Role of Cloud Computing in Big Data Analytics Using MapReduce Component of Hadoop

Kanchan A. Khedikar
Department of Computer Science & Engineering
Walchand Institute of Technoloy, Solapur, Maharashtra, India.

Komal V. Kumawat
Department of Computer Science & Engineering
Walchand Institute of Technoloy, Solapur, Maharashtra, India

Abstract- Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (U.S. National Institute of Standards and Technology (NIST))[1].

Cloud services are popular because they can reduce the cost and complexity of owning and operating computers and networks. Since cloud users do not have to invest in information technology infrastructure, purchase hardware, or buy software licenses, the benefits are low up-front costs, rapid return on investment, rapid deployment, customization, flexible use, and solutions that can make use of new innovations [2].

Big data analytics and cloud computing are Two IT initiatives. Both technologies continue to evolve. Organizations are moving beyond questions of what and how to store big data to addressing how to derive meaningful analytics that respond to real business needs. As cloud computing continues to mature, a growing number of enterprises are building efficient and agile cloud environments, and cloud providers continue to expand service offerings.

This paper gives introduction to cloud computing and Big data, types of cloud computing such as private, public and hybrid cloud. It also gives brief introduction about services use for cloud computing like SaaS, PaaS, IaaS and HaaS. It also explains how to manage big data using Hadoop.

Keywords – Cloud computing, Big Data, Hadoop, Mapreduce, HDFS

I. Introduction

Cloud infrastructure and services are growing significantly. Cloud computing is a paradigm where tasks are assigned to a combination of connections, software and services accessed over a network which shows in Cloud computing conceptual diagram. (See figure.1)

This figure shows the combination of connection and different services which form a cloud. Many users can access services to share various computing resources e.g., servers, mobile applications, storage, e-mails, networks, and hardware service.

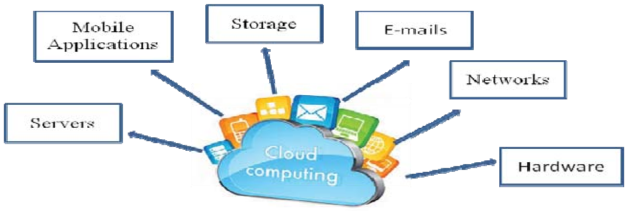

Figure 1. Cloud computing conceptual diagram

Cloud services are typically made available via a private cloud, community cloud, public cloud or hybrid cloud.

- *Public Cloud*: Services provided by a public cloud are offered over the Internet and are owned and operated by a cloud provider. e.g online photo storage services, e-mail services, or social networking sites. However, services for enterprises can also be offered in a public cloud.
- *Private cloud*: the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party. They are expensive and are considered more secure than Public Clouds.
- *Community cloud*: the service is shared by several organizations and made available only to those groups. The infrastructure may be owned and operated by the organizations or by a cloud service provider. For example, all the government agencies in a city can share the same cloud but not the non government agencies.
- *Hybrid cloud*: It is a combination of different methods of resource pooling. for example, combining public and community clouds [1].

Common deployment models for cloud computing include platform as a service (PaaS), software as a service (SaaS), infrastructure as a service (IaaS), and hardware as a service (HaaS).

- *SaaS (Software as a Service)*: This is the most popular form of cloud services. Sometimes referred to as "software on demand," The service provider offers software to support the service on demand. The software is built by the service provider while the end users can configure it to suit their needs. The clients (end users) however, cannot change or modify the software.
- *PaaS (Platform as a Service)*: Offers a platform to clients for different purposes. The user data may be increased or decreased as per the requirement of the applications. PaaS offerings may include facilities for application design and development, testing, deployment and hosting as well as application services such as team collaboration, web service integration, database integration, security, scalability, storage, persistence, state management, application versioning etc.
- *IaaS (Infrastructure as a Service)*: In the IaaS model, a client business will pay on a per-use basis for use of equipment to support computing operations including storage, hardware, servers, and networking equipment. Services available to businesses through the IaaS model include disaster recovery, compute as a service, storage as a service, data center as a service, virtual desktop infrastructure, and cloud bursting, which is providing peak load capacity for variable processes. Benefits of IaaS include increased financial flexibility, choice of services, business agility, cost-effective scalability, and increased security.
- *HaaS (Hardware as a Service)*: The HaaS model allows the customer to license the hardware directly from the service provider which alleviates the associated costs Vendors in the HaaS arena include Google with its Chromebooks for Business, CharTec, and Equus.

Cloud delivery models offer exceptional flexibility, enabling IT to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of valuable external data sources and applications provided in public clouds. The availability of cloud based solutions has dramatically lowered the cost of storage, amplified by the use of commodity hardware even on a "pay as-you-go" basis that is directed to effectively and timely processing large data sets. The big data could be delivered in form of "as -a -service". Google BigQuery is only one example of applying big data solutions in a cloud based platform [3].

Using cloud infrastructure to analyze big data makes sense because:

1. Investments in big data analysis can be significant and drive a need for efficient, cost-effective infrastructure.
2. Big data may mix internal and external sources
3. Data services are needed to extract value from big data.

Three major reasons to use cloud computing for big data technology implementation are hardware cost reduction, processing cost reduction, and ability to test the value of big data. The major concerns regarding cloud computing are security and loss of control. Companies that can extract facts from the huge volume of data can better control processes and costs, can better predict demand and can build better products. Dealing with big data requires two things that is Inexpensive, reliable storage; and New tools for analyzing unstructured and structured data [4]. Big data could be interpreted as a complex data infrastructure. New powerful data technologies and management solutions are needed. This will be directed to improve the decision making processes and forecasting through application of advanced data exploratory studies, data mining, predictive analytics and knowledge discovery. The main key characteristics that define big data are volume, velocity, variety and Veracity which could be also considered an additional characteristic. The related big data models are presented in figure 2.

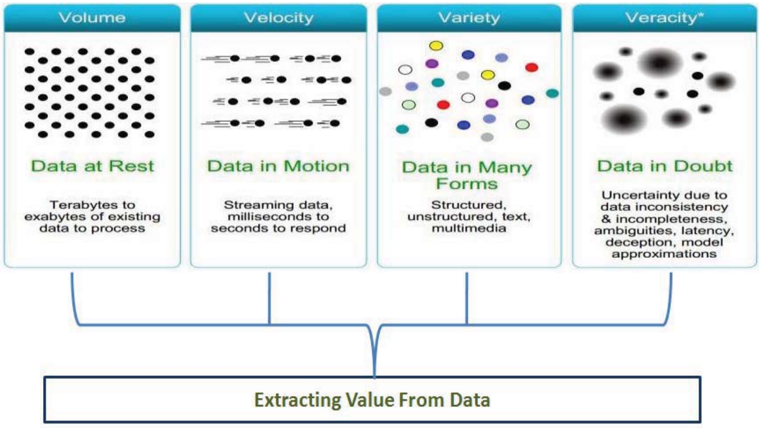
Volume
Velocity
Variety
Veracity*
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
Extracting Value From Data

Figure 2. Big Data Characteristics [3]

Hadoop offers enterprises a powerful new tool for managing big data. Hadoop is a powerful open source software platform. Hadoop is an Apache project. All components are available via the Apache open source license. Hadoop is written in Java. It doesn't maintain indexes or relationships; you don't need to decide how you want to analyze your data in advance. It breaks data into manageable chunks, replicates them, and distributes multiple copies across all the nodes in a cluster so you can process your data quickly and reliably later. Hadoop introduces many components like MapReduce, HDFS, Hive, H-base, Pig, Chukwa, Avro, ZooKeeper etc.[2]. Hadoop's MapReduce and HDFS use simple, robust techniques on inexpensive computer systems to deliver very high data availability and to analyze enormous amounts of information quickly. Information is given in Table 1.

Table-1 Hadoop Components [1 & 5]

Component	Developer	Description
MapReduce	Yahoo !	Distributed computation framework
HDFS	Yahoo !	Distributed file system
H-Base	Powerset (Microsoft)	Distributed Scalable Data Store (Based on Google's Bigtable)
Pig	Yahoo !	Framework for Analyzing Large Data Set (Pig Latin - At Yahoo > 60% of Hadoop usage is on Pig)
Hive	Facebook	Data Warehouse Framework (allows querying of large data sets stored in Hadoop - HiveQL)
ZooKeeper	Yahoo !	A Service for Maintaining Configuration Information, naming, providing Distributed Synchronization and Providing Group Services
Chukwa	Yahoo !	Data Collection and Analysis System
Avro	Yahoo ! & Cloudera	Data serialization system

II. Experimental Work Carried Out

The approach expressed in this paper enables us to develop Parallel Genetic Algorithm (PGA). Cloud computing framework could be used for this purpose. This has been recognized as one of the latest computing paradigm where applications, data and IT services are provided over the Internet. A plugin to data mining packages OlexGA is used for developing classification. Hadoop technology along with its components can be used to store big data and process computations at a faster rate. We concentrate on developing a system that increases processing speed and capability to process huge amount of data. This contribution has an excellent demand in today's academic, medical, scientific and regular business. Document categorization is needed in the above fields; therefore it has a generalized application for commercial approach. We try to parallelize GA to improve the processing speed. We use Hadoops MapReduce and HDFS (Hadoop Distributed File System) framework approach.

- *MapReduce Framework*:

MapReduce programs break problems into Map and Reduce phases. The Map Phase handles all of the parallel computation, and the Reduce phase handles all of the sequential computation. The programming model of MapReduce takes a set of input key/value pairs, and produces a set of output key/value pairs. Figure 3 explain MapReduce framework. The output key/value pairs of the map phase are sorted by their key and each reducer gets a particular key and a list of all the values belonging to that key. Hadoop automatically does this sorting and routing of information between the many machines in the cloud [6].

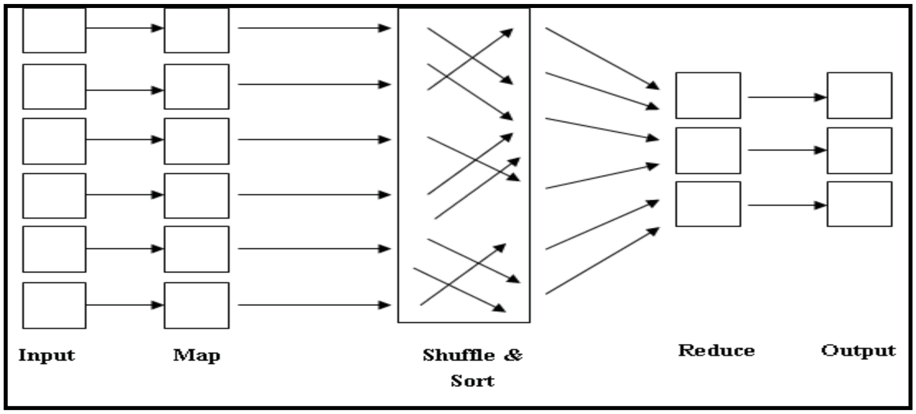

Figure 3. MapReduce Framework

- *HDFS*

HDFS is the distributed file system that is available with hadoop. It highly fault-tolerant and provide high throughput access to application data and is suitable for applications that have large data sets. MapReduce task uses HDFS to read and write data. HDFS includes a single NameNode and multiple DataNodes. For HDFS setup there is a need to configure NameNode and DataNodes and then specify the DataNodes in slaves file. When we start the NameNode, startup script starts the DataNodes. HDFS has a master/slave architecture.

- *Basic System Architecture*

The basic architecture for the proposed system is shown in Figure 4. The blocks used in this architecture involved have been elaborated as below.

A. *Master Node*: Master node functions to upload the document to HDFS. Main scheduling is performed by this master node. The reducer at master node calculates fitness value and provides a category to documents.

B. *Slave Node*: Slave node takes a document from HDFS. The mapper at slave node generates the chromosomes and sends to master node for further operation.

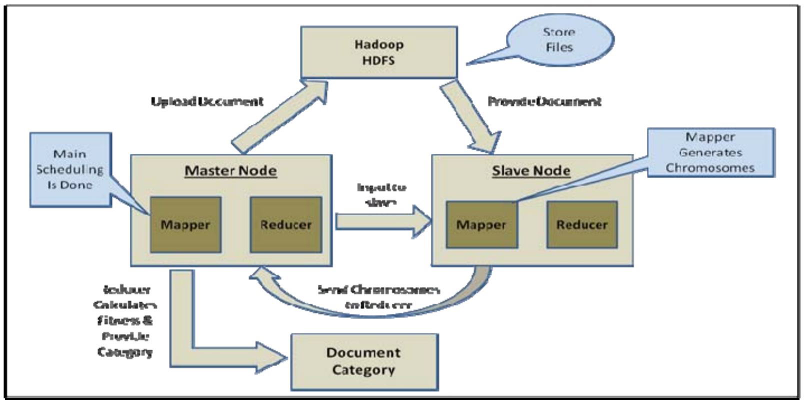

Figure 4. Basic Architecture [7]

C. *Hadoop HDFS*: It accepts the documents from master node and store as a file. For further calculation HDFS provides a file to slave node [7].

Our implementation is related to finding a category of documents which helps in many fields like Education, Medical, IT Industries, and Government etc. This model has two phases named as Train and Test. Train model provides a classifier which is input to test model which provides category of document. The detail methodology is explained in the sections below:

A. *Train Model*

The Train model takes documents and keywords from Hadoop HDFS which then provides to Parallel Genetic Algorithm (PGA). PGA gives a population which is given to an Input Format. The Input Format splits data and provides it to the mapper. There are number of mappers, they generate chromosomes and pass to reducer. Reducer is responsible for calculating fitness. Then output of reducer is forwarded to Output Format after that new population is obtained which is given to PGA and finally the output is directed to a classification model. The Classifier is output of train model.

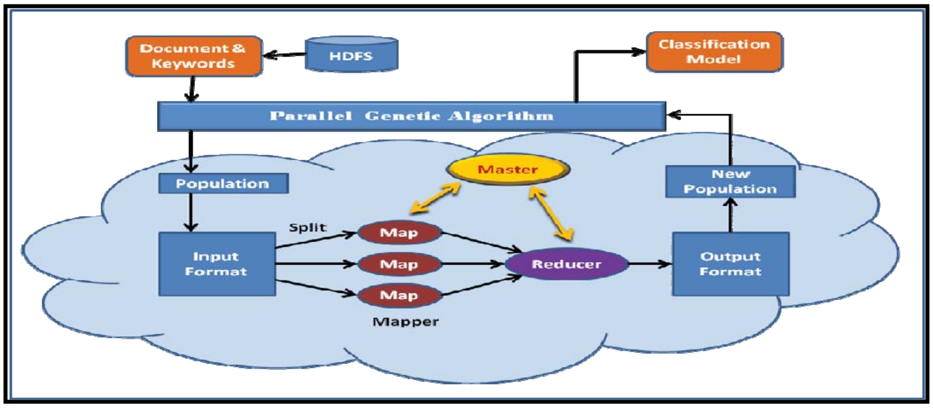

Figure 5. Train Model [8]

B. *Test Model*

Output of train model i.e. classifier, keywords and test document are given to test model. Test model then finds Chromosomes with maximum fitness and returns the category of fittest chromosome. Final output of test model will show the category of a given text document [8].

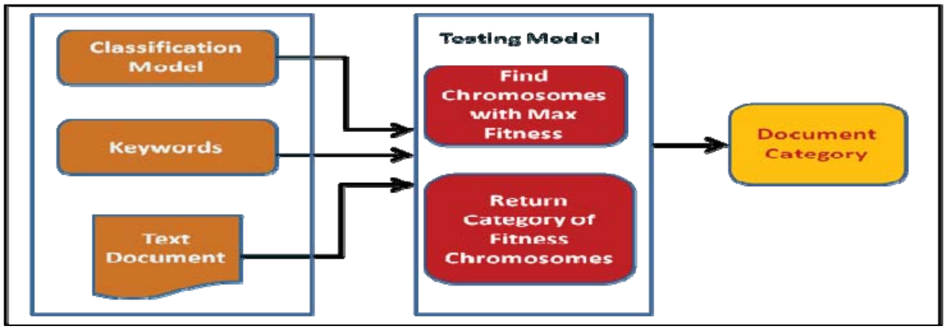
Classification Model
Keywords
Text Document
Testing Model
Find Chromosomes with Max Fitness
Return Category of Fitness Chromosomes
Document Category

Figure 6. Test Model [8]

III. Results

The implementation involves collecting data from many reputed journals like IEEE, ACM and Science Direct. Once this is done a text document is available. This stored into hadoop HDFS. In this implementation we parallelize GA to improve the processing speed. We use map and reduce components of hadoop for fitness evaluation. Based on these data entries, results are generated. Figure 7 shows the results obtained using single node. It takes maximum time to find document category. But results obtained by using two or three nodes will give better results in terms of time. This is known as parallelization genetic algorithm using hadoop mapreduce framework.

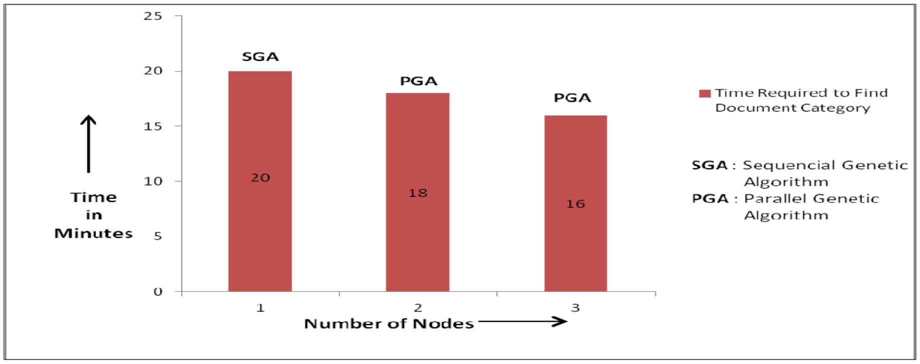

Figure 7. User Model Result

VI. Advantages

- Reduced Cost: Cloud technology is paid incrementally, saving organizations money.
- Increased Storage: Organizations can store more data than on private computer systems. i.e On demand storage and computing for your stuff
- Highly Automated: No longer do IT personnel need to worry about keeping software up to date.
- Flexibility: Cloud computing offers much more flexibility than past computing methods.
- More Mobility: Access “your stuff” from anywhere (on any device)
- Allows IT to Shift Focus: No longer having to worry about constant server updates and other computing issues, government organizations will be free to concentrate on innovation.

- Less Maintenance: Hardware, applications and bandwidth are managed by the provider.
- Continuous Availability: Public cloud services are available wherever you are located and it is free.
- Scalability: Pay only for the applications and data storage you need.
- Elasticity: Private clouds can be scaled to meet your changing IT system demands.
- Expert Service: Expedient's cloud computing services are continuously monitored and maintained by our onsite staff of expert data centre technicians.

V. Disadvantages

- Cloud computing will only be possible if there is strong internet connection.
- Cloud computing might not work in areas where internet connection is weak. Although there are applications that might be work with simple dial-up connectivity
- The application could easily go down when there is too many data to be processed
- Your data is now in someone else's hands
- Your data is visible to the cloud service provider.

VI. Conclusion

Cloud computing is a powerful new abstraction for large scale data processing systems which is scalable, reliable and available. This paper took in to account basic idea about cloud computing, how it can be relate with big data which is managed by Hadoop. Cloud computing is an emerging area of offering affordable services to users of all sizes in the form of software, platform and or infrastructure. Services offered are simpler and scalable.

Cloud computing enables small to medium sized business to implement big data technology with a reduced commitment of company resources. The processing capabilities of the big data model could provide new insights to the business pertaining to performance improvement, decision making support, and innovation in business models, products, and services. Benefits of implementing big data technology through cloud computing are cost savings in hardware and processing, as well as the ability to experiment with big data technology before making a substantial commitment of company resources [10].

As organizations continue to increase the amount and values of collected data formalizing the process of big data analysis and analytics becomes overwhelming. Dealing with big data requires inexpensive, reliable storage and new tools for analyzing unstructured and structured data. This is done by hadoop. Apache Hadoop is a powerful open source software platform. Hadoop's MapReduce and HDFS is used to deliver very high data availability and to analyze enormous amounts of information quickly. Hadoop offers a powerful new tool for managing big data.

REFERENCES

[1] Kanchan A. Khedikar and Dr. Mrs. S. S. Apte. "Latest Technology In Networking: Cloud Architecture", in ICETT 2010

[2] Introduction to cloud computing.pdf downloaded from www.priv.gc.co on 5/3/14

[3] Towards Big Data Mining and Discovery.pdf by Irina Neaga, Yuqiuge Hao.

[4] What Is Hadoop? Managing Big Data in the enterprise. "An Updated Forecast of Worldwide Information Growth Through 2011" IDC, March 2008.

[5] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop Distributed File System, 978-1-4244-7153-9/10/$26.00 ©2010 IEEE

[6] Kanchan Sharadchandra Rahate(Khedikar) and Prof. L.M.R.J. Lobo. "Modified Classification Technique Encountering Parallelization of Genetic Algorithms" International Journal of Latest Trends in Engineering and Technology (IJLTET), ISSN: 2278-621X, Vol. 2 Issue 4 July 2013.

[7] Kanchan Sharadchandra Rahate(Khedikar) and Prof. L.M.R.J. Lobo. "Fitness Evaluation in Parallel Environment using MapReduce" International Journal of Computer, Information Technology and Bioinformatics (IJCITB), Vol. 1, issue 6.

[8] Ms. Kanchan Sharadchandra Rahate (Khedikar) Prof. L.M.R.J. Lobo "A Novel Technique for Parallelization of Genetic Algorithm using Hadoop" International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013. ISSN: 2231-5381

[9] http://www.expedient.com/products/cloud-computing/advantages.php downloaded on 13/3/2014.

[10] Bernice M. Purcell, Big Data Using Cloud Computing, OC13030