrid scheme that cumulates aggregated witnesses together with Bloom filters. Determinately, they utilize an offline pre-computing strategy and parallel execution to further reduce online proof generation time[6].

Gupta Anita et. al. (2015), This paper addresses the challenges of cloud computing and Astronomically immense data analytics. On cloud sundry security issues that can be faced are of networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, and concurrency control and recollection management. Moreover the challenges of security in cloud computing environments can be categorized into network level, utilizer authentication level, data level, and generic issues. Network interconnecting the system should be secure. Mapping of the virtual machines to the physical machines has to beperformed very securely. Networ Mathewk protocols, network securitysuch as distributed nodes, distributed data and Internode communication are the main challenges at the network level. After time δt it calls the optimization function to select the path and send RREP. Optimization function uses the individual node's battery energy; if node is having low energy level then optimization function will not use that node[7].

Chris Sweeney et. al. , This paper describes the Hadoop image processing frame work in details which can be used to build application containing large set of images[8].

III. THE PROPOSED MODEL FOR SEARCHING AN IMAGE USING HIPI

Hadoop Image processing Interface, An image processing library[8]. It is designed in such a way that it can be utilize with Apache Hadoop Mapreduce. Apache hadoop is one type of framework designed for distributed processing of large database across clusters of computers. It uses simple programming language. HIPI provide efficient image processing with high throughtput using mapReduce style parallel programs mostly executed on a cluster of computer. It also provides a solution for storing a lage collection of image storage on Hadoop distributed file system which can be further used for efficient distributed processing. The other advantage of HIPI is it provides integration with openCV, contain many image processing algorithm[9]. Working of the HIPI is shown below:

A. Step 1

The primary input object to a HIPI program is a Hipi Image Bundle(HIB) . A HIB is an accumulation of images represented as a single file on the HDFS. The HIPI distribution includes several subsidiary implements for engendering HIBs, including a MapReduce program that builds a HIB from a list of images downloaded from the Internet.

The HIPI program take input as a HIPI Image Bundle (HIB). HIPI image bundle is large collection of images these images are represented as a one file on the Hadoop Distributed File System. The HIPI have many tools for creating HIPI image bundle like MapReduce program which made a HIB from a collection of images downloaded from the web.

The HIPI program is divided into a number of stages in the first stage based on variety of user condition like a spatial resolution or using a image meta data criteria images get filtered in a HIB based this stage is known as culling stage fig1. This functionality is available in the culler class.

Fig1: Hipi image Bundel to culling stage

B. Step 2

Once the images are culled no need to decode the full image which saves the processing time. After the culling stage many images are survived these images are given to individual map task fig2. mapping is done in such a way that it maximize the data locality, which is a corner stone of Hadoop Map Reduce programming Module. To achieve this HIB Input format class is used. In the final stage individual images associcated with HIPI image header is given to the mapper as object derived from HIPI image abstract base class.

Fig2: Survived Images given to mapper

C. Step 3

The records given by the Mapper are accumulated and transmitted to the Reducer according to the built in Map Reduce shuffle algorithm that is used to minimize network traffic. . Determinately, the user defined reduce tasks are executed in parallel and their output is aggregated and in directed to the HDFS.

Fig3: Final output by reducer using Shuffle algorithm

The main advantage of HIPI is it provide support for Open source computer vision library (OpenCV)[9]. For example image classes extended from RasterImage such as ByteImage and FloatImage, require to be converted to OpenCV, Java Mat Objects which uses routines in the OpenCV Utils class. The open CV mat Writable class provides a wrapper for Java Mat Classes in the OpenCV which can be used as a value object or key in the Map Reduce programs.

IV. METHODOLOGY

To search the probe image from a large Image database initially Cloudera quick start VMs is to be setup as it contain single node apache Hadoop cluster, example data, queries, scripts, and Cloudera Manager to manage the cluster. This will help to getting commenced with all the implements needed to run image processing utilizing Hadoop[10]. But to install the Cloudera first VMWare Fusion must be installed. Below is the steps given:

1. Install the VMWare Fusion on Mac OS which allows run different operating system like Windows , NetWare, Linux etc on Virtual Machine along with Mac OS.
2. Install Cloudera Quickstart VM 5.4.x or cloudera can directly be installed on the Ubuntu.
3. Then Open VMWare Fusion and the open Cludera
4. Next we have to start the HIPI to run the MapReduce job on Apache Hadoop as Cloudera is already having Hadoop 2.6 installed which is needed for running HIPI
5. Install Apache Ant.
6. Install HIPI either by cloning latest HIPI distribution from GitHub and build from source or by downloading a precompiled JAR.
7. Next download Apache Hadoop tarball and untar it this is also needed to build HIPI.
8. Build HIPI binaries then build HIPI using ant make sure it build tools and examples.
9. Now keep your map reduce program ready for building the data to be searched on the Hadoop.
10. Add this program to any .xml file & create a jar file of it.
11. Now run this map reduce program create a .hib file on HDFS file from the image saved with HIPI using hibiimport tool and this will be input to the map reduce program.
12. Next start running the map reduce program.
13. Further we have to make ready OpenCV for , build it on Linux which will create a jar containing Java binding and all the open CV libraries this can be used to build the searching programs.
14. Combine HIPI with OpenCV , HIPI uses Hipi Image Bundle Class to represent collection of images on HDFS and Float Image for representing the image in the memory. This float image must be converted into OpenCV mat format for Image processing. ConverFloatImagetoOpenCvMat function is used for this purpose.
15. Mapper and Reducer Task for searching
    a. Load Open CV native library
    b. Create a cascade classifier.
    c. Convert HIPI float Image to OpenCV mat Image.
    d. Search and Find the matched Image.
    e. Write the closest match found.
16. Reducer
    a. Count the number of File processed
    b. Count the Top closest search.

Output these images found after the search.

V. CONCLUSION AND FUTURE WORK

A HIPI based image retrieval method is proposed in this paper. Initially VM ware Fusion is installed for Mac machine. The proposed methodology which clearly described that when HIPI is used with openCV it will be very easy to implement image processing tasks having BIG image data. Further paper shows that Search task is also optimized with the help of Hadoop MapReduce for images. Moreover two different method of installing the cloudera is presented so that both machintosh and Intel PC's can be tested with cloudera. Our future work will to test the system using the proposed methodology and show the results.

VI. REFERENCES

[1] Seyyed Mojtaba Banaei, Hossein Kardan Moghaddam (2014), " Hadoop and Its Role in Modern Image Processing", Open Journal of Marine Science, 239-245.

[2] Luigi Mascolo , Marco Quartulli , Giovanni Nico , PietroGuccione , Igor G. Olaizola (2015),"Optimised Data Structures For Large Scale Content-Based Geo-Indexing", IEEE,IGARSS,ISBN: 978-1-4799-7929-5, Pages 1488-1491.

[3] Qinghua Lu, Zheng Li, Maria Kihl, Liming Zhu, And Weishan Zhang(2015), "A Conceptual Framework for Big Data Analytics Applications in the Cloud",IEEE Translations and content mining, Vol. 3,ISSN: 2169-3536 Page 1944-1954.

[4] Saibharath S and G. Geetha kumara (2015),"Preprocessing of Evidences from Cloud components for effective forensic analysis", IEEE, International Conference on Advances in Computing, Communications and Informatics (ICACCI),ISSN: 978-1-4799-8792-4, Page 395-399.

[5] Baaskar Hari ,Sujitha and Praveen(2015), "Analysing Cloud Simulation Results Using Big Data Analytics Model" IEEE International Conference ICCCI, Coimbatore, INDIA, ISBN: 978-1-4799-6805-3,Page 08 – 10.

[6] Jingyu Zhou, Jiannong Cao, Bin Yao, and MinyiGuo (2015), "Fast Proof Generation for Verifying Cloud Search", IEEE 29th International Parallel and Distributed Processing Symposium, ISSN: 1530-2075, Page 504-513.

[7] Gupta Anita ,Abhay M. and P. M. Khan(2015), "Challenges of Cloud Comp.& Big Data Analytics",IEEE, ISBN:978-9-3805-4416-8, Page 1112-1115.

[8] Chris Sweeney, Liu Liu, Sean Arietta, Jason Lawrence(2011), "HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks", University of Virginia

[9] Timofei Epanchintsev, and Andrey Sozykin, "Processing Large Amounts of Images on Hadoop with OpenCV", 1st Ural Workshop on Parallel, Distributed, and Cloud Computing for Young Scientists (Ural-PDC), At Yekaterinburg.

[10] Frank Rischner, "Setting up a Hadoop Cluster with Cloudera Manager and Impala",University of St. Thomas.