UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 7, July 2017

Hadoop-Based Word Count Simulation on Amazon Cloud

K. Madasamy[1], M. Ramaswami[2]

Research Scholar (Part-Time), Department of Computer Applications, Madurai Kamaraj University, Madurai, India[1]

Associate Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai, India[2]

Abstract: Processing very large amounts of data with the traditional conventional database systems are no longer able to handle such a data in an effective manner and practically now obsolete. Due to the introduction of new automated systems and Internet of Things (IoT), accumulation of massive size of the data through heterogeneous sources with unstructured or semi-structured form is quite obvious. To cope up with the current scenario of big data processing needs, Hadoop MapReduce is being the widely preferred choice among many organizations. With the recent growth of cloud computing paradigm, on-demand distributed and parallel data intensive processing is much cheaper and easier on the cloud. The objective of this research paper is to measure the execution time on different sizes of text files by performing a simple MapReduce simulation on the word count program which is very popular in the Big data and Text mining arena. Also, an improved version of the word count program has been designed and variants of word count programs have been tested and simulated on Amazon EC2 Cloud environment. A comparative study of both methods has been carried out and critically reviewed.

Keywords: Big Data Analytics, Hadoop, Hadoop Distributed File System (HDFS), MapReduce, Parallel and distributed Processing , Amazon EC2, Cloud Computing.

I. INTRODUCTION

In recent years, with the hasty development of social networking, Internet of Things (IoT), digital city initiatives, sensors, satellite and other large-scale network applications, accrues huge volume of data which are inherently unstructured or partially structured. The data collected from these heterogeneous sources lead to many shortcomings in processing and detailing the relevant information from these sources of data is a sheer challenging task. To solve this problem, Google, Amazon and other companies [1] introduced the concept called "Cloud Computing" in 2006, for the utilization of network services everywhere and anytime on demand. It affords a simple right of entry to a shared resource pool such as computing facilities, storage devices, applications, etc. Cloud computing enables users to avoid expensive and complex provisioning of infrastructure and software, instead allowing them to utilize the infrastructure, platform and software offered by commercial or open source cloud initiatives and these resources can be engaged or released according to their traffic load [8]. In the meantime, pay-as-you-consume cloud computing model can improve the quality of services while reducing operating and maintenance costs [2]. Cloud computing refers to services that are offered by cluster having 1000 to 10000 client machines[6]. Cloud computing delivers computing resources as a service. It may be Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and Storage as a Service (STaaS) etc. Fig.1. shows the different services offered by any typical cloud.

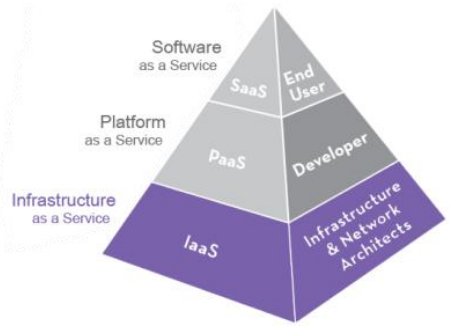

Fig.1. Typical services offered by cloud

Hadoop has become a typical open-source cloud platform based on MapReduce (MR) with Big Table and Google File System (GFS) proposed by Google [4]. Due to its features of scalability, easy to deploy, and high efficiency, it has been

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2017

accepted and well used in both industry and academia. Despite of Hadoop, some novel distributed platforms, e.g., Apache Storm [3, 17] and Spark [4], have also been proposed and widely accepted and applied to process Big data [15,6]. This paper presents a cloud based MapReduce application that allows users to upload large amount of data for computing word count.

The rest of the paper is structured as follows: Section 2 discusses the literature survey. In section 3, the basic concept of MapReduce framework has been briefly discussed. Section 4 describes an overview of Amazon Web Service and the important tools like PuTTY and WinSCPfor cloud computing. Section 5 explains the different versions of word frequency count programs. In section 6, a detailed result analysis of the proposed work has been carried out. The conclusion of the research work is presented in section 7.

II. LITERATURE SURVEY

This survey refers the crux of the research articles published in the various journals in the field of Big data analytics in relation to cloud computing. The Big data analytics and cloud computing are gaining popularity and they will open new vistas in their respective fields. Hadoop happens to be an ideal and easy framework for dealing with Big data. In this literature survey, an attempt is made to focus on some research articles which throw light on the nub issues in connection with Big data analytics over the cloud platform and it has been presented here.

The issue about cloud computing over Map-Reduce has drawn more attention by the public to a great extent for the past years. Many researches about the related issues are conducted to evaluate the implementations, as well as to improve the performance. Li et al. [13] have built the corresponding non-preemptive priorityM/G/1 queuing model for differentiated QoS requirements of cloud computing resources. Maggiani et al. [11] introduced cloud computing visibly in diverse point of views such as storage, security etc. For cloud computing, [14] proposed an architecture implemented on InterGrid system which aims to afford an execution environment for running applications on top of the interconnected infrastructure.Experiments showed that the load can be balanced between distributed sites and have validated that a bag-of-tasks application can run on cloud providers by using virtual machines. Alternatively, cloud computing is also applied to the education field, as described in [5]. Wei et al. [18] have done a comparison between Hadoop, which is an implementation of Map-Reduce paradigm with a system that was developed earlier by Agrawal's group at Ohio State, FREERIDE (Framework for Rapid Implementation of Data mining Engines) by taking three data mining algorithms, which are K - Means Clustering, Apriori Association Mining, and k-Nearest Neighbour search, and a simple data scanning application, Word-Count. The Results showed that, the performance of Hadoop can become better based on the factors like maximum number of concurrent Map per node and number of Reduce. Also, the performance of Hadoop is improved with the increasing size of data set. Cloud systems regularly present persistent storage with the goals of scalability, high availability, low latency and low cost. Amazon S3 [2] is the storage service offered on Amazon's Cloud service, allowing the storage of arbitrary objects up to 5 terabytes organized into buckets.Windows Azure [13] offers blob storage as well as relational storage and non-SQL and others. Some cloud environments often also offer their own persistent storage. Alberto Fernandez et.al [21] have presented a detailed study on big data with cloud computing. They proposed many alternatives to MapReduce programming model for large scale data processing needs. Madasamy et.al [22] have presented a paper on the performance evaluation of word frequency count in single node and multinode setup under Hadoop environment. They have concluded that multinode configuration yields comparatively better results as the input text file size scales.

III. MAPREDUCE FRAMEWORK

Hadoop has two main components, namely Hadoop Distributed File System (HDFS) and MapReduce. The MapReduce [9, 10] framework was first proposed by Google in 2004, and since then it become de facto standard for large scale data processing model in the cloud paradigm. MapReduce provides a simplified programming model, hiding the details of parallelization, fault tolerance [16], data distribution and load balancing in any big data processing requirements. The abstraction for the programming model is based on the map and reduce primitives present in Lisp and many other functional languages.The high degree of parallelism combined with the simplicity of the programming framework and its applicability to a large variety of application domains is the strength of the MapReduce programming framework [4,11].

The degree of parallelism depends on the input data size and is achieved by dividing the workload across a large number of machines. MapReduce has the advantage of handling large data sets, so it is suitable for cloud computing platform [13]. A MapReduce computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user writes a Map and Reduce function that operates on these pairs.

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2017

Map function processes a single input key/value pair and produces a set of intermediate key/value pairs. The MapReduce framework groups together all intermediate values with the same intermediate key I and passes them to the Reduce function. The Reduce function processes a key I and the set of associated values for that key. These values are merged together to form a new, possibly smaller, set of values. Figure 2 shows the block diagram of MapReduce framework.

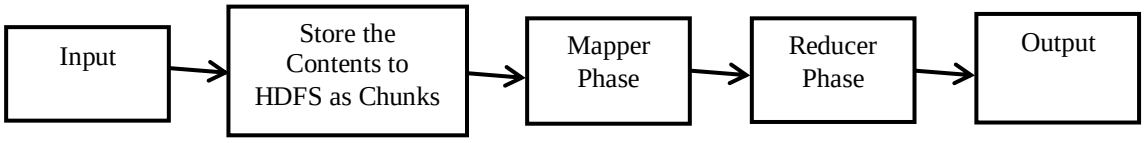

Figure 2. Block diagram of MapReduce Framework

IV. AN OVERVIEW OF AMAZON WEB SERVICES

Amazon Web Service (AWS) offers a wide spectrum of services including computing services, storage services, database services and many more [23]. All these services are mainly used in various applications that includes log analysis, data mining, data analytics and bioinformatics [7]. The clients from various domains are utilizing billions of Amazon Elastic MapReduce clusters to fulfill their computing demands with Amazon EC2 web services. AWS also offers MapReduce service which is used to process vast amounts of data effectively and quickly over the cloud setting. It uses Hadoop, which distributes the data across diverse Amazon EC2 instances automatically after configuring the essential number of instances.

The fundamental piece of Amazon web service is the Amazon Elastic Compute Cloud (EC2) that allows users to run their applications on the virtual machines. EC2 provides a scalable web service to organize the applications that can boot through Amazon Machine Image(AMI) so that a virtual machine can be created. The Amazon Instance contains any software where the users can launch, create and terminate as per their demands and the payment is applicable for hourly basis or monthly basis and therefore it is termed as elastic. The geographical location of the instance will be in control with the EC2 users and that helps to reduce latency and redundancy level of the instances. A resizable compute capacity is provided with the Amazon EC2 and web scale computing is even made easier for the developers. A whole set of computing resources is made to run on Amazon cloud environment. Amazon EC2 reduces the time to boot new server instance in minutes and allows fast scaling capacity based on the user requirements. Amazon EC2 provides efficient developer tools in order to build flexible application and reduce failure scenarios. The resizable computing capacity in Amazon EC2 eliminates hardware upfront investments and this aids the user to develop applications very faster. Moreover, the virtual servers launched within the Amazon EC2 also provides network features, storage and security [12].

A. Tools for Cloud Invocation

To execute the proposed task and to balance the cost constraints, **"C4 –large"** category of provisions from the Amazon EC2 Cloud service have been hired. Red Hat Enterprise Linux (RHEL) Server 7.2 version was installed and in which, Apache Hadoop 2.7.3 was manually configured with 1 Master node (running with a Name node, Secondary Name node and a Resource manager) and with 2 slave nodes (running with a Node manager and a Data node) . The **"C4-large"** instances are the latest generation of Compute-optimized instances, featuring the highest performing processors. It includes high frequency Intel Xeon E5-2666 v3 (Haswell) processors optimized specifically for EC2 . Also, it has two Virtual CPU's with 3.75GB Primary memory. The storage hired is 10GB. Further tools like **PuTTY** and **WinSCP** have been used and their details are given as under:

- **PuTTY**

PuTTY is a very versatile open-source tool for remote access to another computer [20]. PuTTY supports numerous network protocols, including SCP, SSH, Telnet, rlogin and raw socket connection. Also, it can be connected to a serial port. PuTTY works by sending typed commands and receiving text responses over a TCP/IP socket like a traditional terminal (TTY), but it uses secure socket (SSH) with public key encryption wrapping the packet payloads. With the help of PuTTY, we can activate our Cloud instances like Master node and the two slave nodes.

- **WinSCP**

WinSCP is an open source free SFTP client, FTP client, WebDAV client and SCP client for Windows [19]. The main function of WinSCP is to transfer files between a local and a remote computer. Further, it offers scripting and basic file manager functionality. The WinSCP software uses cryptographical methods, integrated in SSH to protect our login

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2017

details and private information. Also it, prevents different types of attacks like password between the local computer and the remote server. In our proposed study, this tool have been mainly used to transfer all our input text files of varying sizes (250 MB, 500 MB, 750 MB and 1 GB) from the local machine to the virtual Hadoop master node created in the amazon cloud environment.

B. **Implementation details**

This section briefly describes the implementation details of the proposed study. Firstly, the remote master node instance and two of its slave instances are activated with the help of PuTTY tool on the Amazon Cloud. Then the input text files of various sizes are uploaded to the Cloud by using the WinSCP . After manually initiating the Hadoop daemon processes, the necessary input text file will be copied into the Hadoop distributed file system. Then the program under examination will be executed, and the time taken for the task will be recorded. The same process will be simulated for all the input text files under study.

V. WORD FREQUENCY COUNT PROBLEM

A. Basic word count: Description

Word count is a simple Python application program that is used to counts the number of occurrences of each word in a given input text file. This kind of program has wide applications in the text mining domain. This program does not count the words with special characters. The word count program will be executed in two stages: Map and Reduce. The Map task maps the data in the file and counts each word in the data chunk provided to the map function. After the Map phase execution is completed successfully, shuffle phase is executed automatically wherein the key-value pairs generated in the Map phase are taken as input and then sorted in alphabetical order. In the Reduce phase, all the keys are grouped together, and the values for similar keys are added up to find the occurrences of a particular word. The Reducer phase takes the output of shuffle phase as input and then reduces the key-value pairs to unique keys with values added up. The algorithm for word count is shown in figure 3.

```
The mapper emits an intermediate key-value pair for each word
in a document.
Class Mapper
    method Map(docid a, doc d)
    for all term t   doc d do
    Emit(term t, count 1)

The reducer sums up all counts for each word.
Class Reducer
    method Reduce(term t, counts [c1,c2...] )
    sum ← 0
    for all count c    counts[c1,c2...] do
    sum ←  sum +c
    Emit ( term t, count sum)
```

Fig.3. Algorithm for Basic word count in MapReduce version

```
The mapper compute word co-occurrence

Class Mapper
Define separator S
method Map(docid a, doc a)
For all term t      doc d do
Emit(term t, Separator S, count 1)

The reducer group by groups multiple word count pairs by word.
Class Reducer
Define Separator S
 method Reduce (pair p, counts [c1,c2...] )
 method Map ( docid a, doc d )
S ← 0
for all term w  doc d do
for all term u      Neighbors(w) do
S ←  S+c
Emit (pair (w,u),Separator S, count 1)
```

Fig.4. Algorithm for Improved word count in MapReduce version

B. Improved word count: Description

The improved word count program also works similar to the word count program, but the difference lies in mapping the given input text. The improved word count program maps the words present beside with the special character, numbers separately. The missing words with special characters in the Word count program are identified and counted using an improved word count program. Figure 4 shows the algorithm for improved word count program.

VI. RESULTS AND DISCUSSION

Many benchmarks have been proposed for evaluating the performance of MapReduce framework. MapReduce is an ideal framework for processing huge volumes of data. In this proposed analysis, the execution time of both the Basic word count and its improved version of word count programs in multinode configuration (1 Master node with 2 Slave nodes) have been tested under cloud environment using Hadoop MapReduce framework. As we have hired 10 GB of storage in the cloud, text files of varying sizes like 250 MB, 500 MB, 750 MB and 1 GB have been taken for the

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2017

simulation study. Table.1 shows the execution time of different size of text files in seconds. Each input text file is executed consecutively for five times and the average time consumed for each run is recorded.

Table.1. Execution time of different versions of Word count programs on Cloud

Size of Data (MB)	**Basic word count- multimode** (in seconds)	**Improved word count- multinode** (in seconds)
250	238	214
500	516	502
750	662	655
1024	704	701

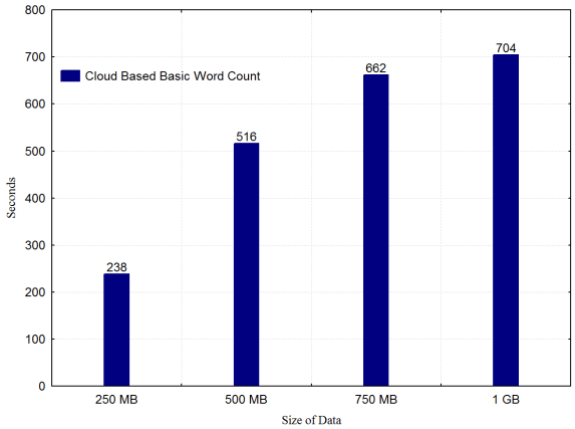

Fig.5. Cloud based Basic word count analysis chart

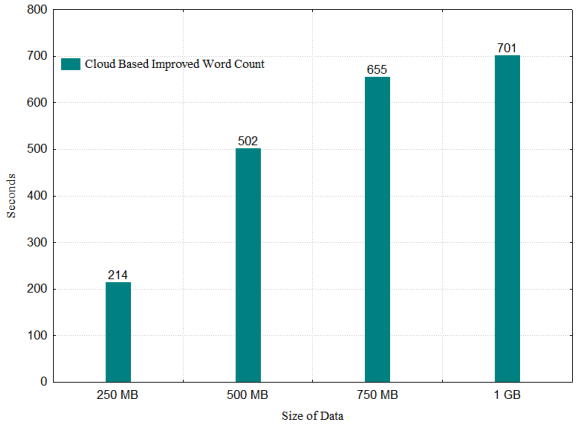

Fig.6. Cloud based Improved word count Analysis chart

Figure 5 and figure 6 depict the results of execution of Basic word count and Improved word count programs in multinode execution respectively. In both the test cases, text files with varying sizes like 250 MB, 500 MB, 750 MB and 1 GB were used. The graphs are plotted as "Size of data" versus "Execution time in Seconds".
By comparing the results shown by figure 5 and figure 6 it is observed that when the size of the data file is increasing, their execution time is also increasing and the Improved word count program consumes lesser execution time than the Basic word count program.

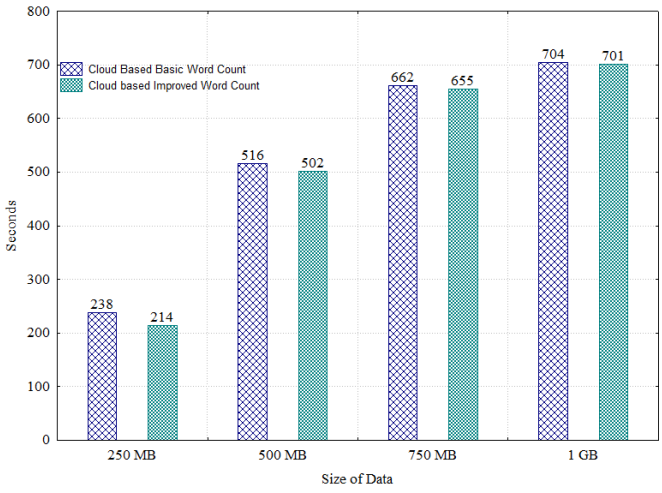

Fig. 7.Comparison of Cloud based Basic and Improved word count analysis chart

Figure 7 shows the comparison of both the Basic word count program and the Improved word count program in multinode execution. The same set of text files have been used. The graph is plotted as "Size of data" versus "Execution time in Seconds". It is understood that the Improved word count program outperforms the Basic word count program in terms of consumption of execution time.

UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 7, July 2017

VII. CONCLUSION

Hadoop MapReduce is an ideal programming framework for processing and analyzing very large data sets in a fast, scalable and distributed manner using commodity hardwares. Many organizations are spending huge costs for the establishment of Big data processing environment. Currently with the dawn of Amazon EC2 Web services, it is possible for anyone to hire the required amount of infrastructure, platform and the services to fulfill their computing needs. In this paper, a simple word count and a slightly improved version of the word count programs have been successfully executed over the cloud with one master node and with two client nodes. Text files with different sizes have been selected for test runs. From the point of considering the execution speed of both the programs, it is observed that, the improved word count outperforms the simple word count program as the size of input data scales. In future, one can configure the present setup with many number of slave nodes by hiring more storage space on the Amazon Cloud Services and can implement the simulation with higher capacity text files which may reduce the execution time and ultimately decreases the cost for processing Big data.

REFERENCES

[1] C. Alexandre Di, A. Marcos Dias De, and B. Rajkumar, "Harnessing Cloud Technologies Fora Virtualized
[2] Distributed Computing Infrastructure", IEEE Internet Computing, 13, pp. 24-33, 2009.
[3] Amad, Amazon. Simple Storage Service (S3). http://aws.amazon.com/s3/.
[4] Amazon Elastic MapReduce, available online at http://aws.amazon.com/elasticmapreduce
[5] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [C] Proc. of the 6th Symposium on Operating System Design and Implementation, San Francisco. 2004.
[6] De Oliveira Branco M: Distributed Data Management for Large Scale Applications. Southampton – United Kingdom: University of Southampton; 2009.
[7] Dean J, Ghemawat S: MapReduce: simplified data processing on large clusters. Commun ACM 2008, 51 (1):107–113.
[8] M. Grant, S. Sehrish, J. Bent, and J. Wang. "Introducing Map-reduce to High End Computing". 3rd Petascale Data Storage Workshop, Nov 2008.
[9] http://readwrite.com/2012/10/15/ why-the-future-of-software-and-apps-is-serverless
[10] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. In Proceedings of the 6th symposium on Operating Systems Design & Implementation, OSDI '04, pages137 – 150, Berkeley, CA, USA, 2004.
[11] Jeffrey Dean and Sanjay Ghemawat. MapReduce: a flexible data processing tool Communications of the ACM, 53(1):72–77, January 2010.
[12] L. LI, "An Optimistic Differentiated Service Job Scheduling System for CloudComputingService Users and Providers", QINGDAO, China, 2009, pp. 295-299.
[13] R. Maggiani, "Cloud Computing is Changing How We Communicate", in 2009 IEEE International Professional Communication Conference, Ipcc 2009, Waikiki, Hi, United States, 2009.
[14] Micc, Microsoft. Windows Azure Virtual Machines. http://www.windowsazure.com.
[15] Namiot, D. On Big data stream processing. Int. J. Open Inf. Technol. 2015, 3, 48–51
[16] Rodriguez-Mazahua, L.; Sanchez-Cervantes, J.L.; Cervantes, J.; Garcia-Alcaraz, J.L.; Alor-Hernández, G. A general perspective of Big data: Applications, tools, challenges and trends. J. Supercomput. 2015.
[17] Stoica I., Conquering Big data with spark and BDAS. In Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), New York, NY, USA, 16–20 June 2014; p. 193.
[18] Toshniwal, A.; Taneja, S.; Shukla, A.; Ramasamy, K.; Patel, J.M.; Kulkarni, S.; Bhagat, N. Storm@ twitter. In Proceedings of the 2014 ACM International Conference on Management of Data(SIGMOD), Snowbird, UT, USA, 22–27 June 2014; pp. 147–156.
[19] J. Wei, V.T. Ravi, and G. Agrawal, "Comparing MapReduce and Freeride for Data-Intensive Applications", in Cluster Computing and Workshops", 2009. Cluster '09. IEEE International Conference on, 2009, pp. 1-10.
[20] https://winscp.net/eng/docs/introduction
[21] https://en.wikipedia.org/wiki/PuTTY.
[22] Alberto Fernandez, Sara del Rio, Victoria Lopez, Abdullah Bawakid, Maria J.del Jesus, Jose M.Benitez and Francisco Herrera, "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks", WIREs Data Mining Knowl Discov 2014, 4:380-409. doi:10.1002/widm.1134.
[23] Madasamy.K, Ramaswami.M, " Performance Evaluation of Word Frequence Count in Hadoop Environment" , International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol.6, Issue 6, June 2017, DOI:10.15680/IJIRSET.2017.0606186.
[24] Amazon Web Services, https://aws.amazon.com/