The Subodh Journal of Recent Trends in Information Technology
Volume 10, Issue-02, June 2019 | ISSN No. 0975-9875

Parallel Processing for Big Data

Apoorva Verma
Assistant Professor
S.S. Jain Subodh PG College, Jaipur

ABSTRACT: The various properties of Big Data like volume variety ,veracity etc and the huge amount of valuable information it contains has resulted into development of various parallel processing systems. At present we are living in an era where data is one of the most precious resources and need to be handled with great systems. As the data being generated is so huge in amount thus it became difficult to be handled by traditional data processing methods. As a result organizations started to spend in development of new technologies and as a result parallel database systems, Map Reduce, Hadoop, Pig, Hive, Spark, and Twister were some of the products being developed. The main focus of this research paper is to scrutinize and figure out the performance of various types of parallel processing for Big Data.

KEY WORDS: Big Data, Parallel processing, Shared memory Architecture, Shared Disk Architecture, Shared Nothing Architecture

1. INTRODUCTION

Back in 90s the data generated was not so huge as a result the traditional DBMS were enough to handle and process the data. But with the emerging technologies the data now being generated changed a lot and thus the traditional systems failed to fulfill the needs. As a result more powerful system was required. For example, 20 petabytes are processed per day by Google; more than one million transactions per hour are processed by Walmart, and these transactions are more than 2.5 petabytes of data; AT&T has a 312 terabytes database which includes 1.9 trillion phone call records [1].Thus to process such a huge amount of data a powerful system was required and thus came into existence parallel processing Systems. In this paper, we presentparallel Database system and its architectures. We analyze the advantages and disadvantages of each Architecture for handling big data problems.

2. BIG DATA

Big data can be defined as extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. The major way to describe Big Data is be describing its 5V: volume, velocity, variety, veracity and value

[1]

1) Volume: Volume can be defined as the amount of data being generated in a unit time. Big data is now being generated in TB's an dPB's.

2) Velocity: Velocity can be described as the frequency or the speed with which data is being generated.

3) Variety: It describes the different types like audio, video, images, gifs and other types of data being generated in a unit oftime.

4) Veracity: Veracity describes the authentication and originality of data.

5) Value: It describes the statistical aspect of the data.

The big data is not about the quantity of data but it is basically about the outcome which you get from that bulk amount of data. You can analyze data from any source and can find the answers for 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making. When you combine big data with high-powered analytics, you can accomplish business-related tasks such as:

- Finding actual reason behind failures, issues and defects in real time.
- Generating coupons at the point of sale based on the customer's buying habits.
- Recalculating entire risk portfolios in minutes.
- Detecting behavior before it affects your organization.

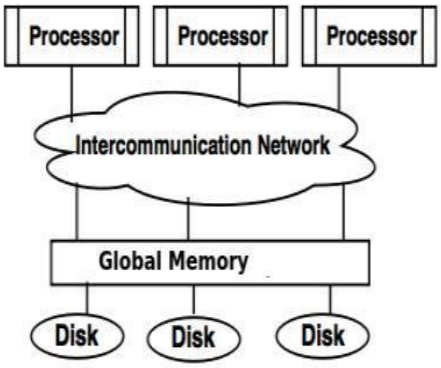
3. PARALLEL PROCESSING

Parallel processing is a type of processing in which Multiple processes are executed simultaneously. It reduces total computational time. The primary

purpose of parallel processing is to enhance the computer processing capability and increase its throughput, i.e. the amount of processing that can be accomplished during a given interval of time. Now we analyze each parallel database architectures for handling big data problems.

1. **Shared memory architecture**:

This architecture consists of various processors which are attached to a global memory via intercommunication channel.

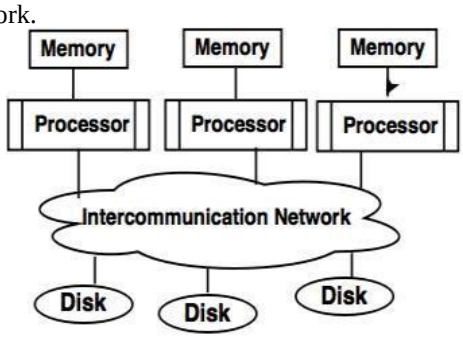


[2]

Shared Memory System in Parallel Databases

2. **Shared Disk architecture**:

In this architecture multiple disks can access multiple processors which in turn have their individual memory and disks are connected to various processors using intercommunication network.

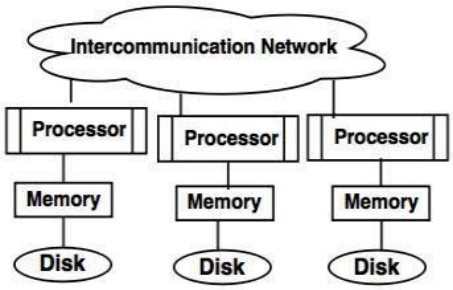


[3]

Shared disk system in Parallel Databases

3. **Shared Nothing architecture**:

In this architecture nothing is shared among the systems is. Each processor has it own individual memory and its own local disk and each processor communicates with each other using intercommunication network.



Shared nothing disk system in Parallel Databases

Comparison

Architecture	Advantages	Disadvantages
Shared Memory System	1. Easy to access data. 2.Memory is shared between processor	1. LackScalability 2. If new processor added would increasetraffic.
Shared Disk System	1.High availability of data 2.Providesincreamental growth	1. If more processors are added slows down the system.
Shared Nothing System	1. Provide unlimited growth 2. Fault tolerance	1. High speed network is required.

CONCLUSION

In above table we put side by side advantages and disadvantages of different types of parallel architecture. Neither shared memory architecture nor shared disk architecture performs very well on data based applications especially on big data if number of processors are increased it becomes difficult to manage at a certain level. Another factor is network which can be considered as limitation for this parallel system to fall down.

So what is the most adaptable system for Big data applications? The answer is shared nothing architecture. The main reason is that more high performance and low cost commodities. Another reason is they provide high scalability and fault tolerance. Thus this paper concludes that shared nothing architecture can be fully implemented into the data based applications which can deal with large scaleproblems.

4. REFERENCES

[1] A. Grama, A. Gupta, . V. Kumar, and . A. Gupta, Introduction to Parallel Computing (2nd Edition). Boston: Pearson,2003.

[2] Survey of Parallel Processing on Big Data by ChengLuo.

[3] Parallel Processing Systems for Big Data: A Survey ByYunquan Zhang, Member IEEE, Ting Cao, Member IEEE, Shigang Li, Xinhui Tian, Liang Yuan, HaipengJia, and Athanasios V. Vasilakos, Senior MemberIEEE.

[4] "Parallel processing & parallel databases," [Online].

[5] https://www.sas.com/en_in/insights/big-data/what-is-big-data.html