RESEARCH ARTICLE OPEN ACCESS

A Review on Scaling Platforms for Big Data Clustering

AMAR S. PIMPALKAR[1]

[1]Department of computer science, St. Francis de Sales College, Nagpur, India.

PROF. MAHENDRA P. DHORE[2]

[2]Pro-Vice Chancellor, Sant Gadge Baba Amravati University, Amravati, India.

ABSTRACT

Undoubtedly, now we have entered in the era of big data. The challenges produce by this big data includes how to store, search and analyse this huge and complex data. A significant challenge for researchers in big data is identifying the right platform for analysing large datasets. Thus, Researchers today are concentrating on scalable clustering techniques that utilize various acceleration platforms to address challenges associated with big data. This paper explores two primary approaches to scaling big data analysis platforms: vertical scaling and horizontal scaling. The platforms that are available for big data analysis in terms of both vertical and horizontal scaling are also included in this study.

Keywords - Big data, Parallel clustering, Big data scaling platform,

I. INTRODUCTION

Clustering is one of the important techniques in data mining. Clustering is the process of dividing a set of objects into groups such that objects having same properties form a single group (cluster) and objects with different properties forms different groups (clusters) [1]. Clustering has been widely used in numerous applications such as market research, pattern recognition data analysis etc. [2]. The main aim was to identify groups that were not known before, which was a desired outcome in various everyday challenges. This could be achieved through different categories of clustering methods such as hierarchical methods, partitioning methods, density-based methods, grid-based methods or other clustering techniques [3].

With the advent of 5G technology, an enormous volume of data is being produced at an unprecedented speed, resulting in a massive quantity referred to as big data. The features of big data, including its large volume, diverse variety of data, high velocity and multivalued data complicate the process of data analysis. Extracting valuable insights from large datasets is a challenging endeavor. Clustering algorithms serve as essential tools in the process of data mining and hold significant importance in the analysis of big data. Clustering methods are mainly divided into density-based, partition-based, grid based, hierarchical and model-based clustering [4].

Clustering with big data presents several significant challenges, mainly from the characteristics known as 3V's of big data —volume, velocity and variety proposed by Gartner analyst Doug Laney [5]. Some of the major challenges are Scalability, Storage and Management, High Dimensionality, Noise and Outliers, Computational Complexity, Data Quality, Dynamic Data etc.

In order to overcome some of these limitations, today's research is moving toward the idea of parallel computing, which is leading to the emergence of what are known as parallel clustering algorithms. As the name suggests, this kind of algorithms divide the big data sets into several small chunks and then carry out actions for each small chunk parallelly on one or more processing devices. The final result is obtained by integrating the intermediate clustering results at the end.

The concept of parallelism aims to improve the speed-up, the throughput, and the scalability of the clustering process so that it becomes effective to meet the challenges of big data [6].

The remaining paper is organized as follows. Section I provides introduction to clustering, big data and need of parallel clustering, An overview of big data characteristics is provided in Section II, In Section III classification and overview of different scaling platforms for big data clustering is covered, some vertical and horizontal scaling platforms are discussed in Sections IV and V respectively and Section VI conclude the research paper with future direction.

II. OVERVIEW OF BIG DATA CHARACTERISTICS

Big data is characterized by three primary features known as the 3Vs: volume, velocity, and variety.

Volume

The name big data itself is related to an enormous size. Volume refers to tremendous amount of data generated through many different sources such as business, social media platforms, networks, human interactions, healthcare, transportation, power grid, intelligent etc [7].

For example, data generated on social media every minute is as follow:

Snapchat: 5,27,760 photos are shared, Twitter: 4,56,000 tweets are sent, Instagram: 46,740 photos are posted, Facebook: 4 million likes are generated, 5,10,000 comments are posted, 2,93,000 statuses are updated and 2,40,000 photos are uploaded, Facebook family of apps: Nearly 7 billion messages are sent.

Volume of big data presents number of challenges including those related to processing, storage, data quality, scalability, the curse of modularity and dimensionality and many more.

Velocity

Velocity refers to the speed at which new data is generated and moves around. Big data is often produced in real time, so it needs to be processed, accessed, and analysed at the same rate [8]. Application logs, business processes, transactional data, networks, social media sites, sensors, mobile devices, data stream etc. are some of the sources of quick data generation. Managing this aspect of big data presents several challenges such as Real-Time Processing, Data Storage, Scalability, Data Quality, Security and Privacy, Complexity of Analysis etc.

Variety

In past, data was gathered solely from databases and spreadsheets. But now a days the data comes in variety of forms like text, numbers, PDF format, emails, audio recordings, social media posts, images, videos, spatial data, biometrics etc. They come from various places viz web, text mining, image mining and so on. These data do not have a fixed structure and rarely present themselves in a perfectly ordered form and ready for processing [9]. This variety of data is classified as follow:

Structured Data: Data that is arranged in an appropriate way usually takes the form of rows and columns making it easily readable and analysable. Now-a-days just 20 percent of the existing data is structured data. For example, Databases, spreadsheets and transactional data.

Unstructured Data: Data that does not follow any defined format or structure, making it more challenging to analyse. Now-a-days 80 percent of the data exists in this form. For analysis of this type of data machine learning techniques are used [10]. Example includes text documents, emails, social media posts, videos, audio files, etc.

Semi-Structured Data: It is a type of data that does not fit precisely into table or spreadsheet format but still has some organizational properties. Such type of data can be analysed easily when compared with unstructured data. XML, JSON, NoSQL databases, logs are some examples of this type of data.

This aspect of big data presents several challenges like data integration, schema evolution, data quality and consistency, complexity in analysis etc.

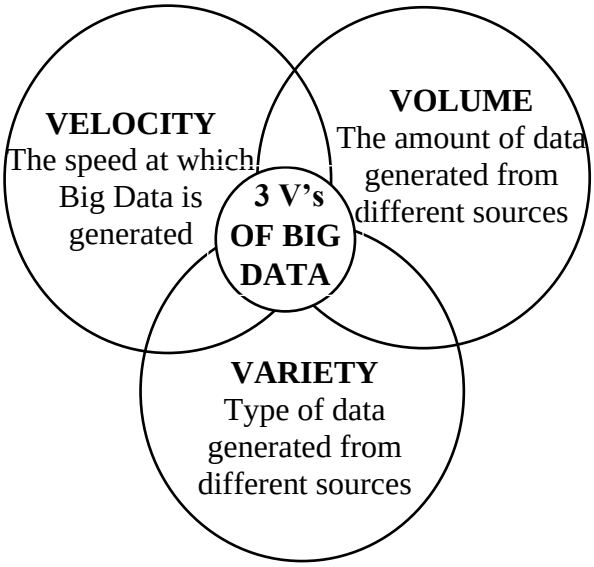

Fig. 1. 3 V's of Big data

III. OVERVIEW ON THE DIFFERENT SCALING PLATFORMS OF BIG DATA CLUSTERING

System scaling refers to the ability of a system to handle an increase in demands by adding additional resources to the system [11]. Multiple big data platforms are available, each with unique features. Selecting the appropriate platform requires a comprehensive understanding of all its characteristics. When discussing a hardware/software parallel system's capacity to use increasing computing resources for the analysis of big data sets, the term "scalable data analysis" is used [12]. According to their scalability, big data platforms are classified into two categories, as shown in figure 2.

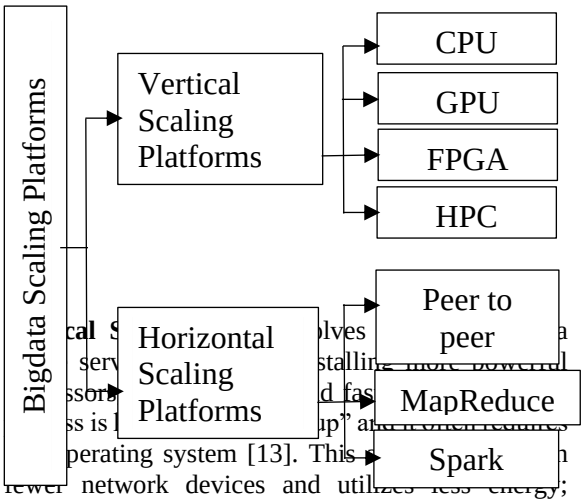

Fig. 2. Bigdata Scaling Platforms

expansion is anticipated.

Horizontal Scaling: Horizontal scaling refers to the process of spreading the workload over numerous servers, which can include standard machines. This approach is commonly known as "scaling out." In this method, several independent servers are combined to enhance overall processing power. Typically, different machines are running different instances of the operating system [14]. Horizontal scaling is a long-term solution as it allows addition of more servers whenever needed.

However, switching from a single, large system to this kind of cluster may be a difficult, yet it is a highly effective solution.

The resources of the computing infrastructure must be increased in order to support both horizontal and vertical scaling platforms. The primary distinction between them is that, horizontal scaling adds more machine resources to your existing system infrastructure whereas vertical scaling approach refers to increasing the machine infrastructure capability. We discuss each of these platforms with their capabilities.

Table1: Vertical Scaling vs Horizontal Scaling

Vertical Scaling	**Horizontal Scaling**
It Involves enhancing the capabilities of a single machine or instance by incorporating additional resources such as CPU, memory or storage.	It involves adding more machines or instances to a pool to handle increased load.
If the upgraded server fails, your entire system might be affected until it is fixed or replaced.	If one instance fails, others can continue to operate, enhancing reliability.
There is a maximum limit to how much you can scale a single machine. Eventually, we will reach hardware limitations.	Easier to scale up and down based on demand by adding or removing instances to or from pool.
Easier to implement as it involves upgrading a single server rather than managing multiple instances.	More instances mean more complexity in terms of load balancing, data consistency and management.
No need for complex data synchronization between multiple servers.	Increased overhead in network communication and data synchronization.
It uses less network hardware and consumes less power.	It uses more network hardware and consumes more power.
It provides a short-term fix, especially if continued growth is expected.	It provides a long-term solution.
More cost-effective initially.	Higher initial cost.

VERTICAL SCALING PLATFORMS

Multicore CPU

A multi-core processor is a technology that has developed through improvements in network and processor technologies. The significant progress in CPU chips over the years has led to the creation of multi-core architectures, which are essential for the computational power needed to handle big data. It is a single chip an integrated circuit which consists of two or more processor cores [15]. Since, each core shares a common memory, the parallelism in CPUs is mainly achieved through multithreading. A big task is divided into several smaller tasks and are assign to different threads. Each thread is executed in parallel on different CPU cores.

Up till a few years back, CPUs were primarily used to speed up the algorithms for large data analysis. The multi-core CPUs offers increase in performance, reduce power consumption and low heat generation [16]. Despite these advantages, there are some drawbacks of multi-core CPUs. The primary limitations of CPUs include their restricted number of processing cores and their dependency on system memory for data access. Since the system memory is confined to several hundred gigabytes, this imposes a constraint on the amount of data that a CPU can handle effectively. When the data size surpasses the available system memory, disk access will become difficult.

GPU (Graphics processing unit)

In the past two decades, GPU has made huge improvement in performance and capability by taking advantage of the parallelism that is a fundamental aspect of graphics processing.

Until the past few years, GPUs were mainly used for graphical operations such as image and video editing, accelerating graphics-related processing etc [17]. A modern GPU architecture consists of an array of processors executing in parallel. GPUs massively parallel architecture, it's recent hardware developments and associated programming frameworks have resulted to the emergence of GPGPU (general-purpose computing on graphics processing units). Compared to a multicore CPU, a GPU has large number of processing cores. In addition to the processing cores, a GPU has high throughput DDR5 memory, which is very fast as compared to standard DDR3 memory. GPU also utilizes two parallel architectures: The single program, multiple data (SPMD) programming model and the single instruction, multiple data (SIMD) parallel architecture.

CUDA is a parallel computing platform and programming model designed and developed by NVIDIA. This platform license to boost computing performance by harnessing the power of the GPU. CUDA greatly simplifies GPU programming [18]. There are many advantages of using GPUs for high-performance computing (HPC) applications. Which includes GPUs are widely available, reasonably inexpensive, and energy-efficient compared to conventional computing devices. GPUs are highly efficient accelerators with hundreds of processing cores. Beside these advantages, the primary drawback of GPU is its limited memory. With a maximum of 24GB memory per GPU, it is very difficult to manage terabyte scale data. When the size of the data surpasses the GPU memory capacity, performance decreases significantly as the disk access becomes the main constraint.

FPGA (Field programmable gate arrays)

FPGA known as Field Programmable Gate Array, is a form of integrated circuit type that can be reconfigured after manufacturing to meet variety of requirements. FPGAs are well known for their versatility and high performance and are used in various sectors, including telecommunications, automotive, aerospace, option pricing etc. [19]. The flexibility of the FPGA is due to its fundamental component called as configuration logic block (CLB). CLB's offers the primary logic and storage capability. The key features of FPGAs include their increased speed, volume designs, programmable functionalities and reduced complexity. They are written in Hardware Descriptive language (HDL). As a result, the development cost is much higher in comparison to other platforms. FPGA is suitable for a wide range of applications, including random number generation, image processing, cryptography, audio and video processing, and the implementation of various algorithms [20]. The key benefits of FPGAs include their flexibility and expandability, increased speed, less power consumption and lower cost of prototypes they provide. Furthermore, FPGA is not suitable for processing large data sets.

High-performance computing (HPC) clusters

An HPCC is a type of distributed computing system that consists of multiple computers or multiple nodes connected through a local area network (LAN), commonly with ethernet technology [21]. It is used to perform complex calculations and to process large datasets more efficiently than a single computer could do. Depending upon the user requirement, they can have variety of disk organization, cache, communication mechanism etc. These systems use high-end hardware which is use to optimize speed and throughput. The initial cost of installing such a system can be very high because of powerful hardware. But fault tolerance in such systems is not challenging since hardware failures are extremely rare. Despite not being as scalable as Hadoop or Spark clusters, they can handle terabytes of data processing. The expense of scaling up such a system is much higher than that of Hadoop or Spark clusters.

HORIZONTAL SCALING PLATFORMS

Peer-to-Peer Network

Peer-to-peer (P2P) networks are decentralized networks in which a peer (node) can act as both a server and a client, i.e. they can both consume and serve resources. P2P networks allow peers to share resources directly with each other, unlike traditional client-server models that rely on a single server to provide resources to several clients.

A peer-to-peer (P2P) network can efficiently function as a horizontal scaling platform, where multiple peers (nodes) share resources and workloads without relying on a centralized server.

Typically, Message Passing Interface (MPI) is the communication system used in such a arrangement to share and exchange the data among peers. An individual node has the capacity to store data instances, and the capability to scale out is practically unlimited. However, a significant challenge in this setup arises from the communication between various nodes. In a peer-to-peer network, sending messages is relatively inexpensive, while consolidating data and outcomes incurs much higher costs.

A significant aspect of MPI is its ability to preserve the state of processes. In contrast to MapReduce, all parameters can be kept locally. Hence MPI is well suited for iterative processing [22]. An additional aspect of MPI is its hierarchical master/slave model. When MPI is deployed in the

master–slave model, the slave machine can become the master for other processes. This capability can be advantageous for dynamic resource management, especially when the slave machines are tasked with processing large amounts of data.

MapReduce

In 2004, Google introduced MapReduce as an open-source system and software designed for handling large data sets. MapReduce is a highly programming model that facilitates processing of large volume of data through parallel execution across large clusters. The MapReduce framework is currently utilized in open-source platforms such as No, MongoDB, and Apache Hadoop. MapReduce is a simplified programming model since all the parallelization, load balancing, communication and fault tolerance are automatically handled by framework operations in MapReduce system [23].

This strategy uses the parallel paradigm, which breaks the principal data into pieces of manageable size and distribute them across numerous processing nodes. MapReduce uses two main user-defined functions: Map and Reduce. The input and output for these functions consist of (key, value) pairs.

In Map phase, each Mapper process split its assigned input independently. It reads the input data line by line, processes it and emits key-value pairs. After the Mappers complete their tasks, the framework collects all the emitted key-value pairs and groups them by key. Each Reducer processes the grouped key-value pairs. It takes all values for a specific key and combines them in some way. The output is typically a new set of key-value pairs.

Spark

Apache Spark is a powerful open-source, parallel processing, flexible and user-friendly data processing engine that is designed for large-scale data processing and analytics. It makes use of a hybrid framework that supports both batch and stream processing capability [24]. Apache Spark framework contains Spark core and upper-level libraries viz Spark SQL, Spark Streaming, Spark MLlib, GraphX and SparkR which helps to perform a wide range of workloads including batch processing, machine learning, interactive queries, streaming processing etc [25]. Apache Spark system stands out by offering language-integrated APIs for developing intricate algorithms in a range of programming languages such as SQL, Java, Scala, Python, and R. The main functions are performed on Spark Core. Existing components are closely integrated with the Spark Core, which offers a single unified environment. It is built on the top of Hadoop and can process data much faster than MapReduce [26].

IV. CONCLUSION

Several data processing platforms that are currently available are surveyed in this paper. Details of these hardware platforms, along with some renowned software frameworks such as Spark and Hadoop are provided. This study could help to select the right platform based on their data/computational requirements. Two primary types of platforms are discussed for managing large-scale data processing. In the vertical scaling platform, we have discussed CPU, GPU, FPGA and HPC whereas in horizontal scaling platform we have focused on Peer-to-peer network, MapReduce and Spark.

This study will serve as an initial step to evaluate the effectiveness of each platform for handling real-world applications. Another track will be to inspect the possibility of merging multiple platforms to solve a particular application problem. For example, attempting the integration of horizontal scaling platforms like Hadoop with vertical scaling platforms such as GPUs.

V. REFERENCES

[1] Qi Xianting & Wang Pan, “A Density-Based Clustering Algorithm for High-Dimensional Data with Feature Selection”. International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), doi:10.1109/iciicii.2016.0038

[2] Ze Deng, Yangyang Hu, Mao Zhu, Xiaohui Huang, Bo Du (2014), “A scalable and fast OPTICS for clustering trajectory big data”, Cluster Computing, 18(2), 549–562, doi:10.1007/s10586-014-0413-9

[3] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Foufou S, Bouras A. (2014), “A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis”, IEEE Transactions on Emerging Topics in Computing, 2(3), 267–279. doi:10.1109/tetc.2014.2330519

[4] Bhadani A. K., Jothimani D. (2016), “Big data: Challenges, opportunities and realities”, In Singh, M.K., & Kumar, D.G. (Eds.), Effective Big Data Management and Opportunities for Implementation (pp. 1-24), Pennsylvania, USA, IGI Global

[5] Laney D (2001), “3D data management: controlling data volume, velocity, and variety”, Technical Report, 949, Gartner

[6] Zineb Dafir, Yasmine Lamari, Said Chah Slaoui (2020), “A survey on parallel clustering algorithms for Big Data”, Artificial Intelligence Review, doi:10.1007/s10462-020-09918-2

[7] Kumar Rahul, R. K. Banyal and Neeraj Arora, “A systematic review on big data applications and scope for industrial processing and healthcare sectors”, Journal of Big Data, volume 10, Article number: 133 (2023)

[8] Cuzzocrea, A. & Moussa R. (2017), “Multidimensional database modeling: Literature survey and research agenda in the big data era”, 2017 International Symposium on Networks, Computers

and Communications (ISNCC), doi:10.1109/isncc.2017.8072024
[9] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, "Survey Understandable Big Data: A survey", Computer Science Review, volume 17, 70-81, http://dx.doi.org/10.1016/j.cosrev.2015.05.002
[10] Tanya Garg & Surbhi Khullar (2020), "Big Data Analytics: Applications, Challenges & Future Directions", 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), doi:10.1109/icrito48877.2020.9197
[11] Mahmoud A. Mahdi, Khalid M. Hosny and Ibrahim Elhenawy (2021), "Scalable Clustering Algorithms for Big Data: A Review", IEEE Access, 9, 80015–80027, doi:10.1109/access.2021.30840
[12] Hadjir Zemmouri, Said Labed and Akram Kout, "A survey of parallel clustering algorithms based on vertical scaling platforms for big data", 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), DOI:10.1109/PAIS56586.2022.9946663
[13] Ahmed Hussein Ali, Mahmood Zaki Abdullah, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics", International Journal Of Integrated Engineering, Vol. 11 No. 6 (2019) 138-150
[14] Dilpreet Singh and Chandan K Reddy (2014), "A survey on platforms for big data analytics", Journal of Big Data, 2(1), doi:10.1186/s40537-014-0008-6
[15] Maradana Durga Venkata Prasad and Srikanth Thota, "Article: A Survey on Dependency of Parallel Clustering Platforms on Clustering Algorithms with Their Clustering Criteria for Big Data", Preprints.org, DOI:10.20944/preprints202312.1201.v1
[16] Xiaohong Qiu, Geoffrey Fox , Huapeng Yuan, Seung-Hee Bae, George Chrysanthakopoulos, Henrik Nielsen, "Parallel Data Mining on Multicore Clusters", Seventh International Conference on Grid and Cooperative Computing, doi:10.1109/gcc.2008.100
[17] Chen J. Y. (2009), "GPU technology trends and future requirements", IEEE International Electron Devices Meeting (IEDM), doi:10.1109/iedm.2009.5424433
[18] Vincent Boyer, Didier El Baz, "Recent Advances on GPU Computing in Operations Research", 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, doi:10.1109/ipdpsw.2013.45
[19] Aidan O Mahony, Bernard Hanzon and Emanuel Popovici, "Review: The Role of FPGAs in Modern Option Pricing Techniques: A Survey", Electronics 2024, 13(16), 3186, https://doi.org/10.3390/electronics13163186
[20] Shubham Gandhare & B. Karthikeyan, "Survey on FPGA Architecture and Recent Applications", 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), doi:10.1109/vitecon.2019.8899550
[21] Manuel López-Martínez, Germán Díaz-Flórez,Santiago Villagrana-Barraza, Luis O. Solís-Sánchez,Héctor A. Guerrero-Osuna, Genaro M. Soto-Zarazúa, and Carlos A. Olvera-Olvera, "A HighPerformance Computing Cluster for Distributed Deep Learning: A Practical Case of Weed Classification Using Convolutional Neural Network Models", Appl. Sci. 2023, 13, 6007, https://doi.org/10.3390/app13106007
[22] Sievert O, Casanova H(2004), "A Simple MPI Process Swapping Architecture for Iterative Applications", The International Journal of High Performance Computing Applications, 18(3), 341–352, doi:10.1177/1094342004047430
[23] Fouad Hammadi Awad and Murtadha M. Hamad, "Big Data Clustering Techniques Challenges and Perspectives: Review", Informatica 47 (2023) 203–218, https://doi.org/10.31449/inf.v47i6.4445
[24] Kijsanayothin P, Chalumporn G, & Hewett R (2019), "On using MapReduce to scale algorithms for Big Data analytics: a case study", Journal of Big Data, 6(1), doi:10.1186/s40537-019-0269-1
[25] Shaikh E, Mohiuddin I, Alufaisan Y & Nahvi I(2019), "Apache Spark: A Big Data Processing Engine", 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), doi:10.1109/menacomm46666.2019.8988541
[26] Big Data Analytics: A Comparative Evaluation of Apache Hadoop and Apache Spark
Sukhpreet Singh, Jaswinder Singh, Sukhpreet Singh, "Big Data Analytics: A Comparative Evaluation of Apache Hadoop and Apache Spark", International Research Journal of Engineering and Technology (IRJET), Volume: 10 Issue: 11