JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 115-121
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1437

Design of High-Performance Computing System for Big Data Analytics

Mahantesh K. Pattanshetti

Department of Computer Science, Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

Abstract: The design of a high-performance computing (HPC) system for big data analytics involves selecting hardware and software components that can handle the massive amounts of data generated by big data applications, while ensuring efficient processing and storage of the data. The hardware components of an HPC system typically include high-end servers with powerful processors, large amounts of memory, and high-speed storage devices such as solid-state drives (SSDs) or hard disk drives (HDDs) configured in a parallel or distributed architecture. In addition, specialized hardware such as graphical processing units (GPUs) or field-programmable gate arrays (FPGAs) can be used to accelerate the processing of certain types of data. The software components of an HPC system include the operating system, middleware, and applications. The operating system should be optimized for HPC workloads, with low overhead and high scalability. Middleware such as MPI (Message Passing Interface) can be used to facilitate communication between nodes in a distributed computing environment. Finally, applications should be designed to take advantage of the parallel and distributed processing capabilities of the HPC system, with efficient algorithms and optimized data structures. This paper proposes a design of high-performance computing system for Big Data Analytics.

I. Introduction

High performance computing (HPC) systems are used in big data analysis to provide the computational power necessary to process and analyse massive amounts of data. Big data analysis often involves complex data mining algorithms, machine learning models, and statistical analysis that require large amounts of processing power and memory systems provide a high degree of parallelism and distributed computing capabilities, allowing data to be processed simultaneously across multiple nodes or servers. This allows for faster data processing and analysis compared to traditional computing systems.

In big data analysis, HPC systems can be used to perform tasks such as data pre-processing, data cleaning, feature extraction, and model training. These tasks require significant computational resources and can benefit greatly from the use of HPC systems. In addition, HPC systems can be used to analyse real-time data streams, such as those generated by IoT devices, social media, or financial transactions. These data streams require rapid processing and analysis in order to identify trends, anomalies, or patterns in the data.

Overall, the use of HPC systems in big data analysis enables organizations to process and analyse massive amounts of data quickly and efficiently, allowing for data-driven decision making and innovation.

Other important considerations in the design of an HPC system for big data analytics include data management, security, and fault tolerance. A distributed file system such as Hadoop Distributed File System (HDFS) can be used to manage the storage and retrieval of large data sets. Security measures such as encryption and access control should be implemented to protect sensitive data. Finally, fault tolerance measures such as redundant hardware and data backups should be put in place to ensure system availability and data integrity in the event of hardware or software failures.

Overall, the design of an HPC system for big data analytics requires careful consideration of hardware and software components, as well as data management, security, and fault tolerance measures. A well-designed HPC system can enable organizations to process and analyze massive amounts of data quickly and efficiently, enabling data-driven decision making and innovation.

II. Literature Survey

Authors has addressed several challenges when collocating Big Data and HPC resources. One of the main challenges is the different types of data storage required by Big Data and HPC applications. Big Data applications typically require distributed file systems such as Hadoop Distributed File System (HDFS), while HPC applications often require parallel file systems such as Lustre [1].

BFC is a high performance distributed big file cloud storage system that is based on key-value store. It provides a scalable and efficient solution for storing large files in the cloud. The system is designed to be fault-tolerant and highly available, with support for data replication and data recovery [2].

Monitoring high performance computing (HPC) systems is crucial to ensure that they are running efficiently and effectively. End users need to be able to monitor the performance of their applications and track the utilization of system resources such as CPUs, memory, storage, and network bandwidth [3].

Sandia National Laboratories (SNL) provides a one-stop High Performance Computing (HPC) user support service to its users. This service is designed to help users access the HPC resources available at SNL and maximize their computing efficiency [4].

In the context of high-performance computing, repeatability and reproducibility are especially important due to the complex and specialized nature of these systems. Researchers must ensure that their experiments and simulations are repeatable and reproducible in order to enable verification, validation, and comparison of results [5].

Agile Condor® is a software platform that allows researchers to run complex simulations and machine learning models on high-performance computing (HPC) resources. The platform is designed to make it easier for researchers to access and utilize HPC resources, which can be difficult and time-consuming to set up and manage [6].

Preparing the HPC workforce requires a collaborative effort between academic institutions, industry, and government agencies. By investing in educational and training programs, promoting diversity and inclusivity, and creating a supportive work environment, we can create a highly skilled workforce that can drive innovation and advance scientific discovery [7].

The trend towards a hybrid era in platforms for big data analytics reflects the need for organizations to balance their data processing needs with their budgetary and security requirements. By leveraging both on-premise and cloud-based solutions, and using open-source platforms, organizations can build flexible and scalable big data analytics solutions that meet their specific needs [8].

The case of noise level measurements at the Roskilde Music Festival demonstrates the potential of IoT and big data analytics in addressing real-world problems such as environmental pollution. By collecting and analysing data from IoT sensors, organizations can gain valuable insights and make data-driven decisions to improve the quality of life for people [9].

The four-layer architecture is designed to handle the challenges of big data analytics by providing a scalable, flexible, and modular framework that can be adapted to different use cases and requirements. It allows organizations to efficiently manage and analyse large amounts of data and gain valuable insights to drive business decisions [10].

The convergence of HPC and BDA presents both challenges and opportunities. To design an HPC system for BDA, it is important to consider factors such as data storage, parallel processing, networking infrastructure, software tools, and skilled personnel [11].

III. Requirements for designing the High-Performance Computing System for Big Data Analytics

Designing a high-performance computing system for big data analytics involves several key considerations. Here are some steps that can help in the design process:

i) Determine the data storage requirements: The first step in designing a high-performance computing system for big data analytics is to determine the storage requirements. Big data sets require a lot of storage, so it is important to select the right storage technology that can handle large data volumes and provide fast access times.

ii) Choose the right computing architecture: After determining the storage requirements, the next step is to choose the right computing architecture. There are several architectures to choose from, including shared memory, distributed memory, and hybrid architectures. Each architecture has its own advantages and disadvantages, so it is important to select the one that best suits the needs of the big data analytics project.

iii) Select the right software stack: The software stack is an important aspect of a high-performance computing system for big data analytics. The software stack should include tools and libraries for data processing, machine learning, and visualization.

iv) Choose the right hardware components: The hardware components are also an important consideration when designing a high-performance computing system for big data analytics. The system should have high-speed processors, large amounts of memory, and fast networking capabilities to handle the data processing needs of the project.

v) Implement efficient data processing algorithms: Efficient data processing algorithms are critical to the success of big data analytics projects. These algorithms should be designed to take advantage of the high-performance computing system's capabilities to process data quickly and accurately.

vi) Implement data security measures: Data security is an important consideration when designing a high-performance computing system for big data analytics. The system should be designed with security in mind and include measures such as encryption, access controls, and audit trails to protect sensitive data.

vii) Test and validate the system: After designing and implementing the high-performance computing system for big data analytics, it is important to test and validate the system to ensure that it meets the project requirements. This may involve benchmarking the system and performing stress tests to evaluate its performance under different workloads.

Understanding the High-Performance Data Analytics

High Performance Data Analytics (HPDA) is an emerging field that focuses on using advanced computing technologies, such as high-performance computing (HPC), parallel computing, distributed systems, and machine learning algorithms to analyse large and complex data sets.

HPDA enables organizations to process, analyse, and gain insights from massive amounts of data, at a scale and speed that was previously impossible. HPDA systems are designed to handle structured and unstructured data, and to support real-time analysis, predictive modelling, and data visualization.

HPDA is used in a variety of industries, such as finance, healthcare, retail, manufacturing, and energy, to identify trends, patterns, and anomalies in data, and to inform decision making, product development, and customer engagement strategies.

One of the key features of HPDA is its ability to combine multiple data sources, including structured and unstructured data, streaming data, and historical data, to provide a holistic view of an organization's operations and performance. This enables organizations to gain a deeper understanding of their customers, products, and business processes, and to identify new opportunities for growth and innovation.

Big Data Analytics is possible with the help of high-performance computing systems using the following techniques:

i) Graph Analytics involves modelling and analysing complex networks, such as social networks or transportation networks, to identify patterns and insights.

ii) Compute Intensive Analytics uses innovative techniques, such as parallel processing, to solve computationally intensive problems, such as machine learning algorithms or simulations.

iii) Streaming Analytics is used to rapidly analyse high-bandwidth and high-throughput streaming data, such as sensor data or financial transactions, in real-time, using new algorithms that can handle the velocity, volume, and variety of the data.

iv) Exploratory Data Analysis is used to analyse massive streaming data sources, such as social media data or website traffic data, to discover new insights and patterns, and to inform business decisions.

High performance computing systems are designed to handle large and complex data sets, and to provide fast and efficient processing capabilities. By using these techniques, organizations can gain valuable insights from their data, improve their decision-making processes, and drive innovation and growth.

IV. Comparison of High-Performance Computing System and Big Data

High-performance computing (HPC) and big data are two different computing paradigms that have unique architectural requirements.

HPC systems are designed to process large amounts of data in a parallel and distributed manner. These systems typically consist of tightly-coupled clusters of servers or nodes that work together to perform complex computations. HPC systems require specialized hardware, such as high-speed interconnects, large memory capacity, and high-performance processors, to support high-speed data processing.

On the other hand, big data systems are designed to handle large volumes of data from various sources, both structured and unstructured. These systems require a distributed architecture, with data stored across multiple nodes, to support efficient data processing. Big data systems typically use commodity hardware, such as clusters of commodity servers, to keep costs low and enable easy scalability.

When it comes to data processing, HPC systems are optimized for performing complex calculations on a smaller set of data, while big data systems are optimized for processing large volumes of data with simple computations. HPC systems are typically used in scientific computing, simulation, and modelling applications, while big data systems are used in business intelligence, data mining, and machine learning applications.

In summary, HPC and big data are two distinct computing paradigms with different architectural requirements. HPC systems are designed for high-speed data processing of complex calculations, while big data systems are designed to handle large volumes of data from various sources using commodity hardware. Figure 1. represents the architectural comparison of HPC and big data analytics.

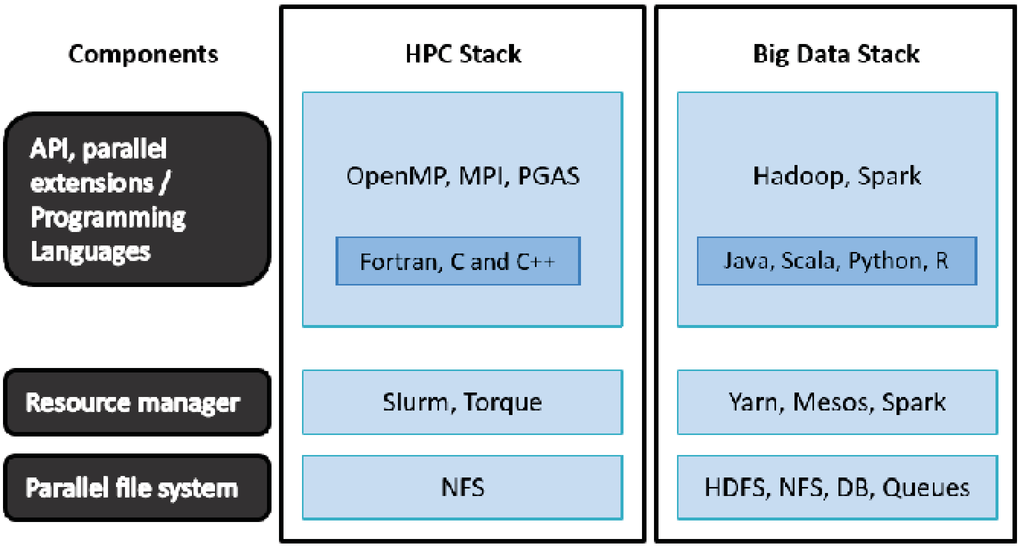

Figure 1: Comparison of HPC with Big Data Analytics

V. Challenges in Big data analytics using High Performance Computing

Big data analytics involves processing large and complex data sets, and it requires high computing capabilities to derive meaningful insights from the data. High performance data analytics (HPDA) aims to combine the capabilities of high-performance computing (HPC) with big data analytics to address the challenges posed by big data. However, big data analytics poses many challenges at both the micro and macro level. At the micro level, there are issues related to statistical modelling of big data, such as how to handle missing data, outliers, and non-normal data distributions. This requires specialized techniques for data pre-processing, feature selection, and model selection.

At the macro level, big data analytics is challenged by the complexities of effective computational prototypes. This involves selecting appropriate hardware and software architectures, designing efficient algorithms for data processing, and developing scalable solutions that can handle the large volumes of data. Additionally, there are challenges related to data storage, data integration, and data privacy and security. To address these challenges, organizations need to invest in advanced technologies and tools for big data analytics, including high-performance computing systems, cloud computing platforms, machine learning algorithms, and data visualization tools. They also need to hire skilled professionals who can design and implement effective solutions for big data analytics.

VI. Need to design high performance computing system for big data analytics

Designing a high-performance computing system for big data analytics is important for several reasons:

i) Data volume: Big data analytics involves processing and analysing large volumes of data, often in the petabytes or even exabytes range. A high-performance computing system can handle this large volume of data efficiently and quickly.

ii) Processing speed: Big data analytics requires processing data in real-time or near real-time. High performance computing systems are designed to process data at a high speed, ensuring quick turnaround times for data analysis.

iii) Complexity of algorithms: Big data analytics involves complex algorithms, which can take a long time to run on traditional computing systems. High performance computing systems are designed to handle complex algorithms and can speed up data processing.

iv) Scalability: As data volumes continue to grow, it is important to have a computing system that can scale up to handle the increased workload. High performance computing systems can scale up easily and efficiently.

Overall, a high-performance computing system is essential for big data analytics to enable businesses to make informed decisions based on data-driven insights.

Conclusion

Indeed, the convergence of High-Performance Computing and Big Data Analytics has become increasingly important in the era of data revolution. The continuous growth of data requires efficient and effective management and processing to extract meaningful insights for various industries and applications. The evolution of data storage technologies and computing models has led to the emergence of new paradigms such as Real Time Analytical Framework, which caters to the need for handling the continuous flow of real data.

However, the convergence of these paradigms also poses several challenges, such as the need to develop efficient data management and computing models that can handle the complex requirements of data-intensive applications. In addition, the evolution of traditional analytical paradigms to cater to the demands of High-Performance Computing and Big Data Analytics requires a sustainable solution that can handle the computational requirements of newer models.

Overall, the convergence of High-Performance Computing and Big Data Analytics is essential for the development of a sustainable solution that can effectively manage and process large amounts of data to extract meaningful insights for various applications.

References

[1] M. Mercier, D. Glesser, Y. Georgiou and O. Richard, "Big data and HPC collocation: Using HPC idle resources for Big Data analytics," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 347-352, doi: 10.1109/BigData.2017.8257944.

[2] T. T. Nguyen, T. K. Vu and M. H. Nguyen, "BFC: High-performance distributed big-

file cloud storage based on key-value store," 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Takamatsu, Japan, 2015, pp. 1-6, doi: 10.1109/SNPD.2015.7176209.

[3] C. L. Moore, P. S. Khalsa, T. A. Yilk and M. Mason, "Monitoring High Performance Computing Systems for the End User," 2015 IEEE International Conference on Cluster Computing, Chicago, IL, USA, 2015, pp. 714-716, doi: 10.1109/CLUSTER.2015.124.

[4] J. A. Greenfield, L. G. Ice, S. E. Corwell, K. Haskell, C. Pavlakos and J. P. Noe, "One stop high performance computing user support at SNL," SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, WA, USA, 2011, pp. 1-6, doi: 10.1145/2063348.2063383.

[5] D. R. C. HILL, "Repeatability, Reproducibility, Computer Science and High Performance Computing : Stochastic simulations can be reproducible too…," 2019 International Conference on High Performance Computing & Simulation (HPCS), Dublin, Ireland, 2019, pp. 322-323, doi: 10.1109/HPCS48598.2019.9188157.

[6] M. Barnell, C. Raymond, C. Capraro, D. Isereau, C. Cicotta and N. Stokes, "High-Performance Computing (HPC) and Machine Learning Demonstrated in Flight Using Agile Condor®," 2018 IEEE High Performance extreme Computing Conference (HPEC), Waltham, MA, USA, 2018, pp. 1-4, doi: 10.1109/HPEC.2018.8547797.

[7] Lathrop, Scott. (2016). A Call to Action to Prepare the High-Performance Computing Workforce. Computing in Science & Engineering. 18. 80-83. 10.1109/MCSE.2016.101.

[8] A. Londhe and P. P. Rao, "Platforms for big data analytics: Trend towards hybrid era," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 3235-3238, doi: 10.1109/ICECDS.2017.8390056.

[9] T. -M. Groenli, B. Flesch, R. Mukkamala, R. Vatrapu, S. Klavestad and H. Bergner, "Internet of Things Big Data Analytics: The Case of Noise Level Measurements at the Roskilde Music Festival," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 5153-5158, doi: 10.1109/BigData.2018.8622406.

[10] J. Y. Zhu, J. Xu and V. O. K. Li, "A Four-Layer Architecture for Online and Historical Big Data Analytics," 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Auckland, New Zealand, 2016, pp. 634-639, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.115.

[11] Bomatpalli, Tulasi & Wagh, Rupali & S, Balaji. (2015). High Performance Computing and Big Data Analytics Paradigms and Challenges. International Journal of Computer Applications. 116. 28-33. 10.5120/20311-2356.