ijecs
open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 9, September 2014 Page No. 8455-8459

Implementing Big Data Management on Grid Computing Environment

Lawal Muhammad Aminu[1]
[1]Umaru Musa Yar'adua University,Faculty of Natural and Applied Science,Department of Mathematics and computer Science,
Dutsima Road, Katsina P.M.B 2218, Nigeria
ameenuida@yahoo.*com*

Abstract: ***With the current advances of today's technology in many sectors such as manufacturing, business, science and web application, a variety of data to be processed continues to witness an exponential rise. This data is referred to as big data. Efficient management and processing of this data poses an interesting but significant problem. To utilize the numerous benefits of grid computing, Big data processing and management techniques should be integrated in the current grid environment. In this paper, the definition, features and requirements of big data platform are explored. Incorporating Hadoop is suggested as it the most commonly used technique in handling Big Data as it offers reliability, ease of use, ease of maintenance and scalability.***

Keywords: Big Data, Grid computing,

1. Introduction

Data size has increased significantly with the advent of today's technology in many sectors such as manufacturing, business, science and web application. Some data are structured, semi structured while others are unstructured and mix with varied types of data such as documents, records, pictures and videos .The size and structure of such data have become extremely dynamic. Application developers are faced with challenges caused by the surge of this data from different sources which are short of structure and schema [1].

The existing technologies such as grid and cloud computing have access to huge amounts of computing power by summing of resources and offering a single system view. These technologies have become an influential architecture that performs large-scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. In addition, an important aim of these technologies is to convey computing as a answer for solving big data issues, such as large scale, multi-media and high dimensional data sets [9].

2. Big Data

The Big data term which is being used now a days is not accurate as it points out only the size of the data not putting too much of emphasis to its other obtainable properties[2]. Data becomes "big data" when it mainly exceeds the existing ability to process it, and deal with it proficiently. These are huge pools of data that are hard to be captured, communicated, aggregated, stored and analyzed. Such datasets have size beyond the ability of classic database software tools to capture, store and manage [3]. Big data can be defined with the following properties associated with it[2]:

2.1.1 Variety

Data being produced is not of single class as it not only includes the conventional data but also the semi structured data from different resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active and passive devices. All this data is entirely dissimilar consisting of raw, structured, semi structured and even unstructured data which is hard to be handled by the obtainable traditional analytic systems.

2.1.2 Volume

The Big word in Big data itself defines the volume. At present, the data is currently in petabytes and is supposed to raise to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is certainly hard to be handled using the obtainable traditional systems.

2.1.3 Velocity

Velocity in Big data is a notion which deals with the speed of the data coming from different sources. This feature is not being restricted to the speed of received data but also speed at which the data flows. For example the data from the sensor devices would be continuously moving to the database store and this quantity won't be small enough. Thus our usual systems are not competent enough on performing the analytics on the data which is continuously in motion.

2.1.4 Variability

Variability considers the inconsistencies of the data flow. Data loads become tough to be maintained particularly with the raise in usage of the social media which normally causes peak in data loads with assured events happening.

2.1.5 Complexity

It is quite a responsibility to link, match, cleanse and transform data across systems coming from a variety of sources. It is also essential to connect and associate relationships, hierarchies and multiple data linkages or data can rapidly twist out of control.

2.1.6 Value

User can be able to run certain queries against the data saved and thus can abstract vital results from the filtered data obtained and can also order it according to the magnitude they need. This information helps these people to establish the business trends according to which they can alter their strategies. As the data stored by different organizations is being used by them for data analytics. It will produce a sort of space in between the Business leaders and the IT professionals the main worry of business leaders would be to just accumulate value to their business and receiving more and more profit not like the IT leaders who would have to concern with the technicalities of the storage and processing. Thus the key challenges that exist for the IT Professionals in management of Big data are:

- The scheming of such systems which would be able to handle such huge quantity of data competently and successfully.
- The second challenge is to filter the most significant data from all the data collected by the organization. In other words we can say adding value to the business.

2.1.7 Examples of Big Data

1. RFID systems generate up to 1000 times the data of conventional bar code systems.
2. 10,000 payment card transactions are made every second around the world.
3. Wal-Mart handles more than one million customer transactions an hour.
4. 340 million tweets are sent per day. That’s nearly 4,000 tweets per second.
5. Facebook has more than 901 million active users generating social interaction data.
6. More than 5 billion people are calling, texting, tweeting and browsing websites on mobile phones.

2.2 Big Data Platform requirements

As with data warehousing, web stores or any IT platform, an infrastructure for big data has distinctive necessities. In considering all the mechanism of a big data platform, it is imperative to bear in mind that the end objective is to easily incorporate your big data with your enterprise data to permit you to carry out deep analytics on the collective data set [5].

2.2.1 Infrastructure Requirements

The requirements in a big data infrastructure span data acquisition, data organization and data analysis[5].

(a) Acquire Big Data

The acquisition phase is one of the key changes in infrastructure from the days prior to big data. Because big data refers to data streams of higher velocity and higher variety, the infrastructure necessary to support the acquisition of big data must convey low, predictable latency in both capturing data and in executing short, easy queries; be able to handle very high transaction volumes, often in a dispersed environment; and sustain flexible, dynamic data structures. NoSQL databases are commonly used to gain and store big data. They are well right for dynamic data structures and are highly scalable. The data stored in a NoSQL database is usually of a high variety because the systems are planned to basically capture all data without categorizing and parsing the data into a fixed schema.

For example, NoSQL databases are frequently used to gather and store social media data. While customer facing applications often change, fundamental storage structures are kept simple. As an alternative to designing a schema with relationships between entities, these straightforward structures often just contain a main key to recognize the data point, and then a content container holding the relevant data (such as a customer id and a customer profile). This easy and dynamic structure allows changes to take place devoid of costly reorganizations at the storage layer (such as adding new fields to the customer profile).

(b) Organize Big Data

In traditional data warehousing terms, organizing data is called data integration. Because there is such a high volume of big data, there is a tendency to sort out data at its early destination location, thus reducing both time and money by not moving around huge volumes of data. The infrastructure essential for organizing big data must be capable of process and maneuvering data in the original storage location; maintain very high throughput (often in batch) to deal with big data processing steps; and handle a big variety of data formats, from unstructured to structured.

Hadoop is a new technology that permits huge data volumes to be ordered and processed while maintaining the data on the original data storage cluster. Hadoop Distributed File System (HDFS) is the long-term storage system for web logs for example. These web logs are turned into browsing behavior (sessions) by running MapReduce programs on the cluster and generating aggregated results on the same cluster. These aggregated results are then loaded into a Relational DBMS system.

(c) Analyze Big Data

Since data is not constantly moved during the organization phase, the study may also be done in a distributed environment, where some data will stay where it was initially stored and be clearly accessed from a data warehouse. The infrastructure needed for analyzing big data must be able to hold deeper analytics such as statistical analysis and data mining, on a diverse variety of data types stored in dissimilar systems; scale to excessive data volumes; deliver faster response times motivated by changes in behavior; and automate decisions based on analytical models. Most significantly, the infrastructure has to be able to incorporate analysis on the mixture of big data and conventional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to offer new perspectives on previous problems.

For instance, analyzing inventory data from a smart vending machine in mixture with the events calendar for the venue in which the vending machine is situated, will state the most favorable product mix and replenishment schedule for the vending machine.

3. Grid computing

Grid computing is a model of distributed computing that uses geographically and administratively different resources. In grid computing individual users can use computers and data transparently, devoid of having to think of location, operating system, account administration and other details. In grid computing the details are abstracted and the resources are virtualized [3].

Grid computing is represented by a number of servers that are interconnected by a high speed network, each of the server plays one or many roles. The two main benefits of Grid computing are the high storage capability and the processing power, which translates to the two main types of grids

3.1 Data Grid

Data grids provide an infrastructure to support data storage, data discovery, data handling, data publication, and data manipulation of large volumes of data actually stored in various heterogeneous databases and file systems. It can also be defined as a system composed of multiple servers that work together to manage information and related operations such as computations in a distributed environment. An In-memory to attain very high performance, and uses redundancy to make sure the resiliency of the system and the availability of the data in the event of server failure [3].

3.2 Computational Grid

A computational grid is a hardware and software infrastructure that provides reliable, consistent, pervasive and inexpensive access to high-end computational capabilities. This grid provides protected access to huge pool of shared processing power appropriate for high throughput applications. The computational grids offer a convenient way to connect many devices which helps in reducing the power consumed and also increases the speed of the systems. Computational grids are used by many organizations example health maintenance, material science collaboratory, computational market economy, government etc [3].

4. Related Studies

In spite of grid computing being beneficial in many ways, current grid infrastructure cannot support big data, expert are yet to find an precise solution for the database to deal with large volumes of data. Although grid computing provides technology to overcome the hardware limitation in term of storage space, processing power and memory capacity, Implementation of big data processing and management using grid computing requires additional techniques for managing the huge data effectively [1].

Big data clustering using grid computing and ant based algorithm was proposed by [1] the framework uses the grid concept to facilitate the storage of data in distributed databases across a wide geographical area while ant-based algorithm is for the clustering of big data. The algorithm‘s basic principle focuses on agents where the agents represent the ants that randomly move around in their environment which is a squared grid with periodic boundary conditions. While ants drifting around in their environment, they pick up the data item that are either isolated or surrounded by different ones. The selected item will be transported and dropped by ants to form a group with a similar neighborhood items base on similarity and density of data items. The probability of picking an element increases with low density and decreases with the similarity of the element. The idea behind this type of aggregation pheromone is the attraction between data items and artificial ants. Small clusters of data items grow by attracting ants to deposit more items. Therefore, this positive feedback leads to the accumulation of larger clusters.

Distributed caching was proposed in [3] to solve the challenge of big data. It provides linear scalability through data grid partitioning. the data is spread out over all servers in such a way that no two servers are responsible for the same piece of cached data. This means that the size of the cache and the processing power associated with the management of the cache can grow linearly with the size of the cluster. When there is a request for cached data, the response can be accomplished with a “single hop” to another server, if the data object is not found in local cache. This means when more servers are added to the grid, the performance of the response does not decrease. The grid allows a number of backups for the caches to be configured, when the primary cache fails, one of the backups will takeover, this provides the failover for clustering technologies.

A Dynamic and Scalable Storage Management (DSSM) architecture for big data management in grid environments was proposed in [4]. It organizes Grid storage devices into a large-scale and geographically distributed storage system to meet the requirements imposed by all kinds of Grid applications. The DSSM divides Grid storage devices into multiple geographically distributed domains to facilitate the data access locality [10],[11]. The architecture consists of two levels. The bottom level adopts multicast to achieve dynamic, scalable, and self-organized physical domains. The method significantly simplifies the intra-domain storage resource management. The upper level is a virtual domain that consists of geographically distributed and dynamic agents selected from each physical domain.

5. Implementing Big Data On Grid Computing Environment.

As earlier stated the main advantages offered by Grid computing are the storage capabilities and the processing power. To effectively process and manage big data in grid environment, modifications are required on the current grid infrastructure. Resources used in big data management and processing should be in cooperated in the grid environment.

The most widely solution to handle Big Data is Hadoop, an open source project based on Google’s MapReduce and Google File System. Hadoop was founded by the Apache Software Foundation. The main contributors of the project are Yahoo, Facebook, Citrix, Google, Microsoft, IBM, HP, Cloudera and many others. Hadoop is a distributed batch processing infrastructure which consists of the Hadoop kernel, Hadoop Distributed File System (HDFS), MapReduce and several related projects [8].

The foundation of Hadoop lies in HDFS (Hadoop Distributed File System), The need for it comes from the fact that Big Data is stored on many machines. HDFS is based on the principle that “Moving Computation is Cheaper than Moving Data”, meaning that it is easier to move the computation where that data to be processed is, rather than moving the data to where the computation is running, this being true especially when the I/O files have a big size [7].

HDFS is a block-structured distributed file system designed to hold big amounts of data, in a reliable, scalable and easy to operate way. The blocks are called chunks and the default size is 64 MB (except for the last block of each file), much bigger than the usual 4 or 8kB block size of most of the block structured file systems. HDFS presents a client-server architecture comprised of a NameNode and many DataNodes. The NameNode stores the metadata for the DataNodes. The metadata comprises the file names, the permissions the replication factor and the location of the chunks of files on the DataNodes. It stores all metadata in memory, so it offers a good speed in terms of operations per second, but this way, the amount of data is limited to the machine's RAM. The NameNode is also responsible with file system operation like opening, closing, moving and renaming files and folders. Also, the NameNode keeps track of the state of the DataNodes by receiving signals called heartbeats from them[8].

The HDFS offers great reliability through the fact that all files are replicated on two or more DataNodes. The default number of replicas is three. The block size and replication factor are configurable per file. The replica scheme is managed by the NameNode and executed by the DataNodes upon the instructions from the NameNode. Unfortunately, Hadoop doesn't support automatic recovery in case of a NameNode failure, but a SecondaryNameNode can be configured (preferably on a separate machine). It doesn't take the place of the NameNode in case of a failure, meaning the DataNodes cannot connect to it, but it performs periodic checkpoints: it downloads current NameNode image and edit log files, it creates a new image that can be uploaded back on the primary NameNode. In order to prevent synchronization bugs that can lead to differences between the information on the NameNode and what really is on the DataNodes, HDFS requests for block reports from the DataNodes. Also, the integrity of the data on the DataNodes is verified using a checksum on every block. In case of failing the checksum, the failing blocks are deleted and replaced by the NameNode[8].

6. Conclusion

Grid computing offers storage capabilities and the processing power for data processing. For the grid to support big data management and processing certain requirements based on big data concept have to be considered. This will help in adopting techniques used for managing big data on the grid environment. Integrating Hadoop in the grid computing environment will offer the desired result. The advantages of using Hadoop, especially HDFS, are reliability (offered by replicating all data on multiple DataNodes and other mechanism to protect from failure), the scheduler's ability to collocate the jobs and the data offering high throughput for data for the jobs processed on the grid. Adding the ease of use, ease of maintenance and scalability combining these two technologies seems like a good choice.

References

[1] Ku Ruhana Ku-Mahamud, "Big Data Clustering Using Grid Computing And Antbased Algorithm" Proceedings Of The 4th International Conference On Computing And Informatics, ICOCI 2013 28-30 August, 2013 Sarawak, Malaysia. Universiti Utara Malaysia.

[2] Avita Katal Mohammad Wazid R H Goudar" Big Data: Issues, Challenges, Tools and Good Practices" Contemporary Computing (IC3), 2013 Sixth International Conference on 8-10 Aug. 2013 Noida

[3] C.Chandhini, Megana L.P" Grid Computing-A Next Level Challenge with Big Data" International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013.

[4] Ajay Kumar And Seema Bawa "Distributed And Big Data Storage Management In Grid Computing"
Arxiv:1207.2867
[Cs.Dc]

[5] An Oracle White Paper June 2013" Oracle: Big Data for the Enterprise"

[6] Elif Dede, Bedri Sendir, Pinar Kuzlu, Jessica Hartog, Madhusudhan Govindaraju" Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on " June 28 2013-July 3 2013, Santa Clara, CA

[7] Apache Hadoop Documentation, http://hadoop.apache.org/

[8] Garlasu D,Sandulesu V,Halcu I,Neculoiu G,Gridoriu O,Marinescu M,marinescu V.A "Big Data Implementation Based on Grid Computing". 11th Roedunet International Conference(ROEDUNET)2013, 17 – 19 Jan,2013.Sinala

[9] Changoing J I,Yu Li,Wenming Oiu,Uchechukwu Awada,Keoiu Li."Big Data Processing in Cloud Computing Environments"Pervasive Systems,Alogorithms and Networks (ISPAN),2012.12th International Symposium,13 – 15 dec,2012,San Marcos,TX

[10] Yuhui Deng , Frank Wang, Na Helian, Sining Wu, Chenhan Liao (2008) "Dynamic and scalable storage management architecture for Grid Oriented Storage device" Parallel Computing 34 (2008) 17-31.

[11] Yuhui Deng and Frank Wang, "A heterogeneous storage Grid enabled by Grid service".ACM SIGOPS operating systems review, Special Issue: File and Storage Systems 41(1) (2007).

Author Profile

Lawal Muhammad Aminu received B.Eng in Electrical and Computer Engineering from Federal University of Technology Minna,Nigeria in 2007 and Master of Computer Science from Universiti Putra Malaysia in 2014.He is a Lecturer at Umaru Musa Yar'adua University.His research interest focus on Computer Networks.