# Implication of Neural Computing on Virtualization Framework for Big Data Analytics

Khin May Win
*University of Computer Studies, Pathein*
*winn.km05@gmail.com*

## Abstract

*Big data has become important as many organizations have been collected massive amounts of domain specific information. This information are useful for problems of national intelligence, fraud detection and medical informatics. In statistical computing, extraction of knowledge from large amount of data has required a lot of efforts. Similarly, big data analysis can be used for impressive decision making in business field by some modification in existing machine learning algorithms. Firstly, this paper will address the relationship between Cloud computing using neural networks, associated with services on the basis of cloud computing, such as the infrastructure of the system. Second, we propose the neural based framework designing take part with different data sources and how to assign distinct data mining tasks. Some popular tools for different mining patterns will also be presented for the better performance of predictive learning on big data. Cloud computing is also considered as a type of resource scheduling.*

**Keywords:** Big data, Neural Networks, data mining, Cloud computing,

## 1. Introduction

As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today. According to the estimation of Lyman and Varian [14], the new data stored in digital media devices have already been more than 92% in last century. Big data Analytics has become feasible as well as recent powerful hardware, software, and algorithms developments; however, the algorithms still need to be fast and reliable [2]. In fact, the problems of data analytics were not suddenly occurred because the creation of data is easier than finding the knowledge data. To clarify several good surveys have been presented recently and each of them views the big data from. The opportunities for new discovering new challenges, e.g how to effectively organize and manage such datasets Datasets are often very large at several GB or more, and they originate from heterogeneous sources. Hence, current real-world databases are highly susceptible to inconsistent, incomplete, and noisy data. Therefore, numerous data preprocessing techniques, including data cleaning, integration, reduction and transformation should be applied to remove noise and correct inconsistencies [9]. In addition to marketing, from the results of disease control and prevention [10], business intelligence [11], and smart city [14], we can easily understand that big data is of vital importance everywhere.

Different definitions of big data have been given by different users and different analyst of big data likeable enough the topic on data analysts and technical practitioners. Fisher et al. [7] pointed out that big data means that the data is unable to be handled and processed by most current information systems and too big to be loaded into a single machine. It also implies that the most data analytics developed for a centralized data analysis process may not be able to be applied directly to big data. Laney [13] presented a well-known definition, called 3Vs: volume, velocity, and variety. The

definition of 3Vs implies that the data size is large, the data will be created rapidly, and the data will be existed in multiple types and captured from different sources, respectively. This definition specifies the most crucial point of big data which can explore the knowledge value from dataset that makes more interesting and challenging.

In statistical perspective, big data is not big just in volume only but it is also big in terms of dimensions. Dimensions are also termed as features or attributes. Including streaming data [13], big data has the unique features of "massive, high dimensional, heterogeneous and complex" which may change the statistical and data analysis approaches [1]. Although it seems that big data makes it possible to collect more data to find more useful information. But, the truth is that more data do not necessarily and contain more ambiguous or abnormal data. From the volume perspective, the deluge of input data makes the bottleneck because it may paralyze the data analytics, Baraniuk pointed out that issue in [1].In addition, from the velocity perspective, streaming data bring up the problem within process duration to handle large quantity of data as input data. From the variety perspective, different types of incoming data have incomplete data and how to handle with input operators. For extracting useful information, there is need for modification in existing machine learning methods for better data extraction and decision making.

The existing machine learning techniques need more effort to be significant from different perspectives. Generally, the field of machine learning is divided into three subdomains: supervised learning, unsupervised learning, and reinforcement learning [16]. Briefly, supervised learning requires training with labeled data which has inputs and desired outputs. Structured, semi-structured, and even entirely unstructured data sources stimulate the generation of heterogeneous, high-dimensional, and nonlinear data. Machine learning algorithms can be used for different mining tasks and data analysis problems because they are typically employed as the "search" algorithm of the required solution. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods.

The advanced one is called distributed learning has been evolved to solve the problem of big data from different sources. One of the machine learning, Neural Computing refers to a pattern recognition methodology. The statistical model, allowing users to access centralized data resources and increased ability to combine data from disparate sources into a single search source. Neural networks have been shown to be very promising systems in many forecasting applications and business applications due to their ability for learning data.

In consideration of the heterogeneity, scalability, real time, complexity, and privacy of big data, we shall effectively "mine" the datasets at different levels during the analysis. Then, it has observed the information from massive amounts of data.

## 2. Related Work

From the data analytical viewpoint, it is essential to secure sensitive data, to protect private information and to manage data quality, exists whether data sets are big or small. However, the specific properties of big data (volume, variety, velocity, veracity) create new types of risks that necessitate a comprehensive strategy to solve the big data considerations. For research, many challenges are significant to develop theories and advanced techniques that can extract knowledge from large, dynamic, multi relational information sources between structured data and human notions and concepts. For effective future prediction, data analysis using statistical modeling techniques may be applied enhances and supports the organization's business strategy.

## 2.1 Analysis on Big data using Traditional Machine learning approaches

The massive amounts of data highlight the current research efforts and the challenges to big data, as well as the future trends. The other is to analyze the connections of machine learning with statistical methods for big data processing from different perspectives. Machine leaning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. Thus, it is a highly inter disciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, information theory, statistics, and predicting demands of customers.

In response to the problems of analyzing large-scale data, quite a few efficient methods [2] such as data sampling, data condensation, density-based approaches , incremental learning have been presented. Some methods can provide accurate prediction and the performance of the data analytics process. The dimensional reduction method (e.g., Principal Component Analysis; PCA[9]) is a typical example that can reduce the input data volume to accelerate the process of data analytics. “Big Data” was coined to capture the profound meaning of this data explosion trend. Most traditional methods or data analytics developed for a centralized process may not be able to apply directly to big data. The other factors such as veracity, validity, value, variability and vagueness were added to make some complement explanation for difficulties of big data analytics.

Datasets with high dimensional features have become increasingly common nowadays. Some are studying on how to reduce the complexity of the input data because even the most advanced computer technology cannot efficiently process with a single machine. In data transformation, data reduction operators can be regarded as the preprocessing processes of data analysis [9] which attempts to extract useful data from the raw data. If the data are a duplicate copy, incomplete, inconsistent, noisy, or outliers, then these operators have to clean them up. If the data are too complex or too large to be handled, these operators will also try to reduce them. If the raw data have errors or omissions, the roles of these operators are to identify them and make them consistent. It can be expected that these operators may affect the analytics result of KDD process. And then, Dawelbeit and McCrindle employed the bin packing partitioning method to divide the input data between the computing processors to handle the high computations of preprocessing on cloud system. The cloud system is employed to preprocess the raw data and then output the refined data [8].

From the perspectives of statistical computation and data mining, Ye et al. [20] presented architecture of the services platform which integrates to provide better data analysis services, called cloud-based big data mining and analyzing services platform (CBDMASP). The design of this platform is composed of four layers: the infrastructure services layer, the virtualization layer, the dataset processing layer, and the services layer.

M. U. Bokhari et al. presented a three layered architecture model for storing and analyzing big data [22]. The three layers are data gathering layer, data storing layer and data analysis & report generation layer. In order to gather and handle the huge volume of big data and a cluster of high speed nodes or severs are kept in the data gathering layer. A recent study shows that some traditional mining algorithms, statistical methods, preprocessing solutions have been applied to several representative tools and platforms for big data analytics. In addition to considering the relationships between the input data, if we also consider the sequence or time series of the input data, then it will be referred to the sequential pattern mining is provided. The results show clearly that machine learning algorithms will be one of the essential parts of big data challenges.

The I/O performance optimization is another issue for the compression method. For this reason, Zou et al. [21] employed the tentative selection and predictive dynamic selection method which support the I/O performance optimization. Scalability is a challenging issue

with the traditional machine learning algorithms. A new field of machine learning called distributed learning has been evolved to solve this problem. Examples of the distributed machine learning algorithms are decision rules, stacked generalization, meta-learning and distributed boosting etc [1].

To speed up the response time of a data mining operator, machine learning [17] and distributed computing [18] were used alone or combined with the traditional data mining algorithms to provide more efficient ways for solving the data mining problem. One of the well-known combinations can be found in [12], Krishna and Murty attempted to combine genetic algorithm and *k*-means to get better clustering result than *k*-means separately done. Alternating direction method of multipliers (ADMM) [7,8] serving as a promising computing framework to develop distributed, scalable, online convex optimization algorithms. Then, a scalable machine learning framework enables the automatic paralleling and distribution of computation applications related to high volumes on such clusters of machines.

## 2.2 Learning for Parallel Computing

By using domain knowledge to design the preprocessing operator which provides the useful information with big data to different users on several machines.Data mining is an exploratory nature and can also be seen as studying data analysis with a special focus on large data collections. Some well-known analysis methods and tools that are used in data mining are, for example, statistics (regression analysis, pattern recognition analysis etc.). Moreover, statistical tests suitable for big data analysis, not only for the computational efficiency but also for the concern of using only part of the data another way to check the validity of the analysis results is to derive interpretable computation models. Dryad presented the new approach in [23], which is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data.The operation structure of Dryad is coordinated by a central program called job manager, which can be executed in clusters through the working nodes.

*All-Pairs*: All-Pairs [11] is a system specially designed for bio-informatics, and data mining applications. All-Pairs is implemented in four sections: system modeling, distribution of input data, batch job management and result collection. This will build a batch-processing submission jobs in partitions, while sequencing them in the batch processing system, and formulating a node running command to acquire data.

Data distribution is changing over time, which needs the learning algorithms need to solve the issue of data are often non-stationary. To surmount these obstacles, parallel computing is significant different from traditional learning methods. So, cluster analysis is used to differentiate objects with particular features and divide them into some categories according to these features, such that objects in the same category will have high homogeneity while different nodes are running. Using the virtual computers, the development of virtualization technologies have made supercomputing more accessible and affordable for big data analytics.

## 2.3 Virtualization Technique : Cloud Computing

Fortunately, some of the machine learning algorithms (e.g., neural network-based algorithms) can essentially be used for parallel computing, which have been demonstrated for cloud computing. Different set of algorithms, which are designed to recognize patterns can provide the better service of reliable data. This means that the interpretation sensory data through different computer nodes for parallel computing. In [12], Kiran and Babu introduced that the framework for distributed data mining algorithm still needs to aggregate the information from different computer nodes. The common design of distributed data mining algorithm is as follows: each mining algorithm will be performed on a computer node (worker)

which has its locally coherent data, but not the whole data.

To develop a meaningful knowledge after each mining algorithm finds its local model, the local model from each computer node has to be integrated into a final model on cloud computing platform. Each rule should use only a few variables and these variables should be partitioned as virtual data storage. Again, Kiran and Babu [12] also pointed out that the communication will be the bottleneck when using this kind of distributed computing framework. So, the development of virtualization technologies with Vector function have made appropriate number of different pair nodes and clusters on supercomputing.

Computing infrastructures that are hidden in virtualization software which support like a true computer with two types of nodes that are a coordinator and workers. It can provide a better performance at Hadoop in terms of the execution time [13]. Because Hadoop supports large memory and storage for data replication and it is a single master node. Hadoop, the architecture of map reduce agent mobility (MRAM) was changed from client/server to a distributed agent.Cluster-oriented approaches additionally suffer from a loss of information and can only determine an appropriate number of rules.

## 3. Structured Cloud based Framework

In this section, we study the virtualization frame work with cloud based optional approach. Big Data and cloud computing technologies are developed with scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques.

**Characteristics**

**Shared Infrastructure**: Through the use of virtualization software, this allows to try to share the physical storage services and operations.

**Dynamic Provisioning**: Provide services the basis of requirements through the use of drivers mechanism.

**Network to Access**: In order to access them via the Internet through a wide number of devices such as computers and mobile

**Managed Metering**: Measure to manage and improve service and provide reports and information that consumers get the services .It allows for the exchange and dissemination services through cloud computing.

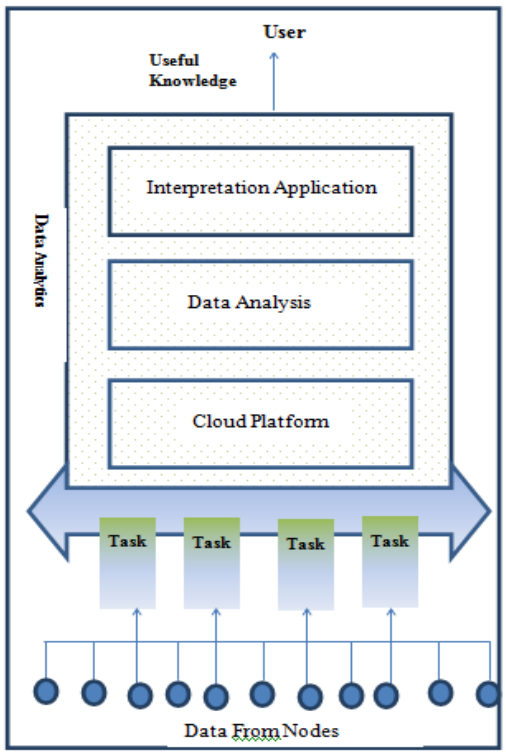


**Figure 1. Big Data Analytics on cloud computing framework**

In this proposed framework data from nodes (clusters of machines) is managed by utilizing network which is a union interface for data access and sharing different mining algorithms ensure data safety and partitioning. So data from different nodes perform the processing by different mining algorithm with distinct tasks. Utilizing the Cloud computing include offering resources when there is a demand on separate storage and supporting data

scheduling. In the data analysis layer, Artificial Neural Network, SVM and Principal Component Analysis are sufficient methods which convert the knowledge from the huge complex data chunks. Cloud computing helps in developing a computational model for all varieties of applications with infrastructure and tools (such as Hadoop).

The cloud environment should provide tools in different action that explore knowledge acquisition data for batch processing which varying client's requirements.

Many applications are hosted in the Cloud,user can get services depending on special demand, software as a service (SaaS), Data as a service (DaaS) and Infrastructure as a service (IaaS).The interaction between the service provided as part of Cloud and consumers is more available. Cloud storage which provides a possible way for storing big data.The time and cost that are needed less to upload and download big data in the cloud environment.

### 3.1 Artificial Neural Network (ANN)

Computing system is based on neural nodes with input layer , hidden layer and output layer using vector matrix function, Each node is connected with others from next layer has particular weight. The Neural network consists of a single layer of weights, where the inputs are connected directly through a series of weights for output and considered the nutrition network forward and calculated total products and inputs in each node to an end.

Computing the activation nodes for the hidden layers, need to multiply the input vector X and weight matrix $\Theta^{1,}$ for the first layer (X* $\Theta^{1}$) and then apply the activation function,

$$\mu_A : X \rightarrow [0,1] \qquad \ldots..(1)$$

$\boldsymbol{\mu_A}$ - a Membership Function (MF), determining the degree, to which *X* belongs to *A*.

The big data is divided into *n* subsets each of which is processed by a computer node (worker) in such a way that all the subsets are processed concurrently, and then the results from these *n* computer nodes are collected and transformed to a another layer node.

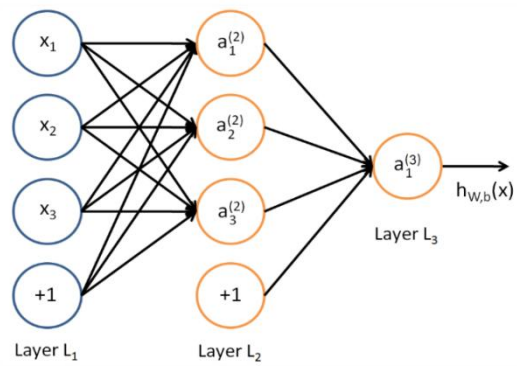


**Figure 2. Neural Nodes with input layer and hidden layer**

$$\left.\begin{aligned} a_1^{(2)} &= g(\Theta_{10}^{(1)} x_0 + \Theta_{11}^{(1)} x_1 + \Theta_{12}^{(1)} x_2 + \Theta_{13}^{(1)} x_3) \\ a_2^{(2)} &= g(\Theta_{20}^{(1)} x_0 + \Theta_{21}^{(1)} x_1 + \Theta_{22}^{(1)} x_2 + \Theta_{23}^{(1)} x_3) \\ a_3^{(2)} &= g(\Theta_{30}^{(1)} x_0 + \Theta_{31}^{(1)} x_1 + \Theta_{32}^{(1)} x_2 + \Theta_{33}^{(1)} x_3) \end{aligned}\right\} \quad \ldots. (2)$$

By multiply the hidden layer vector X and weight matrix $\Theta^{1}$, for the second layer (A* $\Theta^{1}$) and then get output for the hypothesis function g,

$$a_1^{(3)} = g(\Theta_{10}^{(2)} a_0^{(2)} + \Theta_{11}^{(2)} a_1^{(2)} + \Theta_{12}^{(2)} a_2^{(2)} + \Theta_{13}^{(2)} a_3^{(2)}) \qquad \ldots. (3)$$

By applying, Sigmoid Function which choose the one with the highest activation (probability) value, transformed data to demanded nodes.

$$\boldsymbol{S(x) = \frac{1}{1 + e^{-x}} = \frac{e}{e^{x} + 1}} \qquad \ldots. (4)$$

From the perspective of platform performance, the parallel processing models improve the performance of the system by using a new larger computing system. Neural nodes were developed a business model for all varieties of applications with infrastructure and tools. Big data applications using cloud computing can support data analytic and development.

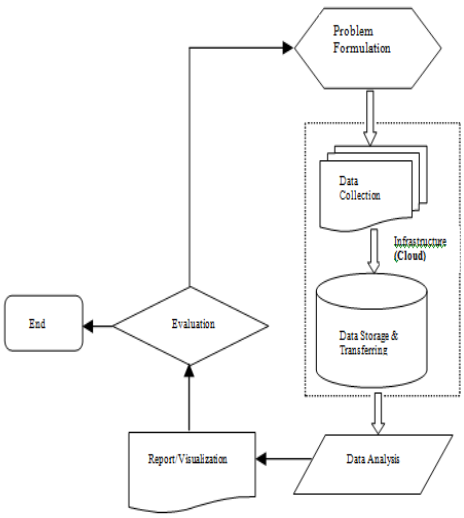


**Figure 3. Work flow for Big Data Computing**

In cloud environment, computing parts which are attached with other cross the network including user part of vision and storage servers. Cloud computing not only solves large applications that may arise in various domains but also enable scaling of different tools from virtual technologies.

## 4. Review on Cloud Computing Platform

Most of the big data analytics frameworks and platforms are using Hadoop. It has been designed for parallel computing via software, Map-Reduce programmed neural-based platform. Hadoop consists of hadoop kernel, hadoop distributed file sharing system (HDFS). Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map function step.

### 4.1. Map Reduce Program Function

*Mapreduce* is programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop and MapReduce works together on framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing. With two primary functions, which specified by users: Map and Reduce.

• Map function: The *Map* function receives a <key,value> pair as input and emits a set of intermediate <key,value> pairs as output.These intermediate <key,value> pairs are automatically shuffled and ordered which will be the input to the *Reduce* function.

• Reduce function: The *Reduce* function obtains the intermediate <key,value> pairs produced in the previous phase and generates the corresponding pair as the final output of the algorithm.

MapReduce reloads the input data from the different source every time, regardless of how much it has changed from the previous iterations. The parallel computing capacity by virtue of cloud computing can improve the efficiency of data acquisition. The studying framework differs in the IT architecture while using the single layer nutrition of big data influences decision-making on advanced IT solution.

## 5. Conclusion

There is a constantly growing demand to keep solutions for different function on big data analytics. In this study, some efficient idea for big data computing using cloud frame work with neural nodes computing is developed. For the computation time, there is no doubt at all that parallel computing is better for big data. Hadoop, and Map-reduce will play the important roles for the big data processing. Adopting the idea of cloud computing, this is to use huge computing and storage resources under concentrated management, providing big data applications with fine-grained computing. The computation resources of the cloud based platform and tasks of data analysis can give better services on big data analytics.

## References


[1] Baraniuk RG. More is less: signal processing and the data deluge. Science. 2011;331(6018):717–9.

[2] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online] https://www.mapr.com/blog/top10bigdatachallengesbigdata

[3] Boyd D, Crawford K. Critical questions for big data. Inform Commun Soc. 2012;15(5):662–79.

[4] C Rudin, KL Wagstaff, Machine learning for science and society. Mach Learn 95(1), 1–9 (2014)

[5] Dean J,Ghemawat S(2008) Mapreduce: simplified dataprocessing on large clusters. ACM 51(1):107–113

[6] D Peteiro-Barral and B Guijarro-Berdinas, "A survey of methods for distributed machine learning", Progress in Artificial Intelligence, Springer, vol. 2, issue 1, pp. 1-11, 2013. DOI:10.1007/s13748-012-0035-5.

[7] Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions.2012;

[8] Ham YJ, Lee H‐W. International journal of advances in soft computing and its applications. Calc Paralleles Reseaux et Syst Repar. 2014;6(1):1–18.

[9] Han J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc.; 2005.

[10] R. Hermon and P. A. Williams, "Big data in healthcare: What is it used for?" 2014

[11] Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: Proceedings of the Interna‐ tional Conference on Contemporary Computing, 2013. pp 404–409.

[12] kranthi Kiran B, Babu AV. A comparative study of issues in big data clustering algorithm with constraint based genetic algorithm for associative clustering. Int J Innov Res Comp Commun Eng 2014

[13] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001.

[14] Lyman P, Varian H. How much information 2003?Tech.Rep,2004.[Online].Available: http://www2.sims.berkeley.edu/research/projects/how‐much‐info‐2003/printable_report.pdf.

[15] Lee J, Hong S, Lee JH. An efficient prediction for heavy rain from big weather data using genetic algorithm. Proceedings of the International Conference on Ubiquitous Information Management and Communication, 2014. pp 25:1–25:7

[16] Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.

[17] "Parallel machine learning toolbox", retrieved fromhttp://www.research.ibm.com/haifa/projects/verification/ml_tool box

[18] Russom P. Big data analytics. TDWI: Tech. Rep ; 2011.

[19] Shirkhorshidi AS, Aghabozorgi SR, Teh YW, Herawan T. Big data clustering: a review. In: Proceedings of the Interna‐ tional Conference on Computational Science and Its Applications, 2014. pp 707–720.

[20] Ye F, Wang ZJ, Zhou FC, Wang YP, Zhou YC. Cloud‐based big data mining and analyzing services platform integrat ing In: Proceedings of the International Conference on Advanced Cloud and Big Data, 2013. pp 147–151. 95. Wu X, Zhu X, Wu G‐Q, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014;

[21] Z.-H. Zhou, "Rule extraction: Using neural networks or for neural networks?" Journal of Computer Science and Technology, vol. 19, no. 2, pp. 249–253, 2004.

[22] M. U. Bokhari, M. Zeyauddin and M. A. Siddiqui, "An effective model for big data analytics", 3rd International Conference on Computing for Sustainable Global Development, pp. 3980-3982, 2016.

[23] IsardM, BudiuM, YuY,BirrellA,FetterlyD (2007) Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Oper Syst Rev 41(3):59–72

[24] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kauf‐ mann Publishers Inc.; 2005.Parallel and Distributed Processing Symposium Workshops, 2014. pp 1228–1237t