International Journal of Information Technology and Management
Vol. 12, Issue No. 1, February-2017, ISSN 2249-4510

An Analysis upon Various Modeling and Evolutionary Optimization of Big Data Analytics

Nitin Kumar Saran*

Research Scholar, SSSUTMS University, Madhya Pradesh

Abstract – Recent advances in technology and the web, as well as and the decreasing cost of data storage, have motivated all kinds of organizations to capture, store and analyze vast amounts of data. The term "Big Data" is often used to describe this phenomenon, meaning information that cannot be processed or analyzed using traditional processes or tools, such as relational databases and data warehouses. We are living in the era of Big Data. Today a vast amount of data is generating everywhere due to advances in the Internet and communication technologies and the interests of people using smartphones, social media, Internet of Things, sensor devices, online services and many more. Similarly, in improvements in data applications and wide distribution of software, several government and commercial organizations such as financial institutions, healthcare organization, education and research department, energy sectors, retail sectors, life sciences and environmental departments are all producing a large amount of data every day.

Keywords: Evolutionary, Optimization, Big Data, Analytics, etc.

- - - - - - - - - - - - - - X - - - - - - - - - - - - - -

INTRODUCTION

The derivation of Big Data is vague and there are a lot of definitions on Big Data. For examples, Matt Aslett defined Big Data as "Big Data is now almost universally understood to refer to the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to limitation of traditional data management technologies". Recently, the term of Big Data has received a remarkable momentum from governments, industry and research communities. In, Big Data is defined as a term that encompasses the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. The term Big Data is basically characterized with 3 Vs:

- Volume, the sheer amount of data generated (i.e. from terabytes to zettabytes),
- Velocity, the rate the data is being generated (i.e. from batch data to streaming data), and
- Variety, the heterogeneity of data sources (i.e. from structured data to unstructured data).

There are several factors that are involved in producing Big Data. One factor is the Internet and communication technology as it has been advanced to enable people and devices to be increasingly interconnected not only some time but all the time. Small integrated circuits are now so economical that people are using in almost every object to make them intelligent which is another reason of generating of mountains of data. The continuous reduction in the prices of storage devices is also a factor for Big Data.

REVIEW OF LITERATURE:

Shadi Ibrahim et.al.(2008) Project says presence of partitioning skew1 causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes In this study, author develop a novel algorithm named LEEN for locality aware and fairness-aware key partitioning in Map Reduce. LEEN embraces an asynchronous map and reduce scheme. Author has integrated LEEN into Hadoop. His experiments demonstrate that LEEN can efficiently achieve higher locality and reduce the amount of shuffled data. More importantly, LEEN guarantees fair distribution of the reduce inputs. As a result, LEEN achieves a performance improvement of up to 45% on different workloads. To tackle all this he presents a present a technique for Handling Partitioning Skew in Map Reduce using LEEN.

H.Herodotou et al. (2009) provides a technique to implement self-tuning in Big Data Analytic systems. Hadoop's performance out of the box leaves much to be desired, leading to suboptimal use of resource,

time and money. This study introduces Starfish, a self-tuning system for big data analytics.

Russom, P. (2011) elaborate, Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage.

According to Kubick, W.R (2012), The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time.

Bakshi, K. (2012) define, The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse. Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions.

OBJECTIVES

The objectives are include –

- To optimize the existing discount simulation algorithm in order to reduce its running time
- To create a model with associated algorithms for a scaling extension of the system's simulation functionality.
- To research methods and techniques, as well as to build tools for easy and efficient processing of very large data sets.

CONCLUSION:

Big data pose new computational challenges including very high dimensionality and sparseness of data. Evolutionary algorithms' superior exploration skills should make them promising candidates for handling optimization problems involving big data. High dimensional problems introduce added complexity to the search space. However, EAs need to be enhanced to ensure that majority of the potential winner solutions gets the chance to survive and mature. In this paper we present an evolutionary algorithm with enhanced ability to deal with the problems of high dimensionality and sparseness of data. In addition to an informed exploration of the solution space, this technique balances exploration and exploitation using a hierarchical multi-population approach. The proposed model uses informed genetic operators to introduce diversity by expanding the scope of search process at the expense of redundant less promising members of the population. Next phase of the algorithm attempts to deal with the problem of high dimensionality by ensuring broader and more exhaustive search and preventing premature death of potential solutions. The algorithm has also been successfully applied to a real world problem of financial portfolio management. Although the proposed method cannot be considered big data-ready, it is certainly a move in the right direction.

REFERENCES:

Bakshi, K. (2012). Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7.

H. Herodotou, H.Lim, G. Luo, N. Borisov, L. Dong, F. B. Cetin and S. Babu. Starfish (2009). A Selftuning System for Big Data Analytics. In CIDR, pages pp. 261–272.

Kubick, W.R. (2012). Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28.

Russom, P. (2011). Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40.

Shadi Ibrahim⋆ _ Hai Jin _ Lu Lu (2008). "Handling Partitioning Skew in MapReduce using LEEN" ACM 51, pp. 107–113

Corresponding Author

Nitin Kumar Saran*

Research Scholar, SSSUTMS University, Madhya Pradesh

E-Mail – chintuman2004@gmail.com