Vol. 3 No. 4 April 2016 ISSN: 2321 – 788X

A SURVEY ON HARDWARE PLATFORMS AVAILABLE FOR BIG DATA ANALYTICS

D.V. Jeyanthi
Assistant Professor, Dept. of Computer Science, Sourashtra College, Madurai

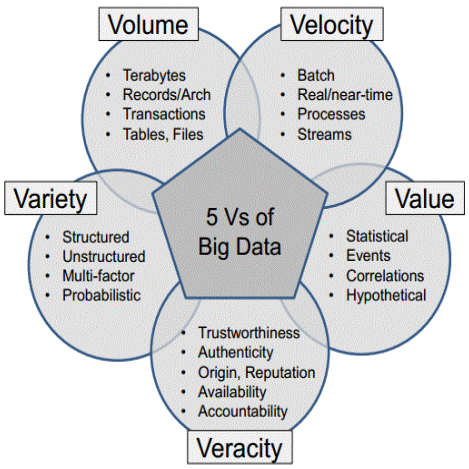
Abstract

This is an era of Big Data. The total digital data in this world is expected to double in less than two years. Big Data is driving radical changes in traditional data analysis platforms and algorithms. This paper provides an in-depth analysis of different platforms available for studying and performing big data analytics. This paper surveys on different hardware platforms available for big data analytics and assesses the advantages and drawbacks of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support. Using a star ratings table, a rigorous qualitative comparison between different platforms is made for each of the six characteristics that are critical for the algorithms of big data analytics. In addition to the hardware, a detailed description of the software frameworks used within each of these platforms is also discussed along with their strengths and drawbacks. Some of the critical characteristics described here can potentially aid in making an informed decision depending on their computational needs.

Keywords: *Big Data Platforms, Big Data Analytics, Hardware Platforms, Horizontal Scaling, Vertical Scaling, big data platforms comparison*

Introduction

Big Data is a collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. The five dimensions of big data are volume, velocity, variety, varsity, value.



Volume refers to the vast amounts of data generated every second like emails, twitter, Facebook, WhatsApp, Hangouts, messages, photos, video clips, sensor data etc. We produce

and share data every second. The data produced are not in terabytes but in Zettabytes or Brontobytes.

Velocity refers to the speed at which new data is generated and the speed at which data moves around.

Variety refers to the different types of data .It holds messages in unstructured and in structured form.

Veracity refers to the messiness or trustworthiness of the data.

Value: There is another V to take into account when looking at Big Data: Value! It is all well and good having access to big data but unless we can turn it into value it is useless. Big data is not just about size. It finds insights from complex, noisy, heterogeneous, streaming, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. The challenges include capture, storage, search, sharing & analysis

Big Data Analytics is the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help to identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the *knowledge* that comes from analyzing the data.

Big Data Requires High-Performance Analytics

To analyze such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics. Using big data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

How Big Data Analytics is Used Today

As the technology that helps an organization to break down data silos and analyze data improves, business can be transformed in all sorts of ways. According to Datamation, today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook.

Another example comes from one of the biggest mobile carriers in the world. France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the

foundation for development projects to improve public health and safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.

Benefits of Big Data Analytics

Enterprises are increasingly looking to find actionable insights into their data. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

Platforms

Powerful Platform is needed to get quick results and the data to be processed. Rate of data transfer is also critical in big data. The choice of hardware/software platform plays a crucial role to achieve one's required goals. To analyze this voluminous and complex data, scaling up is imminent. In many applications, Analysis tasks need to produce results in real-time and/or for large volumes of data. It is no longer possible to do real-time analysis on such big datasets using a single machine running commodity hardware. Continuous research in this area has led to the development of many different algorithms and big data platforms. Scaling is the ability of the system to adapt to increased demands in terms of processing. There are two types of scaling horizontal scaling and vertical scaling.

Horizontal Scaling

Horizontal Scaling involves distributing work load across many servers. Multiple machines are added together to improve the processing capability. It involves multiple instances of an operating system on different machines. The Advantages are a) Increases performance in small steps as needed (b) Financial investment to upgrade is relatively less and (c) can scale out the system as much as needed. The Drawbacks of Horizontal Scaling are Software has to handle all the data distribution and parallel processing complexities and Limited number of software are available that can take advantage of horizontal scaling.

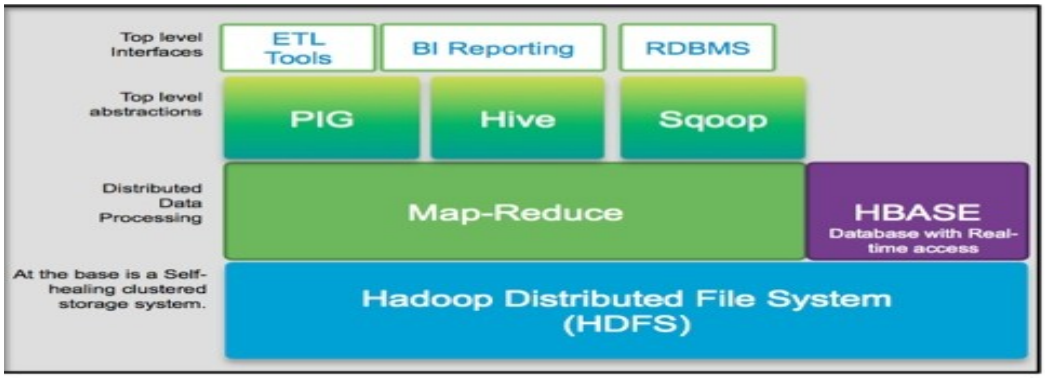
Some Horizontal Scaling Platforms are Peer to Peer Networks, Apache Hadoop, and Apache Spark

Peer to Peer Networks involves millions of machines connected in a network. The network architecture used here is Decentralized and Distributed Network Architecture. Message Passing Interface (MPI) is the communication scheme used and each node capable of storing and processing data Scale is practically unlimited (can be millions of nodes). The main Drawbacks of this peer to peer network is the Communication is the major bottleneck. Broadcasting messages is cheaper but aggregation of results/data is costly. It has Poor Fault tolerance mechanism.

Apache Hadoops is an Open source framework for storing and processing large datasets with High Fault Tolerance and designed to be used with commodity hardware. It

consists of two important components. **HDFS** (Hadoop Distributed File System) which is used to store data across cluster of commodity machines while providing high availability and fault tolerance. **Hadoop YARN** a Resource management layer which Schedules jobs across the cluster.

Hadoop Architecture



Hadoop MapReduce

Hadoop MapReduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks. Basic data processing scheme is used in Hadoop. It includes breaking the entire scheme into mappers and reducers. Mappers read data from HDFS process it and generate some intermediate results. Reducers aggregate the intermediate results to generate the final output and write it to the HDFS. Typical Hadoop job involves running several mappers and reducers across the cluster by Divide and Conquer Strategy.

MapReduce Wrappers

It provide better control over MapReduceCode and aids in code development. Popular MapReduceWrappers includes Apache Pig which provides SQL like environment developed at Yahoo. It is used by many organizations including Twitter, AOL, LinkedIn and more. Hive Developed by Facebook. Both these are intended to make code development easier without having to deal with the complexities of MapReduceCoding

Spark

A Next generation paradigm for Big Data Processing. It's developed by researchers at University of California, Berkeley. Used as an alternative to Hadoop Designed to overcome disk I/O and improve performance of earlier systems. It allows data to be cached

in memory, eliminating the disk overhead of earlier system. Supports Java, Scala and Python Can yield upto 100x faster than Hadoop MapReduce

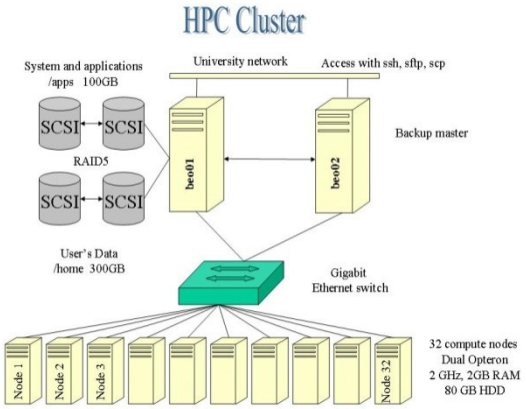
Vertical Scaling

Vertical Scaling Involves installing more processors, more memory and faster hardware typically within a single server. It involves single instance of an operating system. The advantages of Vertical Scaling are (a) Most of the software can easily take advantage of vertical scaling (b) Easy to manage and install hardware within a single machine. The Drawbacks are (a) It requires substantial financial investment (b) System has to be more powerful to handle future workloads and initially the additional performance goes to waste and (c) It is not possible to scale up vertically after a certain limit.

Some Vertical Scaling Platforms are High Performance Computing Clusters (HPC), Multicore Processors, Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGA).

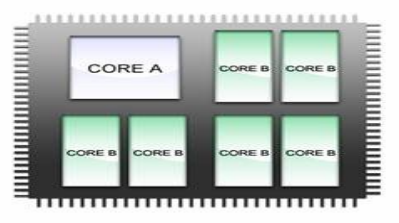
High Performance Computing (HPC) Clusters

HPC's are also known as Blades. These are supercomputers with thousands of processing cores. These can have different variety of disk organization and communication mechanisms. It contains well-built powerful hardware optimized for speed and throughput. Fault tolerance is not critical because of top quality high-end hardware. Not as scalable as Hadoop or Spark but can handle terabytes of data High initial cost of deployment Cost of scaling up is high MPI is typically the communication scheme used.



Multicore CPU

Multicore CPU has dozens of processing cores. Number of cores per chip and number of operations a core can perform has increased significantly. Newer breed of motherboards allow multiple CPUs within a single machine. Parallelism achieved through multithreading Task has to be broken into threads.

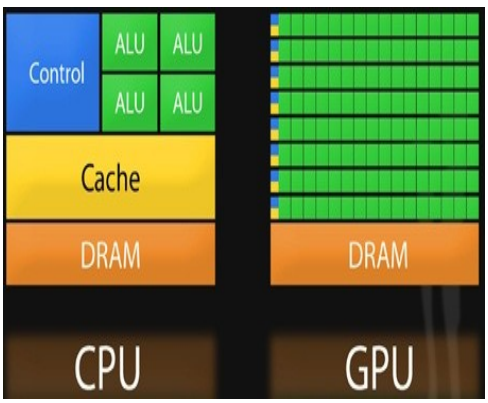


Graphics Processing Unit

It's a Specialized hardware with massively parallel architecture .Recent developments in GPU hardware and programming frameworks has given rise to GPGPU (general purpose computing on graphics processing units). It has large number of processing cores (typically around 2500+ currently). It has it's own DDR5 memory which is many times faster than typical DDR3 system memory. NvidiaCUDA is the programming framework which simplifies GPU programming. Using CUDA, one doesn't have to deal with low-level hardware details.

CPU vs GPU Architecture

- Development in CPU is rather slow as compared with GPU.
- Number of cores in CPU is still in double digits while a GPU can have 2500+ cores.
- Processing power of a current generation CPU is close to 10 Gflops while GPU can have close to 1000 Gflopsof computing power.
- CPU primarily relies on system memory which is slower than the GPU memory.
- While GPU is an appealing option for parallel computing, the number of softwares and applications that take advantage of the GPU is rather limited.
- CPU has been around for many years and huge number of software are available which use multicore **CPUs**



Field Programmable Gate Arrays (FPGA)

These are highly specialized hardware units. This is Custom built for specific applications. This can be highly optimized for speed. Due to customized hardware, development cost is much higher/ Coding has to be done in HDL (Hardware Description Language) with low level knowledge of hardware Greater algorithm development cost. This is suited for only certain set of applications.

Comparison of Different Platforms

Following characteristics are used for comparison: System/Platform dependent and Applications / Algorithm dependent. In System / Platform dependent the metrics used for comparison are Scalability, Data I/O Performance and Fault Tolerance. In Application/Algorithm dependent the metrics used for comparison are Real-time Processing, Data size support and Support for iterative tasks. Comparison is done using the star ratings. 5 stars correspond to highest possible rating 1 star is the lowest possible rating.

Platforms (Communication Scheme)	System/Platform			Application/Algorithm		
	Scalability	Data I/O Performance	Fault Tolerance	Real-Time Processing	Data Size Supported	Iterative Task Support
Peer to Peer (TCP/IP)	★★★★★	★	★	★	★★★★★	★★
Virtual Clusters (MapRedce/MPI)	★★★★★	★★	★★★★★	★★	★★★★	★★
Virtual Clusters (Spark)	★★★★★	★★★	★★★★★	★★	★★★★	★★★
HPC Clusters (MPI/Mapreduce)	★★★	★★★★	★★★★	★★★	★★★★	★★★★
Multicore (Multithreading)	★★	★★★★	★★★★	★★★	★★	★★★★
GPU (CUDA)	★★	★★★★★	★★★★	★★★★★	★★	★★★★
FPGA (HDL)	★	★★★★★	★★★★	★★★★★	★★	★★★★

Scalability – is the ability of the system to handle growing amount of work load in a capable manner or to be enlarged to accommodate that growth. It is the ability to add more hardware to improve the performance and capacity of the system.

Data I/O Performance is the rate at which the data is transferred to/from a peripheral device. In the context of big data analytics, this can be viewed as the rate at which the data is read and written to the memory (or disk) or the data transfer rate between the nodes in a cluster

Fault Tolerance is the characteristic of a system to continue operating properly in the event of a failure of one or more components

Real-Time Processing is the system's ability to process the data and produce the results strictly within certain time constraints

Data Size Supported is the size of the dataset that a system can process and handle efficiently

Iterative Task Support is the ability of a system to efficiently support iterative tasks. Since many of the data analysis tasks and algorithms are iterative in nature, it is an

important metric to compare different platforms, especially in the context of big data analytics

Conclusion

It is observed that Vertical Scaling Platforms such as GPU and MultiCoreCPU are faster than horizontal scaling platforms such as Hadoopand Spark. We also observed that Horizontal Scaling Methods are more scalable. For example Vertical Scaling Platform GPU cannot scale with a data bigger than 90K features and MultiCoreCPU cannot handle more than 70K features. Hadoop and Spark were able to process a data with 1.3 million features. We find that GPU outperforms MultiCoreCPU by spawning very high number of threads. We also see that for both iterative and non-iterative scenarios, Spark yields better timing than Hadoop.

References

1. Demchenko, Y.; Grosso, P.; de Laat, C.; Membrey, P. "Addressing big data issues in Scientific Data Infrastructure", *Collaboration Technologies and Systems (CTS), 2013 International Conference on,* On page(s): 48 – 55
2. http://journalofbigdata.springeropen.com/
3. Journal of Big Data 20152: 21 DOI: 10.1186/s40537-015-0030-3
4. [CCC2011a]Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.