IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Resource Optimization for Big Data processing using MapReduce Framework: Survey

Parul Chhabra[1], Surender Singh[2]

Dept of Computer Science and Engineering, Om Institute of Technology & Management, Hisar, India[1, 2]

Abstract: Big Data is a new area for researchers and it supports various application domains i.e. Medical, Finance, Sales & Marketing, Education etc. Big Data deals with the processing of large data volume. There are different challenges which are associated with the data processing i.e. storage of large scale data, memory management, energy consumption, job scheduling, process management and security etc. There is need to explore each issue and its possible solutions. Researchers have already developed some solution for Big Data processing and in this paper, we will explore them.

Keywords: Hadoop, Big Data, Energy Consumption, HDFS, MapReduce.

I. INTRODUCTION

Big Data deals with the large volume of data which may exist in various forms i.e. Text, Image and Multimedia etc. Applications are generating the large volume of data in a small time interval and it is quite complex and time consuming to analyze this data for decision making. Traditional data processing applications are not suitable for handling this data, so researchers introduced the concept of Big Data and developed some frameworks to process the large scale data efficiently by considering the common issues which are following:

- How much data must be selected from a large scale data block?
- Storage Space is essential or not
- Security
- Data Validation

BIG DATA PROCESSING APPLICATIONS

Apache Hadoop

It supports operations in distributed environment. It consists of various modules:

1. Hadoop Common – shared libraries ;
2. Hadoop Distributed File System (HDFS) – a distributed file-system to store large volume data.
3. Hadoop YARN – responsible for Resource management, process scheduling and user application management.
4. Hadoop MapReduce –Programming interface to process large scale data.

Hadoop Merits

It supports:
Scalability for data processing
1. Low cost data analysis
2. Robust Fault Tolerance

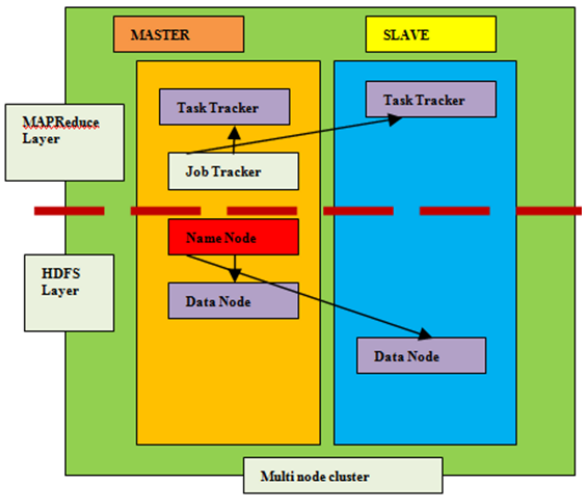

Figure: HADOOP Architecture

- **Data Node:** stores data in HDFS
- **Name Node:** It keeps the track of files stored at HDFS
- **Job Tracker:** It keeps the track of process of MapReduce, in a given cluster
- **Task Tracker:** in a given, it handles the jobs related to Map, Reduce and Shuffle operations by interacting with Job Tracker.

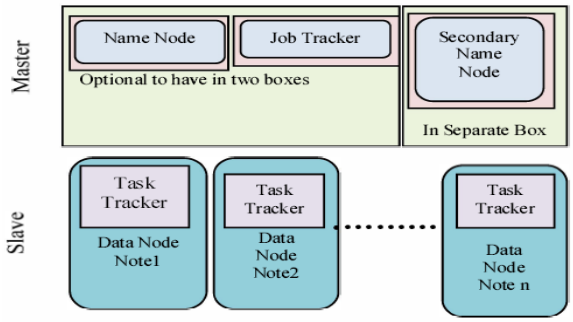

Figure: Hadoop Components

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Hadoop Distributed File System

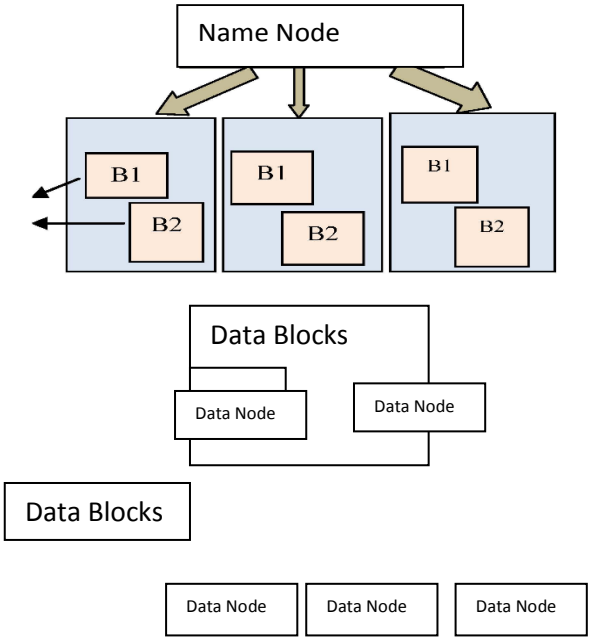

Figure: Data Blocks for HDFS

HDFS is a storage module which deals with Name Node (NN) and Secondary Name Node (SNN) which keep the data track for Data Node (DN). DN performs I/O operations and interacts with NN.

MapReduce system

MapReduce

It is a framework which introduces programming language support for parallel data processing. It can be used to resolve performance issues related to load balancing, error/fault tolerance and Data transmission etc. It supports following file systems for data storage:

- Google File System (GFS)
- Hadoop Distributed File System (HDFS)

Key value pairs are used for data processing which operates in independent environment. MAP processes the key value pair and its output is forwarded to Reducer to generate final output. [16][17]

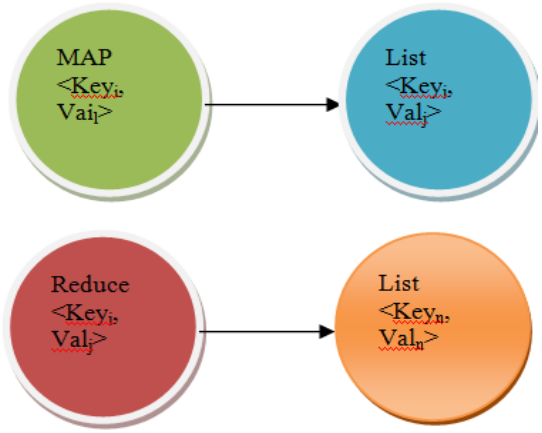

Figure: MAP Reduce Key pair

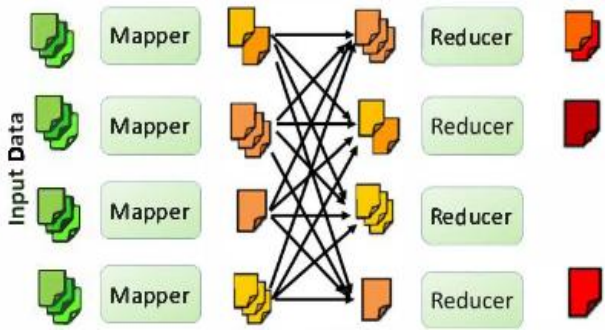

Figure: MAP Reduce Data Processing [17]

II. LITERATURE SURVEY

Lin Wang et al. [2] tried to enhance performance of MapReduce applications using VM assignment and traffic engineering. They focused on the energy consumption by exploring the attributes of MapReduce applications and developed a framework that uses virtual machines for assignments. Local search based on energy-efficient routing problem is resoled using greedy heuristic called GEERA. Simulations results show that combination of the virtual machine assignment and the traffic engineering can enhance energy efficiency of network in MapReduce systems. Authors claim that proposed scheme can optimize energy at large scale as compared with the traffic engineering based solutions.

Lena Mashayekhy et al. [4] focused on the energy consumption by MapReduce scheduling methods and proposed heuristic based methods, called energy-aware MapReduce scheduling algorithms which can conserve the power during application execution phase. They used Hadoop cluster for analysis execution time of different workloads with various benchmarks applications i.e. Tera-Sort, Page Rank, and K-means clustering etc.Results show that proposed methods can ensure the optimal job schedules and can reduce the energy consumption upto 40% as compared to ordinary job scheduling jobs. Proposed work can be extended to provide the energy conservation support for distributed schedulers for multiple MapReduce jobs.

Eunhyeok Park et al. [5] investigated the memory optimization for big data processing to reduce the energy consumption .They developed a function called memory fast-forward (MFF) which can process the graph computations with optimal memory requests. Simulation results show that MFF unit can reduce 54.6% energy consumption due to low memory traffics. Proposed work can be extended to support large scale systems i.e. multi-GPU.

SungYe Kim et al. [6] extended MapReduce method for iterative clustering schemes. The proposed methods works in Intel integrated GPU (having multi-node cluster environment.). HiBench benchmark suite was used for simulation purpose and results show that integrated GPU

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

performs well in terms of energy consumption as compared to ordinary CPU based methods. They also explored the correlation between number of input/output operations and energy optimization. Authors claim that proposed scheme is a generalized method and can be used for other applications with different system configurations. Andrea Acquaviva et al. [7] developed a method which can analyze and construct the information regarding energy consumption of each user, called ESA which can produce the energy consumption signatures in distributing and scalable manner to predict the energy consumption level over a certain time period. It can compare latest and previous conducted energy consumptions having same conditions. Authors did a real time analysis to show its performance in terms of accurate decision making about power consumption. Proposed method can be extended to work with the social platform.

Thi Thao Nguyen Ho et al. [8] proposed a framework to enhance the memory consumption for data intensive applications by obtaining the data values from data stream of the applications. They focused on the processing of the data as per their sensitivity, in cloud environment. Proposed scheme can be extended for data value characterization algorithm.

Dapeng Dong et al. [9] developed a Content-aware, Partial Compression (CaPC) for text using a dictionary-based method which can replaced the original text with specific symbols for compression purpose. They used a set of real-world datasets and several classical MapReducejobs on Hadoop. Analysis results show its performance interms of size reduction upto30% and performance enhancement of I/O jobs upto 32% but it is only suitable for large scale datasets i.e. social media, web pages, and serverlogs etc.
Lena Mashayekhy et al. [10] explored energy-aware scheduling of MapReduce jobs and used a greedy method, called Energy-aware MapReduce Scheduling Algorithm (EMRSA) which identifies the map assignments for scheduling w.r.t. energy conservation and executes the mapreduce applications in real time environment. For simulation analysis, they used large Hadoop cluster to find out the energy consumption of various MapReduce benchmark applications and results show its performance in terms power conservation as compared to ordinary make span minimization algorithms. Proposed scheme can be extended to provide the support for multiple MapReduce jobs.

Karthi Duraisamy et al. [11] have presented an energy efficient multicore architecture with the support of wireless infrastructure MapReduce applications. Use of Wireless interconnections optimize the power consumptions and also enhance the performance by minimizing time penalties. Analysis results show that proposed scheme can achieve an average of 33.7% energy-delay product (EDP) savings as compared to standard baseline non-VFI mesh-based system and execution time penalty does not exceed from 3.22%.

Eugen Feller et al. [12] analyzed Hadoop performance using traditional model of collocated data and compute services. Data Separation and compute services provides more stiffness in environments where data locality might not have a considerable impact such as virtualized environments and clusters with advanced networks. They also did analysis of energy efficiency of Hadoop on physical and virtual clusters using various specifications. Analysis results show that performance on physical clusters is significantly better as compared to virtual clusters. Performance is degraded due to separation of services depends on data to compute ratio. Application completion progress correlates with energy consumption which is application specific.

III. PROBLEM FORMULATION

Huge data collection introduced the concept of Big data that consists of various data types i.e. Text, Images, Multimedia etc. Traditional data processing applications cannot process it at large scale so there are lot of following issues related to Big Data analysis:

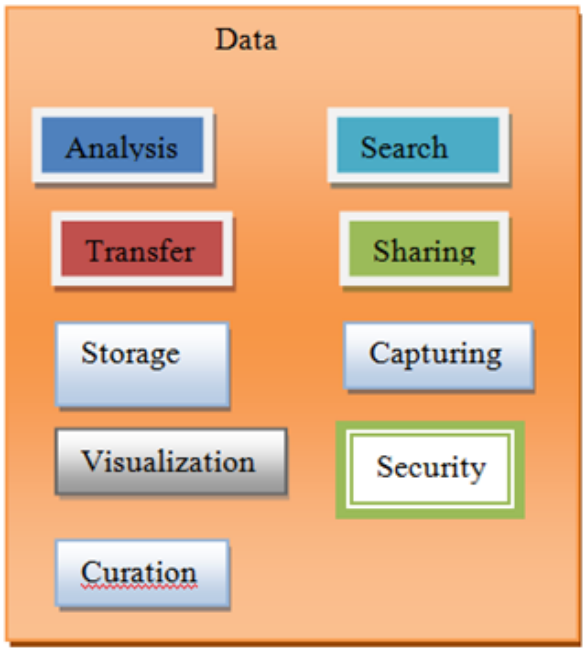

Figure: Data Analysis

Big data analysis is becoming an essential part of decision making processes in business, healthcare, public service, and security sectors. Due to the constant increase of data volume, analytic platforms, such as Hadoop and Spark, are under constant pressure, working at the limits of computation power, network bandwidth, and storage capacity. As per the growing demand for big data analytics, development of energy efficient and high performance MapReduce platform is essential. There is need to explore the energy consumption ratio of MapReduce Applications and to find out an optimal method for energy consumption during execution of the MapReduce Applications.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

IV. CONCLUSION

In this survey paper, energy efficient solutions for Big Data applications were explored. It can be observed that energy can be optimized using various methods i.e. MapReduce scheduling algorithms are commonly used to conserve the energy during job execution, Hadoop clusters may distribute the jobs to multiple schedulers, Memory optimization can also reduce the energy consumption, Optimize Text compression for large scale data may reduce the resource requirements, virtual clusters can execute the processes in a isolated environment and each process consumes small amount of energy.

Finally it can be concluded that energy consumption depends upon the process execution and scheduling. It raises the need to explore other components also which requires energy conservation. This study can be utilized to develop an energy efficient solution for Big Data processing.

REFERENCES

[1]. https://en.wikipedia.org/wiki/Big_data
[2]. Lin Wang, Fa Zhang, Zhiyong Liu, "Improving the Network Energy Efficiency in MapReduce Systems", ICCCN, IEEE-2013, pp.1-7
[3]. https://hadoop.apache.org/
[4]. Lena Mashayekhy,MahyarMovahedNejad, Daniel Grosu, Quan Zhang, WeisongShi,"Energy-Aware Scheduling of MapReduce Jobs for Big Data Applications", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26 (10), IEEE-2015, pp.2720-2733
[5]. Eunhyeok Park, JunwhanAhn, Sungpack Hong, SungjooYoo, and SungguLee,"Memory Fast-Forward: A Low Cost Special Function Unit to Enhance Energy Efficiency in GPU for Big Data Processing", Design, Automation & Test in Europe Conference & Exhibition, IEEE-2015, pp.1341-1346
[6]. SungYe Kim, Jeremy Bottleson, JingyiJin, PreetiBindu, Snehal C. Sakhare, Joseph S. Spisak, "Power Efficient MapReduce Workload Acceleration using Integrated-GPU", Big Data Computing Service and Applications, IEEE-2015, pp.162-169
[7]. Andrea Acquaviva, Daniele Apiletti, Antonio Attanasio, Elena Baralis, Lorenzo Bottaccioli, "Energy signature analysis: knowledge at your fingertips", IEEE International Congress on Big Data, IEEE-2015, pp.543-550
[8]. Thi Thao Nguyen Ho, Barbara Pernici, "A Data-Value-Driven Adaptation Framework for Energy Efficiency for Data Intensive Applications in Clouds", IEEE Conference on Technologies for Sustainability (SusTech), IEEE-2015,pp.47-52
[9]. Dapeng Dong, John Herbert, "Content-aware Partial Compression for Big Textual Data Analysis Acceleration", International Conference on Cloud Computing Technology and Science, IEEE-2014, pp. 320-325
[10]. Lena Mashayekhy, Mahyar Movahed Nejad, Daniel Grosu, Dajun Lu, Weisong Shi, "Energy-aware Scheduling of MapReduce Jobs", IEEE-2014, pp.32-39
[11]. Karthi Duraisamy, Ryan Gary Kim, Wonje Choi, Guangshuo Liu, Partha Pratim Pande, Radu Marculescu, Diana Marculescu, "Energy Efficient MapReduce with VFI-enabled Multicore Platforms", Design Automation Conference (DAC), ACM/EDAC/IEEE-2015, pp.1-6
[12]. Eugen Feller, Lavanya Ramakrishnan, Christine Morin, "On the Performance and Energy Efficiency of Hadoop Deployment Models", IEEE-2013, pp.131-136
[13]. https://en.wikipedia.org/wiki/Apache_Hadoop
[14]. http://doctuts.readthedocs.io/en/latest/hadoop.html
[15]. Sindhu P Menon, Nagaratna P Hegde, "A Survey of Tools and Applications in Big Data",IJSC0, IEEE-2015, pp.1-7
[16]. Jiamin Lu, Jun Feng, "A SURVEY OF PARALLEL PROCESSING TECHNOLOGIES WITH MAPREDUCE", CCT-IEEE-2014, pp.146-155
[17]. Gore Sumit Sureshrao , Ambulgekar H. P. , "MapReduce-Based Warehouse Systems: A Survey ", ICAETR -IEEE-2014, pp.1-8