Asian Journal of Information Technology 15 (3): 418-423, 2016
ISSN: 1682-3915

Hybrid Big Data Image Processing for Creating Photographic Mosaics Using Hadoop and Scalding

[1]U. Mehraj Ali, [1]A. John Sanjeev Kumar and [2]Anantha Kunar
[1]Department of Computer Applications, Thiagarajar College of Engineering, Madurai 15,
[2]Department of Computer Science and Engineering, Thiagarajar College of Engineering,
Madurai 15 Tamil Nadu, India

Abstract: Hadoop has become the standard framework for administering big data process in the industry. Though, hadoop is predominantly applied over textual data, it can also be applied over binary data which includes images. This study deals with a framework, E-scalding which is capable of performing intensive image processing problem, i.e., creating photographic mosaic over big data using cloud storage. Microsoft azure is the cloud storage which acts as the storage repository for the image dataset. Image fusion technique is used as an optimization technique to enhance the resultant mosaic cell. Finally, the performance across the methods such as hadoop API based implementation (MapReduce), conventional scalding method and E-scalding with image fusion method is compared. The time taken to generate mosaic cell is estimated using the results obtained from E-scalding with image fusion method.

Key words: Hadoop, cloud computing, big data, image processing, photographic mosaic and time estimation

INTRODUCTION

Cloud computing: In the recent days, the technology trends suggest a drastic growth in the field of computing. Once considered as an auxiliary technology (Clouds) has now evolved as a trending technology and has affected the market big time. An estimated 52% of the data stored in the Cloud storage is left unsecured (Cuzzocrea *et al*., 2011). This promotes the idea of "Never" loosing data or rather "Ever" shared data. By the use of Cloud Computing technology, the data and resources are connected together via public or private network. This trend cut short the time taken to design the working module which can be scaled up as per the user requirements. This module should ensure the security of data contained within and optimize the data sharing feature. The module should also be compatible to the nature of its placement (either in public or as a private network). Based on the usage and need of the customers, the services provided in Cloud are categorized as:

- Infrastructure as Service (IaaS)
- Platform as Service (PaaS)
- Software as Service (SaaS)

Infrastructure as a Service (IaaS): Infrastructure as a service provides the Cloud technology infrastructure such as the servers storage networks and many on-demand services which the users had to buy separately before the advent of cloud technology. Buying such resources and services did not guarantee a bug free installation. The installation of the resources was also a complex task.

IaaS offers the cloud services as public and private. The public clouds are used in the situations where the resources are to be shared and run in the self-serviced deployment stations on the internet. Whereas, the private clouds maintains its own infrastructure with the help of acquired resources and provides virtualization (which is one of the many benefits in private cloud) in its private network. Additionally, certain cloud providers provide hybrid clouds (public cloud combined with the services of private cloud) as add-on feature too.

Platform as a Service (PaaS): Similar to SaaS, the platform as services brings out the applications to the central storage for the users to benefit out of this technology (Day *et al*., 2005). The differing factor from the SaaS is that, in software as service the software's where stored whereas in PaaS, the software development tools and platform is set for the users and software developers. This enables the users to design and code the software with the comfort of Clouds acting as the backbone.

Corresponding Author: U. Mehraj Ali, Department of Computer Applications, Thiagarajar College of Engineering, Madurai 15, Tamil Nadu, India

Software as a Service (SaaS): Software as service is a technology in which the software's are accessible in the internet with the help of Cloud technology (Bhattacharya *et al*., 2009). In this type of technology, the authenticated provider installs the software data in the cloud storage and he holds the exclusive rights for the distribution of the software. As Clouds provide value added service in terms of sharing, the provider may choose the license type which maybe a subscription or the type where the clients pay only for the resources they have used or it may be totally free. This trend helps both the user and provider to use their resources wisely in such a way that the software runs smoothly on any machine (either old PC's or in workstations). The growing trend of SaaS proves that the public users are realizing the value of Cloud technology.

Microsoft Azure: Azure is a Cloud service provider which provides services with the combination of both infrastructure (IaaS) and platform (PaaS) services which allows the clients to choose from a wide variety of resources and services and test them in the Cloud. The azure services prove to be efficient and economical in storing and backing up the data. A survey reports that 57% of the fortune 500 companies have invested on azure Cloud services. Azure provides enterprise solutions at an efficient and cost effective method. The services offered by azure are optimized as per the customer needs. The azure Cloud storage is well known for its secured methods of storing the customer data and ensures that only the authenticated users are allowed to view and edit the data in the database. The server performance is increased by reducing the hit time to fetch data from the database.

IaaS in azure provides trends in which the data and especially blob data can be stored with ease in the Cloud storage. The blob data type is predominantly used to store image files in the database. The infrastructure provided by the azure Cloud storage helps the blob data to be stored securely. Depending on the nature of user, he or she may opt resources to handle the data stored in the cloud to process and analyze the data.

Big data: For the past few decades, world around us is growing more potential. Innumerous organizations are moving fast towards the technologies which promise to store and secure data. The process of maintaining such large volumes of data and gaining methods to implement security is a key task at hand. This need for storing data securely in a storage provides a healthy competition between organizations which offer such service. The capability to relate personal information on consumer choices with data from tweets, evaluations and social media posts covers a wide range of opportunities for the organizations to analyze the needs of their consumers and foresee their needs and optimize the resources used. This model is termed as big data.

"Data" is always been a key area to all technological advancement. The optimized data storage trends and the analysis have always helped applications to achieve the best results when they are deployed. The data storage always plays a major role in the development of an application. The term big data discussed and referred in this article is "images". The images are stored using the blob data type in the Cloud storage (What is blob data type?).

Starting from the database to the data warehouse, the data has grown in size (from tera byte to yota byte) and the data storage technology needs an upscale technology to optimize its storage needs. This lead to the rise of big data in 2011 with the motive to store large volumes of data and analyze it as per the requirement of the user. Big data provides a great value to the organizations which are willing to adopt this practice but still the process has a considerable number of challenges for the implementation of big data analytics. This trend of working with the specialist in the field of big data analytics helps the organizations to gain insights to the latest trends, understand customer's needs and the future demands of either new products or the marketing trends. But, these strategies are usually expensive and lack suppleness.

Hadoop: Over the past 10 years, analytics in big data has largely influenced the decision support factor in the corporates and numerous innovative techniques have been industrialized to empower and process large volumes and different types of data. The highly omnipresent among these variant technologies is Apache Hadoop (Dean and Ghemawat, 2008) which is a freely available software framework which is used for storing and processing large volumes of data which may be available in clusters and these data can be scaled up which may reach petabytes. With Hadoop as the center, a network of technologies developed, which included: workflows, e.g., Oozie (Wiley *et al*., 2011), databases, e.g., HBase, (Sweeney, 2012) machine learning engines, e.g., Mahout, log analysis tools, e.g., Apache Flume (Almeer, 2012) and productivity frameworks.

Literature survey: Hadoop makes use of the MapReduce technique as its central computational model. This prototype enables the programmer to describe a computation that is simple to allocate data throughout a big data cluster, however it is based on comparatively low level perceptions of mappers and reducers. Hence, this

makes it complex to employ transformations that need several MapReduce execution steps which depends on mutual results. Numerous frameworks have been created to facilitate simpler definition of such complicated data processing pipelines and thus to enlarge productivity margins of the emerging MapReduce applications. The trending productivity frameworks include: Apache pig, Cascading and Crunch (Szul and Bednarz, 2014; Almeer, 2012; Barga *et al*., 2012).

A few frameworks are developed using programming languages which aid functional programming model and try to deliver even more intangible functional-like method for articulating hadoop computations which strictly looks like native functional concepts of the considered language. These frameworks are Scala-based Scoobi, Scrunch and Scalding and Clojure based Cascalog (Szul and Bednarz, 2014; Buyya *et al*., 2009; Barga *et al*., 2012) (the latter two are wrappers on Cascading).

Scala (Crunch, 2013) oriented frameworks are many in number because of the comparatively higher adoption of Scala and its adaptability. Though, there are significant alterations in what way they internally translate conceptual pipelines into a definite string of MapReduce jobs, the APIs they reveal are fairly related.

Scalding (Buyya *et al*., 2009) is a part of the Scala library which aids to specify the Hadoop MapReduce jobs. Scalding is a technology which is built over Cascading (Almeer, 2012) a Java library that removes low-level hadoop details. Scalding offers a field precise language to denote the results of MapReduce computations look related to Scala's collection API (Chen *et al*., 2013). Though hadoop is primarily used for treating textual data, there are many instances of working with hadoop for administering binary data which includes images in astronomy (Sakr *et al*., 2011), life sciences and other areas. In most of these applications, hadoop is used to run minor excruciatingly parallel jobs (map-only) that employ the related independent conversion to a cluster of input images.

This research work studies the features of the scalding framework such that it can be useful in creating a photo mosaic by using E-scalding technique. The image fusion is applied on the resultant image and it is run on hadoop with the help of a large set of input images. It also compares performance and time taken for the E-scalding operation over the standard hadoop API.

MATERIALS AND METHODS

Creating mosaic and image analysis: A pictorial mosaic is termed as an image (photograph) which is segmented into square sections such that each of the segmented sections are swapped with the help of another image which is equivalent to the target photo. A mosaic is produced by segmenting the image into a grid of smaller sized sub-images and then scanning the large dataset of images for a best match for each sub-image. The best match functions may include histogram comparison, color comparison or pixel-by-pixel comparison between the source image and the dataset images. The scanning of the dataset images for the best match to the source sub-image can be critical big data issue when the size of dataset images grows. The mosaic creation can be done by three methods whose workflow is described below.

Pseudo code:

- Get main image and directory of images
- Calculate values by setting the standard size for thumbnail based on original size of the source image
- Get all library images
  - Check file extension (img, png, jpeg)
  - Check if the image read is not grayscale image
  - Resize library images into thumbnails
- Calculate thumbnail average distance from the source image tile:

$$\text{Distance} = \sqrt{\left(a^2 + b^2\right)}$$

Where, "a" and "b" are the pixels in the source image and library image:

- For each tile of image find closest matching thumbnail
- Take mapping of thumbnails and create photomosaic in the source image
- Apply Image fusion to the resulting photomosaic to optimize the image quality

$$\sum = \bar{x}2\left(\left(X - Y\right)^2\right)$$

$$y = 10 \times \log\left(\frac{\left(d^2\right)}{\Sigma}\right)$$

MapReduce method: The MapReduce approach involves two phases namely Map phase and Reduce phase to create mosaic images. In the Map phase, the source image is segmented into sub-images and the distance (likeliness) between the sub-images of the source image is compared with the images in the dataset.

In the reduce phase, the image in the dataset which has minimum distance (likeliness) is considered as the best fit for the targeted sub-image in the source image and it is replaced.

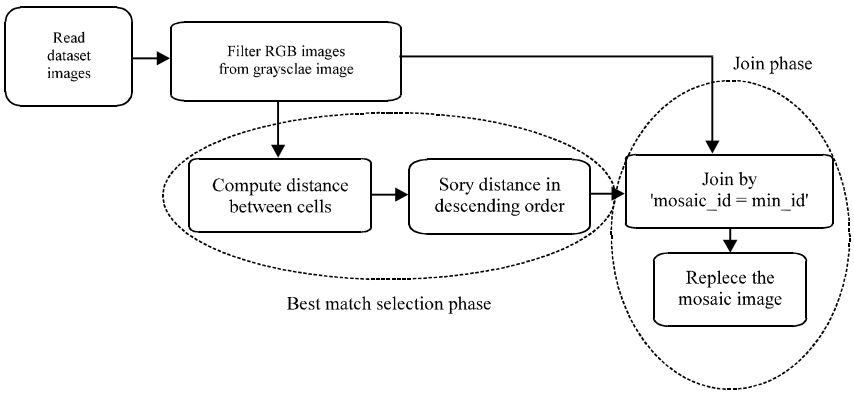

Fig. 1: Workflow of the E-scalding (Enhanced Scalding) method

Scalding method: The MapReduce method has its own disadvantages when the method is performed over Hadoop environment where the Map phase involves the process of computing the distance (likeliness) between the source sub-image and the images in the dataset. This distance has to be evaluated using all the images in the dataset which becomes a complex issue when big data is considered.

But, in the case of conventional Scalding method the mosaic cells of the source image is checked if there are any grayscale images. The distance is computed between the individual mosaic cells and the images in the dataset. Finally, the dataset image with the least distance to the targeted mosaic cell is used to replace the target mosaic cell.

E-scalding with image fusion method: The enhanced scalding approach includes reading the dataset images from the cloud storage and then filtering the RGB images before forwarding the dataset images to the scalding method. In this way, the computation time of searching the matching images to the mosaic cell in the dataset is reduced considerably. The image fusion is the process of clubbing two or more relevant images as a single image. Image fusion is used along with E-scalding as an optimization technique to enhance the resultant mosaic cell.

The E-scalding approach similar to the scalding method proves to be the better solution to create mosaic images while using big data in hadoop environment. The workflow of the E-scalding (Enhanced Scalding) is shown in Fig. 1. The steps to create mosaic's using the scalding framework is described as follows:

- The source images are filtered out in such a way that only the RGB images and the images whose dimension is greater than the mosaic cell are chosen

Fig. 2: Input image

- The "Best Match Selection" phase starts by computing the distance between the mosaic cell and the images in the dataset
- The images are rearranged based on the computed distance
- The "Join" phase includes the mapping of the best fit dataset image to the mosaic cell. The images are mapped using the ID field

The phases of E-Scalding along with Image fusion method can be described using a sample image as follows.

Step 1: The input image is first brought in to the E-scalding framework. The input image is initially checked if it is a grayscale image and the size of the source image is also checked (Fig. 2).

Step 2: The next step is the photographic mosaic creation which involves processing the source image in such a way that the image is mapped into equal sized mosaic cells where the individual mosaic cells represent a unique

Fig. 3: Image after splitting into mosaic cells

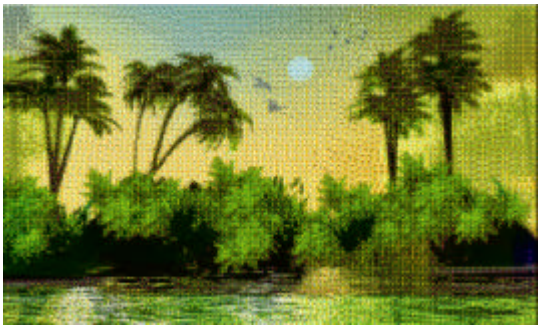

Fig. 4: Image after applying image fusion

sub-image themselves as shown in Fig. 3. The mosaic cells which are obtained by processing the source image are marked with unique ID's and the distance with the dataset images are computed.

Step 3: The final step is the image fusion step where the resultant mosaic cell is optimized and the best fit is found from the image dataset. During the image fusion phase if more than one images in the dataset is found to be the best fit then the images are combined together and replaced in the target mosaic cell (Fig. 4).

RESULTS AND DISCUSSION

The MapReduce and scalding (scalding an E-scalding with image fusion) techniques have used the image datasets ranging from 10MB to 6GB on a 6-node Hadoop cluster. Each node had the configuration of 4x Intel(R) Xeon (R) CPU E5-2660 2.20 GHz CPU, 500GB SAS HDD, 64GB RAM and running on Ubuntu 12.04 with Claudera CDH4.2.0 (hadoop-0.20). 64MB was used as the default split size which enabled the map tasks to be executed in parallel even for the large image dataset. The image dataset contained JPEG encoded images with the resolution of 75×75.

Figure 5 shows the time comparison between the three methods used to generate mosaic cells, namely,

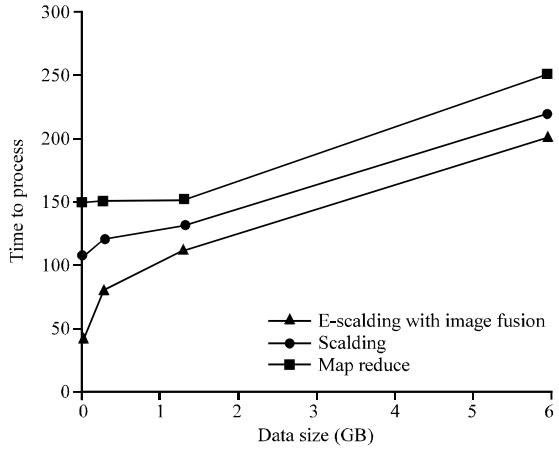

Fig. 5: Time to generate mosaic

Table 1: Results obtained after executing E-scalding with image fusion

Tile images	E-scalding with image scalding method	Fusion method
1	35.36	25.09
2	35.67	25.45
3	36.15	25.37
4	41.62	30.09
5	37.53	25.60
6	35.77	23.80
7	35.95	23.81
8	35.79	23.82
9	35.88	23.78
10	35.80	23.79

MapReduce, scalding and E-scalding with image fusion. The image dataset ranging from 9MB to 6GB is scanned to compute the time taken to generate mosaic cells. It can be observed that the scalding method takes less time compared to the MapReduce approach whereas, E-scalding with image fusion technique takes lesser time to process the mosaic cells over a large dataset of images and hence, it can be concluded as a better solution among the three methods which are used to generate mosaic cells.

The results show that the performance of the E-scalding method with Image Fusion is by average 68% higher than the Scalding method. Table 1 shows the output obtained while performing E-scalding with image fusion technique over 10 tile images. Time estimation is computed using the results obtained after performing E-scalding with Image Fusion technique.

Figure 6 shows the performance graph plotted between the E-scalding method with image fusion and scalding method for 10 tile images from the dataset. The graph clearly depicts that the E-scalding method along with the image fusion optimization provides higher performance when compared to the scalding approach.

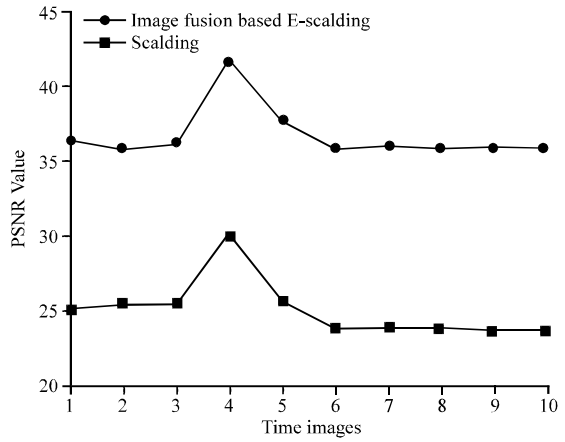

Fig. 6: Performance comparison (image fusion based E-scalding and scalding)

CONCLUSION

This research work offers an application, E-scalding to the hadoop framework to solve big data image processing issue in creating mosaic cells. The image fusion acts as the optimization technique to the process done on the mosaic cells by the E-scalding method. The E-scalding method proves to be advantageous over Hadoop's MapReduce model and the conventional Scalding model. The comparison between various methods shows that E-scalding method performs better and takes relatively less time to generate photographic mosaic cells.

RECOMMENDATIONS

The future research will focus on the execution functions including code optimization, tuning the dataset cluster and integration of image processing techniques with high performance big data image datasets. Thus, the big data image processing over hadoop will play pivotal role in overall image processing domain.

REFERENCES

Almeer, M.H., 2012. Cloud hadoop map reduce for remote sensing image analysis. J. Emerg. Trends Comput. Inf. Sci., 3: 637-644.

Barga, R.S., J. Ekanayake and W. Lu, 2012. Project daytona: Data analytics as a cloud service. Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE), April 1-5, 2012, IEEE, Washington, DC., USA., pp: 1317-1320.

Bhattacharya, I ., S. Godbole, A. Gupta, A. Verma and J. Achtermann *et al.*, 2009. Enabling analysts in managed services for CRM analytics. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1,2009, ACM, New York, USA., ISBN: 978-1-60558-495-9, pp: 1077-1086.

Buyya, R., C.S. Yeo, S. Venugopal, J. Broberg and I. Brandic, 2009. Cloud computing and emerging IT platforms: Vision, hype and reality for delivering computing as the 5th utility. Future Gener. Comput. Syst., 25: 599-616.

Chen, S., T. Bednarz, P. Szul, D. Wang and Y. Arzhaeva *et al.*, 2013. Galaxy+ Hadoop: Toward a collaborative and scalable image processing toolbox in cloud. Proceedings of the ICSOC 2013 Workshops on Service-Oriented Computing, December 2-5, 2013, Springer International Publishing, Gewerbestrasse, Switzerland, pp: 339-351.

Cuzzocrea, A., I.Y. Song and K.C. Davis, 2011. Analytics over large-scale multidimensional data: The big data revolution!. Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, October 24-28, 2011, ACM, New York, USA., ISBN: 978-1-4503-0963-9, pp: 101-104.

Day, M.Y., T.H. Tsai, C.L. Sung, C.W. Lee and S.H. Wu *et al.*, 2005. A knowledge-based approach to citation extraction. Proceedings of the IRI-2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005, August 15-17, 2005, IEEE, USA., ISBN: 0-7803-9093-8, pp: 50-55.

Dean, J. and S. Ghemawat, 2008. MapReduce: Simplified data processing on large clusters. Commun. ACM, 51: 107-113.

Sakr, S., A. Liu, D.M. Batista and M. Alomari, 2011. A survey of large scale data management approaches in cloud environments. Commun. Surv. Tutorials IEEE., 13: 311-336.

Sweeney, C., L. Liu, S. Arietta and J. Lawrence, 2011. HIPI: A Hadoop Image Processing Interface for Image-Based Mapreduce Tasks. Master's Thesis, University of Virginia, Charlottesville, VA., USA.,

Szul, P. and T. Bednarz, 2014. Productivity frameworks in big data image processing computations-creating photographic mosaics with hadoop and scalding. Proc. Comput. Sci., 29: 2306-2314.

Wiley, K., A. Connolly, S. Krughoff, J. Gardner and M. Balazinska *et al.*, 2011. Astronomical Image Processing with Hadoop. In: Astronomical Data Analysis Software and Systems XX. Evans, I.N., A. Accomazzi, D.J. Mink and A.H. Rots (Eds.). Seaport World Trade Center, Boston, Massachusetts, USA., pp: 93-96.