International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 3 Issue 7, June 2016

ISSN: 2394-4404

Comparative Analysis Of Big Data Analytical Techniques Using Mapreduce

Deepa Sirurmath

Department of Computer Engineering,
Ramrao Adik Institute of Technology,
Navi Mumbai, India

Rajashree Shedge

Department of Computer Engineering,
Ramrao Adik Institute of Technology,
Navi Mumbai, India

Abstract: ***Big Data is becoming an emerging topic related to database which is growing tremendously every single day; this has formatted to form wide variety of structured and unstructured data. The traditional methods of using database mining have become difficult to handle such humongous amount of data, which is the main concept behind Big Data. Big Data Analytics is the new information management approach which has come closer to find meaning to such unregulated data.The need of Big Data Analytics is to make humongous amount of data recognizable with correlations, inherent patterns, and anomalies. There are different techniques of analysis like cluster analysis, machine learning, association rule analysis, classification analysis, regression analysis. Hadoop which supports MapReduce technique is the leading platform for analyzing big data. In this paper we compare four different techniques and its algorithms using MapReduce framework of Big Data Analytics on Hadoop which are parallelized K-Means algorithm, Genetic algorithm (one max), Naive Bayes classifier, and Apriori algorithm. By com-paring we find which of these algorithms gives better efficiency and speed to be used for analysis of big data.***

Keywords: Big data, Hadoop and Mapreduce, Analytical meth-ods.

I. INTRODUCTION

Big Data is growing enthusiasm amongst the researchers, organisations, and educational field. With the amount of un-precedented flood in the data today, there is immense pressure on the data processing environment.[1] Big data term is used for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications. Decisions that were previously were based in guesswork, or on the painstakingly constructed models of reality, can now be made based on the data itself [2].

Big data can turn high volume, high velocity, structured or unstructured, heterogeneous, often noisy and high-dimensional data into something one can understand and relate. Major point in Big Data is of data extraction which provides such information that a user should be able to relate it. In such situations, the knowledge extraction process has to be very efficient and close to real time, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such big data.

A way to manage such huge amount of data, Hadoop is the technological build to Big Data. Hadoop Distributed File System (HDFS) and MapReduce programming model is used for storage and retrieval of the big data. The tera bytes size file can be easily stored on the HDFS and can be analysed with MapReduce.

The Computerized analytical method's like artificial intelli-gence, natural language processing, data mining, and predictive analytics are employed to analyze, contextualize and visualize the data. By this wide spectrum of data being collected the datasets currently produced in research, engineering and other fields makes manual analysis infeasible.

Automatic analytical methods such as Clustering analy-sis, Classification analysis, Genetic analysis, Machine learn-ing methods are required to cope with the data sets pro-duced[3][4].A detail study of which analytical technique is

ISSN: 2394-4404

more efficient with different size of data sets and parameters is compared.

The rest of the paper is organized as follows: Section II presents a survey of domain related works. Section III gives an insight brief idea about Analytical methods of Big Data, A comparative study and analysis on big data analytical techniques presented in Section IV followed by the conclusion in Section V.

II. LITERATURE REVIEW

The main characteristic of "big" data is that it contains more information, interesting patterns than "small " data. Data becomes big data when its volume, velocity and variety exceeds the ability of our systems to store, analyze and process it. With the volume and faster flow of data the method of mining to derive knowledge is getting tougher by the organizations to provide an analytic solutions. Speed has become the barrier in the method of data processing compared to traditional methods of data mining which concerned with minimal data sets.

A Big Data processing framework: It showcases a framework in which Big Data mining platform is focused on low level data access and computing. Challenges based on security and information of Big Data application domains and knowledge, which concentrates on high-level semantics, user privacy issues, actual mining algorithms.

A. BIG DATA CHARACTERISTICS

Big Data is mainly characterized by three "V" words which gives an abstract view of the concept. Figure 2.2 below

Figure 1: Big Data Processing Framework

generalises the the three characteristics : Volume, Velocity and Variety [5][6].

Volume is the amount of ever-increasing data volume that is being created every day by an individual or an organisation.

Velocity is the frequency and speed at which data is being generated, captured and shared.

Variety is new data types those from social, machine and mobile sources. New types include content, location or Geospatial, hardware data points, log data, radio frequency identification (RF-ID), social, text and web.

B. BIG DATA ANALYTICS

Analysis means to do something with the data i.e to understand the data. Accumulating it is different but understanding and coming to a decision about that accumulated data is the analysis process. The knowledge that comes from analyzing that data is what Big Data Analytics is.

Big data and analytics, together it represents a new information management approach that has been designed to derive intelligence and insights from unstructured data. Analytics is the process of examining large amount of data, from a variety of data sources having different formats, to provide a deep understanding of data which can enable decisionsin real or near real time. Many analytical concepts such as artificial intelligence, natural language processing, data mining and predictive analytics are used to analyze, contextualize and visualize the data. Big data analytical approaches can recognize inherent patterns, correlations and anomalies from wide range of data formats and sources which create such data.

The tera bytes size files can be easily stored on the HDFS and can be analyzed with Mapreduce. HDFS and mapreduce help in storing large number of files and retrieve information from these files. In this report the study is based on Hadoop by applying a number of files as input to the system and then analyzing the performance of different analysis techniques and the behaviour of parallel mapreduce on Hadoop system.

C. HADOOP AND MAPREDUCE

The tera bytes size files can be easily stored on the HDFS and can be analyzed with Mapreduce. HDFS and mapreduce help in storing large number of files and retrieve information from these files. In this report the study is based on Hadoop by applying a number of files as input to the system and then analyzing the performance of different analysis techniques and the behaviour of parallel mapreduce on Hadoop system.

Hadoop is an open source software for implementation of MapReduce, a powerful tool designed for deep analysis and transformation of very large data sets which is inspired by Googles MapReduce and Google File System [7]. It enables applications to work with thousands of nodes and petabytes of data. Hadoop comes with its default distributed file system which is Hadoop distributed file system [8].

Hadoop design contains the Name hub, information hubs, optional name hub, Task tracker and occupation tracker. Name hub kept up the Metadata data about the piece put away in the Hadoop disseminated record framework. Records are put away in squares in an appropriated way. The Secondary name hub takes the necessary steps of keeping up the legitimacy of the Name Node and redesigning the Name Node Information time to time. Information hub really stores the information. The Job Tracker really gets the employment from the client and split it into parts. Work Tracker then doles out these split occupations to the Task Tracker. Assignment Tracker keeps running on the Data hub they bring the information from the information hub and execute the undertaking. They ceaselessly converse with the Job Tracker.

ISSN: 2394-4404

Hadoop system is made of Name Node, Secondary Name Node, Data Node, Job Tracker, and Task Tracker are on the same system. The User can submit their job in the form of MapReduce task. The data Node and the Task Tracker are on the same system so that the best speed for the read and write can be achieved. Mapreduce programming model of hadoop is based on Merge Sort. The input sets are key/value pairs, and produces a set of output key/value pair.

map(k1; v1) list(k2; v2)

reduce(k2; list(v2)) list(v2)

Map when used in the Hadoop Distributed File System(HDFS) first Maps all the requested file blocks in the HDFS then Reduces them according to the required result which is shown in Figure 2.

D. RELATED WORK

Qing Liao, Fan Yang, Jingming Zhao [9] have proposed an improved parallel K-means algorithm which is one of the most well known clustering algorithms. In which they combined two strategies, one of using rule of Distance Measure which is better than the traditional Euclidean distance and second of using Initial Centroid Selection which is better than the Initial Centroid Random. So, By Distance Measure and Initial Centroids Selection strategies the parallel Kmeans algorithm can achieve stability and high processing speed than the traditional ones.

Noor Elaiza Abd Khalid, Ahmad Firdaus Ahmad Fadzil, Mazani Manaf [10] and Dino Keco, Abdulhamit Subasi [11] proposed genetic algorithm which is implemented on MapReduce model [10] and how parallel Genetic algorithm with One Max problem [11] can improve the performance of implementations of GA on Hadoop cluster. The main concept they have covered is that of constant number of map reduce tasks and constant load per nose in cluster. They have proposed that using the concept of Hadoop Distributed File System (HDFS) the each node having its own set of population performs better by convergence fitness and faster IO footprint.

Bingwei Liu, Erick Blasch, Yu Chen, Dan Shen and Genshe Chen [12] have proposed that Machine learning technologies, such as Naive Bayes Classifier (NBC) to achieve fine-grain control of the analysis procedure for a Hadoop implementation. Additional modules was were implemented on hadoop to run NBC. Parallelization was made possible by the help of MapRe-duce, which also improved in fault tolerance, data distribution and load balance. The result showed the classification accuracy, the computation time and the throughput of the system. In this they proved that by increasing input data size worked well for Hadoop whose main functionality is working for larger data sets.It performed well compared to smaller datasets.

Xin Yue Yang, Zhen Liu, Yan Fu [13] have proposed an improved parallel Apriori algorithm which depends on Association standard. By Using Hadoop,they have looked at the execution of the enhanced parallel Apriori calculation utilizing diverse hubs of Hadoop. It was proved that parallel algorithm scaled well for larger datasets keeping the speedup and by increasing the size of data nodes. Apriori was proved to be easily applied for Big Data analysis which takes full advantage of Hadoop can provide.

III. ANALYTICAL METHODS

Following are few algorithmic techniques which are used in filtering and analysis process:

A. CLUSTERING ALGORITHM

Clustering based algorithms divide data into groups (clusters). The resulting groups have to provide meaning to the data. However clustering itself is a difficult action and is based on many assumptions and contexts. As a result many differing clustering algorithms have been developed. K-means algorithm is the most well-known clustering algorithms that has been frequently used on variety of problems. However, its processing performance has usually encountered a bottleneck if used to deal with massive data. Since MapReduce has the most popular cloud computing parallel framework, which can handle massive data.So, the bottleneck of handling massive data can be solved by developing parallel algorithms, thereby improving the stability and processing capacity of the algorithms[14][9].

B. K-MEANS BASED CLUSTER ANALYSIS

K-means used in MapReduce after traditional partition is done and cluster is formed using centroid.

Initial

- ✓ Input data setx1, x2, x3,...,xn.Then split the whole dataset to sub datasets such as split1, split2, split3...splitn. The sub datasets are formed into <Key, Value> lists. And these <Key, Value> lists input into map function.
- ✓ Select k points arbitrarily from the sub datasets as initial

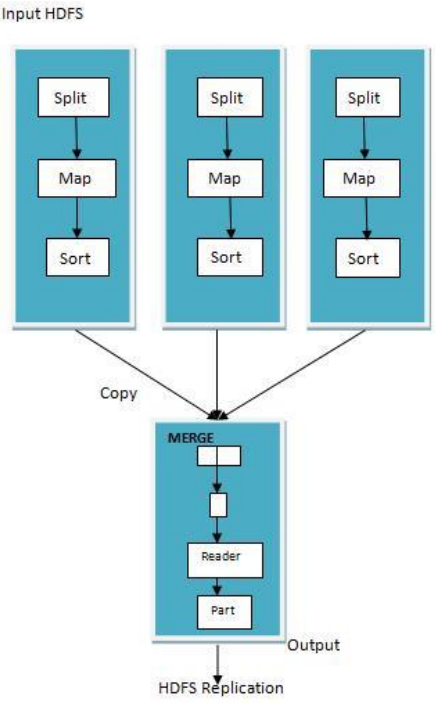

Figure 2: Mapreduce Data Flow with a single reduce task

clustering centroids.

MAPPER

- ✓ Update the clustering centroids if needed. Calculate the distance between the other data points and k centroids.
- ✓ Arrange each data into the nearest cluster until all the data have been processed.
- ✓ Output <ci, xj> pair. And ci is the centre of the cluster xj.

REDUCER

- ✓ Read <ci, xj> from Map stage. Collect all the data records. And then output the k clusters and the data points.
- ✓ Calculate the average of each cluster which is chosen as the new cluster centre.
- ✓ Calculate the difference between the new centroids with the original centroids in the same cluster. If the value is smaller than the threshold or the number of iterations of the algorithm has reached the maximum, the algorithm will stop. Then output those k clusters and the data points. Otherwise, the new cluster centroids are used to update the original centroids. Return to map stage, and continue the algorithm until convergence.

Euclidean separation technique is selected as default which has more exactness and steadiness between the articles and the centroid in each cluster. Later focus is to concentrate on enhancing the execution by diminishing the quantity of cycles and handling speed. It can be demonstrated that our enhanced procedure likewise accomplished better exactness. By utilizing Distance Measure Strategy[15]h group.

which calculates the distance between objects and centroids.So, for parallel K-means algorithm we pick the Euclidean separation methodology as the default which has more precision and solidness between the items and the centroid in every bunch.

Different initial centroids may always lead to different clustering results and different efficiency, thus [9] proposed the initial selection strategy which gives better k centroid selection.

The overall result of using the Distance Measure Strategy and the Initial Centroid Selection with parallel K-means algorithm gives a better average accuracy and improved number of iterations with the larger datasets and parallel K-means algorithm in cluster analysis achieves higher processing speed and stability.

C. GENETIC ALGORITHM

Genetic algorithm depicts the biological process of reproduction. It is a heuristic optimization method. It is a evolution based approach which requires more optimization. Such huge amount of population cant not be implemented on single machine, it wont even resolve problems which is why parallel implementaion is done[11].

MapReduce based parallel genetic algorithm is explained in [16] which was the first implementation, in this they have used one map reduce phase for one generation of genetic algorithm and for each generation new MapReduce phase is executed, it uses chain of map reduce actions for each generation. HDFS is used as a data transfer between each generation of GA. There was a downfall in this approach because of high IO footrpint due to large population of GA was stored in HDFS after every generation.

[11] In this they proposed an improved model which showed that most of the processing has been moved from reduced phase to map phase. This reduced the amount of IO footprint because all processing data was stored in memory instead of HDFS.

D. GENETIC ALGORITHMS BASED ON MAPREDUCE

BY use of ONE MAX problem [17] (or Bit counting) is a simple problem consisting in maximizing the number of ones of a bit string.It can be described as finding an string $^{!} = f\ x$

$x_1; x_2; ::::x_N$ g; withx_i f0,1g that maximizes the following equation:

$$F(x) = \sum_{i=1}^{N} x_i$$

Hadoop cluster using 10 nodes[11], the convergence and scalability of genetic algorithm with constant number of map reduce tasks and load per node in cluster for large number of datasets is performed. These are the parameters of genetic algorithm crossover, mutation, population, number of mappers/reducers, and number of iterations.

The comparison of nodes using population of large datasets up to 105 variable problems [16]. The two aspects of all nodes using same population and each node having its own population has been studied. The Hadoop based MapReduce framework for parallelization of genetic algorithm shows that when each node has its own set of population has better convergence of fitness because it has multiple mappers, which are working on different population, which finds solution much earlier and also the IO footprint is reduced because the data is not written on to the HDFS.

E. MACHINE LEARNING

Machine learning utilizes a self-learning framework. The framework is set up with chronicled data. The preparation sets instruct the framework how to unmistakable interruptions from ordinary client conduct. After the training phase the skilled system is used to detect intrusions. We are interested in using machine learning to distinct normal from irregular behaviour.

Naive Bayes Classifier (NBC) is a machine learning technology. It is used mainly to evaluate scalability of large datasets[12]. It predicts salient features from real time data by" learning" from training datasets.

The NBC categorises the entire tasks in form of Jobs, which is done after problem is simplified -

- ✓ *ALGORITHM 1:* Training job All reviews which are training set are fed into this job to produce a model for all unique words with their frequency in positive and negative review documents respectively.
- ✓ *ALGORITHM 2:* Combining job This job is used to model and test reviews are combined to a intermediate table with all necessary information for the final classification.

- ✓ *ALGORITHM 3:* Classify job This job classifies all reviews simultaneously and writes the classification results to HDFS.

The results statistics include accuracy, computation time and through put of the system. The true positive and true negative increase with respect to dataset size, while the false positive and false negative decrease. As, the dataset size increases, the accuracy gradually grows up showing that NBC is stable when the dataset increases.

F. ASSOCIATION ALGORITHM

Association rules are widely used in data mining technique which is used to find the relationship between the data sets in the database; they automatically extract useful hidden infor-mation from the datasets which are massive, noisy, and vague [10].

Apriori algorithm [13] is one of the association algorithms which are used when dealing with massive data. The MapReduce model which gives a parallel outline example to streamlining application advancements in disseminated situations to enhance the Apriori calculation as to acquire a superior parallel execution.This model can split a Large problem space into small pieces and automatically parallelize the execution of small tasks on the smaller space [18]. We utilize it to decrease the correspondence overhead without taking into record the synchronization operations between nodes.

Apriori algorithrm works on two basis: First it generates all frequent itemsets, Second it generates confident based association rule from the frequent itemsets. MapReduce is used in deploying the first step in parallel, then store the data in HDFS.

Hadoop components perform storage, job execution and workflow information storage. Files ares used to replace the database to store datasets.The datasets in files are divided into small sections. Each line is a transaction, each item in a line is separated by white space. The map function is executed on each of these data sections. Firstly, the candidate sub item are put out with the counts number after the execution of map function, then the frequent itemsets are found after the execution of map function.

Each frequent item is generated through one execution of map and reduce function which are stored on HDFS by splitting the datatsets into smaller sections and then transform them to data nodes. Map function is executed on data sections which gives <key, value> pairs for each data. The framework helps in bringing all pairs having same item and then passe the list of values to candidate items. The reduce function sums all the values and produces a count for the candidate item as a one-time synchronization. The main advantage of Map-Reduce is that is exchanges the count between nodes and not the data.

ALGORITHM OF PARALLEL APRIOIRI:

Input: I (data in HDFS), minsup(minimum support thresh-old),

Output: O, frequent itemsets

METHOD:

- ✓ L1 =find frequt 1 itemsets(D); (k=2; Lk1 ! = ;k++) f
- ✓ Ck = candidate gen (O k 1- , minsup);
- ✓ for each row of data t I f //scan I for counts
- ✓ Ct = map() ; //get the subset of the candidate itemsets
- ✓ g
- ✓ Ok = Reduce() ; // get the subset of the frequent itemsets
- ✓ g
- ✓ return O = O k [Ok ;

Speedup criterion is used to measure algorithms lead. To Increase the speedup , computation and transfer is performed parallel. In case of more than one data node, speedup increases with the increase of number of data. This can prove the high efficiency of the parallel Apriori based on MapReduce. Smaller sized datasets performance turned proved to be lower because

Methods	Clustering Algorithm Analysis[9]	Genetic Algorithm Analysis[11]	Machine Learning Analysis[12]	Association Algorithm Analysis[13]
Type of algorithm used	Parallel K-Means algorithm	Parallel Onemax based algorithm	Parallel Naive Bayes algorithm	Parallel Apriori rule algorithm
Type of result produced	Data is clustered with better centroid value	Data produces better population	Data is classified with sentiment	Data is more clearly associated
Input dataset sizes	10MB	Up to 5MB and less	Up to 2MB	Up to 5MB
Number of Data Nodes used in Mapping	10 nodes	10 nodes	Has separate workflow controller	5 and less
Accuracy met with MapReduce Framework	High accuracy	Average accuracy	Average accuracy	Average accuracy
Performance in computation	Cost time is less	Good	Better with increase in data size	Better with increased data nodes
Efficiency	Good	Better, depends on the initial population	Average, depends on the virtual cluster	Significant increase in efficiency due to parallelization
Speed	Good	Average	Good	Average

Table 1: Comparison Of Big Data Analytical TechniqueS

of extra communication time, occupying a large proportion compared to the total execution time.Dataset determines the performance of a parallel algorithm.

IV. COMPARATIVE STUDY AND ANALYSIS

A. COMPARISON OF BIG DATA ANALYTICAL TECHNIQUES

We have used four different techniques for Big Data Analysis; amongst these four techniques we selected the best

used algorithm for the study.

All four algorithms are improvised to support big data sets by paralleling the algorithms which can be implemented based on MapReduce framework of Hadoop. The Genetic algorithm depends on population of individual, small population size often leads to genetic drift and optimal solution. So, to overcome this if more computing resources are made available the performance can increase with larger population. Machine learning technique which can learn from the data sets can be used mainly for sentiment analysis, by more filtering of data with any intelligent system the analysis can be more prcised. Association rules are mainly designed to operate on transaction based database, which uses the breadth first search and forms set of rules to measure. Cluster analysis in which K-means is the widely used algorithm also known as filtering algorithm, it is easy to implement and has higher efficiency because the data points do not vary throughout the computation and hence there is no need of re-computing at each stage. The centroids play an essential role, the closer the centroids the better the stability and processing speed.

V. CONCLUSION

There are many techniques proposed for Big Data Analytics. K-means based analysis gives a good result for unstructured data with better efficiency and accuracy rate. Naive Bayes Classifier performs well in sentiment analysis with good performance speed. Genetic based analysis is good if the population is initially known with fitness function. Apriori also has good efficiency for analysis. By our analysis K-means and Naive Based Classifier are better techniques for Big Data Analytics.

REFERENCES

[1] Tekiner Firat and Keane A John "Big Data Framework", IEEE Interna-tional Conference on Systems, Man and Cybernetics, Pages 1494-1499, 2013.
[2] Agarwal Divyakant, Bernstein Philip and Bertino Elisa,"Challeneges and Oppurtunities with Big data",
[3] Wu Xindong, Zhu Xingquan, Wu Gong-Qing and Ding Wei, "Data Mining with Big Data", IEEE Transactions on Knowledge and data Engineering, Vol 26, Pages 96-105, January 2014.
[4] Hilbrich Marcus, Weber Matthias and Tschuter Ronny, "A Walk-Through on Methods from Different Perspectives", International Con-ference on Cloud Computing and Big Data, Pages 373-380, 2013.
[5] Emerging Technology Series "Big Data Analytics in Health: Emerging Technology Series. Pages 1-68, Available at "https://www.infoway-inforoute.ca/index.php/programs-services/emerging-technology", 2013.
[6] Whitworth and N Jeffrey. "Applying Hybrid Cloud Systems to Solve Challenges Posted by the Big Data Problem, M.Tech Dissertation, Uni-versity of North Carolina at Greensboro. UMI-ProQuest,No. 1551299, 2013.
[7] Qing Liao, Fan Yang, and Jingming Zhao, "All improved parallel K-means Clustering Algorithm with MapReduce, ICCT, Pages 764-768, 2013.
[8] Whitworth and N Jeffrey. "Applying Hybrid Cloud Systems to Solve Challenges Posted by the Big Data Problem, M.Tech Dissertation, University of North Carolina at Greensboro. UMI-ProQuest,No. 1551299, 2013.
[9] Qing Liao, Fan Yang, and Jingming Zhao, "All improved parallel K-means Clustering Algorithm with MapReduce, ICCT, Pages 764-768, 2013.
[10] Martin Hahmann, Gunnar Schrder and Phillip GrosseData, "Analytics Methods and Techniques",\https : ==wwwdb:inf:tudresden:de=misc=WS1112=FK=01$_d$ata$_a$nalytics:pdf"; 2014:
[11] Keco, Dino, Subasi and Abdulha. "Parallelization of Genetic Algorithms using Hadoop MapReduce. Southeast Europe Journal of Soft Comput-ing." Pages 56-58, 2012.
[12] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, "Scalable Sentiment Classification of Big Data Analysis Using Naive Bayes Classifier", IEEE International Conference on Big Data, Pages 99-103, 2013.
[13] Xin Yue Yang, Zhen Liu and Yan Fu," MapReduce as a Programming Model for Association Rules Algorithm on Hadoop", 3rd International Conference on Big Data (IEEE Bigg Data) Pages 99-102, 2013.
[14] A Fahad, Alshatri, N.Tari, Z. Alamri and A.Bouras, "A Survey of Clus-tering Algorithms for Big Data: Taxonomy and Empirical Analysis", Emerging Topics in Computing, IEEE Transactions on (Volume: 2 , Issue: 3), Pages 267-279, 2014.
[15] Pun WK Daniel and Ali ABM Shawkat. "Unique Distance Measure Approach for K-means Clustering Algorithm,"TENCON 2007 IEEE Region 10 Conference." Pages 65-68, 2007.
[16] Verma, Abhishek; Xavier Liora; Goldberg, E David; and Roy, H. Cambell, "Scaling Genetic Algorithms using Map Reduce." Ninth In-ternational Conference on Intelligent Systems Design and Applications. Pages 13-18, 2009.
[17] VJ Schaffer and L Eshelman "On crossover as an Evolutionary Viable Stratergy . In R. Belew and L.Booker, editors, Proceedings of the 4th International Conference on Genteics algorithms, pages 61-68. Morgan Kaufmann, 1991.
[18] Ekanayake J, Pallickara S and Fox G. "Mapreduce for Data Intensive Scientific Analyses", Proceedings of 4th IEEE International Conference on eScience, Pages 277-284, 2008.