ISSN: 2321-7782 (Online)
Volume 3, Issue 6, June 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Image Recognition System in Hadoop

K. AshaRani[1]
Asst.Professor
Department Of Computer Science Engineering
GPREC, Kurnool - India

S. Subbalakshmi[2]
Asst.Professor
Department Of Computer Science Engineering
GPREC, Kurnool - India

Abstract: ***Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. It helps to manage and process a huge amount of data efficiently. Hadoop has become a widely used open source framework for large scale data processing. Hadoop Map Reduce frame-work that allows processing of extremely large video files or image file on data nodes. By using this, Here we propose smarter way for object or image recognition , where we use a CCTV camera to be fixed at various junctions, which automatically captures the image of the person and checks the observed image with the database .This aims at processing these input images in real-time using low-end computers. The input image received from the user will direct the node-head to search for a particular object within the stored files and return the output with the files in which the object will be found. The input storage may be extended to terabytes of storage files without degrading the systems performance.***

Keywords: Bigdata, Hadoop, HDFS, MapReduce.

I. INTRODUCTION

An "**Image Recognition System"** is used to automatically identify or verify a person or an object from a digital image or a video frame from a video source. One of the ways to do this is by comparing selected image features from the database.This used for retrieving the corresponding images from the database based on their feature of images which derived the image itself. The retrieval of the image is based on the content of an image and it is more effective than the text based which is called content based image retrieval that are used for a various applications like vision techniques of computer .

Traditionally, search of the images are using text, tags or keywords or annotation assigned to the image while storing into the databases. Where as if the image which is stored in the database are not uniquely or specifically tagged or wrongly described then it's insufficient, laborious and extremely time consuming job for searching the particular image in the large set of databases. For getting most accurate result Image Recognition Systems are used which searches and retrieve the query images from the large databases based on the image content which is derived from image itself by using image processing techniques.

To store such colossal amount of data instead of using a simple Client Server architecture it'll be better to use architecture where in the data exhibits the property of logical independence. A system where in the data must be distributed on a large number of workstations so that it may reduce the burden of analysis on a single machine. Video processing is very well suited to distributed system implementation. Processing in the Hadoop is inherently distributed. Hadoop supports parallel running of applications on large clusters of commodity hardware. Big data is being generated by everything around us at all times. Every digital process Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.

II. BIG DATA

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term "Big Data is believed to be originated from the Web search companies who had to query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties:

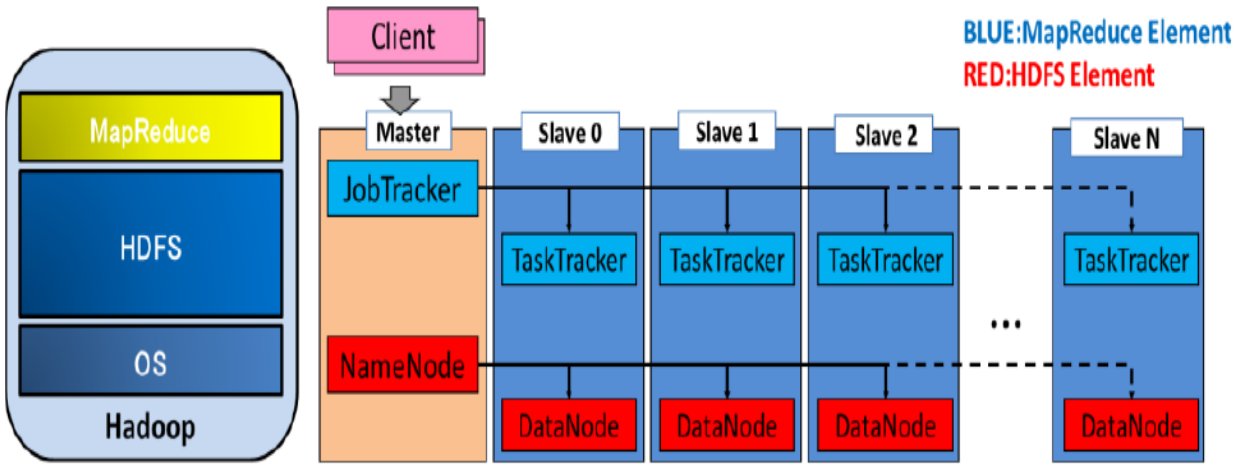
a) Volume: Many factors contribute towards increasing Volume - storing transaction data, live streaming data and data collect from sensors etc.,

b) Variety: Today data comes in all types of formats – from traditional databases, text documents, emails, video, audio, transaction s etc.,

c) Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

III. HADOOP

Hadoop, which is a free, Java-based programming framework supports the processing of large sets of data in a distributed computing environment. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. Hadoop framework has two main sub components – Hadoop Distributed File System (HDFS)and Map Reduce. There are two elements of MapReduce, namely JobTracker and TaskTracker, and two elements of HDFS, namely DataNode and NameNode.



Hadoop with HDFS and MapReduce

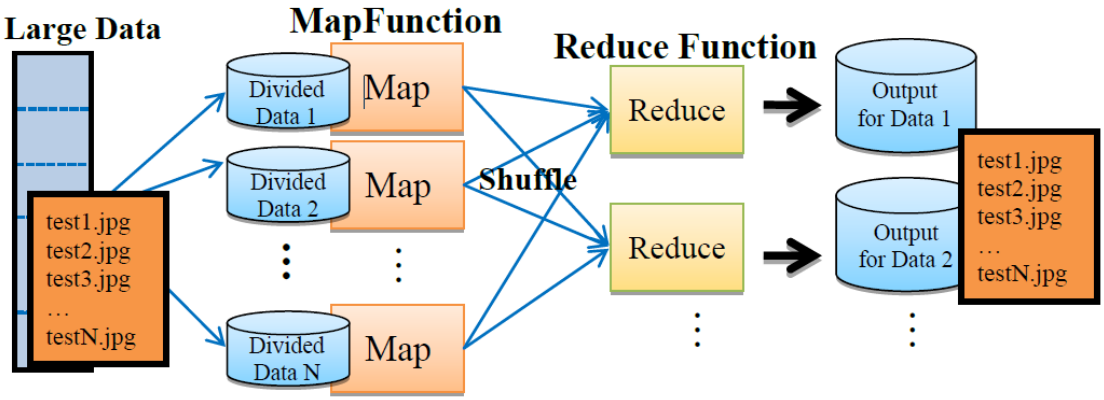
Hadoop Distributed File System (HDFS)

HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures. In this proposal HDFS is used to collect data from various digital sources such as cc cameras at hospitals, airports, and traffic junction's e.t.c.

Map Reduce

Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the

reduce tasks. Generally the input and the output of the job are both stored in a file-system. Here we implement logic to compare input image with the image present in the HDFS by using image processing algorithms.



Process performing the Map and Reduce Phases

The Map and Reduce functions of MapReduce are both defined with respect to data structured in key-value pairs. In short, we can perform distributed processing by creating key-value pairs in MapReduce form. However, for unstructured data such as video data, it can be assumed that it is more difficult to create key-value pairs and perform the processing than processing structured data

Video database processing is performed by splitting the data in a video database and creating key-value pairs. For example, the frame number can be used as a key for a video frame. In the case of parallel processing of a video frame, the video frame is divided into multiple parts, and the part numbers can be the keys (identifiers) for these different parts. Sorting is carried out using the key number, and joining separated frames or separated parts are performed by the Reduce function.

IV. OBJECTIVES

Main objective is

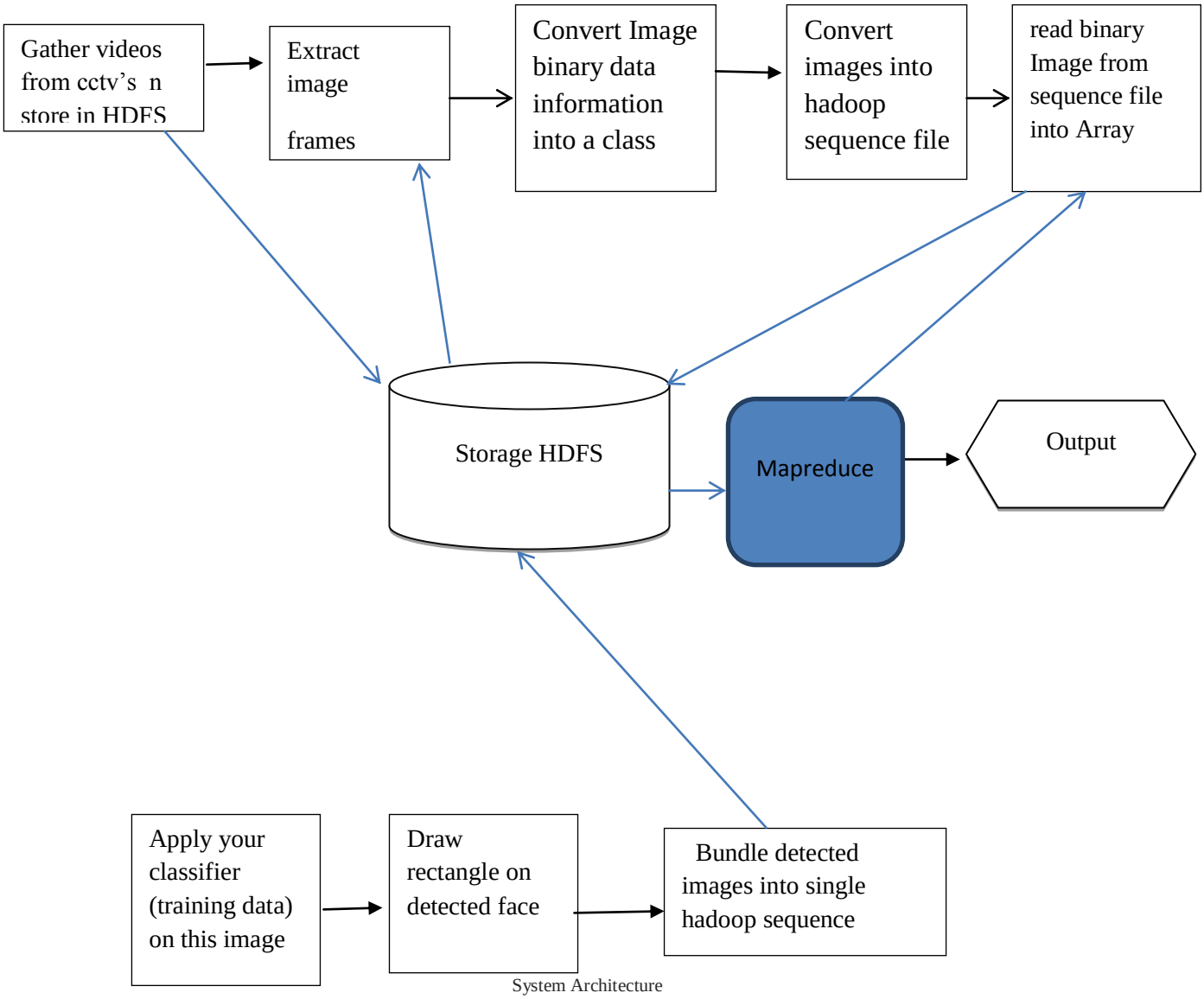
- Assume there is a very large video data base. Giving a set of video frames or image, we hope to find it from that database. The idea is simple but it is very useful in different aspects. The key point of this project is building application with high scalability. When database is increased, the application can still handle it as we are using HDFS framework for storage.
- And we implement logic in mapreduce by using some image processing algorithms so that it tries to find a face with in large database of faces. The system returns a possible list of faces from the database.
- It can be used for identifying malicious persons, most wanted persons, criminals, missing objects etc.

V. MOTIVATION

Earlier, the images were stored with associated labels and searching was done on the basis of these labels. But, such a method is prone to many errors. If the images are wrongly annotated, irrelevant images might be retrieved. Moreover, it is a laborious task to search for an image which has been assigned a wrong label.

As From the study of various sources, it was concluded that in today's era the amount of digital images are growing in a very explosive manner. The storage requirement for storing these images is also increasing from gigabyte to petabyte. Searching and retrieval of particular images from the massive database is not possible when the images in the database are wrongly annotated and described. For getting the correct image, during the search, Image recognition system can be used to search and retrieve the images from the massive collection of images. The query image is compared with database images. This can improve efficiency of searching and retrieval.

VI. ARCHITECTURE



System Architecture

This system provides the application interface for users. The system incorporates two phases – 1) Upload the images and 2) Search the image from the database. Both modules are deployed on Hadoop cluster. For uploading phase, the data is inputted for storage of HDFS and Query Image is searched from the database.

Upload Images

System allows user to upload one or more images at a single point into the system. In this proposed system, adding the image into the databases has sub processes. The user uploads the images on database of Hadoop which is called HDFS.

Searching of Images

Like uploading phase, system provides GUI for user to search and retrieve the images from the database by the query image. When a user wishes to search a particular image, he/she has to provide the search image path. The processes included in upload phase are same for search process.

Framework

HDFS is also used for storing images. In HDFS storage, data is stored in the format of text. Data is broken down into smaller pieces (called blocks) and these chunks are distributed throughout the Hadoop cluster. Hadoop provides us the facility to read/write binary files. As a result, anything which can be converted into bytes can be stored into HDFS. This provides the

scalability that is desirable for large data processing.

First after gathering videos from CCTV's, and store it in HDFS database.

1. Write one mapreduce algorithm to extract video sequence level information that is present only in the first chunk of large video file and store it as text file in HDFS.
2. Write custom recorder to read data upto end of particular Group of pictures in video file using previous text output file.
3. Implement LIBMPEG tool in custom RecordReader to extract Image frames.
4. Convert Image binary data information into a class that implement writtable like Text class.
5. Convert images into hadoop sequence file
6. Write a MapReduce Program that read binary Image from sequence file into Byte Array in map Function.
7. Convert Image in Byte array to IplImage (Image Object in Opencv)
8. Apply your classifier (training data) on this image
9. Draw rectangle on detected face
10. Bundle detected images into single hadoop sequence file
11. Store the sequence file to HDFS.
12. Extract Detected Images from sequence file.

VII. CONCLUSION

An **Image recognition system** is used to automatically identify or verify a person or an object from a digital image or a video frame from a video source. One of the ways to do this is by comparing selected image features from the database.

We propose a system with four aspects. One is uploading, second is search query image. Third is the query image search and fourth is to store all encrypted image value to database which provides security. Images are growing through the various digital devices and these images are added to the image databases and internet for various applications. These images need to be stored and retrieved in effective and efficient manner. The searching time is the most important for any search method while searching it in large datasets of images.

References

1. Sitalakhmi and S. kulkarni "MapReduce neural network framework for efficient content based image retrieval from large datasets in the cloud" IEEE Transactions on Hybrid Intelligent Systems, pp. 63 – 68,2012.
2. Ryszard S. Choras "Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems" International Journal of Biology and Biomedical Engineering, Vol. 1, 2007.
3. Oberoi Ashish, Bakshi Varun, Sharma Rohini and Singh Manpreet "A Framework for Medical Image Retrieval Using Local Tetra Pattern," International Journal of Engineering Science & Technology, Vol. 5, pp.27, Feb2013.
4. Muneto Yamamoto and Kunihiko Kaneko, "Parallel Image Database Processing With MapReduce And Performance Evaluation In Pseudo Distributed Mode," International Journal of Electronic Commerce Studies,Vol.3, No.2, pp.211-228, 2012
5. Apache Avro™ 1.7.5 Hadoop MapReduce Guide
6. http://avro.apache.org/docs/current Hadoop Image Processing Interface Java doc http://hipi.cs.virginia.edu/documentation
7. Parallel Image Database Processing With Mapreduce And Performance Evaluation In Pseudo Distributed Mode. International Journal of Electronic Commerce Studies Vol.3, No.2, pp.211-228, 2012 doi: 10.7903/ijecs.1092
8. S. Chris, L. Liu, A. Sean, and L. Jason, HIPI: A hadoop image processing interface for image-based map reduce tasks, B.S. Thesis. University of Virginia, Department of Computer Science, 2011.

AUTHOR(S) PROFILE

K. AshaRani received B.Tech degree in Computer Science and Engineering from JNT University, Hyderabad and M.Tech degree in Computer Science from JNT University, Anantapuram. She is working as Assistant Professor in Computer Science & Engineering Department in G. PullaReddy Engineering College, Kurnool. Her research interest includes BigData, Cloud Computing, Computer Network Security. She presented 3 papers in International Journals.

S. Subbalakshmi received B.Tech degree in Computer Science and Engineering from Sri Krishnadevara University, Anatapuramu and M.Tech degree in Computer Science from JNT University, Anantapuramu. She is working as Assistant Professor in Computer Science & Engineering Department in G.PullaReddy Engineering College, Kurnool. Her research interest includes wireless mesh networks, Computer Network Security. She presented 3 papers in International Journals.