A Big Data Framework: to Unite Both Organized and Unorganized Data in Economic Outlets

[1] Swaroopa C.K [2]Maruthi G.B [3]Saleem Malik S
[1]P.G Scholar in VLSI and Embedded Systems. [2,3] Assistant Professor.
[1]Sridevi Institute of Technology, Manglore. [2,3]KVGCE, Sullia.

Abstract **— Nowadays large and various types data being produced and processed, web and web 2.0 played a vital role in producing these types of data. This has led to the revolution of big data. Huge amount of data which includes organized and unorganized are produced in economic outlets. Processing these data could help an investor to make an informed investment decision. In this paper, a framework has been developed to incorporate both organized and unorganized data for data optimization. Data Optimization consists of three processes: *Asset selection, Asset weighting and Asset management.* This framework proposes to achieve the first two processes using a 5-stage methodology. The stages include shortlisting stocks using Data Wrapping Analysis (DWA), incorporation of the qualitative factors using text mining, stock clustering, stock ranking and optimizing the data using optimization heuristics. This framework would help the investors to select appropriate assets to make data, invest in them to minimize the risk and maximize the return and monitor their performance.**

Keywords- Data performance. Data Optimization, Big Data, Hadoop,, Economical Outlets, Stock Selection

I. INTRODUCTION

Big data applications running on different machines are generating large volumes of data and these data volumes are only expected to increase in future. Since this massive data is key to scientific discovery, the ability to rapidly store, move, analyze, and visualize data is critical for scientists' productivity. Yet the growth of data volume imposes numerous requirements on I/O performance, which results in I/O bottlenecks in current machines. Furthermore, the enlarged gap between computational power and I/O performance further worsens I/O performance. These two folds combined togetherre-sult in undesirable situations, where I/O bottlenecks devastate the efficiency and scaling of scientific simulations and associated data analyses and visualizations. Scientists are forced to wait a substan- tial portion of the simulation runtime writing data to the storage [1]or simply forgo writing out data in order to keep total I/O within reasonable bounds. These trends will be speeded up in the planning for exascale computing platforms in which the attainable I/O performance is further exacerbated by increased contention on shared resources [3]on those platforms.

To improve I/O performance, in-situ data analytics has emerged as an effective way to substitute for the traditional offline data analytics and overcome the increasingly severe I/O bottleneck for sci-entific applications running at the petascale and beyond. Through processing data before placing it on disk, in-situ analytics can re-duce I/O costs (both in time and in power), extract and deliver valuable insights from live simulation output in time, and gain im-proved end-to-end performance. The utility of processing is demonstrated by its wide use by leading scientific applications, like the Metagenomics [2], the Cosmology Simulation [6,7], and the Maya [12].

Data performance has been proven to be a very effective data analytics method in improving end-to-end performance in the wide-area networking domain, in which the large size data are compressed and transferred with the worst networks bandwidth [4,6,9]..Typical performance include point selection, subsetting, cutting planes, global data sum-maries, and feature extraction. Such performance often return only sub-sets of the original data and can significantly reduce the amounts of data that have to be moved and/or further processed for display. Modern visualization systems exploit early data reduction within their visualization engine to optimize visualization performance [5].

In economic, data is defined as the collection of assets. Assets range from stocks and bonds to real estate. With the seminal work of Markowitz [6], data optimization has been a topic of research. Data optimization is the investment decision-making process to hold a set of economical assets to meet various criteria of the investors. In general terms, the criteria are maximizing return and minimizing risk. In this paper, the scope of the work is limited to stock analysis. Data optimization consists of three major steps: asset selection, asset weighting, and asset management. In this paper, a framework is proposed to integrate unorganized and organized data to make an informed investment decision.

II. RELATED WORK

Data analytics has been widely used in many data-intensive applications to reduce data size. Data optimizaton and performance is useful because it helps minimize storage utilization, re-duce networks bandwidth and energy consumption in hardware. However, most of optimizatons and performance techniquies are not transparently operated at the I/O layer, but rather explicitly worked at the appli-cation level [12,13,9]. In many recent studies [11,12], optimizaton is applied on big data research to reduce transfer latency and improve response time. All of them concentrate on the study on optimizaton al-gorithm and data saving on parallel file system. Our work differs from these optimizaton studies, as it does not focus on the study of optimizaton algorithms. The exploration of optimizaton selec-tion is conducted in the paper. We investigate the

effects of data optimizaton on the whole I/O path rather the file system to find the best location to place the optimizaton. We show that the com-pression improves network end-to-end transmission time, network bandwidth, and consequently the I/O performance with the well-planned placing. The area of data optimizaton in cloud computing has received many attentions in recent literature. A recent study [3]investi-gated the effects of optimizaton in MapReduce clusters. This study focused on increasing I/O performance in order to reduce clus-ter power consumption. Our work focuses on improving I/O per-formance to achieve better end-to-end application performance. Another data optimizaton studies focus on applying the data com-pression into grid computing. The highly parallel-distributed data management in grid has the different requirements on data trans-mission with HEC. Such research is designed for use in Grid com-puting environments. So, the existing data optimizaton methods cannot be applied on HEC directly. There are a number of related efforts that provide solutions to data visualization and analysis in the HPC community [1,13,35]. They all use the data sever/client architecture supporting visualiza-tion by permitting remote users to acquire images via their visual client from visual servers located on large-scale machines placed 'next' to simulation. The problem of how to move data from the simulation to the visualization machine remains to be unsolved. Out work addresses it. In addition, related work on performance-driven visualization [18]considers static data performance rather than ad hoc, dynamic performance enabled by FlexAnalytics. There are also contribu-tions on data indexing, which we have not yet leveraged. Instead, we flexibly place analytics among I/O path to improve both simu-lation and data access performance. Additional input to out work can be derived from out previous work on adaptive optimizaton and FlexPerformance [35,37].

III. PROPOSED SYSTEM

The proposed framework for data optimization can be explained using 5-step process: (*a) Data Wrapping Analysis (DWA), (b) Validation of selected stocks, (c) Stock clustering, (d) Stock ranking, and (e) Optimization*. All listed firms at a particular stock exchange are considered as the initial input to the framework and the output would be a set of stocks that would maximize the return and minimize the risk. The abstract framework for data optimization is shown in figure 1. DEA is used to narrow the sample space of firms by identifying the efficient firms. In order to validate these firms as potential candidates for data optimization, the latest information about the company is retrieved and processed from online news articles and tweets using text mining to the sentiments about the company in current context. The validated efficient firms are clustered into different groups to aid the diversification of data. This is further followed by ranking of the stocks within each cluster and followed by asset weighting using optimization algorithms

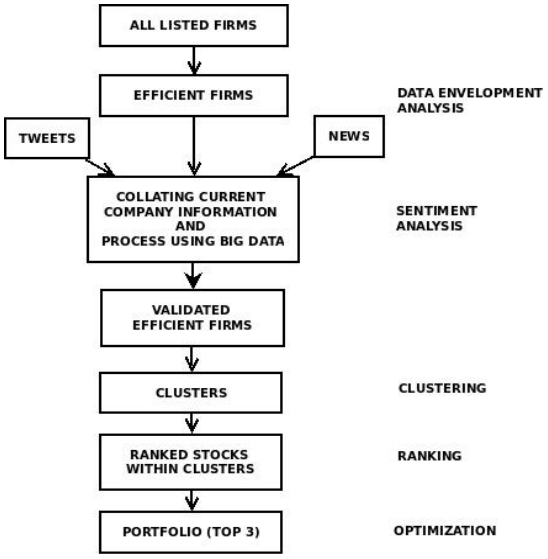

Fig 1:Proposed Framework

IV. STAGES OF PROPOSED FRAMEWORK

A)Data Wrapping Analysis (DWA) is a non-parametric linear programming that calculates the efficiency score of a Decision-Making Unit (DMU) based on a given set of inputs and outputs.The DMUs with score 1 are considered to be efficient [7]. Apart from its applicability in the discipline of manufacturing, DEA can be used for stock selection. In this case, stocks form the DMU. Based on the previous studies [8, 9], four input parameters, namely, *total assets, total equity, cost of sales* and *operating expenses* and two output parameters, namely, *net sales and net income*, can be considered. These data can be obtained from standard databases like *Bloomberg*. The stocks with efficiency score 1 are considered for next stage.

B)Hadoop Framework for Sentiment Analysis: This stage involves processing of frequently generated unorganized data using Hadoop MapReduce. This step complements the previous stage. Events like election, change of management and announcement of dividends have an effect on the market sentiment, which is not captured using the quantitative analysis. As first step, online news articles and tweets of the efficient firms are retrieved. Tweets can be obtained through Twitter API but it is limited to 1500 tweets. The ease-of-use, scalability, and failover properties make Hadoop MapReduce a popular choice for processing big data efficiently [10]. Tweets and news articles are processed using text mining to obtain the positive and negative sentiments about the firm. Hadoop MapRaduce infrastructure quickens the distributed text mining process.

C)Stock Clustering: In this stage, the correlation coefficients of the returns of the stocks are calculated. The stocks are assigned to different clusters based on these correlation coefficients. The greater the number of clusters more is the diversification. The objective for number of clusters and quality of clusters is to maximize similarity within the cluster and to minimize the similarity between the clusters. Various clustering algorithms like k-means clustering or Louvian clustering [11] can be used. This process reduces

the data risk through diversification of stocks [12]. These resulting clusters can consist of firms with similar business activity or size in real sense.

D) Stock Ranking: The appropriate stocks from each cluster should be selected. The stocks in each cluster can be ranked using Artificial Neural Network (ANN) [13]. Till the previous stage, only the internal factors of the firms were considered. At this stage, external factors like Gross Domestic Product (GDP) growth rate and interest rate can be considered [12]. ANN is a model for information processing that consists of three layers: *input, hidden and output layer* respectively. The inputs for ANN can be GDP growth rate and interest rate and the outputs can be future return on investments. This results in ranking the stocks within a cluster

E)Optimization: Previous stage leads to an assumption that the investor might choose the top stocks from each cluster. But the question that still remains is: How much to invest in each stock? Previous study [12] considered simple (equal) stock weighting method, a primitive method, to allocate the resources among the stocks. Hence the ranked stocks should be optimized to maximize returns and to minimize risk. Markowitz's mean-variance model can be used at this stage [6]. Various optimization heuristics like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and Genetic Algorithm (GA) can be used. The distribution of the stocks in a data will be formed at the end of this stage. Top 3 performing data will be suggested to the investor.

V. DESIGN AND IMPLEMENTATION

We profile end-to-end latency with 222MB float data transfer-ring between two nodes with three data optimization algorithms – lossy [11], bzip2 [1], and gzip [4]. The profiling results can help us to evaluate the selected optimization algorithm. Three optimization algorithms emphasize the different requirements on the compression speed and optimization ratio. Bzip2 and gzip are lossless data optimization. Bzip2 is an implementation of the Burrows–Wheeler algorithm, which is more effective on data optimization ratio. Gzip is based on the DEFLATE algorithm, which provides better com-pression speed and worse optimization ratio than bzip2 [2]. We check the performance of lossy algorithm described in [11]. The lossy optimization is better than lossless on speed and ratio. How-ever, information loss constraints its wide use.

The testing nodes are equipped with 2.13GHz Intel Xeon Pro-cessor with 4MB L3 cache and 12GB memory. The nodes con-nected with 16GB infiniband networks. Table3lists the profiling information of three optimization algorithms: lossy, bzip2, and gzip. Lossy optimization will be introduced in experimental sec-tion.To evaluate the behavior of optimization with real-world use cases, we measure end-to-end latency on varying transfer band-width. The optimization and deoptimization have been exclusively executed on one single processor..For lossy com-pression, the (de)optimization accounts for over 95% of end-to-end latency. If we compress the data in parallel, there are at least 4 processors for lossy optimization, 63 processors for gzip, and 74 processors for bzip2 to catch up the data transfer with 1.034Gb/s bandwidth. When the transfer bandwidth reduces to 0.258Gb/s, the proportion of (de)optimization in latency starts to decrease. The latency of data optimization and transfer with gzip is bet-ter than non-optimization after the available bandwidth less than 0.016Gb/s. The measured results verify our quantitative model. In our case, all three optimizations are not effective method to opti-mize data transfer when the bandwidth is larger than 264.19Mb/s. It is similar with optimization analytics. We profile end-to-end latency with 222MB float data transferring between two nodes with three different visualization performance – statistics, similarity, and combination. Fig 2. Shows the implementation of proposed framework

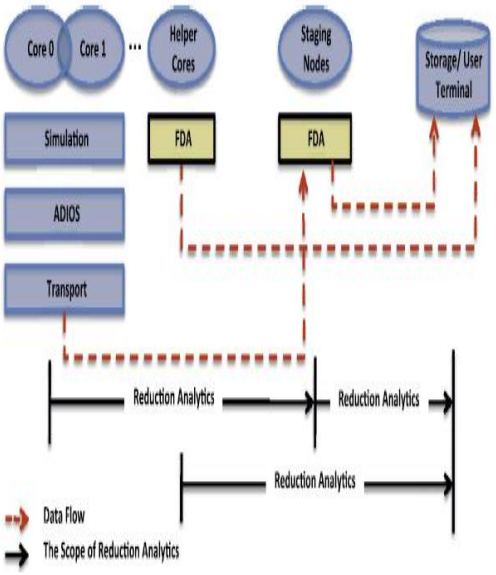

Fig 2: Implementation of Proposed Framework

Three performance have different workload for comput-ing resource requirement. Statistic performance(ang_vel) is a lightweight performance, which requests the minimal value of angle velocity in the data set. Because the return results are very small data set, the statistic performance can significantly increase the transmission band-width online. We inject controlled additional network traffic be-tween two nodes variability, resulting in available bandwidth vary-ing from 20MB/s to 120MB/s. These two performance stand for heavy load and joint performance, which request much more computing resource for data filtering. So, such performance are good for data reduction with high available transfer bandwidth. 20MB/s available bandwidth is a special case in the profiling. Although similarity and combination performance are heavy load op-erations, the low available bandwidth still needs data analytics to improve I/O performance. Because the data performance is the operation which is explicitly specified by user terminal, the performance cannot be applied on the output data in general cases.

VI. EXPECTED OUTCOMES

Apart from letting the investors make an informed investment decision, this framework would be useful to the

naive investors as well. The expected outcome of the framework is to generate data in conformance with the criteria of the investors: *minimize the risk* and *maximize the return*. The first stage of DWA would result in a group of the potential stocks for data optimization. The current information related to those resulting companies is analyzed using text mining to obtain the sentiment about the company. Sentiment analysis gives the qualitative aspect of the firms. The stocks resulting from the previous stage are grouped into different categories using the correlation co-efficient of the returns. This leads to the diversification of the stocks. The stocks are ranked to select appropriate stocks from each group. At the end of the entire process, top three data suggestions would be provided to the investors so that they can select one of those three.

VII. CONCLUSION

The proposed framework tries to integrate both organized data from database (stock price, balance sheet data etc) and unorganized data from online news articles and tweets. Consideration of qualitative factors (Management of firms, etc.) along with quantitative factors (economical ratios) provides better alternatives for formation of data. The assets are well-diversified using k-means clustering and ANN. Top three data suggestions obtained using optimization heuristics gives flexibility to the investors to choose the appropriate assets suiting their risk profile. The proposed model can be applied to any stock market data. The utility of the framework is limited only to stock analysis and investments. The parameters suggested for DWA is based on the previous studies. Instead of variance as risk measure, other risk measures like Value-at-risk and downside risk measures can be used in the final stage of the framework.

REFERENCES

[1] D.E. O'Leary. Artificial Intelligence and Big Data. IEEE Computer Society, 96-99, 2013.

[2] S. Chaudhuri.How Different id Big Data? IEEE 28th International Conference on Data Engineering, 5, 2012.

[3] H. Topi. Where is Big Data in Your Information Systems Curriculum? acmInroads, Vol. 4. No.1, 12-13, 2013.

[4] IBM, Big Data at the Speed of Business, What is big data, Online available from http://www 01.ibm.com/software/data/bigdata/

[5] S. Alsubaiee, Y. Altowim, H. Altwaijry, A. Behm, V. Borkar, Y. Bu, M. Carey, R. Grover, Z. Heilbron, Y.-S. Kim, C. Li, N. Onose, P. Pirzadeh, R. Vernica, J. Wen. ASTERIX: An Open Source System for "Big Data" Management and Analysis (Demo). Proceedings of the VLDB Endowment, Vol 5, No. 12, 1898-1901, 2012.

[6] C. Okoli, K. Schabram. A Guide to Conducting a Systematic Literature Review of Information Systems Research. Sprouts: Working Papers on Information Systems, 2010.

[7] B. Kitchenham, S. Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report EBSE-2007-01, 2007.

[8]Merriam-Webster.Online available from http://www.merriamwebster.com/dictionary/definition

[9] H. Suonuuti. Guide to Terminology, 2nd edition ed. Tekniikan sanastokeskus ry, Helsinki, 2001.

[10] G. Jung, N. Gnanasambandam, T. Mukherjee. Synchronous Parallel Processing of Big-Data analytics Services to Optimize Performance in Federated Clouds. IEEE 5th International Conference on Cloud Computing (CLOUD), 811-818, 2012.

[11] X. Qin, H. Wang, F. Li, B. Zhou, Y. Cao, C. Li, H. Chen, X. Zhou, X. Du,, S. Wang. Beyond Simple Integration of RDBMS and MapReduce -- Paving the Way toward a Unified System for Big Data analytics: Vision and Progress. Second International Conference on Cloud and Green Computing (CGC), 716-725, 2012.

[12] D. Zeng, R. Lusch. Big Data Analytics: Perspective Shifting from Transactions to Ecosystems. Intelligent Systems, IEEE, Volume 28, Issue 2, 2-5, 2013.

[13] A. Aboulnaga, S. Babu. Workload management for Big Data analytics. IEEE 29th International Conference on Data Engineering (ICDE), 1249, 2013.