ISSN: 2455 – 5428

International Journal of Current Research and Modern Education

Impact Factor 6.725, Special Issue, January - 2017

International Conference on Smart Approaches in Computer Science Research Arena

On 5th January 2017 Organized By

Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

# USING THE IDLE CYCLES OF CPU IN A CLUSTER OF COMPUTERS TO WORK ON BIG DATA

**Dr. S. Umadevi* & R. Kiruba Nagini****

* Assistant Professor, Department of Economics, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

** I MBA, Department of Management Studies, Pondicherry University, Pondicherry

**Cite This Article:** Dr. S. Umadevi & R. Kiruba Nagini, "Using the Idle Cycles of CPU in a Cluster of Computers to Work on Big Data", International Journal of Current Research and Modern Education, Special Issue, January, Page Number 35-37, 2017.

**Abstract:**

Big Data is the data available in unimaginable size with different organizations, who operate online and sometimes offline. The organizations are trying to utilise this data to tap out the dormant markets lying niche with the heterogeneous population. This challenge is supported by the Internet architecture that has distributed algorithms, shared architecture and tools for data visualization. This paper deals with a process model to utilise the idle cycles in a cluster of CPUs available, to download the big data available on the memory slots available and to analyse those chunks of data using distributed algorithms, being resident in the shared architecture of the transaction processing systems.

**Key Words:** Big Data Analysis, Shared Architecture, Parallel Algorithms, Distributed Algorithms, Data Visualization Tools & Grid Computing

## Introduction:

The virtual and online business organizations are working with a minimal workforce that carryout the business worldwide. But during the process of acquiring their clientele, they collect data about the customers and store them for informing them about the next business deals – the market: make it- with them. For so long a time, these data were not looked at as a treasure from which a new business can be developed, or the new products developed. The new business that can be developed was done by marketing and selling the available products with new offers or discounts to the existing customers. But the new product development started when the organizations started looking into the type of data that were being collected as a source of new business.

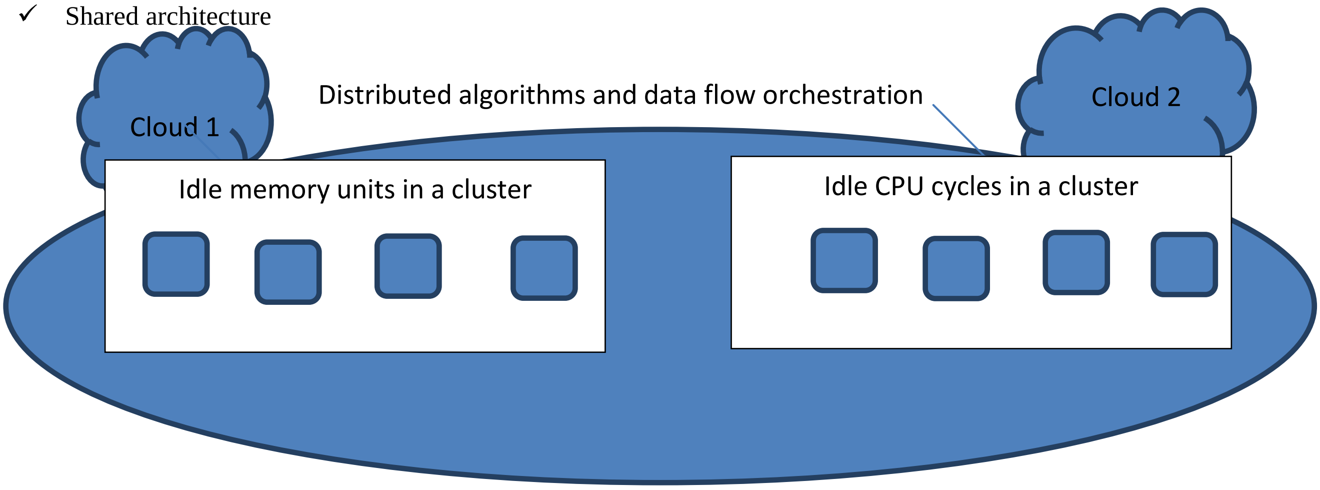
## Data Collection and Profile Creation:

At the customer's point, they provide data to the organizations with which they deal a business, like the residing location, age, date of birth, wedding anniversary, number of members in a family, number of dependents, children, gender, physical and health conditions, range of products sought for, range of price sought for, the frequency of visiting the portal, etc. This is called the Creation of Profile of a customer. Subsequently, the organization retains its customers by introducing the other available products and satisfying them. The profiles collected and stored are now analysed for any scope of a new product to be launched for the customers. At this point, the data is analysed using statistical techniques and some results are arrived at. These results are visualized for a better understanding. Now, certain projections from the similar profiles provide scope for development of new product. For e.g., if a cluster of 10000 profiles collected on a day has almost 5000 persons with new born children or with young children, say, below the age of 10, then the organization plans for the launch of a new baby product! It extends the logic to every segment of population it projects. This is how the virtual business organizations find new markets with the Big Data available with them!

## The backbone of the Big Data Analytics:

The backbone of Big Data Analytics comprise of the following:

- ✓ Data with the organization
- ✓ Distributed algorithms
- ✓ Shared architecture


Cloud 1
Distributed algorithms and data flow orchestration
Cloud 2
Idle memory units in a cluster
Idle CPU cycles in a cluster

The Shared architecture has different entities to be shared, like memory and processing units. Hadoop presently has the Map Reduce architecture to deal with breaking down of data and reducing them to meaningful chunks. It produces the values as a combination of <key, value> tuples. A little modification to this tuple is suggested to enable a better knowledge of the location of the data chunk. Distributed algorithms are the ones which are divided into small functions, stored and called from various points of actions and executed. A modification to this logic is also suggested for a better execution of the distributed nature of such algorithms. Data with the organizations are heterogeneous so that they are to be clustered to form homogeneity in it. This homogeneity is brought out by the statistical or mathematical models applied to the data.

**The New Approach to Data Storage and Analytics:**

The current approach of Hadoop in Java, which makes use of Map Reduce, produces the results in the form of tuples, <key, value>. For example, if the term “age<45” is a key, then “5200” may be the value. So, the <key, value> tuple will be <age<45, 5200> meaning that profiles of people reveal that 5200 of a particular record fed to the MapReduce haspopulation with age<45.

**Extending the Tuple:**

Now, this paper tries to add another parameter to it – the free slot where the data is downloaded for analysis. So, the tuple is now extended as <key, value, slot-id>. Consider a cluster of computers within a location having idle CPU slots and free memory slots. The memory slots are provided with a slot-id. The data getting downloaded and stored into these slot-ids are processed or analysed by the idle CPU slots in the systems in the cluster. The address of the location of the memory slot is decided by the default values of the location of the customer. In case the customer does not reveal the location, the location of the system from which it is produced is set as the default value for slot-id. Once this is done, this value is appended to the <key, value> tuple, as <key, value, slot-id>

**Why the slot-id?**

From the data collected, we try to set the slot-id as the identifier for a location-specific product. When the slot-id is identified, a comparative study can provide a data about the specific location, the generic taste of a location on certain products, and the percentage of people shopping online. If the organization aims at expanding the customer base in that particular area, this kind of slot-id details will definitely provide a clear picture about the customers. If it aims at expanding the market share and not increasing the customer base, then it can concentrate on the type of goods to be sold to the people. The concept of “Content Marketing” developed sometime before and this new technique of adding the slot-id to the tuple will further rip down the population to betargetted.

**The Role of the Shared Architecture:**

The resources available in the shared architecture are completely modifiable, and should be compatible with the modifiable architectures. Recently, cloud architectures suggest a better practice for micro-services based distributed streaming and task/ batch data pipelines. These applications should be able to create, unit-test, troubleshoot and manage micro service applications in isolation; should be able to build data pipelines rapidly using the out-of-the-box stream and task/batch applications; consume micro service aplications as maven and dockerartifacts; scale data pipelines without interrupting data flows; orchestrate data-centric applications on a variety of modern runtime platforms including Cloud Foundry, Apache YARN, Apache Mesos, and Kubernetes; take advantage of metrics, health checks, and the remote management of each micro service application. The distributed algorithms for the downloading and storing of data using the idle CPUs and idle memory:

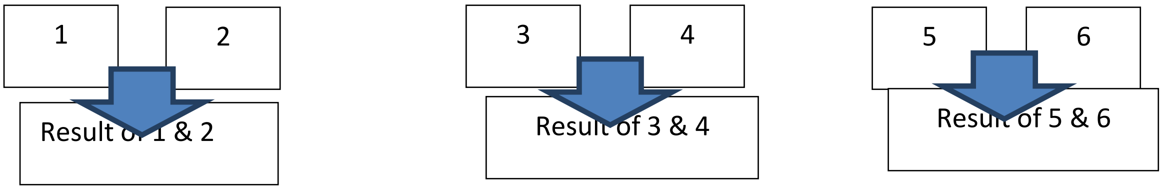
The types of distributed algorithms deployed for downloading and storing the data into the idle memory slots and analysing by the idle CPU cycles may be

- ✓ Ring algorithm
- ✓ Tree algorithm
- ✓ Polling algorithm or
- ✓ Echo algorithm.

A hashing function can be created to select the technique to be followed in downloading and storing the data. Similarly, a hashing function can be formulated to select the analysis of data downloaded. The problems in distributed algorithms would be failure in detection of data to be downloaded, non-availability of CPU cycles when data is downloaded, non-availability of memory slots when CPUs are idle, non-availability of the concerned function or subroutine to analyse data, and so on.

**Using Parallel Algorithms in the Process:**

In case of closely related memory units or closely located memory units, parallel algorithms may be used. If the data is really big to be handled at a single site, then distributed architecture can be made use of. But in case of a data that can be handled at a single site, it is better to follow the parallel algorithm.



The process of using the idle CPU cycles and the shared memory at a cluster of computers for downloading and storing data, and analysing it using the distributed or parallel algorithms can be utilised very well for the Big Data Analysis by the virtual business organizations.

**Conclusion:**

Big Data Analytics is a booming sector in which the available data in an organization can be analysed, visualized and be used for better decision making and product innovation. In the current scenario any business organization is viable to lose its market at anytime and so has to indulge itself in continuous research process to update themselves and to create a benchmark. Then comes the real challenge – to sustain the clientele and generate a new customer database, which is based on the existing customer database. At points of analysis, the idle CPU cycles and shared memory architecture can be used latently on a distributed architecture or parallel architecture. This analysis produces a tuple with <key, value, slot-id> which can be used to start the analysis. Technically this slot-id can be used to tap the location of the customer and generate new product necessities. This tuple can be extended to use the time taken to download and store, find the operational efficiency of the algorithms and so on.

**References:**

1. https://www.infoq.com>presentation
2. https://cloud.spring.io>spring-cloud-data-flow
3. Nancy Lynch, “Distributed Algorithms”, 1e, 1996
4. Joseph Jaja, “Introduction to Parallel Algorithms”, Pearson Education India, 1992.
5. “A Scalable Asynchronous Distributed Algorithm for Topic Modeling”, Hsiang Fu Yu, Cho-JuiHeigh, Hyokun Yun, S.V.N. Viswanathan, Inderjit S. Dhilon, International World Wide Web Conference Committee