# KNOWLEDGE OF THE COMPRESSION APPROACH BASED ON SCALABLE DATA FRAGMENT RESEMBLANCE FOR PROCESSING BIG DATA IN THE CLOUD

**REVATHI DURGAM**
M.Tech
Assitant professor
AVN Institute of Engineering and Technology
revathidurgam@gmail.com

## Abstract

*Big data is the data with large volume and high speed. Large detection data is generated in many systems in different applications, such as research and large organizations. Cloud is the platform where data can be stored, examined and calculated in the cloud. Because data is large, data processing is required, such as data compression. Data compression techniques require high scalability and efficiency for processing large volumes of high-speed data. In this document, we propose the technique for data compression. This technique is based on the on cloud requirement of data compression and the calculation of the similarity between data fragments. In this method, in the first place, big data are divided into different fragments and then compressed according to the fragments. Here, the other algorithm known as Jaccard is used to find the similarity between the data. Thus, this algorithm is compared with the similarity algorithm of existing systems.*

## 1. INTRODUCTION

## BIG DATA:

In 2009 a new FLU virus was discovered.That is bird flu and Swine Flu. This disease is spread quickly. By these diseases, public health agencies are feared. In 1918 Spanish Flue that had infected half a billion people and killed tens of millions. Worse, no vaccine against the new virus was readily available. The only hope public health authorities had been to slow its spread. But to do that they needed to know where it already was. In the United States, the Center for Disease Control and Prevention (CDC) requested doctors inform them of new FLU cases. Some people might feel sick, but they wait before consulting the doctor. So exact information about the FLU cases will take to reach to Central organizations took time, It is a splash between health officials and computer scientists, but they are not showing interest in it.

In this situation more people search about the flu and their preventing medicines. Google gets nearly 3 million client queries. They know that every client wants to details about the flu and their medicines. So google done 450 million different mathematical terms in order to test search terms, comparing their predictions against actual flu cases from the CDC in 2007 and 2008. It is an only one area where big data, making a big difference. Whole business divisions are being reshaped by bid data also. Purchasing plane tickets are a decent sample.

Etzioni is one of America's foremost computer scientists. He sees the world as a series of big-data problems ones that he can solve. And he has been mastering them since he graduated from Harvard in 1986 as its first undergrad to major in computer science. One day he booked a plane ticket to

attend his brother's marriage. He booked ticket one month early. He thought that if he purchase a plain ticket before then the cost will also less. When he get in the plane eagerly asked his beside persion about the price of the ticket. His ticket cost is less than the Etzioni. That persion booked his ticket retcently. He got angry. But he is a computer scientist that's why he realize that there is a big data problem.

From his roost at the University of Washington, he began a huge number of big-data organizations before the expression "big-data" got to be known. He helped form one of the Web's first web crawlers, MetaCrawler, which was propelled in 1994 and gobbled up by InfoSpace, then a ma¬jor online property. He helped to establish Netbot, the first major comparison-shopping site, which he sold to Excite. His startup for extracting importance of content records, called ClearForest, was later obtained by Reuters.

Now a days online social media are increasing rapidly. Every hour people uploading 10 million images in Facebook. Per day they are clicking like button and leave a comment 3 billion times. In Google and YouTube 800 monthly users uploading every second. The video length is approximately one hour length. In 2012 twitter had exceeded 400 million tweets a day. Like this every day online storage was increasing rapidly.

Before big data our its enough to use the normal process to mine the data. But nowadays it is impossible to mine the data easily by using the normal process. Because millions of data increasing in every social sites. Online shopping also having the big data problem. Cosumers are showing interest to purchase items online.

Maintaining the large amount of data is not so easy. So we should use big data process to perform actions like minig data, maintain data. In big data we should consider four Vs. They are

1. Volume:
It refers to the vast amount of data generated every second.
2. Velocity:
It refers to speed of data generated and moves around.
3. Variety:
It refers the different types of data we can use. Previous days we consider only structured data. Infact 80% of data is unstructured. With big data we can easily analise and bring data of different types such as messages, chatting, images, video and audio.
4. Veracity:
It refers to the trustworthyness of the data.

For big data processing we should use Hadoop. Apache hadoop is an open source software framework for storage the large scale processing of datasets. Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

Nowadays big data usage was increasing rapidly. To maintain and process the data perfectly we should use Big Data Techniques.

## 2. RESEARCH WORK:

### 2.1 Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility:

With the significant advances in information and communication technologies (ICT) in the last half century, there is an increasingly perceived view that computing will one day be the fifth utility (afterwards) water, electricity, gas and telephony). This computer utility, like the other four existing utilities, provide the basic level of the computer service considered essential to meet the daily needs of the general community. To offer this vision, a series of computer paradigms have been proposed, of which the last is known as cloud computing. Therefore, in this document, we define cloud computing and provide architecture to create clouds with the allocation of resources targeted to the market by exploiting technologies like virtual machines (VM). We also provide information on market-based resources Management strategies that include customer and computational service management risk management to maintain the allocation of resources oriented to the Service Level Agreement (SLA). Furthermore, we reveal our first reflections on the interconnection of clouds for the dynamic creation of global cloud exchanges and markets. So we present some representative Cloud platforms, especially those developed in industries, together with our current work towards the realization of resource allocations oriented to the cloud market how it was done in the Aneka company cloud technology. Furthermore, we highlight the difference between High Workload of Performance Computing (HPC) and workload of Internet-based services. We also describe a metanegotiation infrastructure to establish global cloud exchanges and markets and to illustrate a case study of exploiting "Storage Clouds" for the provision of high-performance content. Finally, we conclude with the need to convergence of competitive IT paradigms to offer our 21st century vision.

### 2.2 MapReduce Algorithms for Big Data Analysis

There is a growing trend of applications that should handle big data. However, analyzing big data is a very challenging problem today. For such applications, the MapReduce framework has recently attracted a lot of attention. Google's MapReduce or its open-source equivalent Hadoop is a powerful tool for building such applications. In this tutorial, we will introduce the MapReduce framework based on Hadoop, discuss how to design efficient MapReduce algorithms and present the state-of-the-art in MapReduce algorithms for data mining, machine learning and similarity joins. The intended audience of this tutorial is professionals who plan to design and develop MapReduce algorithms and researchers who should be aware of the state-of-the-art in MapReduce algorithms available today for big data analysis.

### 2.3 Big Data Processing in Cloud Computing Environments:

With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Effective management and analysis of large-scale data poses an interesting but critical challenge. Recently, big data has attracted a lot of attention from academia, industry as well as government. This paper introduces several big data processing technics from system and application aspects. First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computing platform, cloud architecture, cloud database and data storage scheme. Following the MapReduce parallel processing framework, we then introduce MapReduce

optimization strategies and applications reported in the literature. Finally, we discuss the open issues and challenges, and deeply explore the research directions in the future on big data processing in cloud computing environments.

## 3. IMPLEMENTATION

For implementing proposed system on cloud we need to go through two essential stages that are generation of data chunk and compression based on these chunks. For implementing proposed system on cloud we need to go through two essential stages that are generation of data chunk and compression based on these chunks.

### 3.1 SYSTEM DESCRIPTION

In the proposed work, the main algorithm, the superior process, the large detection is used data. Therefore, some features of the large data tracking will be studied and analyzed. To carry out the compression, the similarity between two You have to define different pieces of data. So, how to define and model the similarity between data fragments is a primary requirement for data compression. After the definition of the previous model of similarity for fragments of data, how to generate those standard data fragments for the future Data compression is also a critical technique we design. There The new compression algorithm is developed and based on ours Model of similarity and generation of standard data fragments.

**Input File:** This is the first module of the system. Here user gives the file as input. User can input any formatted file as input. Once user inputs the file will transmitted for generating chunks in next module.

**Chunk Generation:** Here we have introduced technique for generating chunks of data. In the introduction we have represented an idea about data chunk which is based on compression. This is theme is not useful for compressing the data. It is like compression of high frequent element. Here difference is that compression of these elements identifies only simple data units. But our data chunk based compression identifies complex partition and pattern during compression process. Similarity algorithms for finding the similarity between the data chunks Here the Jaccard similarity algorithm is used to finding the similarity between the data chunks and the data stream. In existing system the cosine similarity algorithm is used for finding the similarity. Following the both algorithms are given.

### 3.2 MapReduce:

MapReduce is a programming model which is used to process the data. It can develop using various types of languages. Hadoop can run the MapReduce programs. MapReduce programs can develop in Ruby, Java, and Python. Developer needs to develop two methods such as Map and Reducer. Map and Reducer methods having the data set in the form of key value pair.

## 4. CONCLUSION:

Our proposal for scalability is shown compression based on similarity of the data fragment improves data Compression performance increases with the loss of data accuracy. The significant compression ratio brought that is space and time cost savings.

## 5. REFERENCES:

*1. R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg and I. Brandic, Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility Future Generation Computer Systems 25(6): 599-616, 2009.*

*2. K. Shim, MapReduce Algorithms for Big Data Analysis, In Proc. of the VLDB Endowment, 5(12): 2016-2017, 2012.*

*3. C. Ji, Y. Li, W. Qiu, U. Awada and K. Li, Big Data Processing in Cloud Environments, 2012 International Symposium on Pervasive Systems, Algorithms and Networks, 2012, pp. 17-23.*

*4. W. Wang, D. Lu, X. Zhou, B. Zhang and J. Wu, Statistical Wavelet-based Anomaly Detection in Big Data with Compressive Sensing, EURASIP Journal on Wireless Communication and Networking, 2013.*

*5. Bharath K. Samanthula and Wei Jiang, Secure Multiset Intersection Cardinality and its Application to Jaccard Coefficient IEEE Transactions on Dependable and Secure Computing.*

*6. C. Yang, X. Zhang, C. Liu, J. Pei, K. Rama mohanarao and J.Chen, A Spatiotem-poral Compression based Approach for Efficient Big Data Processing on Cloud, Journal of Computer and System Sciences (JCSS). vol. 80: 1563-1583, 2014.*

*7. L. Wang, J. Zhan, W. Shi and Y. Liang, In cloud, can scientific commu-nities benefit from the economies of scale? IEEE Transactions on Parallel and Distributed Systems 23(2): 296-303, 2012.*