International Journal of Engineering Research and General Science Volume 4, Issue 3, May-June, 2016
ISSN 2091-2730

An Alternative Approach to Big Data Computation

Prof. Parul Wadhwa

Asstt. Prof. (ECE) NHCE,Bangalore

parulwadhwa5@gmail.com

Abstract— The advancements in the field of Information Technology and Sensor Devices have significantly increased the amount of data being captured, stored and processed every day. Moreover, social networking and faster internet access from mobile devices have further accelerated the speed with which the data is captured and transmitted. The computation of this vast amount of complex data is currently one of the biggest challenges faced by the IT industry.

Heterogeneous Computing and In-Memory Databases have emerged as two famous high performance computing (HPC) solution for big data computation. This paper discusses both In-Memory databases and heterogeneous computing and proposes a solution by combining the best features of the same.

Keywords— Big data, big data analytics, big data computation, cloud computing , in-Memory databases, distributed computing, parallel processing .

INTRODUCTION

The amount of data [2] in our world has been increasing day by day and analyzing this huge amount of data is bringing forth new challenges apart from providing new arenas of productivity growth, innovation, and consumer surplus. Besides the data-oriented workers, business personnel in every sector will have to deal with the implications of big data. The rapid increase in the volume and detail of information being captured by enterprises, the rise of multimedia, social Medias and vast exploration of things will lead to exponential growth in data for the coming future.

WHAT IS BIG DATA?

Big data [1] refers to a large collection of data whose size and volume renders it difficult to be processed using traditional data processing systems. The various problems encountered with big data are its storage, searching, analysis, processing as well as its sharing and transfer. Big data is formed from the analysis of a single large data set which leads to even bigger data set as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to establish business trends. According to [2] companies gather a huge amount of data about their customers, operations and suppliers. It is becoming evident that besides from other factors, data is a quintessential part for growth and development of an enterprise.

THE NEED FOR BIG DATA COMPUTATION

Big data [1] refers to a large collection of complex data whose size and volume renders it difficult to be processed using traditional data processing systems. The various problems encountered with this vast amount of data are its storage, analysis, replication, processing and transfer. Big data is formed from the analysis of a single large data set which leads to even bigger data set allowing correlations to establish leading business trends.

Today, companies gather and analyse this huge amount of data about their customers, operations and suppliers by means of ERP or CRM systems. These information systems enable companies to understand customer demand, mitigate market risk and enhance its business operations [2]. Therefore, it is evident that processing data into valuable information has become an integral and inseparable part of every business enterprise and industry for growth and development in today's globally competitive world. Following are benefits of big data and its computation [3].

- Big Data computation makes information observable at a much higher rate.

- With business intelligence, big data analytics can substantially improve decision making process of a company as it can provide more accurate forecasting information and spotting trends.
- Big data enables focus strategy by aiming 'narrow segment' of customer with precisely built products and services.
- Big data can be used to improve the next generation of products and services.

CHALLENGES IN BIG DATA COMPUTATION

There has been a significant increase in computing power and storage capabilities of IT systems. However, the need to store and process big data has outpaced the computing capabilities of existing IT systems. Furthermore, big data sizes are rapidly and constantly increasing over the past few years, ranging from a few dozen terabytes to many petabytes of data in a single data set. It is expected that by 2020, there would be around 50 billion networking devices on the internet (such as smart phones, Laptops, RFIDs, Intelligent Appliances etc) [4]. Therefore, computation of big data becomes a real challenge for IT industry today.

SOLUTIONS

The possible solution in order to address the big data challenge can be the combination of In-Memory databases and computation of big data on cloud. These have been discussed in details as below:

1. IN-MEMORY DATABASES (IMDB)

In-Memory Database is one of the most efficient ways to address the big data computation requirements [10]. It is different from traditional database in a way that it stores data directly into computer's memory. Therefore, when data is needed, it is already available in memory and can be accessed quickly. Regular falling price of random access memory (RAM) has made it possible for big companies to now afford In-Memory Databases with very large RAMs. In-Memory database platforms such as SAP HANA, Oracle's TimesTen can support up to 100 TB of uncompressed data [9].

In case of In-Memory Databases, the execution algorithms are simpler and execute fewer instructions as compared to the old disk storage approach. In addition, seek time also gets eliminated. Hence, in-memory databases provide better performance, which is around 100 times faster in comparison to traditional databases [10]. By using In-Memory Databases, real time applications, such as telecommunications network equipments or location based mobile advertising networks, can get many benefits. Moreover, In-Memory Databases have gained a lot of significance in the data analytics field.

2. HETEROGENEOUS CLOUD COMPUTING

Cloud computing or distributed computing is another way to compute big data. In cloud computing [7] a large number of inter connected computers, share the load and run the same program simultaneously (parallel processing). Therefore, with high computing power, which is generated from interconnected computers mesh, it becomes possible to manage vast amount of both structured and unstructured data [8]. Moreover, mesh structure of cloud computing provides scalability, which makes it suitable to handle increasing rate of data generation. Depending upon the budget, infrastructure and software requirements cloud computing provides three types of service models: Software as a service (SaaS), Infrastructure as a service (Iaas) and Platform as a service (PaaS)[11].

To achieve scalability as well as performance, big data computation needs to efficiently utilize both parallel processing and algorithm [8]. The current popular big data processing models are General Purpose GPU and Map Reduce Algorithm.

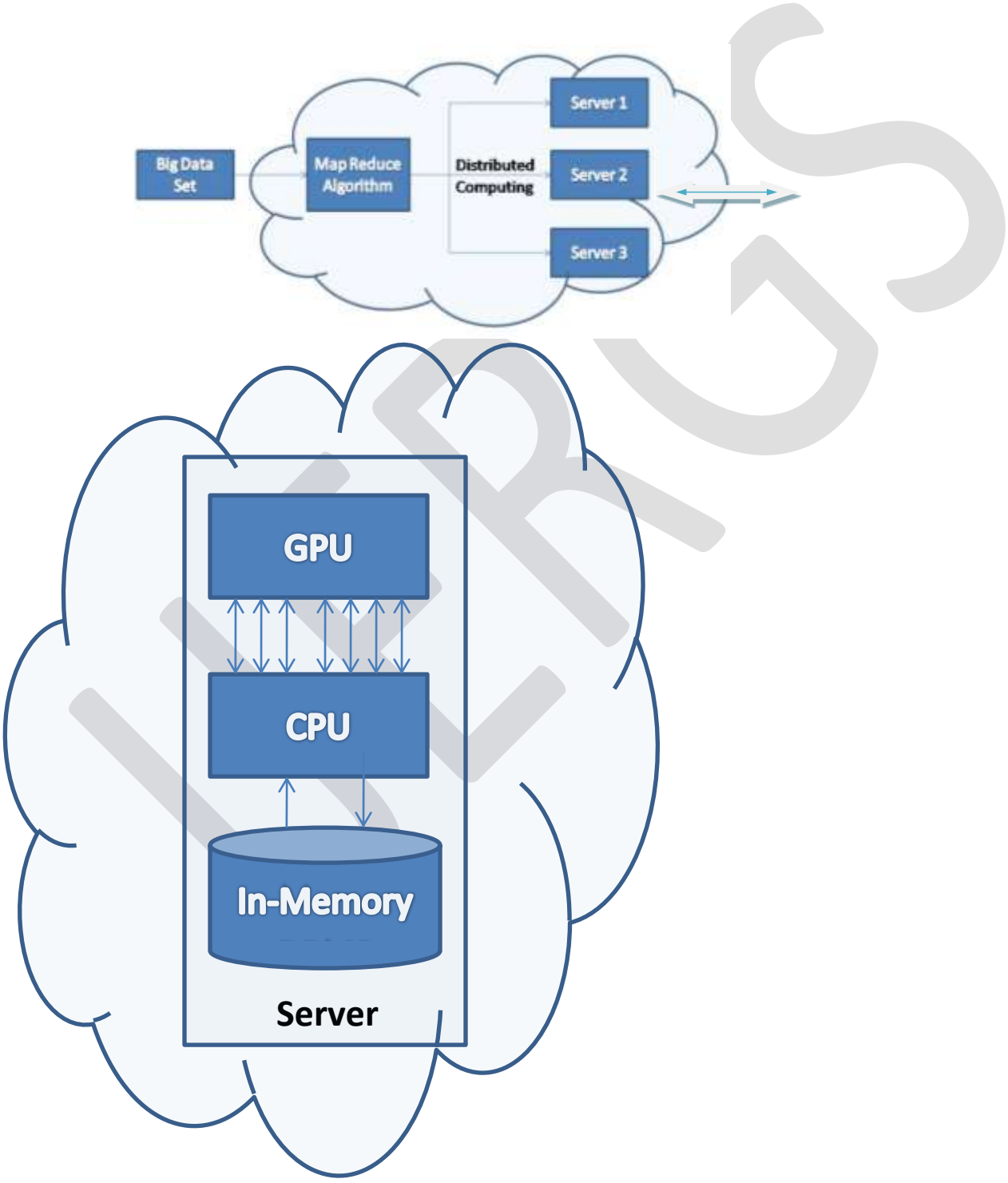
2.1 GPU Computing: GPU computing is a very powerful combination of CPU with application intensive Graphics Processing Unit, to accelerate general purpose engineering computations. A CPU consists of multi cores which are optimized for serial processing whereas a GPU consists of thousands of more efficient small cores which are designed for parallel processing [5]. The main code runs on CPU with computer intensive load being offloaded to more dedicated GPU.

2.2 Map Reduce Algorithm: Heterogeneous cloud computing combines computing solutions offers by different vendors, varying in their computing powers and even algorithms, to achieve a more effective big data computation. However for successful implementation of heterogeneous computing different vendor's need to use a common standard. MapReduce algorithm, proposed by Google, is one of the most efficient and widely accepted programming model for big data computation. It not only optimizes the

processing algorithm, but also hides the complex details related to data distribution, storage and load balancing [12]. Therefore, industry wide programmers are successfully able to use this algorithm for big data computation.

Also, MapReduce algorithm does not use any index or schema, hence its integration with DBMS to retain for performance is one of the major challenge. However, many leading industry vendors are able to overcome this challenge. For example, Apache HadoopDB combines scalability feature of Map reduce algorithm with performance of DBMS to provide efficient hybrid solution [8].

CONCLUSION



As shown in the figure, a combination of GPU based parallel processing, map reduce algorithm and In-Memory databases can be a possible solution for big data computation. This will combine the scalability, performance and cost effectiveness of heterogeneous cloud computing with extremely fast processing speed of In-Memory Databases. However, integration of these separate systems might be a challenge and therefore top players of the industry (such as Intel, SAP and Apache Software Foundation) might collaborate to collectively develop big data computation solutions.

In the era of big data, the exposure and problem solving approach that I have gained through this research paper will serve as foundation to face big data challenges in upcoming IT industry.

REFERENCES:

[2] "*Hadoop: The definitive guide*" by Tom Guide.

[3] Big Data: The Next Frontier for Innovation, Competition & Productivity. Jun2011, preceding p1-143. 147p. 32 Charts.

[4] Mckinsey & Company Quaterly, *"Business technology : Big Data. You have it, now use it"*.

[5] Dave Evans "The Internet of Things :How the Next Evolution of the Internet Is Changing Everything", Cisco White Paper, April 2011.

[6] Tesla, Nvidia: "*What is GPS Computing*".

[7] *"Introduction to Cloud Computing*" by William Voorsluys, James Broberg, and Rajkumar Buyya.

[8] Mariana Carroll, Paula Kotze, Alta Van Der Merwq *"Securing Virtual and Cloud Environments in Cloud Computing and Services Science"*, Research and Innovations in the service Economy, 2012.

[9] "*Big Data Processing in Cloud Computing Environments*", College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China.

[10] "IBM and SAP create the worls's largest SAP HANA system", IBM Solutions.

[11] Chris Preimesberger "In-Memory Databases Driving Big Data Efficiency: 10 Reasons Why", eWeek Feb 2013.

[12] Peter Mell and Timothy Grance *"The NIST Definition of Cloud Computing*", NIST, US Department of Commerce.

[13] Jimmy Lin and Chris Dyer "*Data-Intensive Text Processing with MapReduce"*, University of Maryland, College Park, Manuscript prepared April 11, 2010.