Role of Parallel Data Processing in Big Data

Mr. S.B. Khandagale[1,] Dr. Gitanjali Sinha[2], Dr. B. T.Jadhav[3]
[1]*Research Scholar, School Of IT, Mats University , Raipur (C.G.)*
[2]*School Of IT , Mats University , Raipur (C.G.)*
[3]*Yashavantrao Chavan Institute of Science ,(Autonomous) Satara , Maharashtra*

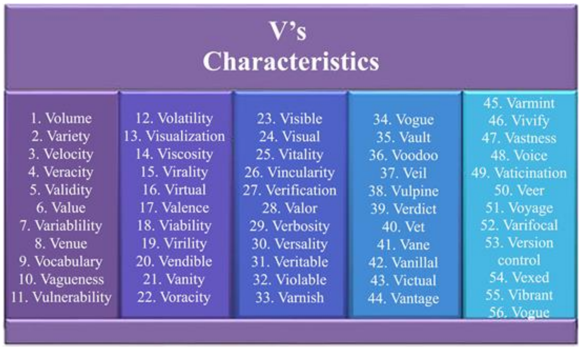
***Abstract*- Parallel processing is a method in computing of running two or more processors (CPUs) to handle separate parts of an overall task. Breaking up different parts of a task among multiple processors will help reduce the amount of time to run a program.**

Parallel processing uses two or more processors or CPUs simultaneously to handle various components of a single activity.

Parallel processing is a technique in which a large process is broken up into multiple, smaller parts, each handled by an individual processor.

INTRODUCTION

BIG DATA


V's
Characteristics

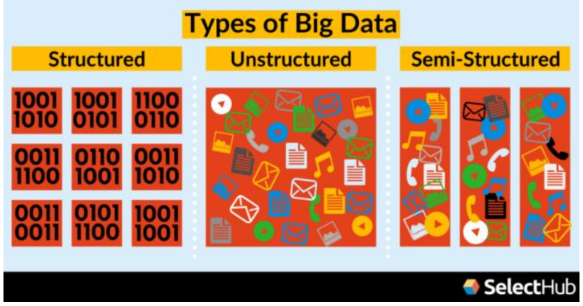
1. Volume
2. Variety
3. Velocity
4. Veracity
5. Validity
6. Value
7. Variablility
8. Venue
9. Vocabulary
10. Vagueness
11. Vulnerability
12. Volatility
13. Visualization
14. Viscosity
15. Virality
16. Virtual
17. Valence
18. Viability
19. Virility
20. Vendible
21. Vanity
22. Voracity
23. Visible
24. Visual
25. Vitality
26. Vincularity
27. Verification
28. Valor
29. Verbosity
30. Versality
31. Veritable
32. Violable
33. Varnish
34. Vogue
35. Vault
36. Voodoo
37. Veil
38. Vulpine
39. Verdict
40. Vet
41. Vane
42. Vanillal
43. Victual
44. Vantage
45. Varmint
46. Vivify
47. Vastness
48. Voice
49. Vaticination
50. Veer
51. Voyage
52. Varifocal
53. Version control
54. Vexed
55. Vibrant
56. Vogue

Fig 1.1

Big data refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around for a long time .

Fig 1.1 shows different characteristics of big data like
1 .Volume:- Store Data
2.Velocity:- Speed of Data generation
3.Variety:- Different types of data
4.Veracity:- Data Accuracy
5.Value:-Useful data
6.Validity:-Data quality
7.Variability:-Dynamic and existing behavior of Data etc.

Types of big data are structured, unstructured, semi structured



Structured Data:
Structured data is the neatly organized data you keep in databases, datasets, and spreadsheets. It's easy for traditional analytics tools to read this data. Organizing unstructured data into structured data is time-consuming, but possible with the right solution.

Unstructured data :
Unstructured simply means that it is datasets (typical large collections of files) that aren't stored in a structured database format. Unstructured data has an internal structure, but it's not predefined through data models. It might be human generated, or machine generated in a textual or a non-textual format

Semi structured data:-
Semi-structured data refers to data that is not captured or formatted in conventional ways. Semi-structured data does not follow the format of a tabular data model or relational databases because it does not have a fixed schema

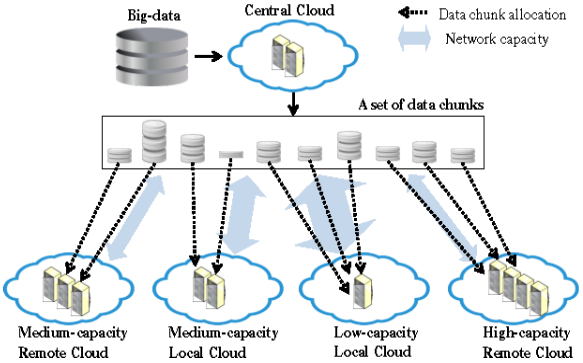
Big data stored all types of data with all characteristics mention in Fig 1.1 It is very complicated task to manage all data

That is Big data has more space required and within time large amount of data must be processed. To manage data processing fast and accurate parallel data processing is used

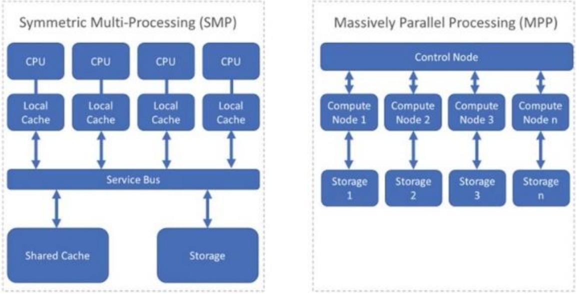
In the big data era, datasets can grow in enormous sizes and therefore it may be impossible to load them into a single machine. In parallel computing, such datasets can take advantage of multiple computer machines in order to load them in a distributed fashion, by partitioning them

PRALLEL DATA PROCESSING

Parallel data processing saves time, allowing the execution of applications in a shorter wall-clock time. Solve Larger Problems in a short point of time. Compared to serial computing, parallel computing is much better suited for modeling, simulating and understanding complex, real-world phenomena



Parallel processing is a computing technique when multiple streams of calculations or data processing tasks co-occur through numerous central processing units (CPUs) working concurrently.

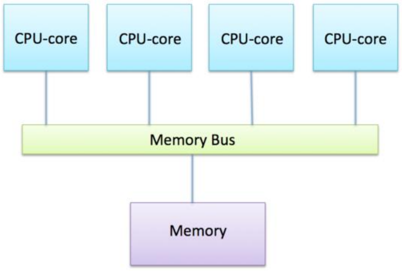


Systems can slash a program's execution time by dividing a task's many parts among several processors. Multi-core processors, frequently found in modern computers, and any system with more than one CPU are capable of performing parallel processing.

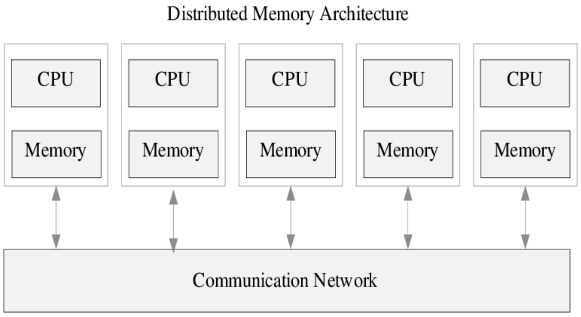
Parallel computing is becoming critical as more Internet of Things (IoT) sensors, and endpoints need real-time data. Given how easy it is to get processors and GPUs (graphics processing units) today through cloud services, parallel processing is a vital part of any micro service rollout

There are 3 distinct architectures. Of parallel data processing

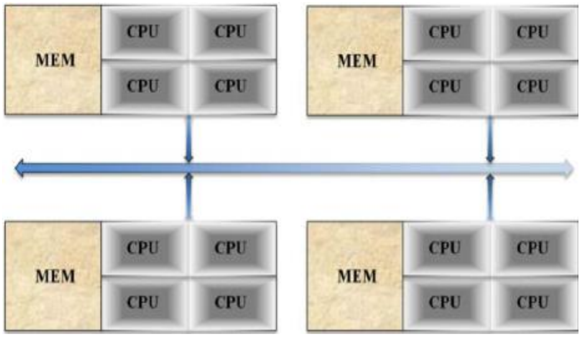
Shared memory parallel computers use multiple processors to access the same memory resources. Examples of shared memory parallel architecture are modern laptops, desktops, and smartphones.



Distributed memory parallel computers use multiple processors, each with their own memory, connected over a network. Examples of distributed systems include cloud computing, distributed rendering of computer graphics, and shared resource systems like SETI [17].



Hybrid memory parallel systems combine shared-memory parallel computers and distributed memory networks. Most "distributed memory" networks are actually hybrids. You may have thousands of desktops and laptops with multi-core processors all connected in a network and working on a massive problem.



Working

In general, parallel processing refers to dividing a task between at least two microprocessors. When processing is done in parallel, a big job is broken down into several smaller jobs better suited to the number, size, and type of available processing units. After the task is divided, each processor starts working on its part without talking to the others. Instead, they use software to stay in touch with each other and find out how their tasks are going.

After all the program parts have been processed, the result is a fully processed program segment. This is true whether the number of processors and tasks and processors were equal and they all finished simultaneously or one after the other.

Following are types of Parallel data processing

1. Single Instruction, Single Data (SISD)

In SISD a single processor is responsible for simultaneously managing a single algorithm as a single data source. A computer organization having a control unit, a processing unit, and a memory unit is represented by SISD. It is similar to the current serial computer. Instructions are carried out sequentially by SISD, which may or may not be capable of parallel processing, depending on its configuration.

2. Multiple Instruction, Single Data (MISD)

Multiple processors are standard in computers that use the Multiple Instruction, Single Data (MISD) instruction set. While using several algorithms, all processors share the same input data. MISD computers can simultaneously perform many operations on the same batch of data. As expected, the number of operations is impacted by the number of processors available.

3. Single Instruction, Multiple Data (SIMD)

Computers that use the Single Instruction, Multiple Data (SIMD) architecture have multiple processors that carry out identical instructions. However, each processor supplies the instructions with its unique collection of data. SIMD computers apply the same algorithm to several data sets. The SIMD architecture has numerous processing components.

4. Multiple Instruction, Multiple Data (MIMD)

Multiple Instruction, Multiple Data, or MIMD, computers are characterized by the presence of multiple processors, each capable of independently accepting its instruction stream. These kinds of computers have many processors. Additionally, each CPU draws data from a different data stream. A MIMD computer is capable of running many tasks simultaneously.

5. Single Program, Multiple Data (SPMD)

SPMD systems, which stand for Single Program, Multiple Data, are a subset of MIMD. Although an SPMD computer is constructed similarly to a MIMD, each of its processors is responsible for carrying out the same instructions. SPMD is a message passing programming used in distributed memory computer systems. A group of separate computers, collectively called nodes, make up a distributed memory computer.

6. Massively Parallel Processing (MPP)

A storage structure called Massively Parallel Processing (MPP) is made to manage the coordinated execution of program operations by numerous processors. With each CPU using its operating system and memory, this coordinated processing can be pplied to different program sections.

Advantages of parallel Data processing

1. Parallel computing saves time, allowing the execution of applications in a shorter wall-clock time.
2. 2Solve Larger Problems in a short point of time.
3. Compared to serial computing, parallel computing is much better suited for modeling, simulating and understanding complex, real-world phenomena.
4. Throwing more resources at a task will shorten its time to completion, with potential cost savings. Parallel computers can be built from cheap, commodity components.
5. Many problems are so large and/or complex that it is impractical or impossible to solve them on a single computer, especially given limited computer memory.
6. You can do many things simultaneously by using multiple computing resources. Can using computer resources on the Wide Area Network(WAN) or even on the internet.
7. It can help keep you organized. If you have Internet, then communication and social networking can be made easier.
8. It has massive data storage and quick data computations.

Disadvantages Of Parallel Data Processing

1. Programming to target Parallel architecture is a bit difficult but with proper understanding and practice, you are good to go.
2. The use of parallel computing lets you solve computationally and data-intensive problems using multicore processors, but, sometimes this effect on some of our control algorithm and does not give good results and this can also affect the convergence of the system due to the parallel option.
3. The extra cost (i.e. increased execution time) incurred are due to data transfers, synchronization, communication, thread creation/destruction, etc. These costs can sometimes be quite large and may actually exceed the gains due to parallelization.
4. Various code tweaking has to be performed for different target architectures for improved performance.
5. Better cooling technologies are required in case of clusters.
6. Power consumption is huge by the multi-core architectures.
7. Parallel solutions are harder to implement, they're harder to debug or prove correct, and they often perform worse than their serial counterparts due to communication and coordination overhead.

CONCLUSION

According to research paper we conclude that data can Parallel Data processing saves money, saves time, solve more complex or larger problems, quick data computations so parallel data processing is back bone of big data

REFERENCES

[1] Quoc-Cuong To, Juan Soto & Volker Markl (2018) “A survey of state management in big data processing systems”, Springar link

[2] Ejaz Ahmed, Ibrar Yaqoob, IbrahimAbaker Targio Hashem, Imran Khan, AbdelmuttlibIbrahim Abdalla Ahmed, Muhammad Imran, Athanasios V.Vasilakos (2017) ” The role of big data analytics in Internet of Things” , Computer network

[3] Ahmed Oussous , Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih (2018) “Big Data technologies: A survey” , Journal of King Saud University - Computer and Information Sciences.