PERFORMANCE COMPARISON OF MAPREDUCE APPROACHES FOR POSSIBLE USES IN INDUSTRY 4.0.

Ashish Batchu
Department of CSE,
SRM University-AP, India,
ashish_batchu@srmap.edu.in

Sricharan Gudi
Department of CSE,
SRM University-AP, India,
sricharan_gudi@srmap.edu.in

Saidhanush Rayapureddy
Department of CSE,
SRM University-AP, India,
saidhanush_r@srmap.edu.in

Haasitha Ambati
Department of CSE,
SRM University-AP, India,
haasitha_ambati@srmap.edu.in

Arnab Mitra
Assistant Professor,
Department of CSE,
SRM University-AP, India,
arnab.m@srmap.edu.in

Abstract:

We are surrounded by a data environment where information is always growing rapidly due to many data generating elements such as sensors, healthcare, online shopping, social media, retail, hospitality, finance, airlines, security surveillance, and so on. Big data refers to the gathering of datasets that are inaccessible by typical database management systems, as well as data that exceeds storage capacity and computing capability. Hadoop is a new technology for analysing data volumes that are quickly expanding in Gigabytes, Terabytes, Petabytes, Zetabytes, and so on. The purpose of the current paper is to explore Hadoop and related technologies, with an emphasis on MapReduce and an examination of university research data sets, as well as a technical comparison of OpenMP vs MPI vs Hadoop.

Keywords: Big Data, MapReduce, Industry 4.0, OpenMP, MPI, WordCount, Hadoop

I. Introduction:

The concept of Industry 4.0 is regarded as an industrial era in which the integration of both vertical and horizontal production procedures, as well as product connectivity, can assist businesses in achieving improved industrial performance . It includes technology such as industrial Internet of Things networks, artificial intelligence, big data, robots, and automation [1]. Through the eyes of practitioners and experts with an in-depth understanding of Industry 4.0 ideas and technologies is discussed in [2]. Industry 4.0, a result of previous industrial periods, involves increased digitisation, automation, and communication through a digital value chain. It includes IoT, cloud computing, as well as digital industrial systems. BMW, Jaguar Land Rover, and Mondelez have all utilised it to improve operational efficiency. Industry 4.0 offers opportunities for innovation and strategic advantage, but a rigorous approach to disruption is still lacking [3] . Big data describes complicated, huge data sets that are utilised for analytics in order to identify hidden patterns and provide important insights to businesses and organisations while also resolving privacy concerns [4] .

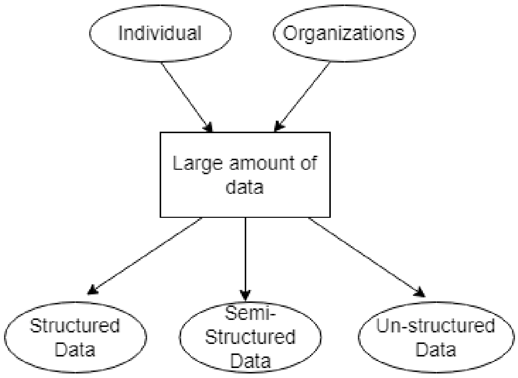

Fig:1 Various types of data.

Structured data, semi- structured information, and unstructured information are the three types of data. Structured data is acquired by sensors or intelligent/automatic sources, whereas semi-structured data is collected via lab instruction manuals, machine logbooks, and other documents. The required data in unstructured data is collected in the form of photos, diagrams, and drawings. All of the data types specified in the preceding data sets are essentially unused.

This article examines big data applications via the frameworks of structuralism and functionalism, examining cutting-edge technologies, analytics methodologies, studies in progress, open research challenges, and prospects, and recommending forthcoming technologies for big data problems [5] . To facilitate the usage of big data platforms in the Industry 4.0 area, the study presents a visual and dataflow-based architectural framework. This framework enables programmers to quickly integrate data mining and machine learning techniques into business process monitoring and improvement operations, thereby overcoming the complexities of traditional big data analytics systems. This strategy intends to increase the industry's embrace of big data technologies [6] . The integration of BD with Industry 4.0 has led to the development of wearable medical devices and sensors, collecting vast amounts of data. This data, known as "Big Data," is crucial for efficient and secure management in integrated industries. The study aims to present a comprehensive survey of studies on BD in healthcare [7] . MapReduce is a programming model for processing large datasets, used by Google for processing over twenty petabytes of data daily, with over ten thousand programs implemented internally [8] . This paper provides an overview of MapReduce research achievements and bibliometric analytics techniques for evaluating the growing body of knowledge on the topic. The article explores MapReduce framework enhancements for huge data sets with concurrent distributed algorithms, as well as future research possibilities [9] . This study investigates the performance of MapReduce on a 100-node in an Amazon EC2 cluster, highlighting five design elements that influence Hadoop's total performance. By fine-tuning these parameters, overall performance can be increased by 2.5 to 3.5 times, bringing Hadoop closer to parallel database systems and showcasing the promise for a freely accessible cloud data processing system [10] . This study provides an affordable MapReduce model for Industry 4.0 based on Cellular Automata, lowering the current 36 CA rules to two for a sustainable implementation. [11] . This research article presents a MapReduce design using Equal Length Cellular Automata (ELCA) for efficient data processing and cost-effective implementation in heterogeneous Cloud architecture [12].

II. Related Works:

Past Works on Industry 4.0:

To properly describe the current Industrial Revolution, one must first comprehend or see the three prior industrial revolutions. Table 1 describes previous industrial revolutions[13].

Industrial Revolution	Era	Restorations
Revolution I	18^{th} & 19^{th} Century.	The development of steam engines and the beginning of mechanical manufacturing.
Revolution II	At the beginning of the 20^{th} century.	Telegraphs, industrial equipment, and mass manufacturing are all introduced.
Revolution III	21^{st} Century.	Digital technology period.

Table 1: Historical industrial revolutions and restorations

The above information discusses many Industrial Revolutions that have revolutionised our modern civilization in a gradual pace. Mechanical production (Revolution I), mass production (Revolution II), and the digital age (Revolution III) were the primary developments for a positive-attitude society. Similarly, the term "Industry 4.0" refers to the fourth Industrial Revolution. It may appear to be a perfect synthesis of previous revolutions, but Industry 4.0 will be far more varied and significant than that.

The so-called fourth industrial revolution is in its early stages, with enhanced information technology providing the essential information throughout manufacturing. AI and IoT technologies enable devices to collaborate with one another in order to improve output.

Important works on big data in 4.0:

The primary advantage of big data in Industry 4.0 is the ability to collect statistics or information from within the system. This type of in-process surveillance ensures that identification, evaluation, and correction are achievable without requiring involvement in the production line during automation and manufacture. Based on previous data from the facility, big data applications are used to forecast any type of failure or equipment overload. It is critical in lowering expenses like maintenance and downtime. This assists industries with many types of real-time manufacturing and assurance data.

After receiving the data, it is filtered, normalised, and structured before being sorted into significant correlations that may be used in scientific research. As with any system, dependable resources for data will yield reliable data for decision-making.

III. Background:

The study of ECA-based MapReduce design in Industry 4.0 focused on a detailed empirical examination of 36 elementary cellular automata (ECA) rules. Finding suitable configurations for an ECA-based MapReduce model while taking into consideration both homogeneous and heterogeneous CA setups was the objective of the work. Uniform transition lengths under various fixed boundary conditions were part of the study, and correlation-based studies were used to evaluate shuffle quality, an important consideration for Industry 4.0 applications[15].

There were two primary components to the empirical analysis. Examining homogeneous and heterogeneous CA structures under various fixed boundary conditions was one component of the empirical investigation. A review of previous research helped to clarify several CA dynamics. The concept of maximal correlation was introduced in another component, which consists of correlation-based studies, it highlighted the importance of shuffling in MapReduce design, and indicated suitable state spaces for MapReduce applications[16,17,18,19].

Following an analysis of the data, it became evident that CA topologies with state spaces of four or more and low correlation values had the potential to be effectively designed MapReduce systems. Two major claims were made with regard to these findings:

Claim 1: For CA-based MapReduce architectures in Industry 4.0, homogeneous CA setups with rules 102 or 153 were considered suitable due to their consistent dynamics across a range of boundary conditions and correlation coefficient values.

Claim 2: A simple matrix-based study was suggested to investigate the characteristics of the CA-based MapReduce architecture, focusing on the selection of additive CA rules 102 and 153. [15,16,17,18].

Research on CA-based MapReduce design in Industry 4.0 benefited greatly from the study's useful data. Future research and applications of CA frameworks in the dynamic field of Industry 4.0 data processing are made possible by the claims made, empirical analysis, and correlation-based evaluations [11].

IV. Proposed Work:

Performance Comparison of MapReduce, OpenMP, and MPI in Word Count:

MapReduce:

MapReduce is intended for large-scale distributed processing. MapReduce provides strong scalability and failure tolerance in the Word Count example. The programming approach uses simple map and reduce processes, making it intuitive. The computational cost of processing tiny jobs in a distributed system, on the other hand, may have an influence on performance, particularly for smaller datasets. MapReduce is ideal for batch processing, but it may be inefficient for iterating algorithms or operations requiring low-latency processing.Furthermore, while MapReduce excels at dealing with large datasets, its strict two-step processing architecture may not be suitable for sophisticated iterative algorithms. Real-time data processing and low-latency needs may necessitate the investigation of alternative distributed computing frameworks such as Apache Spark, which provides in-memory processing for enhanced performance in certain cases. The suitable framework is chosen based on the specific requirements and features of the data handling operations at hand.

OpenMP (Open Multi-Processing):

OpenMP is a parallel programming approach for shared memory. OpenMP is ideal for multi-core systems in the Word Count example. It has high scalability for moderate-sized datasets, exploiting shared-memory parallelism efficiently. When faced with bigger data sets that surpass the capability of a single system, it may experience scaling problems. OpenMP is simple to program and is appropriate for activities that can be complemented across several threads on a single node.When faced with distributed computing scenarios or jobs that need communication between various nodes, OpenMP's effectiveness suffers. Because of its simplicity and ease of usage, it is ideal for applications with modest parallelism that may be effectively parallelized within a single system. As computing architectures change, OpenMP remains a crucial tool for leveraging parallelism on shared-memory systems, particularly in the context of current multi-core processors.

MPI (Message passing interface):

MPI is a distributed-memory parallelism protocol that is commonly used in high-performance computing clusters. In the Word Count example, MPI is capable of handling huge datasets dispersed across nodes efficiently. MPI has high scalability and is ideal for jobs that need communication between nodes. MPI programs, on the other hand, frequently include more complex code than OpenMP programs since developers must explicitly control data distribution and synchronisation. MPI shines in situations where data cannot fit into a single machine's memory, making it appropriate for large-scale parallel execution. In addition, because MPI allows for precise control over communication and synchronization, it is well-suited for applications with complex dependencies and irregular processing patterns.

While MPI succeeds in high-end computing environments, its complexity may result in increased development and maintenance expenses. As technology advances, hybrid techniques combining MPI with other parallel programming models are becoming more investigated in order to find a compromise between performance and accessibility of development in distributed computing.

Map reduce Algorithm :

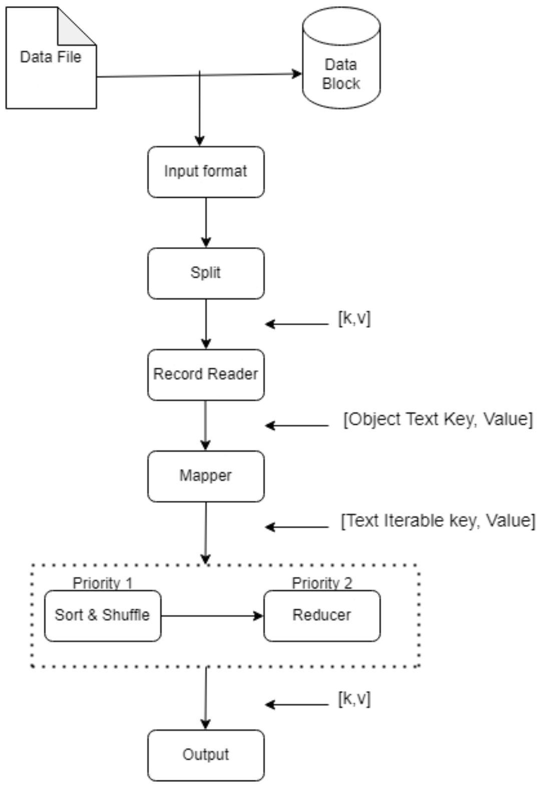

Fig 4.1 : Flowchart of the MapReduce process

Step1: Input queries should be in the form of.JAR files including Driver, Mapper, and Reducer code.

Step2: The mapper tasks are assigned by Job Tracker by monitoring the company's logic from the.JAR file on all accessible task trackers.

Step3: When all the task trackers have completed the mapping operations, they provide the same condition to Job Tracker.

Step4: The entire set of task monitors complete the mapper phase, after which the job tracker begins the sort and shuffle phase on all mapper outputs.

Step5: If the sort and shuffle process is completed, the job tracker begins the reduction phase on all accessible task trackers.

The MapReduce algorithm tokenizes text, reorganises and sorts subsequent key-value pairs for each word, and decreases word counts across all documents. This algorithm is used in distributed computing systems to solve the Word Count issue, where it is executed over numerous nodes to process massive data volumes.

Performance of Map reduce:

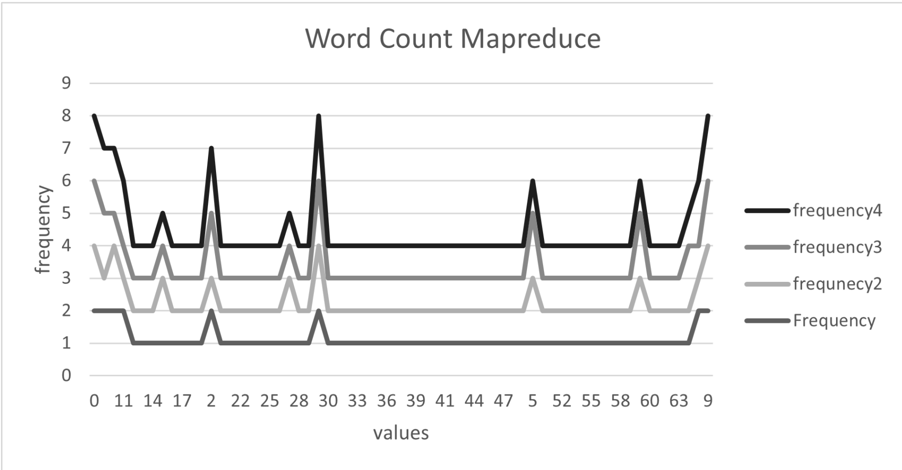

Fig 4.2 : Graphical representation of four cycles from the dataset

The graph is plotted between the values and frequency in which they are repeated from each cycle. The above graph is taken for four different cycles, It gives a good understanding of the dataset. The study investigates state spaces for homogeneous Equal Length Cellular automata (ELCAs) under various fixed boundary conditions at automaton sizes 8, 10, and 12. The findings were reported in the Journal of Cellular Automata and MaxCA. The ELCAs obtained at automaton size 7, 8 for homogeneous CAs were found to be identical to the state spaces described in the earlier research. The research design, comprehension, and presentations are novel and have not been cited previously[11]. The dataset is saved locally, cleaned, and then uploaded to HDFS. To identify the research field, the WordCount module is written in Java, and the data has been tokenized using the MapReduce technique. The top 14 keywords with the most occurrences are chosen from each dataset. The output is saved to HDFS before being transferred back to the local file systems.

Correlational Analysis:

Each framework displays distinct traits in a correlational analysis of MapReduce, OpenMP, and MPI for parallel computing. MapReduce, which is designed for distributed processing, excelled at processing large-scale data analytics workloads while remaining fault tolerant and scalable. OpenMP, a shared-memory multiprocessing API, provides parallel programming efficiency for shared-memory architectures, which makes it suited for jobs that may be complemented within a single machine. MPI is a message-passing standard designed for distributed memory systems, allowing communication between cluster nodes. While

MapReduce excels in large-scale data processing, OpenMP and MPI address various parallelization demands, with OpenMP focusing on shared-memory parallelism and MPI on distributed-memory parallelism. The framework of choice is determined by the precise requirements and characteristics of the computer task at hand.

V. Results:

In the context of the Word Count problem, MapReduce outperforms OpenMP and MPI, particularly in terms of scalability, fault tolerance, and simplicity of programming. MapReduce excels in handling huge datasets in terms of scalability by effectively allocating workloads across a cluster of processors. The framework's intrinsic fault tolerance, achieved by data replication, ensures that operations continue even when hardware fails, adding to greater reliability. Furthermore, MapReduce has a high-level interface that accelerates programming, thus rendering it more accessible to a broader variety of users and allowing for the relatively easy development of parallelized solutions. This feature is especially useful for activities like Word Count, when simplicity and effectiveness are critical. While OpenMP is appropriate for shared-memory parallelism on a single machine and MPI succeeds in distributed-memory environments, MapReduce's advantages lay in its capability to handle data-intensive processes with distributed datasets, making it the preferred choice for scenarios requiring large-scale processing.

Conclusion:

The research looks into Hadoop and similar technologies, with a focus on MapReduce and university research data sets. It examines the rising data environment and compares OpenMP, MPI, and Hadoop, emphasising the growing relevance of big data and the need for better database management solutions. A Word Count MapReduce implementation entails mapping input words to key-value pairs, shuffling and sorting intermediate data, then reducing to provide a list of unique words and their counts. This distributed paradigm enables parallel processing across a cluster, increasing the scalability and efficiency of word counting for large datasets.The shuffle and repeating of succeeding words structure has been described as a unique way for enabling effective scalable data handling for Hadoop MapReduce in big data applications. It has been documented for one or more files on a single machine, as well as several files on multiple systems using a distributed file system. Both options can yield a comparable output file from the same or unique file after Word Count and correct shuffling in a phrase in a paragraph or entire file. It is scalable and efficient for better word processing applications such as many identical words, word combinations, and comparable phrases in file(s), including remote file systems.

References:

[1] Dalenogare, L. S., Benitez, G. B., Ayala, N. F., & Frank, A. G. (2018). The expected contribution of Industry 4.0 technologies for industrial performance. *International Journal of production economics*, *204*, 383-394.

[2] Benitez, G. B., Lima, M. J. D. R. F., Lerman, L. V., & Frank, A. G. (2019). Understanding Industry 4.0: Definitions and insights from a cognitive map analysis. *Brazilian Journal of Operations & Production Management [recurso eletrônico]. Rio de Janeiro, RJ. Vol. 16, no. 2 (June 2019), p. 192-200*.

[3] Papadopoulos, T., Singh, S. P., Spanaki, K., Gunasekaran, A., & Dubey, R. (2022). Towards the next generation of manufacturing: implications of big data and digitalization in the context of industry 4.0. *Production Planning & Control*, *33*(2-3), 101-104.

[4] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

[5] Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231-1247.

[6] Gokalp, M. O., Kayabay, K., Akyol, M. A., Eren, P. E., & Koçyiğit, A. (2016, December). Big data for industry 4.0: A conceptual framework. In *2016 international conference on computational science and computational intelligence (CSCI)* (pp. 431-434). IEEE.

[7] Karatas, M., Eriskin, L., Deveci, M., Pamucar, D., & Garg, H. (2022). Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. *Expert Systems with Applications*, *200*, 116912.

[8] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

[9] Hashem, I. A. T., Anuar, N. B., Gani, A., Yaqoob, I., Xia, F., & Khan, S. U. (2016). MapReduce: Review and open challenges. *Scientometrics*, *109*, 389-422.

[10] Jiang, D., Ooi, B. C., Shi, L., & Wu, S. (2010). The performance of mapreduce: An in-depth study. *Proceedings of the VLDB Endowment*, *3*(1-2), 472-483.

[11] Mitra, A. (2021). On the capabilities of cellular automata-based MapReduce model in industry 4.0. *Journal of Industrial Information Integration*, *21*, 100195.

[12] Mitra, A., Kundu, A., Chattopadhyay, M., & Chattopadhyay, S. (2018). On the exploration of equal length cellular automata rules targeting a mapreduce design in cloud. *International Journal of Cloud Applications and Computing (IJCAC)*, *8*(2), 1-26.

[13] Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2021). Significant applications of big data in Industry 4.0. *Journal of Industrial Integration and Management*, *6*(04), 429-447.

[14] Sahane, M., Sirsat, S., & Khan, R. (2015). Analysis of Research Data using MapReduce Word Count Algorithm. *Internl. Journal of Advanced Research in Computer and Commn. Engg*, *4*.

[15] Mitra, A., Kundu, A., Chattopadhyay, M., & Chattopadhyay, S. (2018). On the exploration of equal length cellular automata rules targeting a mapreduce design in cloud. *International Journal of Cloud Applications and Computing (IJCAC)*, *8*(2), 1-26.

[16] Teodorescu, H. N. (2015). On the regularities and randomness of the dynamics of simple and composed CAs with applications. *Romanian Journal of Information Science and Technology, Romanian Academy*, *18*(2), 166-181.

[17] Mitra, A., & Teodorescu, H. N. (2016). Detailed Analysis of Equal Length Cellular Automata with Fixed Boundaries. *Journal of Cellular Automata*, *11*.

[18] Mitra, A. (2016). On the selection of cellular automata based PRNG in code division multiple access communications. *Studies in Informatics and Control*, *25*(2), 218.