A Big Image Data Distributed Processing Frame Work in Static and Dynamic Image Cloud Processing

[1]Shaik Asha,[2]Shaik Shakeera,
[1]Assistant Professor, [2]Assistant Professor
[1]Department of Computer Science and Engineering , [2]Department of Computer Science and Engineering
[1]KG Reddy College of Engineering and Technology, Moinabad, Hyderabad, India
[2]KG Reddy College of Engineering and Technology, Moinabad, Hyderabad, India

Abstract: In this paper suggesting a functional processing frame work nominated Image Cloud processing (ICP) to powerfully subsist with the data in Image processing field. The image processing algorithms to increase greater proficiency, it focuses on giving a general structure to the algorithms that can be executed in parallel to achieve the goal The ICP framework consist of two processing systems i.e.; Static ICP (SICP) and Dynamic ICP (DICP). SICP is handling the image information pre-stored in the distributed system; DICP is initiate for dynamic information. To Manage SICP, two new data representations named as P-Image and Big-Image are implemented to coordinate with Map Reduce to attain more optimized marshalling and higher coherence. DICP is enact into a parallel processing schedule working with the standard technique of the proper framework. Image Net dataset are used to authorize the capacity of ICP framework over the traditional state-of-the-art methods, both in time efficiency and quality of results.

Keywords: **Big data, Image processing, MapReduce, Distributed system, Cloud computing**

1. INTRODUCTION:

In Present days everywhere widely uses the Image processing due to its inclusion applications in different areas, like engineering, industrial manufacturing, military, and health, etc. Large data amount comes and triggers severe constraints on data storage and processing coherence that calls for urgent solution to relieve limitations. Web age and search engine started to develop and boom, most real business Web sites such as Google, Gmail, Twitter, Face book, etc. All these are deals with millions of users' requests for image storage, indexing, querying and searching within average time. The benefit of big image information is that the two small robotic feet were part of a larger robot, which would be used for space missions. People are constantly seeking clarity about things that they don't understand. Clarity can be defined as the ability to understand something without doubts or questions. Businesses and organizations should make it a priority to not only explain to their employees how to perform their duties, but it's also just as important to explain why. By doing so, people will be able to understand the direction in which a business is going and can align their objectives accordingly with the job. The ability to see and understand the big picture clearly is a key benefit because people can see just how their responsibilities support the organization. This paper analyze a novel fruitful distributed framework is named as ICP (Image Cloud Processing) which is dedicated to offering a valid and productive model for vision. The core design of ICP is used to utilize the affluent global computing resources provided by the distributed system, it implements effective parallel processing[1]. The sophisticated distributed processing mechanism contains two comprehensive perspectives: 1) efficiently processing those static big image data already stored in the distributed system, such as the task of image classification, image retrieval, etc. that do not demand immediate response to the users .2) time processing dynamic input needs to be processed immediately and return an immediate response to the users, especially for the requests from the mobile terminal, e.g., the image processing software in the users's mobile phone. These two processing mechanisms are SICP and DICP, where S and D denotes Static and Dynamic respectively.

2. RELATED WORK:

Big data has successfully become a central theme applied to large-scale computing problems in modern years. Researches are based on parallel computing and distributed system, it have been carried out Big Table. Map Reduce is capable of processing large data amount in a parallel distributed manner across numerous nodes. There are three main processing phases in Map Reduce: the Map phase, the Shuffle phase and the Reduce phase. In Map phase, the input data is distributed across the mapped machines, where each machine then processes a subset of the data in parallel and produces some < key, value > pairs for each data record. Second on is Shuffle phase, In these gained < key, value > pairs are repartitioned and sorted within each partition so that values corresponding to the same key can be grouped together into a values set v1, v2, ... Final phase is Reduce phase, each reducer machine processes a subset of the < key, v1, v2, . . . > in parallel and writes the final results to the distributed file system.

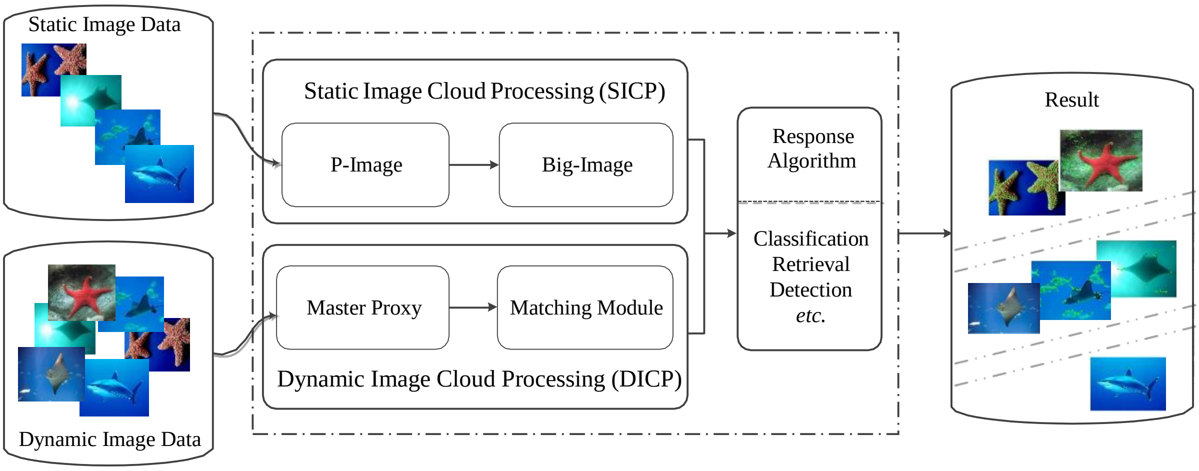


Fig 1: Frame work of ICP.

Static image data is stored in the distributed system; by gaining the image data representation (P-Image and Big-Image) response algorithms can be called to process the static image data. Dynamic image data is mainly distributed in users mobile terminals; it followed by a Master Proxy and Matching Module, related response algorithms will be called according to the input parameters[3].

P-Image:

P represents pure in P-image which consists of the imperative information, it includes filename, pixel value and width height of underlying image, all these are gotten by translating the fundamental information.

- The actual compressed variant of real image is P-Image, which just contains the important picture data acquired by decoding the initial input.
- P-Image incorporates the filename, the pixel values, and the width height of the underlying picture.
- The majority of the image processing algorithms in computer vision depend on pixel data.
- P image consist of data that will not be lost hence time utilization will be incredibly diminished by keeping away repeated operations. Once consisted in P-Image, these data would not get lost and in this way, time utilization will be incredibly diminished by keeping away from the repeated decoding operations.
- Finally there will be two dimensional lattices to keep pixel values.

Image processing utilizes Hadoop to extract the SIFT(Scale-invariant feature transform) features and produce inverted index files .Hadoop employs to achieve image feature extraction and SVM training. Even though both methods have effectively improved the time efficiency of image processing by the simple use of Hadoop, it lack adequate performance in processing big image data due to their ignorance of the fact that Hadoop is mainly developed for massive text processing. It is very hard to show the advantages of Hadoop without any additional designs in processing large-scale images. Other alternative solution is devoted to enhance the ability of Hadoop in processing images namely text processing[4]. The key point of this method is first to convert the image data to binary data stream and then process these image data using the built-in data type of Hadoop (*e.g.* Binary Writable).

Image processing algorithms are based on this method to employ the surrounding pixel points around the central one. Whereas the traditional serialized processing of image data does not support this operation. The performances of Hadoop have been successfully improved by implementing the customized image data interface. This approach has showed their progress in making the related image I/O formats distinguished by Hadoop. Different algorithms are required to implement the conversion among distinct formats of image data. These methods neglect the parallelism and efficiency of the image processing algorithms based on Hadoop platform.

3. SYSTEM OVERVIEW:

ICP framework contains two complementary processing mechanisms. They are SICP (Static Image Cloud Processing) and DICP (Dynamic Image Cloud Processing). SICP is aimed at processing those large-scale image data that have been stored in the distributed system. Static images are decoded first to maintain the required information as their corresponding P- Images which will be then stored in the data file contained in Big- Image. When the image processing is required, just need to index the index file also stored in Big-Image to find the demanded P-Images which provide the necessary image information. Using the needed image information, implement the related image processing algorithms aimed image classification, retrieval, detection, etc… DICP is designed for the dynamic requests from the clients and immediately must be able to return the results. Master Proxy accepts the client's processing request and delivers it to the Master working in the inherent mechanism that the traditional distributed system. Similarly the Master Proxy transmits the accessible parameters (e.g. the image filename and file extension) defined from the requests to the Matching Module in which these parameters would be matched with the set before hand according to real applications. If it is successfully matched, the related response algorithms will be called and use the information provided by the inherent Master-Slave mechanism of the conventional distributed system to accomplish corresponding image processing operation. From the working mechanism of SICP and DICP, it is clear that SICP is fit for processing the big image data stored in the distributed system, while DICP is more applicable when millions of mobile terminals simultaneously make a request of image processing and demand for immediate response.

4. The Model of P -Image and Big-Image

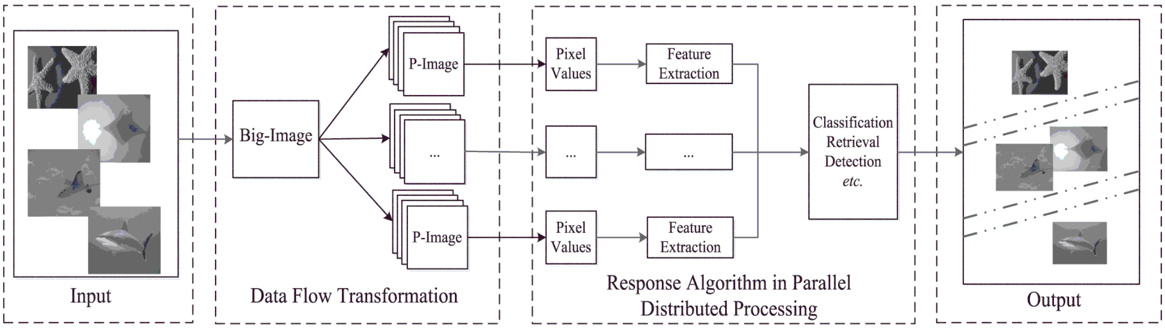


Fig 2: The Image data flow of P-Image and Big-Image

Image processing methods are based on a single node to decode the images and store all of the gained image information in the memory. Based on this perspective the image scale will be restricted to a low level due to the insufficient memory space. When the processing is finished, the image data stored in memory will be lost. Big-Image is effectively avoid the queuing delay.Users can set a threshold controlling the size of Big-Image according to the real applications. If the size of Big-Image is bigger than the threshold, then a new level of file can be designed to store multiple Big-Image files. The index structure is similar to that of Big-Image indexing P-Image.

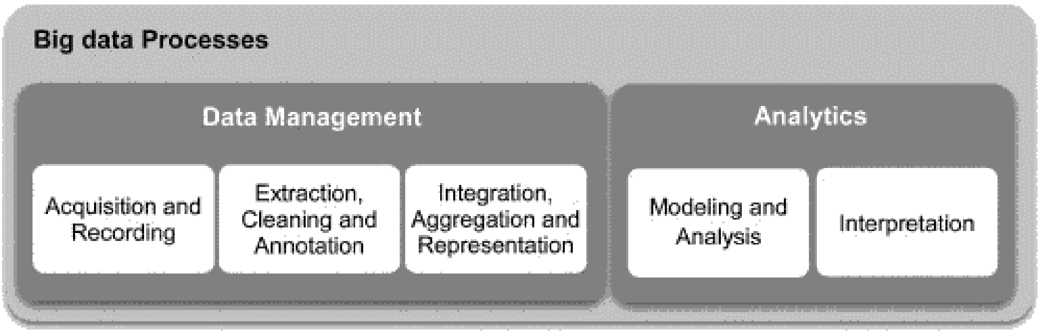

Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Fig 3: Big data process

Apache Hadoop is a distributed computing framework modeled after Google Map Reduce to process large amounts of data in parallel. Once in a while, the first thing that comes to my mind when speaking about distributed computing is EJB. EJB is de facto a component model with remoting capability but short of the critical features being a distributed computing frame work, that include computational parallelization, work distribution, and tolerance to unreliable hardware and software. Hadoop on the other hand has these merits built-in. Zoo Keeper modeled on Google Chubby is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and group services for the Hadoop cluster. Hadoop Distributed File System (HFDS) modeled on Google GFS is the underlying file system of a Hadoop cluster. HDFS works more efficiently with a few large data files than numerous small files. A real-world Hadoop job typically takes minutes to hours to complete, therefore Hadoop is not for real-time analytics, but rather for offline, batch data processing. Recently[6], Hadoop has undergone a complete overhaul for improved maintainability and manageability. Something called YARN (Yet Another Resource Negotiator) is at the center of this change. One major objective of Hadoop YARN is to decouple Hadoop from Map Reduce paradigm to accommodate other parallel computing models, such as MPI (Message Passing Interface) and Spark.

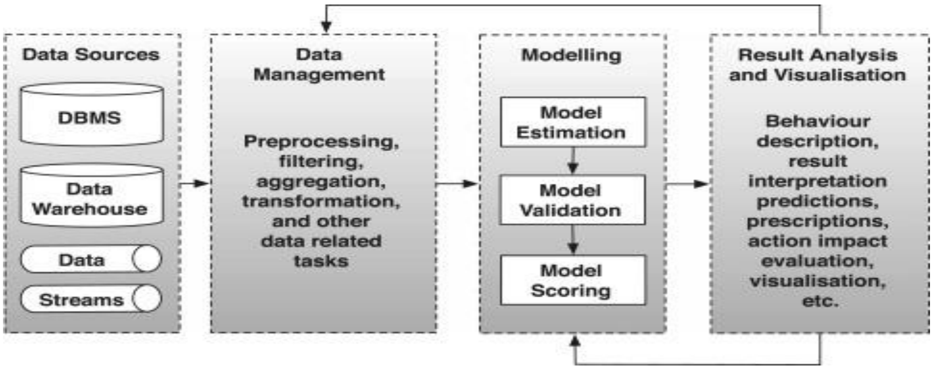

Data Sources
DBMS
Data Warehouse
Data
Streams
Data Management
Preprocessing, filtering, aggregation, transformation, and other data related tasks
Modelling
Model Estimation
Model Validation
Model Scoring
Result Analysis and Visualisation
Behaviour description, result interpretation predictions, prescriptions, action impact evaluation, visualisation, etc.

Fig 4: Big Data Computing and clouds

4.1 Static Image Cloud Processing (SICP)

SICP is used in processing the large-scale image data that have been stored in the distributed system. Static images are decoded first to maintain the required information as their corresponding P- Images which will be then stored in the data file contained in Big- Image. When the image processing is required, just need to index the index file also stored in Big-Image to find the demanded P-Images which provide the necessary image information. Using the needed image information, implement the related image processing algorithms aimed image classification, retrieval, detection, etc…

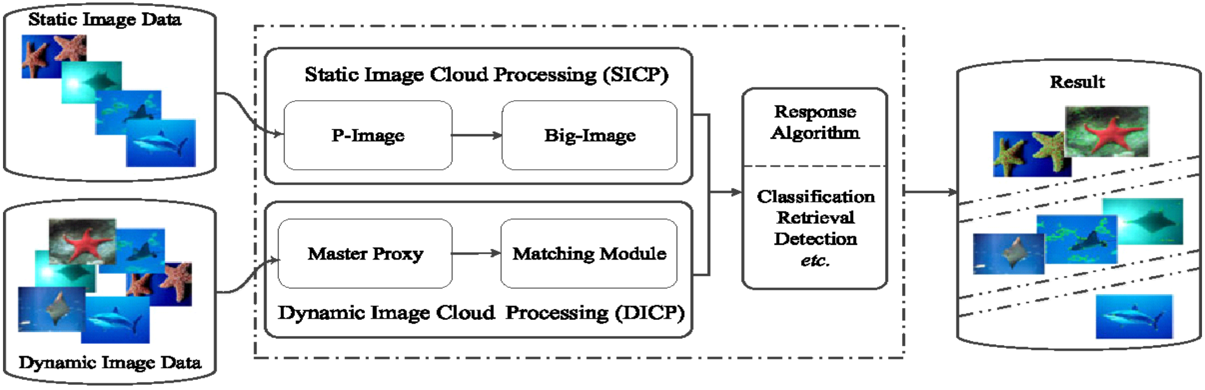


Fig 5: image data processing

4.2 Dynamic Image Cloud Processing

DICP is designed for the dynamic requests from the clients and immediately must be able to return the results. Master Proxy accepts the client's processing request and delivers it to the Master working in the inherent mechanism that the traditional distributed system. Similarly the Master Proxy transmits the accessible parameters (e.g. the image filename and file extension) defined from the requests to the Matching Module in which these parameters would be matched with the set before hand according to real applications. If it is successfully matched, the related response algorithms will be called and use the information provided by the inherent Master-Slave mechanism of the conventional distributed system to accomplish corresponding image processing operation. From the working mechanism of SICP and DICP, it is clear that SICP is fit for processing the big image data stored in the distributed system, while DICP is more applicable when millions of mobile terminals simultaneously make a request of image processing and demand for immediate response.

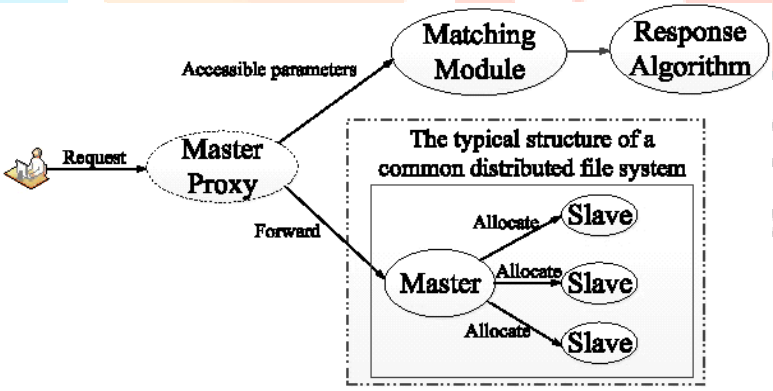


Fig 6: The Model of Dynamic Image Data Processing

5. EXPERIMENTS

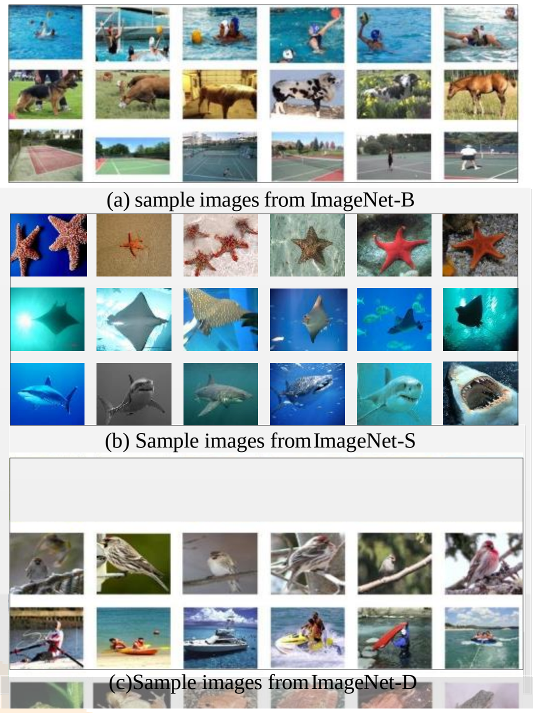
This section provides comprehensive experimental evidence for the performance of our proposed ICP: 1) to validate the efficiency of Big-Image over the traditional small image files when acting as input; 2) to verify the time efficiency of SICP when processing large-scale static image data; 3) to prove the stability and pressure resistance of DICP when processing the dynamic input .

Experimental Environment

In this paper representative results achieved on the challenging ImageNet [8] dataset running on Hadoop-1.0.3 cluster of two IBM minicomputers, each of which equips with 16-core 2.2GHz IBM CPU and 30GB memory space. Since the hardware ar- chitecture of IBM minicomputer is ppc64 bits, we used Java6 SDK also released by IBM to perform better compatibility. The operating system of the two minicomputers is SuseLinuxEnter- prise11. In order to bring the Hadoop cluster into full play, we set mapre.map.tasks as 24 and mapred.reduce.tasks as 8, both of which are key setup parameters of Hadoop. In our cluster, the number of Map Node is 8.

ImageNet Dataset

ImageNet [7] is a large-scale image dataset aiming to provide researchers an easily accessible image database and it is organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a synonym set or synset. There are more than 100,000 synsets in WordNet, majority of which are nouns (80,000+). In ImageNet, approximately 1,000 images are provided to illustrate each synset, and images of each synset are quality-controlled and human-annotated. In its completion, ImageNet will offer tens of millions of cleanly sorted images for most of the synsets in the WordNet hierarchy.



(a) sample images from ImageNet-B

(b) Sample images from ImageNet-S

(c)Sample images from ImageNet-D

Fig 7: Sample images from ImageNet.

6. CONCLUSION

This paper elaborates an effective distributed processing framework named ICP aiming to efficiently process the large- scale image data without compromising the results quality. ICP contains two types of processing mechanism, *i.e.* SICP and DICP, to achieve effective processing on the static big image data and the dynamic input, respectively. Collaborating with MapReduce, P-Image and Big-Image play the key roles of SICP to boost the time efficiency. Relying on the two newly proposed structures, time efficiency would be greatly improved by utilizing SICP to process large-scale images stored in the distributed system when compared with traditional methods based on a single node. If the new-coming image files need to be processed urgently, DICP allows for immediate response without any delay to avoid undermined problems. Extensive experiments have been conducted on ImageNet dataset to validate the efficiency of ICP. From the desirable results, we believe that big image data processing is, a promising direction, which calls for endeavor in infrastructure, computing framework, modeling, learning algorithm, applications, and all walks of life.

REFERENCES

[1] R. Hong, Y. Yang, M. Wang, X. Hua, “Learning Visual SemanticRelationships for Efficient Visual Retrieval,” IEEE Transactions on Big Data, vol.1, no.4, pp.152-161, 2015.
[2] K. Huang, C. Wang and D, Tao, “High-Order Topology Modeling of Visual Words for Image Classification,” IEEE Transactions on Image Processing, vol.24, no.11, pp.3598-3608, 2015.
[3] X. Tian, Y. Lu, N. Stender, L. Yang, D. Tao, “Exploration of Image Search Results Quality Assessment,” IEEE Transactions on Big Data, vol.1, no.3, pp.95-108, 2015.
[4] F. Wu, Z. Wang, Z. Zhang, Y. Yang, J. Luo, “Weakly Semi-Supervised Deep Learning for Multi-Label Image Annotation,” IEEE Transactions on Big Data, vol.1, no.3, pp.109-122, 2015.
[5] Y. Yang, F. Shen, H. T. Shen, H. Li, X. Li, “Robust Discrete Spectral Hashing for Large-Scale Image Semantic Indexing,” IEEE Transactions on Big Data, vol.1, no.4, pp.162-171, 2015.
[6] Y. C. Wang, C. C. Han, C. T. Hsieh, et al, “Biased Discriminant Analysis With Feature Line Embedding for Relevance Feedback-Based Image Retrieval,” IEEE Transactions on Multimedia, vol.17, no.12, pp.2245- 2258, 2015.
[7] J. S. Xu, Q.Wu, J. Zhang, F. Shen and Z. M. Tang, “Boosting Separability in Semisupervised Learning for Object Classification,” IEEE Trans. Circuits and Systems for Video Technology, vol.24, no.7, pp.1197 - 1208, 2014.
[8] L. Dong, J. Su and E. Izquierdo, “Scene-oriented Hierarchical Classification of Blurry and Noisy Images,” IEEE Trans. Circuits and Systems for Video Technology, vol.21, no.5, pp.2534-2545, 2012.
[9] L. Dong and E. Izquierdo, “A Biologically Inspired System for Classification of Natural Images,” IEEE Transactions on Image Processing, vol. 17, no. 5, pp.590-603, 2007.
[10] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, D.H.J.Epema, “Performance analysis of cloud computing services for many-Tasks scientific computing,” IEEE Trans. Parallel and Distributed Systems, vol.22, no.6, pp.931-945, 2011.
[11] Y. Lin, F. Lv, S. H. Zhu, and M. Yanget al, “Large-scale image classification: fast feature extraction and SVM training,” CVPR, pp.1689-1696, 2011.
[12] J. Su, L. Dong, P. Ren, and E. R. Hancock, “Hypergraph matching based on marginalized constrained compatibility,” ICPR, pp.2922-2925, 2012.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol.60, no.2, pp.91-110, 2004.

[14] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, vol.51, no.1, pp.107-113,2008.

[15] X. Y. Zhang, L. T. Yang, C. Liu, J. J. Chen , "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud," IEEE Trans. Parallel and Distributed Systems, vol.25,