Big Data Analytics and High Performance Computing

Dhawal Chaturvedi
Indiana University
2679 E. 7th St, Apt. C
Bloomington, IN 47408, USA
dhchat@iu.edu

ABSTRACT

This paper provides an introduction to Big Data and High Performance Computing and tries to find how they are related to each other. We describe what exactly is Big Data and High Performance Computing. We then describe what technologies are in use in these respective fields and technology that can be used to combine them.

KEYWORDS

i523, hid204, Big data, High Performance Computing, SPIDAL

1 INTRODUCTION

Data is growing faster than ever, and at the same time, it is becoming obsolete faster than ever. The challenge is to how quickly and effectively one can analyze the data and gain insights that can be useful to solve problems. High Performance Computing plays an important role in running predictive analytics, especially when time is of crucial importance. In this paper, we analyze the ecosystem of the two data-intensive applications. We discuss the important features of the two fields, and then compare the functionality of the two paradigms.

2 BIG DATA

The quantity of computer data generated is growing exponentially in this world for many reasons. Retailers are building vast databases of recorded customer activities. Organizations working in logistics, financial services and health-care are also capturing more data. Social media is creating vast quantities of digital material. Big data is a term used for a combination of structured and unstructured data which has a potential to be mined for information.[6] It is often characterized by 3Vs : the enormous **Volume** of data, the **Variety** of data and the **Velocity** at which data is processed.

Here, Volume poses both the greatest challenge and the greatest opportunity as big data could help many organizations to understand people better and allocate resources more effectively. Big Data velocity also raises a number of issues as the rate at which data is flowing into many organizations is exceeding the capacity of their IT systems. In addition, user increasingly demand data to be streamed to them in real-time and delivering this can prove quite a challenge. Finally, the variety of data-types to be processed are becoming increasingly diverse. Today not only text documents, but audio, video , photographs are all equally important source of data.[6]

Recently Big data has been connected with terms such as data analytics, predictive analytics or any other kind of analytics which helps an organization to predict the user behavior so that they can improve their business. Data sets have been growing so rapidly mainly due to increasing number of ways data can be collected such as smartphones, your internet history or even your search history on a website.

3 HIGH PERFORMANCE COMPUTING

High-Performance Computing (HPC) is the use of parallel processing for running applications efficiently and quickly.[7] This term is especially used for computing architecture having capacity of more than a teraflop operations per second. It involves a lot of distinct computer processors working together on a complex problem. The complex problem is divided into smaller parts and distributed among the processors which are inter-connected using an architecture which is either massive centralized parallelism, massive distributed parallelism or something else entirely.

Massive Centralized Parallel computing refers to a computer architecture in which several high processing nodes are connected via a fast local area network. All these pseudo independent nodes are coordinated by a central scheduler. All the processors are connected to a single piece of memory. It is essentially a bigger version of a multi-core processor. It used to be the most common type of HPC architecture 15 years ago, but we don't see much of them anymore. This type of architecture is quite expensive and doesn't really scale. [7]

Massive Distributed Parallel computing refers to a computer architecture in which several high processing nodes are interconnected but with a more diverse administrative domain. It is a more opportunity based architecture in which the resources are allocated on the basis of their availability instead of having a centralized scheduler. The way these different nodes communicate with each other is standardized through a library called Message Passing Interface(MPI).[7]

Almost every Super Computer these days is a hybrid of Distributed and Shared memory in some way. Each node will be a shared-memory system. The network connecting these nodes will be some sort of topology. Along with the architecture, the way code is written needs to get optimized as well. Parallel computing is the key to increase the performance of Super Computing. Ideally, if you have T processors, you would like your program to be T times faster. But thats not the case. This is because not all parts of a program can be successfully split into T parts which can be processed in parallel. Splitting up the program might even cause additional overheads such as communication.

HPC is typically used for scientific research or simulation and analysis of an environment through computer modelling.HPC brings together several computer technologies such as Computer Architecture, algorithms together to solve these high process demanding problems.

4 BIG DATA AND HIGH PERFORMANCE COMPUTING SOLUTIONS

4.1 Amazon Web Services

Amazon Web Services(AWS) provides a variety of tools which are not only capable of handling huge amount of data but also provides technology and techniques for working productively with data at any scale. Another advantage of using AWS for big data analytics is the low cost at which amazon provides these tools. There is no capital investment required, no subscription requirements. Along with this, the ease with which you can configure these services is incredible. Anyone with a basic knowledge of command line can configure these tools with ease. Some of the major analytics tools provided by AWS are Amazon S3, Amazon Kinesis, Amazon DynamoDB, Amazon RedShift and Amazon Elastic MapReduce.
Amazon S3 is an object storage built to store and retrieve any amount of data from anywhere such as web sites and mobile apps, corporate applications etc. It is the only cloud storage solution with query-in-place functionality, allowing you to run powerful analytics directly on your data at rest in S3.[1] Amazon Kinesis is real-time streaming and processing for BigData. It is a highly-durable buffer that can handle all that work-load on the front-end as well as on the back-end with the help of series of EMR nodes which can give you an almost realtime analytics.[2]

Amazon DynamoDB is a NoSQL Database with high throughput and low latency for both read and write operations. It is a fully managed cloud database and supports both document and key-value store models. Amazon RedShift is a petabyte scale data warehouse which is massively parallel with over 1000 nodes running at a time.

4.2 Apache Hadoop Framework

The most widely known technology that helps to handle large-data would be a distribution data process framework is Apache Hadoop. It is an open-source framework used for processing huge datasets using a Map-Reduce model. It is based on a master-slave architecture where low-end commodity hardware is interconnected using ethernet. The framework broadly consists of 2 components, the storage part known as Hadoop Distributed File System(HDFS), and the processing part known as Map-Reduce.[8] The Master node split large files into smaller parts and distributes them across the slave nodes. After this, it sends the same code to every node which is used to process the data.
In the Map step, the slave nodes applies the map function to the data and stores the output temporarily. In the Shuffle step, slave nodes reshuffle data between them on the basis of key-values pairs such that data belonging to particular key is located on the same node. After this, slave nodes work process the respective keys in parallel. This results in increased efficiency as all the nodes are working in parallel independently. In the end, the MapReduce system collects the Reduce output from each node and combines it to produce final result.
MapReduce is useful in a wide range of applications, including distributed pattern-based searching, distributed sorting, web link-graph reversal, Singular Value Decomposition(SVD) and other Machine Learning algorithms.[8]

4.3 Hybrid of Hadoop and HPC

There has been convergence at many levels between HPC and Hadoop even though they were originally created to fulfill completely different purpose. HPC was designed for high-end, parallel computing jobs whereas hadoop was designed for cheap data storage and computing jobs.
There has been research going on offering a scale of comparison for different data-intensive computing fields, including blending the best of both computing paradigms using a hybrid of MPI and Hadoop. ”The goal is to successfully bring the two data-intensive computing paradigms together to share the developments versus ”reinvent the wheel” on either side”[4]. Machine Learning is another area which will have a lot to gain by this hybrid of HPC and Big Data as most of the ML algorithms are based on Linear Algebra which is a common HPC problem. if we run K-Means on MPI and Hadoop, MPI gave out better results than Hadoop. But the second generation Hadoop frameworks such as Spark gave out significantly better performance as they are adopting techniques such as effective collective operations which were previously only found in HPC architecture [5].
Another approach that has been proposed to converge these 2 systems is running Hadoop on top of HPC. However, a lot positives of Hadoop such as higher cluster Utilization are lost in this approach. Furthermore, Hadoop2(YARN) is capable of implementing both HPC applications and data- intensive applications but it still needs work [4].

4.4 Scalable Parallel Interoperable Data-Analytics Library (SPIDAL)

Many of the currently available commercial environments are more shifted towards the data-intensive paradigm. To make these environments work with HPC, there is need for HPC to look towards JAVA to run its codes as most of these commercial environments use JAVA whereas HPC has traditionally preferred C,C++. In the last few years, development has been done in this domain and SPIDAL JAVA has demonstrated significant performance gains when running on clusters upto 3072 cores[3].The developer friendly Java interface in SPIDAL Java will help to integrate it with other big data platforms such as Apache Hadoop, Spark, and Storm in future.[3]

5 CONCLUSIONS

Big Data Analytics and High Performance Computing are quite similar paradigms even though they were built for completely different purpose. In the next few years, its not unrealistic to believe that hadoop jobs to be processed on high end super computers instead of low end commodity infrastructure it presently runs on. This will not only help the Big Data industry but also other fields such as Machine Learning which certainly requires high end computing architecture.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.
The author would also like to thank Mr. Aditya Tandon for proof reading this paper.

REFERENCES

[1] Amazon. 2017. Big Data Analytics and High Performance Computing. Web Page. (Oct. 2017). https://aws.amazon.com/S3/ HID: 204.

[2] Amazon. 2017. Big Data Analytics and High Performance Computing. Web Page. (Oct. 2017). https://aws.amazon.com/kinesis/ HID: 204.

[3] Saliya Ekanayake, Supun Kamburugamuve, and Geoffrey Fox. 2016. SPIDAL Java: high performance data analytics with Java and MPI on large multicore HPC clusters. (04 2016), 3 pages.

[4] Shantenu Jha, Judy Qiu, André Luckow, Pradeep Kumar Mantha, and Geoffrey Charles Fox. 2014. A Tale of Two Data-Intensive Paradigms: Applications, Abstractions, and Architectures. *CoRR* abs/1403.1528 (2014). http://arxiv.org/abs/1403.1528

[5] André Luckow, Mark Santcroos, Ole Weidner, Ashley Zebrowski, and Shantenu Jha. 2013. Pilot-Data: An Abstraction for Distributed Data. *CoRR* abs/1301.6228 (2013). http://arxiv.org/abs/1301.6228

[6] Wikipedia. 2016. Big Data Analytics and High Performance Computing. Web Page. (June 2016). https://en.wikipedia.org/wiki/Big_data HID: 204.

[7] Wikipedia. 2016. Big Data Analytics and High Performance Computing. Web Page. (June 2016). https://en.wikipedia.org/wiki/HPC HID: 204.

[8] Wikipedia. 2016. Big Data Analytics and High Performance Computing. Web Page. (June 2016). https://en.wikipedia.org/wiki/Apache_Hadoop HID: 204.

A BIBTEX ISSUES

Warning–entry type for "Spidal" isn't style-file defined

–line 90 of file report.bib

Warning–page numbers missing in both pages and numpages fields in DBLP:journals/corr/JhaQLMF14

Warning–page numbers missing in both pages and numpages fields in DBLP:journals/corr/abs-1301-6228

(There were 3 warnings)

B ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

B.1 Writing Errors

Short form of verbs is for spoken language. Do not use them in scientific writing. example: can't is incorrect, use cannot

B.2 Citation Issues and Plagiarism

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

It is better to use more scientific resources as opposed to commercial web sites and Wikipedia.

Reference N. 7 (https://en.wikipedia.org/wiki/HPC) does not refer to a single article.

The citation mark should not be in the beginning of the sentence or paragraph, but in the end, before the period mark. example: ... a library called Message Passing Interface(MPI) [7].

Put a space between the citation mark and the previous word.

re