Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data Analytics: Image Enhancement Based Approach

Goutam Datta, Shikha Gupta, Deepti Sharma
Dept. of Computer Science and Engineering, Advanced Institute of Technology and Management, Palwal, Haryana, India

Abstract— Big data is the collection of large set of structured, unstructured and semi structured data. In big data most of the data are unstructured in nature. If the data happens to be an image, specifically, then it becomes an important role in enhancing the quality of distorted, noisy , blur image so that it becomes an easier task of map reduce function to operate on such data to get correct (key, value)pair which may eventually be exploited for accurate statistical data analysis. In this paper, first we discuss the big data and its challenges. Next we discuss the storage aspect of big data and how mapreduce works and finally image enhancement techniques that can be helpful during mapreduce job to obtain correct statistical analysis.

Keywords— Big Data, Hadoop, MapReduce, Image Enhancement

I. INTRODUCTION

Big data is big which can vary from typically terabytes or even peta bytes. It is collected from various sources e.g. POS(point of sale), mobile internet, social networking sites like twitter, facebook etc. Big data can be traditional database, it can be text data, image data, audio data, video data etc. Big data goes on increasing as new data flows in.

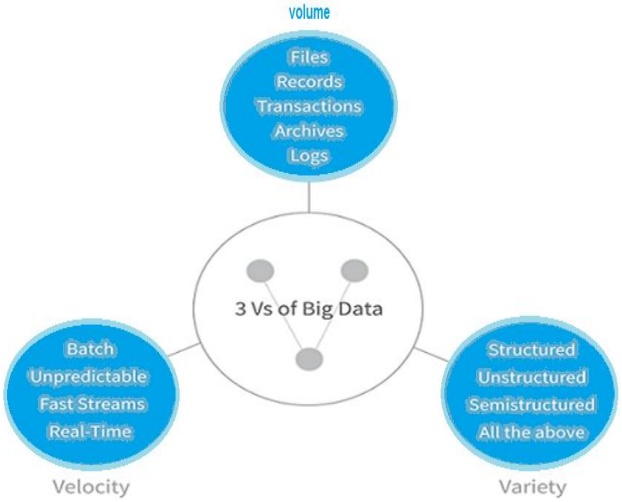

Fig. 1

1.1. Big Data Technologies : Following are some of the well known technologies used in Big Data:
Hadoop, Pig, Hbase, Dremel, No SQL and Mapreduce.

- **Hadoop**: Apache Hadoop is a software framework for storing, processing, and analyzing big data.

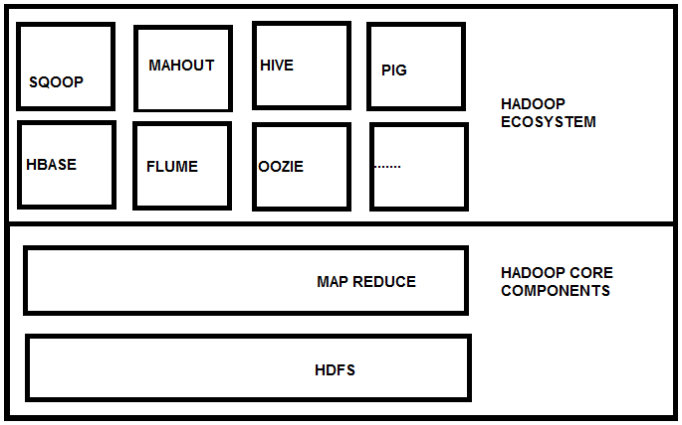

Fig. 2

Hadoop is invented and named by Doug Cutting after his son's elephant toy. Hadoop ecosystem includes multiple projects which are HDFS, MapReduce, Hive, HBase, Pig and others. The main advantage of Hadoop is its ability to efficiently handle unstructured data. Since about 80 to 90 % of Big data are unstructured data hence it has gained a lot of popularity.

Two most important components of Hadoop are Hadoop Distributed File System(HDFS) and Mapreduce. HDFS provides distributed data storage capabilities and Mapreduce which is a parallel programming framework. HDFS breaks down the processing data into smaller pieces called blocks and stores them across various nodes of a Hadoop cluster. Unlike relational databases which depend on defining schemas to store structured data, HDFS provides no restrictions on the type of data and can easily handle unstructured data too. HDFS allows schema less storage of data which makes it more popular in the context of Big Data in conjunction with NoSQL.(refer jigsaw academy's begineers guide to Analytics).

MapReduce: MapReduce is the heart of Hadoop which processes the data stored on the different nodes in distributed manner. MapReduce consists of two functions Map() and Reduce(). As the map job executes, the documents are first sent to the mapper that will count each unique word for each document : a list of key value pairs is thus created with the word as the key and its count as the value. Refer Fig. 3 which says that for the first text document the mapper produced the result would be like this:

(I,1) (Like,1)(Programming,1)

The list of (Key/value) pairs generated by all mapper tasks are then processed by the reducer that basically aggregates the (key/value) pairs from each mapper to finally produce a list of all the words and the summed counts from the three ,mappers, producing a final result which typically looks like this:

(I,1) (Like,1) (Programming,3)

In order to do Hadoop programming at the MapReduce level one need to work with Java API. And since Hadoop framework is developed on Java platform, MapReduce programming using Java language is more suitable by its design. However, for those who are not good in programming or not having knowledge of java programming, they can do Hadoop programming in terms of Pig,Hive. Even using Hadoop streaming components it is very easy to build and run MapReduce activity with any general purpose programming languages such as Perl,C++, Pythone, Ruby etc.[4]

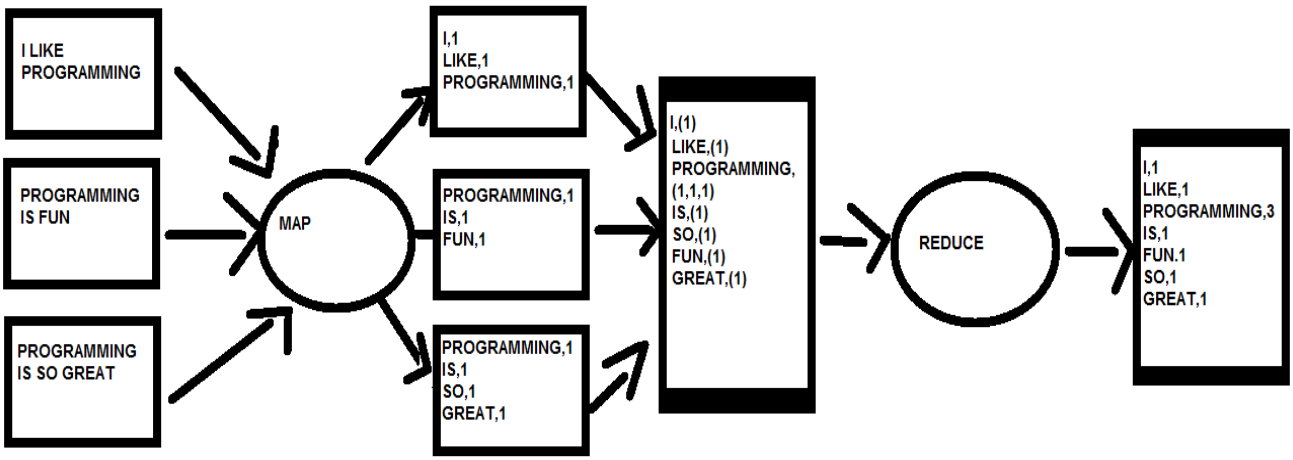

Fig. 3

Pig: Pig is a highly interactive and script based environment for executing MapReduce jobs on the data stored Hadoop clusters. Pig is very much similar to how SQL works with Relational Database Management Systems.

Pig is an easier way to write MapReduce queries. It is very much similar to Pythone and allows more efficient and shorter code to be written as compared to java.

Hive: Since most of the people are familiar with SQL and hence to make their life simpler people at facebook decided to create Hive which is an alternative to Pig. Hive enables code to be written in Hive Query Language(HQL). It is very much similar to SQL.

NOSQL: NoSQL refers to the databases that do not follow any tabular structure i.e. the data is not organized in any proper rows and columns format. Image data, audio data, video data, spatial data, email data, text from social media are some of the examples of unstructured data. There are a number of NoSQL data base technologies that are well suited for specific data problem. Some of the NoSQL databases are Hbase, MongoDB, Cassandra, CouchDB etc.

Mahout: Machine learning algorithms like clustering, classification etc. can be performed with the help of Machine learning algorithms on Hadoop distributed database.

Impala: It is a technology that is mainly used for analytics on big data. Impala is developed and promoted by Cloudera.

II. IMAGE ENHANCEMENT TECHNIQUES

Image enhancement techniques usually have one of these two goals:

i) To improve the subjective quality of image for human viewing.
ii) To modify the image in such a way as to make it more suitable for further analysis and automatic extraction of its contents [1].

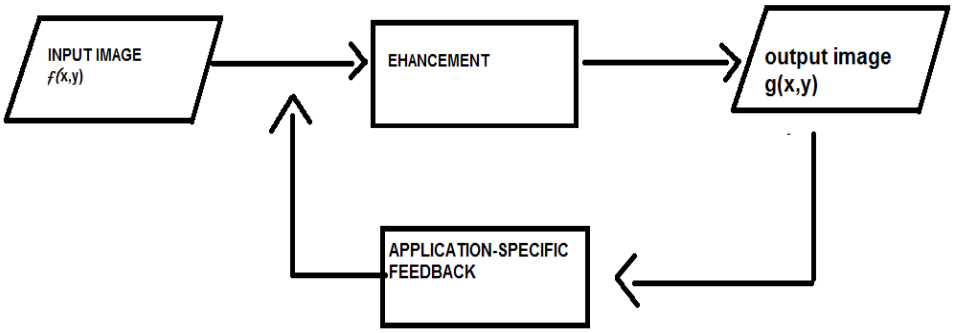

Fig. 4

In the first case , the ultimate goal is an improved version of the original image, whose interpretation can be left to a human expert-for example an enhanced X –ray image that will be used by a medical doctor to evaluate the possibility of a fractured bone. In the second scenario the goal is to serve as an intermediate step toward an automated solution that will be able to derive the semantic contents of the image- for example by improving the contrast between characters and back ground on a page of text before it is examined by an OCR algorithm.Image enhancement methods improve the detectability of important image details or objects by man or machine. It is important to mention that image enhancement algorithms are goal specific. It is typically an interactive process in which different techniques and algorithms are tried and parameters are fine tuned, until an acceptable result is obtained(Fig. 4) .

One of the popular image enhancement techniques is histogram equalization. The histogram of a monochrome image is a graphical representation of the frequency of occurrence of each gray level in the image. The data structure that stores the frequency values is a 1D array of numerical values, h, whose individual elements store the number(or percentage) of image pixels that correspond to each possible gray level. Each individual histogram entry can be expressed mathematically as $h(k)=n_k=\text{card}\{(x,y)|f(x,y)=k\}$ where k=0, 1, ….,L-1 is the number of gray levels of the digitized image and card {…} denotes the cardinality of a set that is number of elements in that set($n_{k)}$. [1]

A normalized histogram can be mathematically defined as $p(r_k)=^{n}_{k}/n$ Where n is the total number of pixels in the image and $p(r_k)$ is the probability percentage of the gray level (r_k).

Histograms are normally represented using a bar chart, with one bar per gray level, in which height of the bar is proportional to the number(or percentage) of pixels that correspond to that particular gray level.[1]

Fig. 5 represents a gray scale image. Its corresponding histogram representation is shown in Fig.6. After applying histogram equalization the image becomes more brighter(Fig.7). Histogram equalization is a technique by which the gray level distribution of an image is changed in such a way as to obtain a uniform(flat) resulting histogram, in which the percentage of pixels of every gray level is the same.[1,5] We can see the difference between the histograms of two images Fig.6 and Fig. 8. Moreover, apart from histogram equalization techniques there may be other available image enhancement techniques that can be applied as per requirement to enhance the quality of an image before feeding the unstructured data to mapreduce for further processing.

Fig. 5

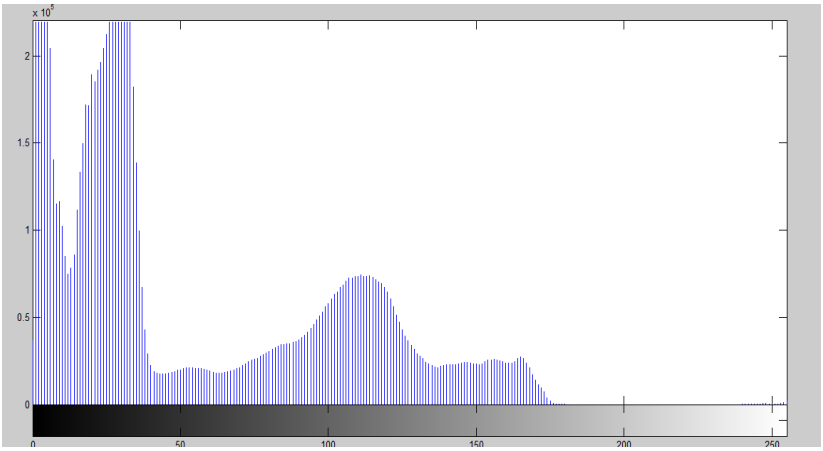

Fig. 6

Fig. 7

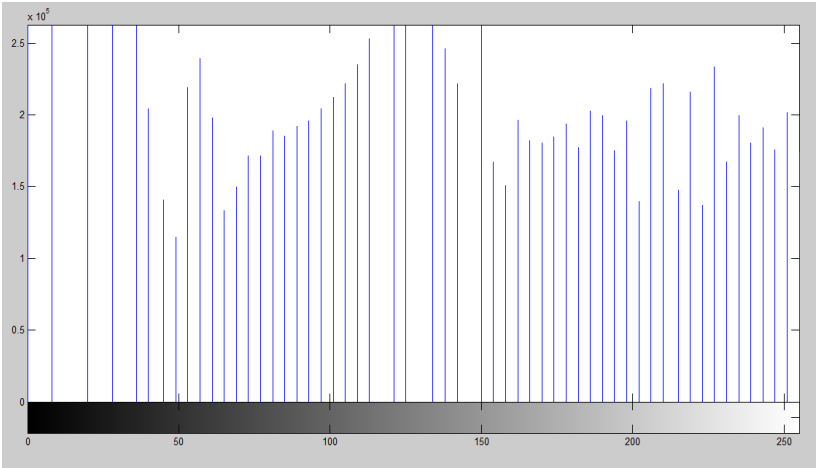

Fig. 8

III. CONCLUSION AND DISCUSSION

In this paper we have discussed the big data and its technologies and more importantly the mapreduce and its role. Mapper function plays a very vital role in generating {key,value}pair by referring various types of data i.e. structured, unstructured and semi structured. In case of unstructured data like image it may get distorted because of noise or any other factors. Enhancing the quality of degraded image so that mapper function can perform satisfactorily and accurately in generating key ,value pair which is again applied as an to the reducer function for further processing is of prime importance. This paper discusses histogram equalization technique as one of the techniques available in image enhancement which may be applied in conjunction with mapreduce function for efficient and accurate operation.

REFERENCES

[1] Oge Marques , "Practical Image and Video Processing using Matlab" , John Wiley and Sons, Hoboken, New Jersey
[2] Cloudera , " Essentials of Apache Hadoop"
[3] Big Data University- Initiative by IBM
[4] Jigsaw Academy- Resources
[5] R.C.Gonzalez, R.E.Woods, "Digital Image Processing", 3rd Edition, Prentice Hall.

