Cost effectiveness in Educational institutions using Cloud Computing

Kalluri Tejaswi

Submitted as part of the requirements for the degree
of MSc in Cloud Computing
at the School of Computing,
National College of Ireland
Dublin, Ireland.

September 2013

Supervisor
Michael Bradford

Abstract

Today, the educational institutions are focusing on innovation and developments in technology. For the purpose of carrying out experiments, huge amount of resources are required. Due to growing need of huge amount of resources, educational institutes have to spend large capital on their infrastructure to fulfill their requirements. Here, Cloud Computing is an excellent solution to the educational institutions especially which are under budget. Cloud Computing is a distributed computing technology which offers hardware and software through internet in the form of services. This paper is investigating if Cloud Computing could reduce the costs by improving the utilization of resources. This paper, mainly focuses on how Cloud Computing can be used in the educational institutions in order to reduce the cost and improve the use of computing resources in teaching and learning. Cost, mainly covers the infrastructure management, utilization of resources and serving the students and staff in and outside the campus. The education system can improve the teaching methodology by using the Cloud Computing and its application which will allow students to have an interactive open learning environment and to speed up their learning process. In this paper, the research has been conducted using the Hadoop technology with MapReduce system and HDFS. A log analyzer and cost analyzer are used to take decision whether Cloud Computing is the best option for educational institutions or not on basics of usage cost.

Keywords: Cloud Computing, Hadoop, MapReduce, HDFS, Cost Analyzer, Log Analyzer

Definitions and Acronyms

- On-premise: The infrastructure established in the local environment
- Log file: The record of all the activities performed on the system
- Log Analyzer: The MapReduce application that analyses the stream of log files
- Cost Analyzer: The java application that calculates the cost of local and cloud infrastructure
- Report file: The analyzed result of the log files from the log analyzer
- Cost Report: The report with cost effective parameters from the cost analyzer
- HDFS: Hadoop Distributed File System
- TCO: Total Cost of Ownership, the estimated cost including the direct and indirect cost of a service or product
- LMS: Learning Management System, a software application process for e-learning and online training

Acknowledgements

Foremost, I would like to extend my outmost gratitude to my project supervisor Michael Bradford for his time, effort and numerous discussion about the thesis and lastly for keeping my spirits up. His encouragement from the beginning inspired me to do my best to complete this thesis.

I would also like to give my equal gratitude to Horacio Gonzalez Velez for accepting my proposal to work on this topic and providing the support to complete this thesis in time. My special thanks to Dr. Pramod Pathank, Vikas Sahni, Alina Madalina Popescu, Robert Duncan, Keith Brittle and all the staff at NCI, for your helpfulness and support through out the course which should be really appreciated.

I heartfully thank Frank McGonagle, owner of WholeSchool software Ltd., for giving me an oppurtunity to work with him. The real time usecases in our discussions really helped me to much analyze the thesis.

I also thank my friends for their continuous support and being good friends throughout the duration. All your support helped me to complete my thesis.

I would like to extend my love and gratitude to my parents, Kalluri Bhaskarao, Malleswari and my brother Vinay deep, I thank them from the bottom of my heart for their continuous support, without them this work wouldn't be possible. I would like to dedicate this thesis to my Mom and Dad.

Declaration

I confirm that the work contained in this MSc project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed .... Date ....
Kalluri Tejaswi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cloud is used as a metaphor for Internet with different characteristics like "ubiquitous, convenient, on-demand network access to shared pool of configurable resources that can be rapidly provisioned and released with minimal management effort or service provider interaction" [16]. As educational institutes are one of the reason for the development of all the sectors, various research have been conducted to improve the IT infrastructure of the educational institutes. From the characteristics of Cloud Computing, it will provide the advantage of better choice and flexibility in the usage of computing resources for teaching and learning purposes. The present IT systems have many drawbacks which fail in better utilization of computing resources in teaching and learning. In order, to provide efficient and effective IT solutions without consuming more capital for the IT infrastructure like computers and network devices, Cloud Computing might be the best solution. In this paper, the role of Cloud Computing is reviewed and the factors that enable Cloud Computing as a solution for educational institutions over the existing IT infrastructure has been measured in the case of an LMS.

Cloud Computing has a capability to provide computational and storage resources as service. Educational sector is one of the largest sectors using IT support for its activities. But, the usage of Cloud Computing is not in a matured state in the educational institutions. The figure 1.1 from the survey of Gartner [10] provides the usage of Cloud Computing in different sectors. The survey mentions that the usage of Cloud Computing in the field of finance and business is greater when compared to all the other sectors. This paper investigates and discusses about how much the educational sectors are spending on IT and how Cloud Computing can be a boon to minimise the cost of IT infrastructure. As the traditional methods are not enough for todays modern environment for learning. Praveen, et. al. [21] provided a comprehensive methodology for facilitating and using the resources of the cloud in the universities and colleges as on-demand computing by using the application of cloud in universities. So, to optimize the usage of the resources by reducing the cost and to improve the quality of education, Cloud Computing in educational sector may be useful. Cloud Computing has a capability to provide flexibility for teaching and learning for different user types. It costs heavily to maintain the traditional IT system. Cloud has brought a simple theme of everything as a service. By using it educational institutions can be highly beneficial. In this paper, the investigation of how Cloud Computing in the educational institutions reduces the cost of software updates, IT resource management and how the resources for the students and staff for learnng and teaching purposes can be provisioned as they needed.

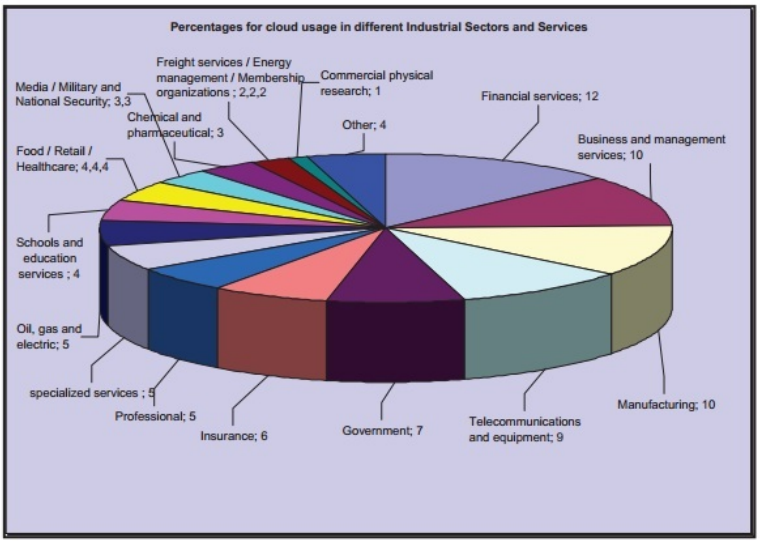


Figure 1.1: Cloud usage in different sectors
[10]

1.1 Thesis Outline

This structure of this thesis is ordered as follows:

Chapter 1: The introduction of how Cloud Computing benefits the educational institutions.

Chapter 2: The backgroud of ICT in educational instituitions and Cloud Computing utilization is described in this chapter.

Chapter 3: The description of the design used to prove the cost effectiveness

Chapter 4: The implementation of the design and installations are provided in this chapter

Chapter 5: The cost effectiveness of local systems and the Cloud is compared based on the usage of the resources

Chapter 6: The conclusion of effectiveness of Cloud Computing in the educational institutions is provided in this chapter.

Appendix: Provides the screenshots of the work.

Chapter 2

Literature review

From the definition of the Cloud Computing and its characteristics, it can be measured that the usage of Cloud Computing in educational institutions is beneficial. This chapter discusses on Cloud Computing in the educational sector in section 2.1, the factors enabling the cloud technology in the educational institutions is discussed in section 2.2, the advantages of cloud technology in the educational institutions is described in section 2.3 and various cloud based applications for the education purposes are provided in the section 2.4.

2.1 Cloud Computing in Education

In this modern computing environment, different learning strategies are being followed in the educational institutions using the technology, which enables the students to grasp the subject quickly and effectively[14]. The new pace for technology in this computing environment is Cloud Computing. Cloud Computing, in general refers to the offering of computing components like hardware and software resources as services [5]. Cloud Computing is highly scalable and uses virtualized resources that can be shared by the users. This network of resources is located anywhere in the world [6]. In [27] clearly describes that Cloud Computing provides various options to the learning environment which are not found in the traditional IT models. By providing the software and resources of the cloud provider with the components of the organization leads to balance system management, lower costs and at the same time helps to improve the quality of teaching. There is difficulty with this model as the resources are available from the single domain the failure may costs the institute heavy.Patel, et. al. has described the Service Level Agreement system of the cloud systems which provides high availability for the data of the organization [19].

Alabbadi et. al. describes the capabilities of Cloud Computing which avoids the regular updates and maintenance of the IT components [4]. Institutes accessing the Cloud Computing services have no burden of buying the commercial software licenses and frequent upgrades and maintenance costs. Glossed over this insight are the compatibility of the data on the upgraded version of the software has to be answered. Google's Cloud Computing services has word processing, spread sheets, PowerPoint, web production, email and other application as Google Apps[1] for free of cost for personal use [22].

Pina, et. al. stated that the Cloud Computing is the new IT [Information Technology] enabled market constructs [20] which has a huge impact on management, in terms of administrative processes. The vision of Cloud Computing is the challenge of business transformation from which the educational sector cannot escape. This transformation to Cloud Computing will help the organization to restructure their IT operations and has to decide which services should be accepted from the cloud provider. This explanation has a few weaknesses that other researchers have pointed out. If transformation to cloud system is made the solution for existing IT systems has to be found. Cloud Computing for academic environment, claims that the existing IT infrastructure can be transformed to private cloud and in need of more resources it can demand from the public cloud making it a Hybrid cloud.The advantage of Cloud Computing is large pool of processing resources on demand. Students or scientists research which take days or hours long to execute on single computer can get speed up and better performance by tasking the cloud to provide processors for a few minutes. Lower costs and on demand resource allocation is the major advantage of Cloud Computing.

In many researches on E-Learning, points that there are number of technical and operational challenges in IT environment for E-learning strategy like IT management, Infrastructure, tools, support services, etc. which could be addressed using Cloud Computing. The main challenge of the Cloud Computing is to enable efficient IT environment in educational sector by providing high access to the resources from anywhere, reducing the cost by using the pay as you use plan and flexibility of scaling the resources as and when required [15] . But the inevitable situation arises when the network fails there is nothing left to be done on the personal system as everything relies on the cloud.

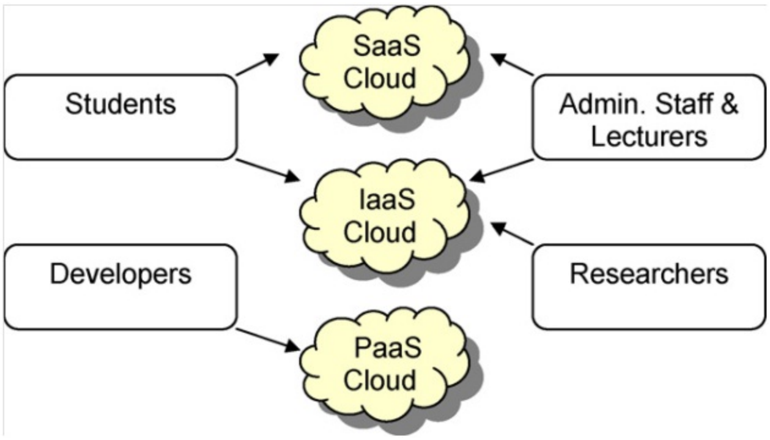


Figure 2.1: Cloud Computing in Education

[15]

From the article [23], it shows that how Cloud Computing can be used in the educational environment to meet the demands of different users like students, developers, faculty and researchers. The major job in the educational institutes is to provide students and staff the software like email accounts, productivity application and hardware like storage and computing. Researchers require some special HPC softwares and can demand more hardware during the research. Developers need special development tools for hosting web applications. All these needs for the educational institutes can be served at low capital expenditure using Cloud Computing technology.

The educational institutes can utilize the Cloud Computing at the maximum level by making an agreement called Service Level Agreement (SLA) with the cloud provider. This SLA will build a relationship between the institution and the service provider to simplify the complex issues in provisioning of resources. An inevitable pressure is to do more with the less making the educational institutions to transform to the cloud. There is a vast increase in the number of educational institutions that are transforming to the Cloud Computing mainly because of economic reasons [23]. Few examples are Washington State Universitys school of Electrical Engineering and computer science, University of California, higher education institutions from UK, Africa [17], U.S and others.

From the research paper [7], the benefits of Cloud Computing in education institutions is clearly summerised as in the following table

Users	Benefits
Students	-The accessibility of the computing resources raises effectively - The integrity and availability of the data with applications and research work increases - The mobility for the use of provided services increases - The client applications and resource usage footprints are minimized - The performance of the application and the computing resources is increased - The capacity of the storage and the computing is increased - The access to the virtual class is made convenient
Administrators	-The process and application delivery are standardized - The management of data and applciations are made provisioned accordingly - The total cost of ownership is minimized from 50-90 % - The inhouse IT infrastructure need is reduced - The management of IT infrastructure cost is reduced including power and cooling costs - Reduces the burden of purchasing the licences of software - The allocation of resources is optimized
Faculty	- The virtual machines can be provided - The delivery of instructions, assignment and materials can be scheduled - Custom images for the specific course can be created - The departments are isolated and elimates the information leakage

Table 2.1: Table: The benefits of Cloud Computing in the Educational sector

2.2 Factors enabling Cloud Computing in educational Sector

In the near future, the IT infrastructure of Education will be running on Cloud Computing environment. This Cloud Computing environment will provide the software and hardware resources for the efficiency in teaching and learning systems. Integrating resources in the Cloud Computing system will enable the high speed data processing with massive demand of resources in the current information age 'information explosion'. The rapid transformation of IT infrastructure to Cloud Computing in various domains set to vast development in the market which also stands for education sector. Cloud Computing has optimized the usage of educational resources, reduce costs, meet demand for green energy; is conducive to centralized management, ease of operation and maintenance, and ultimately will enhance information security. The table 2.2 shows how the cost is reduced in the organizations with the use of Cloud Computing [7] by considering the different requirements for computations.

Costs Direct or Indirect	On-premise	Cloud
hardware:server, end points	\$5500	-
server OS, Client access licence	\$1500	-
Backup hardware and software	\$2000	-
Auxillury server equipment	\$500	-
Installation or migration costs	\$4000	\$3000
Total	\$13500	\$3000

Table 2.2: Cost Variation between on premise and cloud services

The factors mainly focused on the transformation of IT infrastructure to Cloud Computing for efficiency in education systems are

2.2.1 Resource Sharing

The cloud resources can be accessed from the simple personal devices which can access the internet. In computing it is possible that a service can store the resources to avoid the reconstruction and upgrade of software and hardware[29]. Here, integrating the resources of the different organizations into a cloud cuts the resource investment. The Cloud is utility based which allows the institutions to demand the resources when required and pay for what they have used. Integration saves the institutions over finance on the IT infrastructure.

2.2.2 Social learning

Social learning in Education plays an important role in knowledge sharing by creating community in which learners can get emotional support. In the social learning environment, trust is the major concern among the learners [3] else the learners will feel helpless while exploring the subject, Social learning can be established among the tutors and students in the form of online exchange, online document editing, on-line using the concept map tool, for example Google collaboration platform [22] . The cloud forms a community where all the users setup a social environment to learn.

2.2.3 Security

Confidential information and private data exists for all the organizations. The key issue is securing this type of information. In the article [12], it has described that the data in the internet can get viruses and trojan attacks. When the users are dealing with the educational resources over the internet. Mayer, et. al., described that, the security of resources is greatly enhanced by storage and mechanism which protects and monitors data [18]. In the Internet world, the managers could only unify data management, load balance, resources allocated, software deployment and the control of the security result in the reduction of investment in human resources. The research results, employees' accounts students' scholastic records etc. are the most sensitive data in the educational institutions for which special attention should be paid for security.

2.2.4 Learning in a network

The modern teaching environment focuses on the student centric active teaching strategy. In the traditional teaching environment, the solution for ideal and individualized learning approach was failed. In the cloud times, different types of services are available to the learner as for their choice which provides different learning methods and contents from various sources. In todays web tools that belongs to cloud services are, iGoogle that allows to personalize the web space, Diigo bookmarks to create a personal theme, Sakai management network courses (Hongyu Zhao 2007) When the web tools are providing different aspects of the cloud computing Several attempts have been made on the distance education, obtained the media that supports the personal learning systems such as text, audio and video training can be obtained from the cloud services [27]. Learners have to just sign in to the browser and personalize their learning environment, which establishes the virtual classrooms through the network at any time and any place to communicate. This doesnt require any large pool of resources or complex software implantation.

2.2.5 Service balancing in Cloud Computing

Educational institutions mainly focus on the hardware and the software components to use, where as cloud computing provides different types of services. Numerous studies have attempted that the service, that provides hardware resources to the institute is Infrastructure as a Service (IaaS). The service that provides tools for development purposes is Platform as a Service (PaaS). The service that provides the applications to run on the local system is Software as a Service (SaaS) [5]. The organization has to choose appropriate service to get the efficient use of cloud computing. The educational institutes that are accessing these cloud computing services, can transform from capital expenditure to operational expenditure. This reduces the burden of upgrades and maintenance costs. In cloud computing there are applications that are of low cost or free of cost like office series. Integrating the on premise infrastructure and consolidating with data centers, is the major task to be considered with the cloud computing model, the users local system runs on the graphical interface of the operating system and run everything on the browser to enjoy the cloud.

2.2.6 Application of Cloud Computing in diversified educational information technology

As the information technology is evolving continuously, applying new technology that is cloud computing to education has to be developed. The new technology improves the environment but arose the issue of interoperability of the systems. It clearly presents that cloud computing in education services shows the direction of diversification. Recently Wholeschool[2] has released the Cloud education service which has the overwhelming feature, the Cloud centric teaching system that enables learning from anywhere. Microsoft and Google have launched their own cloud computing platform and enabling the E-learning strategy using cloud systems. Cloud computing also enables to create their own network platform, rent network space, not only eliminates the need to purchase a large number of hardware devices, but also to eliminate the trouble of maintaining the system, in line with our current status of information technology in education, which also contributed to the diversified development of educational information[14].

2.3 Advantages of Cloud Computing in Educational Sector

In educational institutions cloud computing offers the opportunity of concentrating more on research and teaching activities than on complex IT implementation. In higher education many universities have already utilized the potential and efficiency of cloud computing. Among them are University of California, Washington State Universitys School of Electrical Engineering and Computer Science, and few of them in higher educations from UK, Africa, U.S and others. North Carolina State University achieved substantially decreasing the software licensing expenses and also reduced the campus IT staff upto 15 employees [28]. Indias Telecom Commission proposal to create US $4.5 billion National Optical Fiber Network (NOFN) which was approved by the Department of Telecom (DoT) which will broaden the countrys existing fiber optic network from the district level to the village level giving the country of about 1.2 billion people services like e-Health, e-Banking, e-Education etc [8]. In Pakistan Aga Khan University found that cloud computing have strengthen security and improve protection against viruses, and resulting in the IT department in reduction of calls upto 66%. The experiment has played an important role in learning with the development of distance education network.

Different students has different capability of learning , by using cloud computing experimental teaching can be implemented to individualize the learning process of the students.As it is nonverbal communication and evaluation functions in the process of teaching it plays unique role in the Experimental teaching. For students basic requirements hands-on practice is possible in experimental teaching where as the distance learning mode cannot meet this requirement. There is a community, Collaborative Learning that creates a support to students and builds trust among them. By combining all the resources of different educational Institutions the cloud will cut the investment resources to single [9]. In cloud computing by using the concept map tool such as Google Collaboration Platforms, students and teachers can implement collaborative learning like online exchange, online document editing etc . By this students can receive some emotional support for self-study between companies. Cloud provides a benefit from the burden of software upgradation, even with the terminal devices which is not capable of accessing high quality of internet, cloud services can be used. This is provided by using the services which stores the resources [26].

2.4 Categories of Applications in Cloud Computing

From the cloud computing perspective, there is no need to host and operate the resources locally in the college and universities which benefits a lot in the computing environment. There are many applications that has been developed in recent years for the educational institutions in cloud computing. Using these applications, the immediate access to the educational resources is provided to the students and teachers. In October 2007, a team of IBM and Google supported the growth of cloud computing in the education by training the students to program cloud applications on the cloud resources for free. A global cloud forum for educators was launched by IBM in 2009, to initiate the growth of cloud computing in educational sector[5]. Many large organizations has dedicated huge amount of resources to the development of cloud computing in the educational field.

One of the best example of educational institute that uses the cloud computing applications is Pike Country School, which has replaced 1,400 workstations by deploying a cloud based virtual desktop solution on the IBM cloud. This has benefited the institute with 60 percent of cost reduction, increased security and reduced software license, overall maintainance cost [11].

From the research and the IDCs survey of IT professionals, it is clear that many educational organizations are showing their interest in transforming the IT into cloud computing. Many application types are required to successfully make the use of cloud computing in education which will reduce the cost and burden of management. Some of the categories of applications required in cloud computing platform for education are

2.4.1 Collaboration Application

From the IDC report[13], 67 percent of survey states that E-mail, chat, file sharing and conferencing such as SharePoint are good fit for the cloud because this collaborative application reduces the cost. Collaborative applications that can be used for education are like Gmail for project management. Skype for free voice conferences and github.com used for project documentation archiving and source code sharing

2.4.2 Web Serving Application

Web Services are the applications that provide users the flexibility to select and use over any computing device from mobile phone to the personal computers. These web service applications reduce the cost of IT expenses and enable to connect the users to their operations quickly and cheaply[25]. Web Service application provide different strategies of computing usage in education system like Instruction level education and online education.

2.4.3 Cloud Backup

Many organizations are willing to move backup offsite to the cloud as to be best protected from the natural disaster, power issues, IT misusing and other serious issues. GitHub.com and sourceforge.com are the best application for the backup of student projects. The large datasets backup in the low network bandwidth makes its challenging[24]

2.5 Problem statement

Most of the papers discussed here, provides different means and methods and argues that cloud computing reduces the cost of ICT in the educational instituitions. But, the investigation with real time experiment of providing the solution with the cost analysis will be helpful for decision making of cost optimization between the local infrastructure and the cloud computing usage. This paper mainly provides the implemented solution which uses the real time statistics to analyze the usage using the Hadoop technology and calculates the cost of the usage in the cloud to show the difference between the costs. This implementation can also be used in the cluster environments where the usage of each server can be analyzed and can be obtained for efficient and effective result.

Chapter 3

Specification

This chapter provides the background research of the work carried out to investigate the cost efficiency of the cloud system in the educational environment in the section 3.1. Different technologies used in the project to analyze the logs of the systems like Flume and Hadoop are discussed in the sections 3.3 and 3.4. The cost analyzer which predicts the cost of the utilization in the cloud and local infrastructure is discussed in section 3.5.

3.1 Background

In order to get the cost effectiveness of Cloud Computing compared with the on premise infrastructure, the actual resource consumption i.e, CPU, Memory and Storage of an application is obtained from the on premise infrastructure and the cost of the same resource consumption in the cloud is calculated. The resource consumption of an application in the on premise infrastructure is always over provisioned where as in the cloud using the characteristic of scalability, the resources can be spined up and spined down as the application requires. For example, the resource utilization of the application in an educational institution is high at working hours and low at night time. By analyzing this with the use of Cloud Computing, the resources can be scaled down which reduces the cost.

As to prove the cost effectiveness of Cloud in the educational sectors, an application server running Moodle is used. Moodle is an open source Learning Management System with large number of registered user base and registered websites. Most of the universities and colleges run Moodle to handle and manage the courses. The usage of moodle as a common platform for the educational institutions can be efficiently utilized in this project using the virtual load testing tool Apache JMeter. The usage is continuously streamed into log files. As the log files are of huge quantity and are streamed continuously, MapReduce provides the efficient way to analyze this large quantity of data. Flume provides a service to stream this huge log files from the application server to the Hadoop systems. In order to compare the cost of the utilization, the resource consumption of an application is obtained from the current and predicted workloads of the server and this logs are moved to the Hadoop master using the Flume.

The current and predicted values from the application server are stored in Hadoop Distributed File System (HDFS) and this values are analyzed using MAPREDUCE to get the most appropriate usage report of the application on the on premise infrastructure. The cost analysis of this report is carried to get the cost effectiveness of cloud computing in the educational institutions when compared to the on premise infrastructure. The log analysis report can also be used to identify the requirement of resources in the cloud to run the application. Once, the requirements are identified the cheapest cloud vendor can be selected from the available cloud vendors in the market. Apart from the resource consumption metrics, other metrics like administrative costs, licensing costs, Operating system costs, Maintenance cost, Electricity and physical infrastructure cost should also be considered. This will eventually increase the expenditure of the educational institutions on ICT. Getting the actual consumption of the resources, the problem of overpaying for the resources can be reduced. Fig 3.1 provides the dataflow diagram of the proposed work for cost analysis.

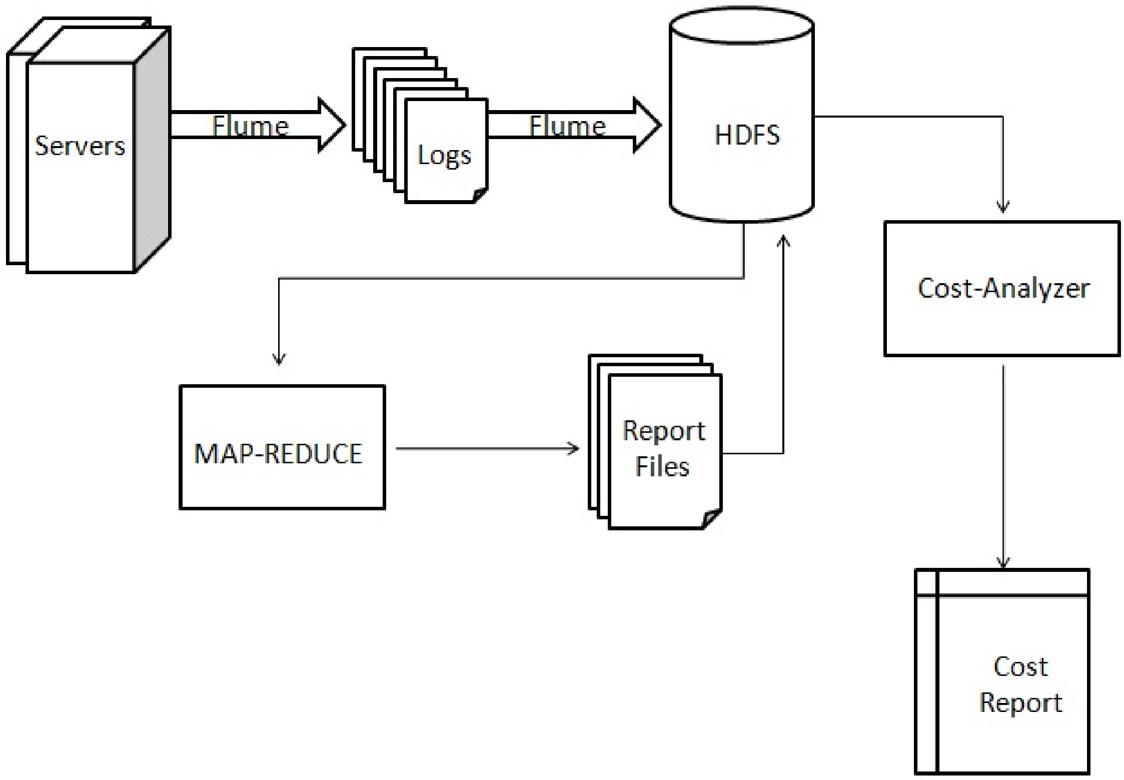


Figure 3.1: Data flow Diagram for Cost analysis

3.2 Moodle Server

The most common server which all the bodies (i.e, Students, Faculty, Administrators) in the educational institutions uses is the Moodle server.As most of the educational institutions are using the Moodle, from the background it is noticed that the utilization of moodle server can be considered to get the common utilization of IT infrastructure in the educational institutions to analyze the cost parameter. As moodle provides different features like

- Online updates
- Assignment uploading
- Examination scheduling
- Files uploading and downloading
- Forums for discussions

- Online exams
- Enrollments

With all the above features, moodle has many external plugin support. This make the moodle server to be in continuous usage in the educational instituitions in all the departments and bodies. As of part of this project, the moodle server is installed and configured and in order to increase the load of the server virtually the open source stress testing tool Apache Jmeter is used.

3.2.1 Setting up virtual users

The main aim of setting the virtual users is to analyze the maximum number of request or users, the server can handle with. It is not possible to increase the load of the server with one client. In order to increase the load virtual requests and users with concurrent usage has to be generated. The stress evaluation softwares like Apache JMeter can be used to load test the servers.

This evaluation software should have the ability to generate the loads on the server automatically using the client scripts. These generated load tests can be used to stimulate the users concurrently. All the responses from the server for the testing results are gathered and an output with statistical graph can also be obtained. The behavior of the servers should be analyzed for the following reasons:

- To point the bottlenecks in the server
- To verify the maximum number of users the server can serve
- To monitor the behaviour of the server at high stress states

JMeter is a load testing software completly built on the Java platform with the multithread framework. The analysis of performance of the server by stimulating the dynamic and static workloads on network or server can be performed using Apache JMeter. It provides ease to run large number of virtual users and requests concurrently on the server by using different load testing scripts. It also provides the support to distributed tests system.

3.3 Flume

Flume is an open source Apache project. It provides a distributed platform for reliable services to move large quantity of data in an efficient way. Flume is mainly used for online analytical applications by providing different failover and recovery mechanisms. Fig 3.2 provides the high level architecture of the flume with different components.

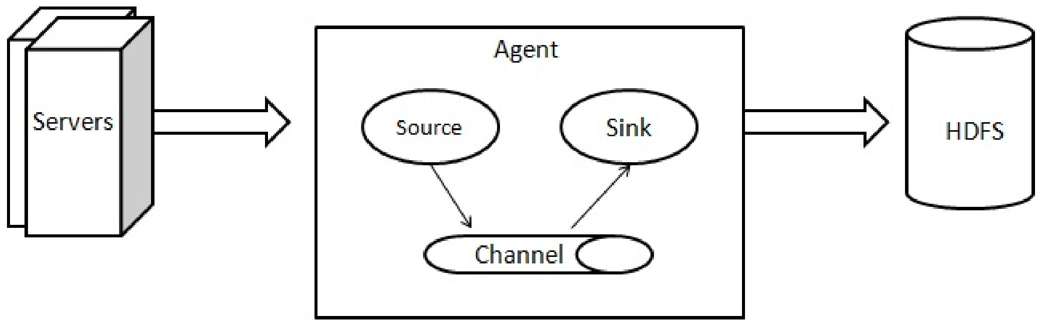


Figure 3.2: Flume Architecture

1. Event: The unit of log data transported using the Flume.
2. Source : The component from which the data enters the flume. There are different types of sources based on the type of data like log4j, syslogs
3. Sink : The component which delivers the data to the destination. There are different types of sinks for the data to stream into destinations like HDFS sinks, avro sinks etc.
4. Channel : The medium between the two major components (source and sink). Events are filled into the channels by the Sources and the sinks empty the channels.
5. Agent : The main collection of sources, sinks and channels. The physical JVM of the flume.
6. Client : The events are produced and transmitted to the source within Agent

The Flume agent is a simple JVM process. All the components as in the Fig 3.2 are hosted by this flume agent, which allows the events to flow from source to the destination end. The Source and Sink runs in separate threads, which produces data in the form of events and collects events from the channel respectively. The Channel (Memory, JDBC, WAL) connects the source and sink by providing the reliability semantics. Flume provides the service to utilise the log data in the most efficient way. The flume provides different services to the users

- Continuously stream the log data to the Hadoop from the multiple sources.
- The large quantity of logs can be collected in the real time

- The source and the sink synchronizes with eachother even when the rate of source and destination differs
- The delivery of the data is guaranteed with all the bytes
- The additional data can be scaled horizontally

3.3.1 Reliability

No data in the flume is lost during the agent operation, because of its high reliability design. The dynamic reconfiguration of the agent without restarting the services is supported which avoids the downtime in the agents. The major advantage with Flume is, the independent structure with no central coordinating system which avoids the single point of failure. With the characteristic of the channel, the load balancing and failure is highly supported in the Flume agents. The horizontal scaling of the Flume is possible with its decentralized design architecture.

3.4 Hadoop

Hadoop is a open source Java framework for scalable and distributed computing environments. Hadoop follows the mechanism of moving the computation to the data instead of moving the huge data sets to the computation. It provides a reliable and efficient way of analyzing the data both in structured and unstructured patterns with the capability of handling large amount of data sets. It runs simple programming model to analyze the large amount of data sets across the clusters of computers. The sub projects of Hadoop that plays a key role in accessing the large amount of data sets in an efficient way are HDFS and MapReduce. In the Hadoop systems, the data gets distributed to all the nodes across the cluster during the load time. The HDFS divides the data files into several chunks and the nodes in the clusters manage all this data files. All the chunks in the HDFS are replicated by default to 3 and can be increased, which provides the fault tolerance mechanism in case of node failures.

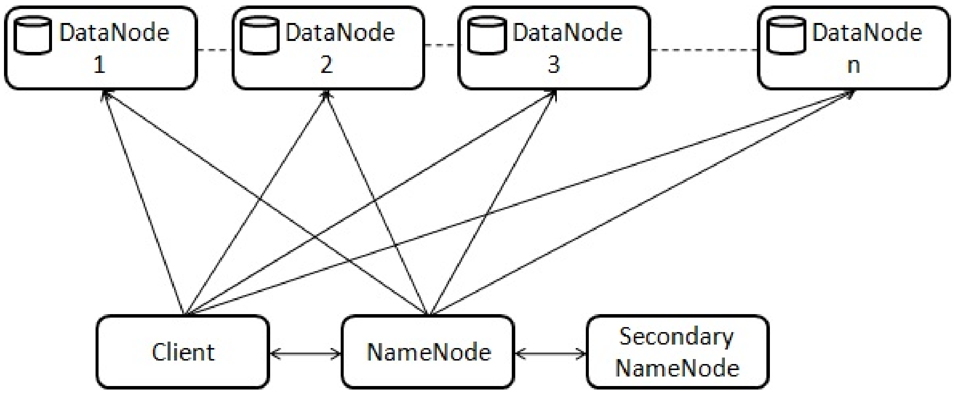


Figure 3.3: Hadoop Architecture

Hadoop consists of master server which controls all the activities of the hadoop cluster. The data nodes works for the master server. The figure 3.3 provides the hadoop architecture. The master server node is named as NameNode and the Slave node is named as data node.

1. NameNode: The file system metadata is managed and controlled by the NameNode. It handles the hadoop cluster with different control services. Only a single NameNode process run on the hadoop file system within the environment.

2. Backupnode: The backup of the NameNode to secure the metadata of the file system is provided by the Backupnode.

3. DataNode: Data operations like storage and retrieval are carried in this process. In a single NameNode, many DataNode processes may run.

4. JobTracker: The distribution and controlling of the jobs in the hadoop environment is handled by JobTracker.

5. TaskTrackter: The Map and Reduce tasks in the datanode is managed by TaskTracker

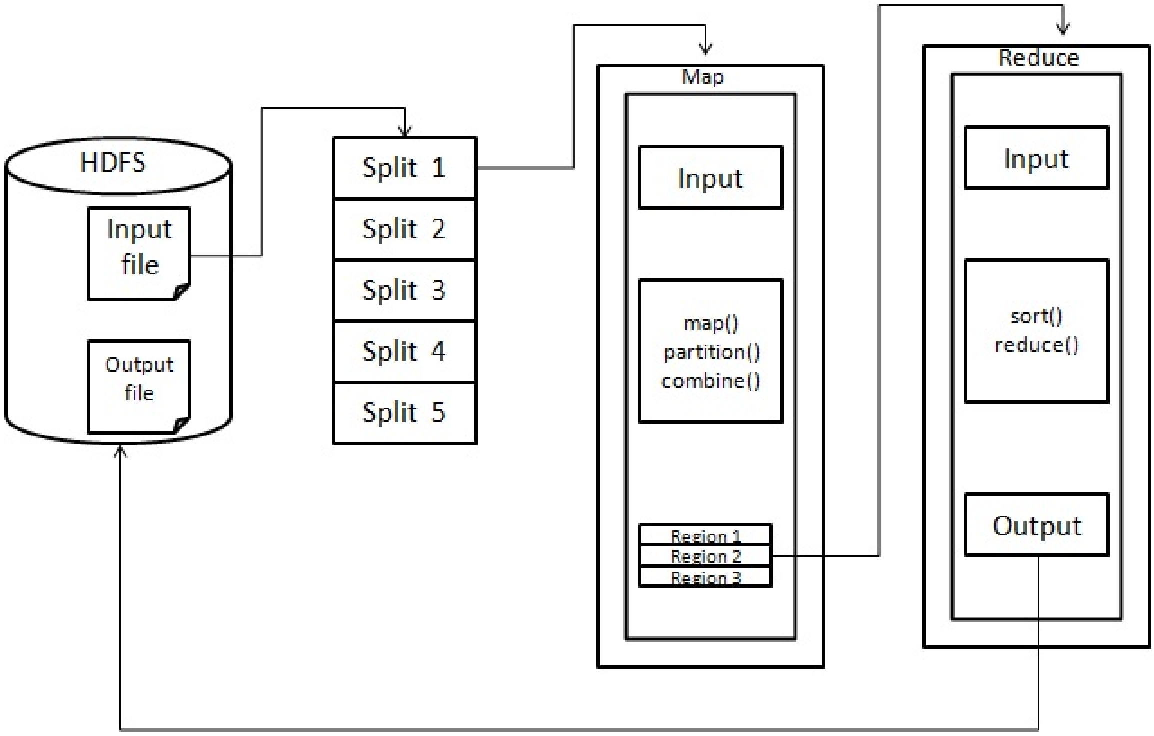


Figure 3.4: MapReduce Architecture

The data is forwarded to the mapper and reducer tasks as shown in the fig 3.4 which processes the records. The mappers and reducers run on the individual nodes where the data records are present.

- The Hadoop system receives the input from a file and splits the content across the Map nodes.
- The Mapper function is executed and an output is generated from each node.
- The generated result from the mapper is represented as a set of key-value pairs
- The key-value pairs are transferred to the reduce nodes as input
- The Reducer function is executed and an output is generated from each node
- The Hadoop system receives the output from each node and aggregates the result set

3.5 Cost Analyzer

Cost Analyzer is a simple application developed on the java platform. It collects the analyzed report file from the HDFS and identifies the specification required in the cloud for that particular utilization from the report file. It calculates the average cost of the on premise infrastructure and the cost of using the cloud for the same utilization . Once, both the costs are calculated they are compared and the cheapest option is provided. It also gives the option of selecting the cloud vendors among AWS and Azure based on the same cost parameter. The static average costs of the clouds are used in the cost analyzer to help with the idea of migrating to cloud or not. It is not using the real time costs of the public clouds.

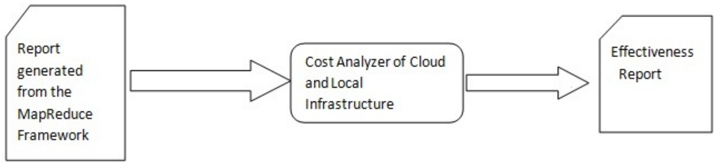


Figure 3.5: Cost Analyzer

Chapter 4

Implementation

Moodle is an open source web application and different educational institutions uses Moodle as online learning sites.As a proof of concept, an application server running Moodle-Learning Management System (LMS) installed on Ubuntu 12.04 server in the Amazon public cloud [fig: 4.1]. Different virtual courses and documents are created and uploaded on this server for testing purpose

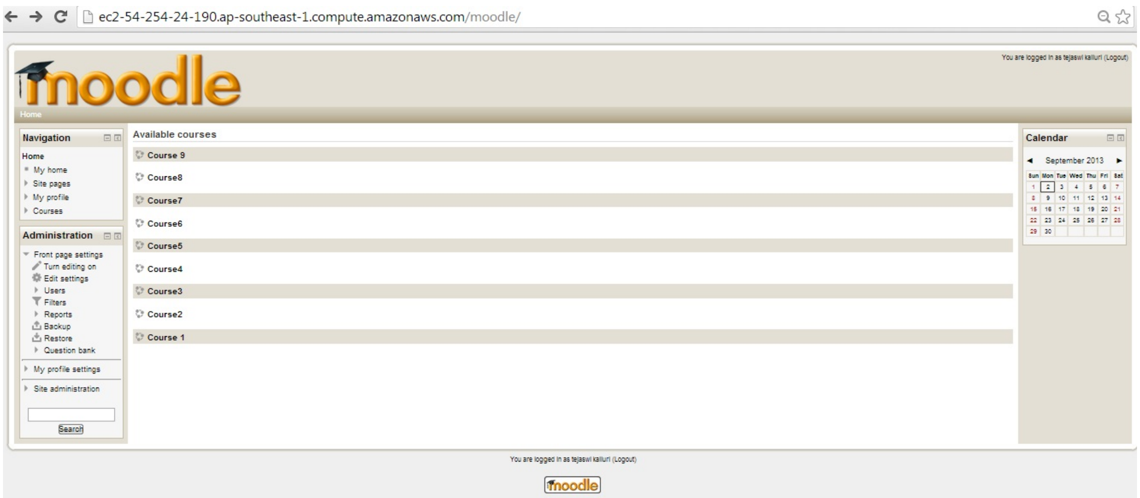


Figure 4.1: Moodle Server

Using Apache JMeter the usage of the web server is recorded with different loads and played with different number of virtual users to increase the load of the server. This Jmeter can virtually create users and virtually manages to use the moodle as the real users. The figures in the Appendix A shows the usage of Apache JMeter for load testing of the application server.

Now, the application server is ready with different number of virtual users. The logs of CPU usage, Memory usage and Storage usage has to be collected. In order to get these logs scripts are generated and executed every minute using crontab.

```
echo   ‘hostname‘ ‘date +%d- %m-%y, %H%M‘ ‘sar 1 59 |tail -1‘ >> /home/ubuntu/thesis/logs/cpu-sar-text
echo   ‘hostname‘ ‘date +%d- %m-%y, %H%M‘ ‘sar -r 1 59 |tail -1‘ >> /home/ubuntu/thesis/logs/memory-sar-text
echo   ‘hostname‘ ‘date +%d- %m-%y, %H%M‘ ‘df -h |head -3|tail -1‘ >> /home/ubuntu/thesis/logs/disk-sar-text
```

The above scripts will generate the logs of CPU, Memory and Storage. It log getting generated by these scripts will consists of the host IP address to determine the load of the particular host with date. These scripts will produce the utilization of every one second. These logs have to be sent to the Hadoop master node using Flume. The application server has to be configured with the flume with the required logs that has to be sent to the Hadoop master. Apache Flume has been downloaded from the official web link of Flume. The Flume has been configured to send the appropriate logs to the Hadoop master and the flume has been configured accordingly for CPU, Memory and Disk.

4.1 Setting up Flume

The flume agent has to be configured in the local environment for data flow. The configuration of flume agent follows the format of Java Properties file. This flume agent configuration file provides the properties of source, channel and sink in an agent and how the data is subjected to flow between them. The components of Flume (Source, Sink and Channel) specifically has a type, name and different set of properties. The agent requires to read the each component and their properties to determine their flow structure.This can be determined by providing the connecting channel of the source and the sinks. Here, is the flume agent configuration file provides all the required properties of each components. Different flume agents can be configured in a single file.

```
flume.conf.cpuagent.sources = apache
agent.sources.apache.type = exec
agent.sources.apache.command = // The log file of the CPU or Memory or Disk
agent.sources.apache.channels = memoryChannel
agent.sources.apache.batch-size = 100

agent.channels = memoryChannel
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 100
agent.channels.memoryChannels.transactionCapacity = 100
```

```
agent.sinks = HDFS
agent.sinks.HDFS.channel = memoryChannel
agent.sinks.HDFS.type = // Type of sink data
agent.sinks.HDFS.hostname = // The IP address of the Hadoop Master
agent.sinks.HDFS.port = 10011
agent.sinks.HDFS.hdfs.fileType = Datastream
agent.sinks.HDFS.batch-size = 100
```

The flume-ng is a shell script available in the bin directory of the flume. It is used to start an agent. To start the agent the name of the agent with the configuration directory and the configuration file name has to be provided.

```
bin/flume-ng agent --conf ./conf/ -f conf/flume.conf.cpu -n agent
bin/flume-ng agent --conf ./conf/ -f conf/flume.conf.memory -n agent
bin/flume-ng agent --conf ./conf/ -f conf/flume.conf.storage -n agent
```

This set up has to be configured in both the moodle and the master server. Once, the flume in both the servers starts, the logs starts moving from moodle to the hadoop master.

4.2 Setting up Hadoop Cluster

A single node Hadoop cluster has been installed on Ubuntu 12.04 in the Amazon public cloud B following the installation procedure

1. Install Java 6 and set the Java environment variable

```
apt-get install openjdk-6-jdk
export JAVA_HOME = /usr/lib/jvm/java-6-openjdk-amd64/
```

2. Download the Hadoop package from the Apache Hadoop mirror list, untar and set the Hadoop environment variable

```
wget http://ftp.heanet.ie/mirrors/www.apache.org/dist/hadoop/core/hadoop ↩
    -2.0.5-alpha/hadoop-2.0.5-alpha-src.tar.gz
tar -xzvf hadoop-2.0.5-alpha-src.tar.gz
export HADOOP_HOME=/usr/local/hadoop
```

3. In conf/hadoop-env.sh, set the Java environment variable path

4. In conf/core-site.xml, set the configuration

```
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/ubuntu/hadoop/hdata/tmp/</value>
</property>

<property>
  <name>fs.default.name</name>
  <value>hdfs://master:9000</value>
</property>
```

5. In conf/hdfs-site.xml, change the value for the replication count.

```
<property>
  <name>dfs.replication</name>
  <value>3</value>
 </property>
```

6. In conf/mapred-site.xml, set the job tracker with its port

```
<property>
  <name>mapred.job.tracker</name>
  <value>master:9000</value>
</property>
```

7. Formatting the Hadoop Distributed File System from Namenode and start all the services of Hadoop (i.e. NameNode, DataNode, JobTracker and TaskTracker)

```
/usr/local/hadoop/bin/hadoop namenode -format
/usr/local/hadoop/bin/start-all.sh
```

Now, the Hadoop system is ready and the flume is sending the log files to the Hadoop master node from the application server. The large amount of logs collected are to be analyzed to get the final report for the usage. This is carried by using the MapReduce programs. The logs are provided to the MapReduce program in the Hadoop system which analyze the large datasets and give the aggregate utilization value. The map function and the reduce function from the mapper and the reducer classes repectively are extended and programmed to split the large amount of datasets and calculate the aggregate value of the logs

```
public void map(LongWritable key, Text value, Context context)
{
  String valueTokens[] = value.toString().split(" ");
  if (valueTokens.length == 10)
  {
    String hostName = valueTokens[0];
```

```
      String date = "";
      try
      {
        date = valueTokens[1].split(",")[0];
      }
      catch (Exception e)
      {
        e.printStackTrace();
      }
      try
      {
        float cpuPercent = 100.0f - Float.parseFloat(valueTokens[valueTokens.length - ↩
              1]);
        moKey.set(hostName + "\t" + date);
        percentVal.set(cpuPercent);
        context.write(moKey, percentVal);
      }
      catch (Exception e)
      {
        e.printStackTrace();
      }
    }
  }
```

```
public void reduce(Text key, Iterable<FloatWritable> values, Context context) throws ↩
      IOException, InterruptedException
  {
    int count = 0;
    float sum = 0.0f;
    for (FloatWritable val : values)
    {
      sum += val.get();
      count++;
    }
    float aggregate = sum / count;
    result.set(aggregate + "%");
    context.write(key, result);
  }
```

These map and reduce functions runs on all the logs i.e CPU, Memory and Storage to get the aggregate utilization of the server. This is the final analyzed report for the server utilization. This report is forwarded to the Cost Analyzer application

4.3 CostAnalyzer

CostAnalyzer is a simple java application. It reads the utilization report obtained from the MapReduce of the Hadoop master and starts to calculate the average cloud cost and compare it with the cost of on premise infrastructure to provide the final cost report [Fig 3.5]. This cost analyzer also gives the appropriate result set to check with different cloud providers such as AWS and Azure.

```
public class CostAnalyzer {
  static double local_cost;
  static double cloud_cost;
  LocalCostCalculator lcc = new LocalCostCalculator();
  local_cost = lcc.avgLocalCost();
  AvgCloudCostCalculator ccc = new AvgCloudCostCalculator();
  cloud_cost = ccc.avgCloudCost();
  ccc.cloudComparison();
  }
```

The cost analyzer calculates the local infrastructure cost by summing up all the parameters as discussed in the background. The function that calculates the local infrastructure cost is as follows

```
public double avgLocalCost() {
    double total_local_cost;
    total_local_cost = (local_number_server * local_server_cost)
        + (local_storage_cost * local_storage_req) + local_os_cost
        + local_application_cost + (local_number_emp * local_emp_sal)
        * local_number_month;
    System.out.println("The cost of Local infrastucture Usage is " +total_local_cost );
    return total_local_cost;
  }
```

Similarly, from the analyzed log report file the requirement of resources in the cloud environment is estimated by the function instanceSelection and the cost of this estimated specification is calculated using the function avgCloudCost as provided.

```
public String instanceSelection() {

    double cpu_utilization = Double.parseDouble(cpu);
    double mem_utilization = Double.parseDouble(mem);
    String instanceType = "";
    if (// Comparison of CPU utilization ) {
    instanceType = type; //Assigning the type of the instance based on the utilization
    if (//Comparison of Memory utilization ) {
```

```
    instanceType = type; //Assigning the type of the instance based on the ↩
        utilization
    }
    return instanceType;
  }
}
```

```
  public double avgCloudCost() {
    double total_cloud_cost;
    SystemSpecAnalyzer ssa = new SystemSpecAnalyzer();
    instance_type = ssa.instanceSelection();
    cloud_server_cost = ssa.instanceCost(instance_type);
    cloud_storage_req = ssa.storageReq();
    total_cloud_cost = ((cloud_number_server * cloud_server_cost)
        + (cloud_storage_cost * cloud_storage_req+2500))
        * cloud_num_months;
    System.out.println("The Average cost of Cloud Usage is " + total_cloud_cost);
    return total_cloud_cost;
}
```

The return values of the average local cost and the average cloud costs are compared with eachother to analyze the better option to select based on the cost parameter

```
public void costComparison(){
        LocalCostCalculator lcc = new LocalCostCalculator();
        local_cost = lcc.avgLocalCost();
        AvgCloudCostCalculator ccc = new AvgCloudCostCalculator();
        cloud_cost = ccc.avgCloudCost();
        if (local_cost > cloud_cost) {
          double percent_diff;
          percent_diff = ((local_cost - cloud_cost) / local_cost) * 100;
        } else {
          double percent_diff;
          percent_diff = ((cloud_cost - local_cost) / cloud_cost) * 100;
        }
    }
```

From the above function, the better option for running the computational resources can be selected. The cost analyzer also provides a feature of selecting the cheapest cloud out in the market with the current market prices for the specification drawn from the program.

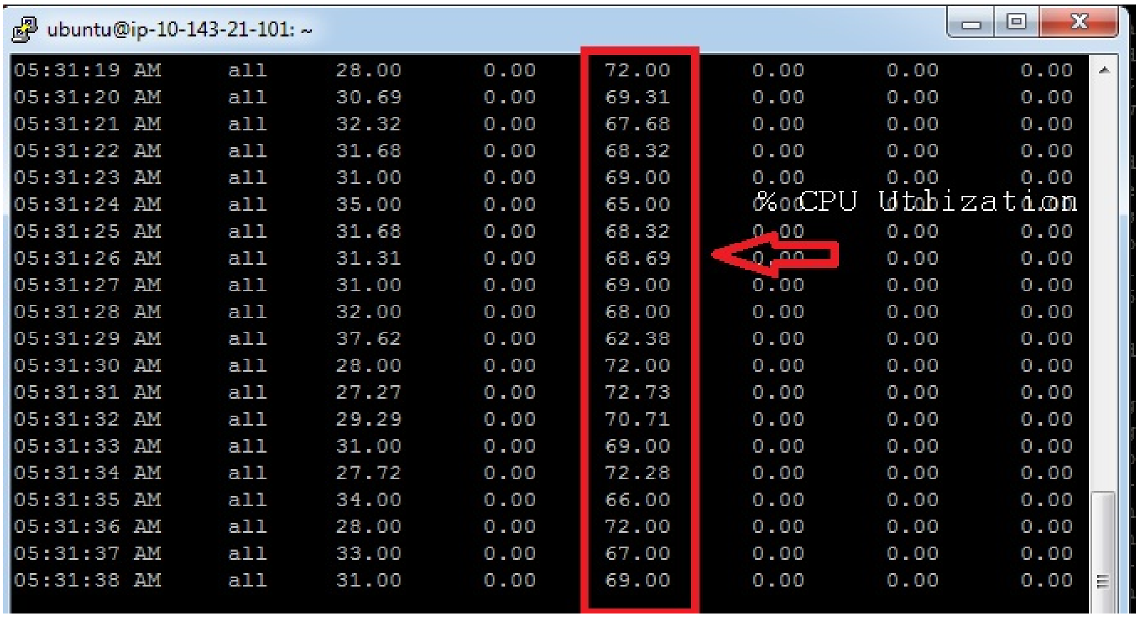
```
public void cloudComparison(){
    AmazonCostCalculator aws = new AmazonCostCalculator();
    double awi = aws.awsCloudCost();
    AzureCloudCost azc = new AzureCloudCost();
    double azi = azc.azureCloudCost();
```

```
  if(awi < azi){
    double percent_diff = ((azi-awi)/azi)*100;
    }else{
    double percent_diff = ((azi-awi)/azi)*100;
  }
}
```

Chapter 5

Evaluation

The usage of computational infrastructure varies continuously. It is required to analyze the areas for cost savings, which cloud computing can bring to an organizations. As a proof of concept, It is considered to analyze the utilization in the local environment and to estimate the cost of that usage in the cloud systems. This will provide the understanding of cost savings when migrated to cloud. To get the utilization of the local environment, a Moodle server is installed as described in the chapter 4, using Apache JMeter the utilization of the server is varied with different number of virtual users. The server utilization in terms of CPU, Memory and Storage are collected. The figures 5.1, 5.2 provides the logs of the cpu and memory utilization of the moodle server.



05:31:19 AM	all	28.00	0.00	72.00	0.00	0.00	0.00
05:31:20 AM	all	30.69	0.00	69.31	0.00	0.00	0.00
05:31:21 AM	all	32.32	0.00	67.68	0.00	0.00	0.00
05:31:22 AM	all	31.68	0.00	68.32	0.00	0.00	0.00
05:31:23 AM	all	31.00	0.00	69.00	0.00	0.00	0.00
05:31:24 AM	all	35.00	0.00	65.00	[illegible]	[illegible]	[illegible]
05:31:25 AM	all	31.68	0.00	68.32	0.00	0.00	0.00
05:31:26 AM	all	31.31	0.00	68.69	[illegible]	0.00	0.00
05:31:27 AM	all	31.00	0.00	69.00	0.00	0.00	0.00
05:31:28 AM	all	32.00	0.00	68.00	0.00	0.00	0.00
05:31:29 AM	all	37.62	0.00	62.38	0.00	0.00	0.00
05:31:30 AM	all	28.00	0.00	72.00	0.00	0.00	0.00
05:31:31 AM	all	27.27	0.00	72.73	0.00	0.00	0.00
05:31:32 AM	all	29.29	0.00	70.71	0.00	0.00	0.00
05:31:33 AM	all	31.00	0.00	69.00	0.00	0.00	0.00
05:31:34 AM	all	27.72	0.00	72.28	0.00	0.00	0.00
05:31:35 AM	all	34.00	0.00	66.00	0.00	0.00	0.00
05:31:36 AM	all	28.00	0.00	72.00	0.00	0.00	0.00
05:31:37 AM	all	33.00	0.00	67.00	0.00	0.00	0.00
05:31:38 AM	all	31.00	0.00	69.00	0.00	0.00	0.00

Figure 5.1: % CPU Utilization of Moodle Server

The marked box represents the CPU utilization of the system and it is generated on the moodle server. As the load varies on the server the utilization also vary and it is depicted in the figure 5.1. These logs are generated for every second.

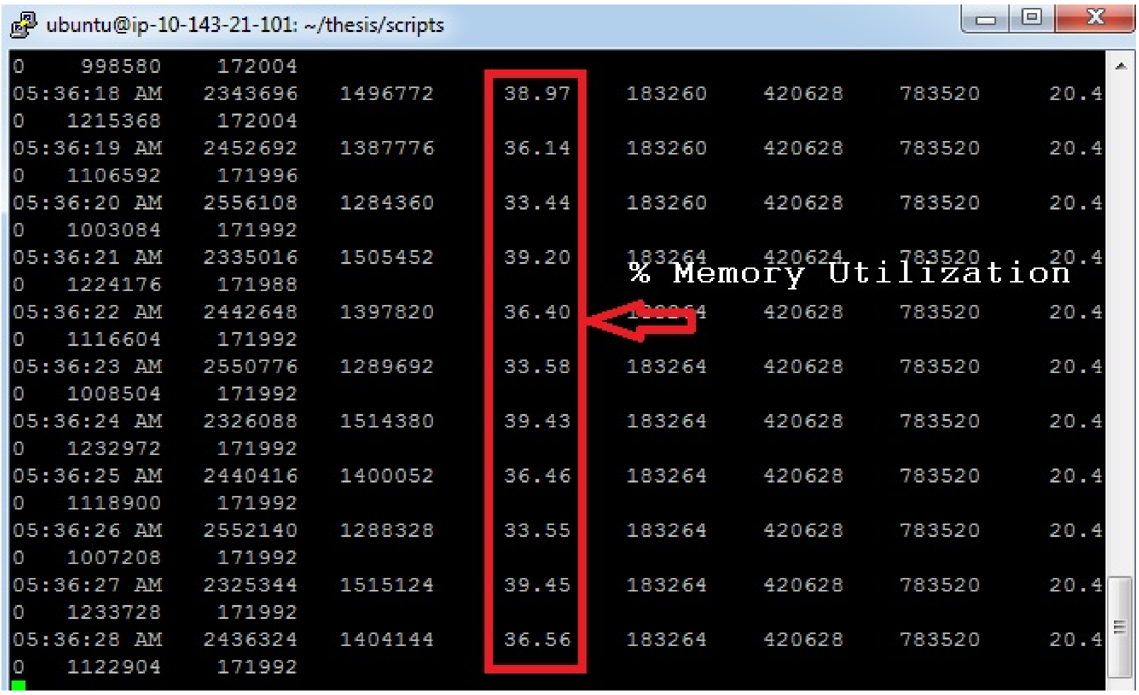
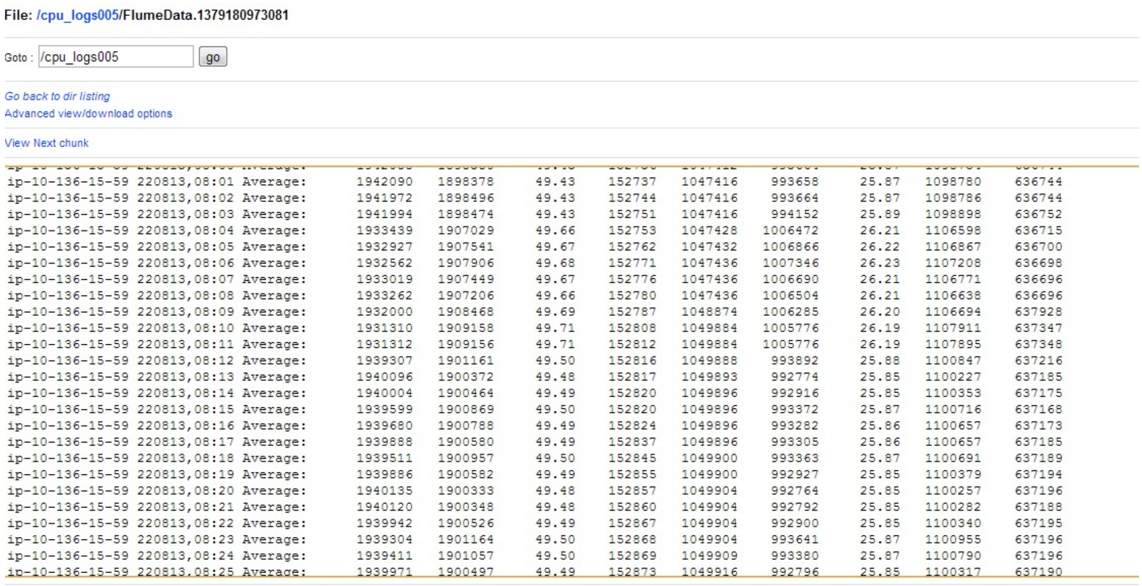


Figure 5.2: % Memory utilization of Moodle Server

In the marked box, the memory utilization of the system and it is generated on the moodle server. As the load varies on the server the utilization also vary and it is proportional to the CPU utilization and it is clearly depicted in the figure 5.2. These logs are generated for every second.

These logs are generated using a script and crontab is used to continuously run the script. The logs are collected in a file and using flume, the logs of CPU and Memory are transfered to Hadoop master node using flume as shown in the figure 5.3.



File: /cpu_logs005/FlumeData.1379180973081

Goto : /cpu_logs005 go

Go back to dir listing
Advanced view/download options

View Next chunk

```
ip-10-136-15-59 220813,08:01 Average:      1942090   1898378     49.43    152737   1047416    993658      25.87   1098780     636744
ip-10-136-15-59 220813,08:02 Average:      1941972   1898496     49.43    152744   1047416    993664      25.87   1098786     636744
ip-10-136-15-59 220813,08:03 Average:      1941994   1898474     49.43    152751   1047416    994152      25.89   1098898     636752
ip-10-136-15-59 220813,08:04 Average:      1933439   1907029     49.66    152753   1047428   1006472      26.21   1106598     636715
ip-10-136-15-59 220813,08:05 Average:      1932927   1907541     49.67    152762   1047432   1006866      26.22   1106867     636700
ip-10-136-15-59 220813,08:06 Average:      1932562   1907906     49.68    152771   1047436   1007346      26.23   1107208     636698
ip-10-136-15-59 220813,08:07 Average:      1933019   1907449     49.67    152776   1047436   1006690      26.21   1106771     636696
ip-10-136-15-59 220813,08:08 Average:      1933262   1907206     49.66    152780   1047436   1006504      26.21   1106638     636696
ip-10-136-15-59 220813,08:09 Average:      1932000   1908468     49.69    152787   1048874   1006285      26.20   1106694     637928
ip-10-136-15-59 220813,08:10 Average:      1931310   1909158     49.71    152808   1049884   1005776      26.19   1107911     637347
ip-10-136-15-59 220813,08:11 Average:      1931312   1909156     49.71    152812   1049884   1005776      26.19   1107895     637348
ip-10-136-15-59 220813,08:12 Average:      1939307   1901161     49.50    152816   1049888    993892      25.88   1100847     637216
ip-10-136-15-59 220813,08:13 Average:      1940096   1900372     49.48    152817   1049893    992774      25.85   1100227     637185
ip-10-136-15-59 220813,08:14 Average:      1940004   1900464     49.49    152820   1049896    992916      25.85   1100353     637175
ip-10-136-15-59 220813,08:15 Average:      1939599   1900869     49.50    152820   1049896    993372      25.87   1100716     637168
ip-10-136-15-59 220813,08:16 Average:      1939680   1900788     49.49    152824   1049896    993282      25.86   1100657     637173
ip-10-136-15-59 220813,08:17 Average:      1939888   1900580     49.49    152837   1049896    993305      25.86   1100657     637185
ip-10-136-15-59 220813,08:18 Average:      1939511   1900957     49.50    152845   1049900    993363      25.87   1100691     637189
ip-10-136-15-59 220813,08:19 Average:      1939886   1900582     49.49    152855   1049900    992927      25.85   1100379     637194
ip-10-136-15-59 220813,08:20 Average:      1940135   1900333     49.48    152857   1049904    992764      25.85   1100257     637196
ip-10-136-15-59 220813,08:21 Average:      1940120   1900348     49.48    152860   1049904    992792      25.85   1100282     637188
ip-10-136-15-59 220813,08:22 Average:      1939942   1900526     49.49    152867   1049904    992900      25.85   1100340     637195
ip-10-136-15-59 220813,08:23 Average:      1939304   1901164     49.50    152868   1049904    993641      25.87   1100955     637196
ip-10-136-15-59 220813,08:24 Average:      1939411   1901057     49.50    152869   1049909    993380      25.87   1100790     637196
ip-10-136-15-59 220813,08:25 Average:      1939971   1900497     49.49    152873   1049916    992796      25.85   1100317     637190
```

Figure 5.3: Logs in the HDFS obtained from flume

Once, the logs gets collected on the master node. It runs the MapReduce to analyse all the logs and calculates the average utilization values for each day and provides the report file as in figure 5.4 . As the logs are generated for every second in the moodle server it is aggregated to get the cost of utilization for one day. The total utilization is summed up and is divided with 24*60*60 (Seconds per day).

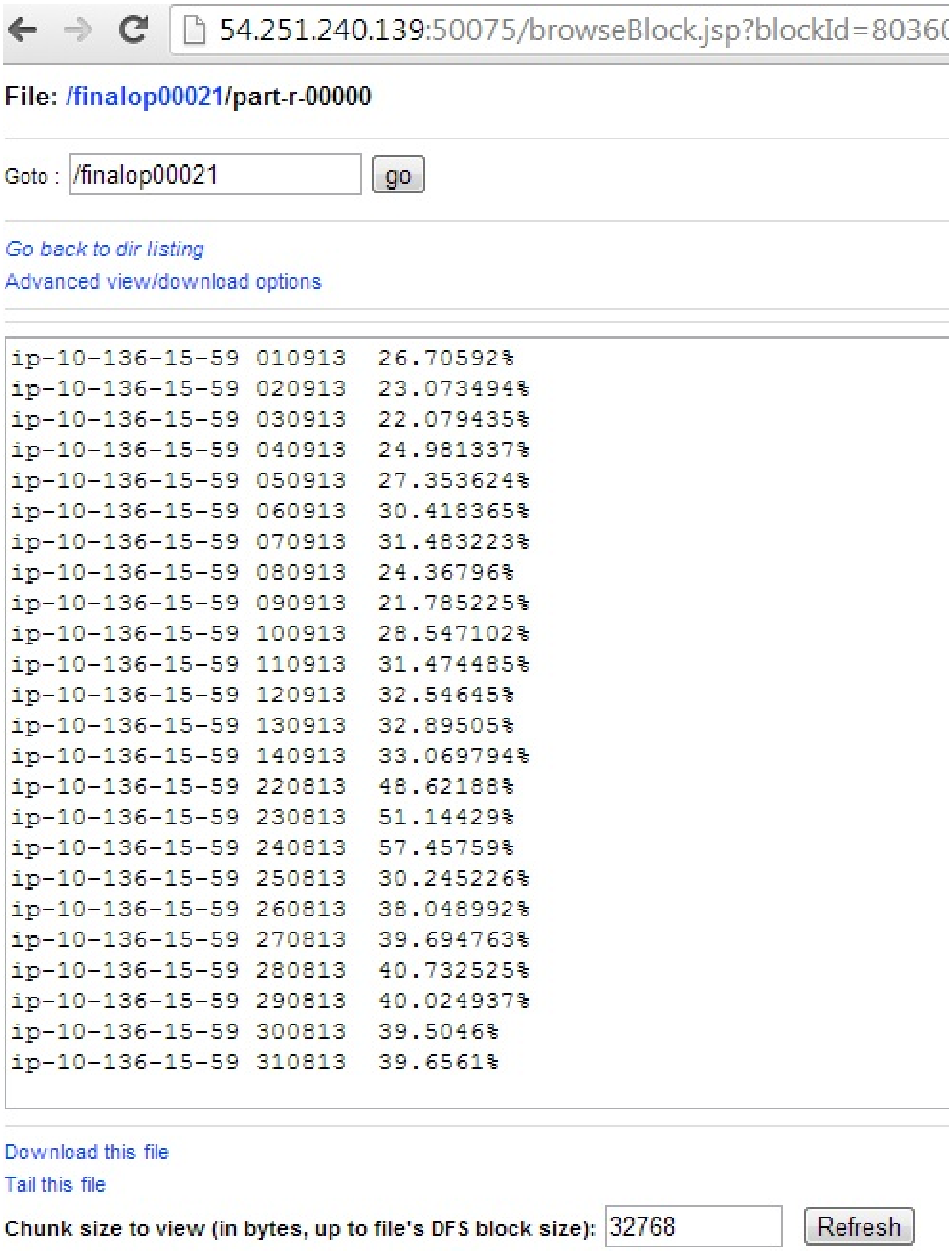


Figure 5.4: Analyzed Log report

As the log files are analyzed by the MapReduce and the log report file is generated the average utlization value of the server resources can be examined. The figures 5.5, 5.6 depicts a graph for the average utilization values of CPU and Memory for a period of one week is obtained from the MapReduce. This CPU and Memory values are obtained when the load on the server is varied with 10-30 virtual users using Apache JMeter.

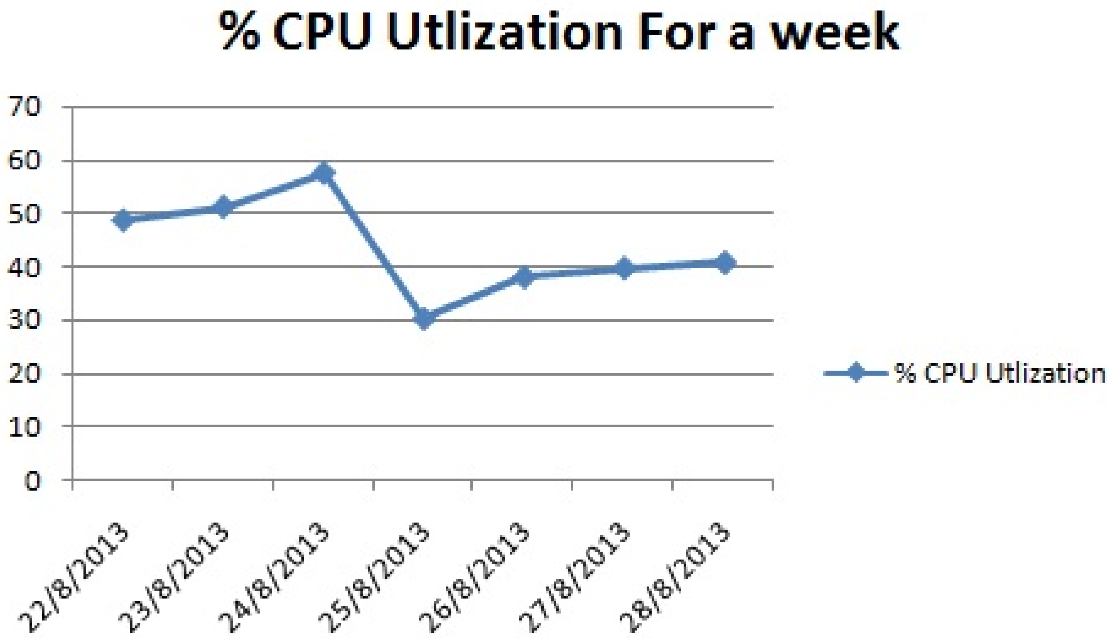


Figure 5.5: CPU utilization of the Moodle Server

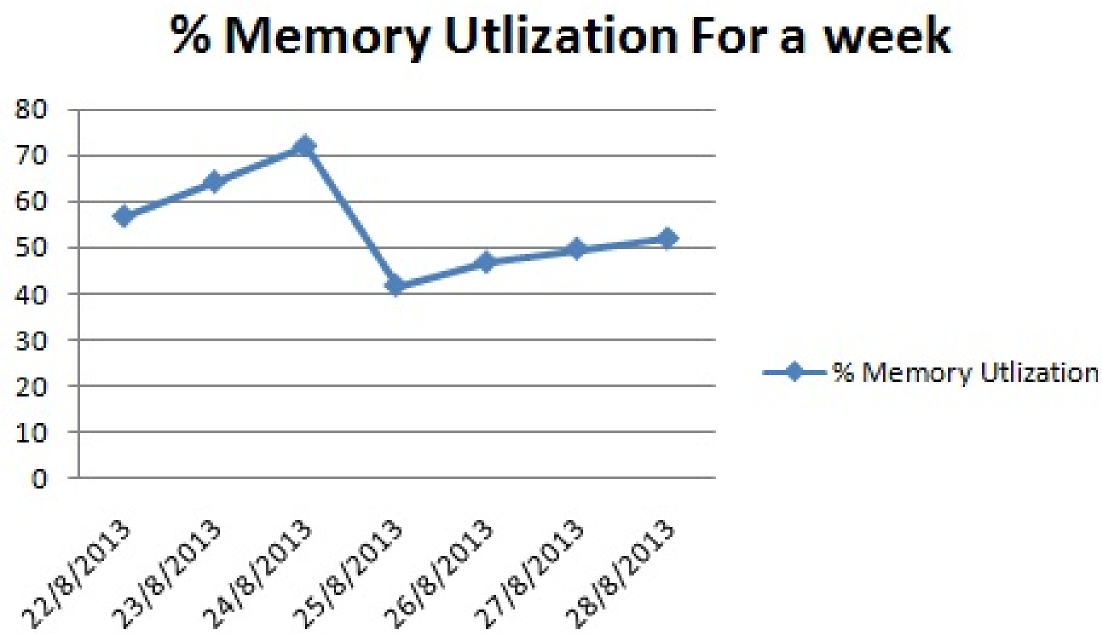


Figure 5.6: Memory Utilization of the Moodle Server

The load on the test moodle server is generated only few days to represent the variation in the utilization of the server. In a period of one week the utilization is calculated and graphed. From the graph it is clear that, the utilization is at peak few days in a week and rest of the week it is low means less number of users utilizing the server. The memory utilization is graphed for the time for the same period, which provides the accurate resource utilization of the server. From the graph, it is clear that the resources are over provisioned and under provisioned depending on the time and load on the server. This over provisioning and under provisioning are proportionally related to the Cost of utilization. Cloud Computing with its characteristic of scalability can avoid the issue of over provisioning and under provisioning by scaling up and scaling down the server count depending on the usage in less time and with no extra set up costs.

The analyzed report [Fig : 5.4] is forwarded to the Cost Analyzer which estimates the utilization and predicts the cost of running the local infrastructure with the cloud cost. It also compares the price of different clouds to evaluate the cheapest as in the figure 5.7.

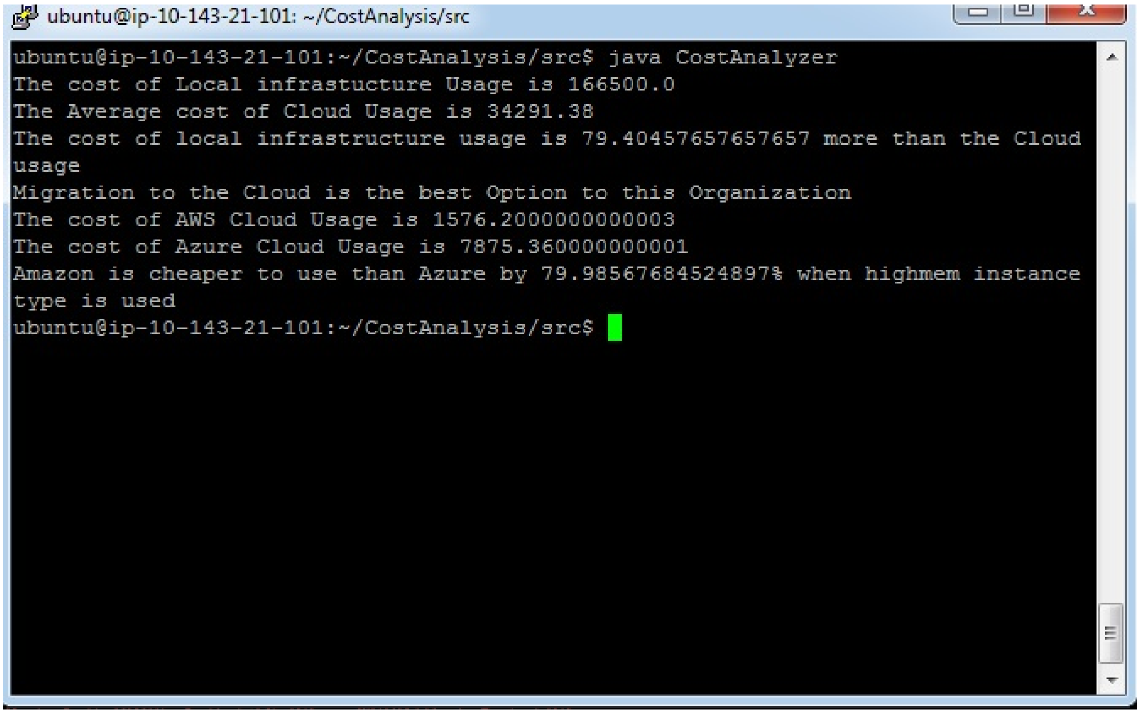
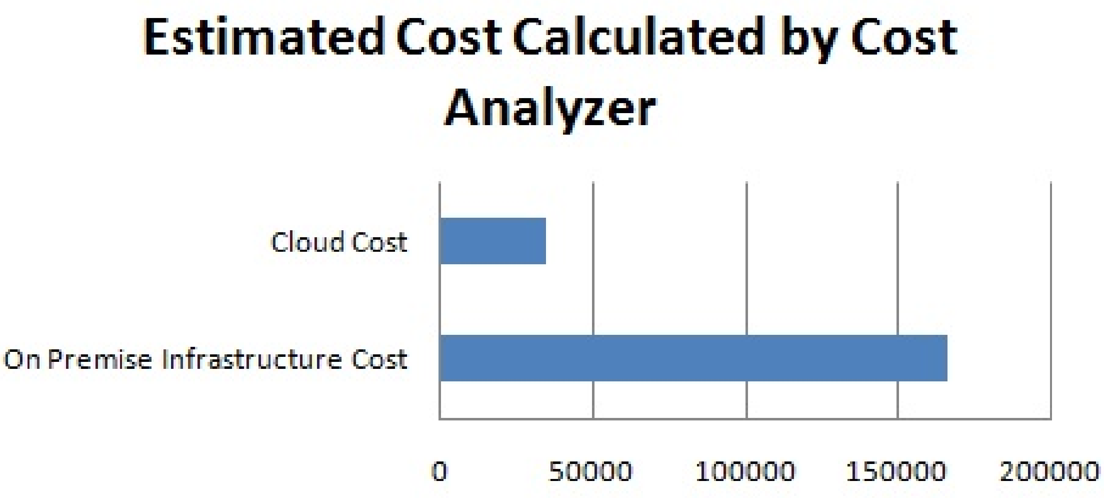


Figure 5.7: Cost Report obtained from Cost Analyzer

The difference in the variation in the costs of cloud and on premise infrastructure is depicted in the graph 5.8 obtained from the Cost Analyzer.



	On Premise Infrastructure Cost	Cloud Cost
Series1	166500	34291.38

Figure 5.8: Difference between Cloud and On premise infrastructure cost

From the analyzed logs of MapReduce, the specification for that utilization report is created and the cost of such specification for the cloud is calculated, which predicts the saving of average of 70% if migrated to the Cloud from the local infrastructure. The tables 5.1, 5.2, 5.3 shows the internal calculations of the Cost analyzer for predicting the cost of local and cloud infrastructure. The table 5.1, describes the approximate cost of on premise computing infrastructure calculated using the VMWare TCO calculator (http://roitco.vmware.com) and used in the Cost Analyzer. VMware provides to a simple TCO calculator to compare the cost of using the non-virtualization solution with a virtualization solution. The cost of the local infrastructure is obtained from the VMware calculator.

On-Premise Computing Infrastructure Cost	
Hardware (servers)	$1500
Hardware (storage) + Software	$12000
Hardware (Security and Networking)	$80000
Administrative Staff	$30000
Electricity Costs	$4000
Room Space	$20000
Others	$2000
Total	$147500
These values are calculated by using VMWare Server TCO Calculator	

Table 5.1: Approximation of On-Premise Computing Cost

From the logs, the specific cloud requirement is estimated and the cost for that requirement on AWS public cloud(http://calculator.s3.amazonaws.com) and Azure public cloud (https://www.windowsazure.com) is calculated taking the values from the AWS monthly cost calculator and Azure Cost calculator. The table 5.2 gives the average cost of using cloud for the same configuration mentioned in table 5.1. The percentage difference between the on premise and the cloud cost as from the calculated values is 79%.

Cloud Computing Cost	
Hardware (servers)	$2800
Hardware (storage)	Included
Hardware (Security and Networking)	Included
Software	$2000
Administrative Staff	$3000
Electricity Costs	$400
Others	$1000
Total	$9200
These values are the average cost of usage of Extra large Instances in Amazon and Azure	

Table 5.2: Approximation of Cloud Cost

The above evaluation shows the cost of using cloud is much cheaper and better when compared to on premise infrastructure for the analyzed average utilization. Choosing the cloud provider is also important in terms of reducing the cost. The table 5.3 depicts the costs of Amazon and Azure instances (Medium, Large, XLarge, CPU intensive Medium and CPU intensive Large) for hour. Such cost analysis is required for effectively reduce the costs in the clouds and get more resources.

AWS and Azure On-demand instance Cost/hour		
Instance Type	AWS	Azure
Extra Small	$0.02	$0.02
Small	$0.091	$0.115
Medium	$0.182	$0.23
Large	$0.364	$0.46
Extra Large	$0.728	$0.92

Table 5.3: Cost Comparison between AWS and Azure for different Instance types

From these evaluation, it is analyzed that the cost of using cloud in the educational instituitions is much cheaper than running on premise IT infrastructure. The cost variance for year is analyzed to be more than 70% cheaper including all the expenses.

Chapter 6

Conclusion

Cloud Computing has clearly made a mark in the ICT sector in terms of cost. From the various services like software, platform and infrastructure any organizations can be benefited a lot for using Cloud Computing. The IT department can be reduced in size as the maintainance, software support and updates are not required while using the Cloud senarios. Students can be much satisfied by making them to use the updated technology in the campus. Despite of various issues in this area, the financial situation of the educational institutions will force them to adobt Cloud Computing inplace of On-premise infrastructure. Many universities have already migrated to the cloud infrastructure and has successfully reduced the expenses of the IT infrastructure. Many oranizations like Microsoft, Amazon, Google have already introduced new services by providing free access and education grants for the educational institutes to attract them to adobt Cloud Computing.

The objective of this paper is to investigate the cost effectiveness of the Cloud Computing system in the educational institutions. The increasing number of users utilizing the IT infrastructure in the educational institutions raised the requirement of huge capital investment. This paper critically discussed about the actual cost of IT infrastructure and translation cost with the Cloud. In order to provide the benefit of Cloud Computing in the educational institutions in terms of cost, the logs of most popular LMS application is used as a test server and logs are analyzed by implementing the log analyzer using parallel programming model. The cost of resource utilization in the local and cloud infrastructure is compared using the Cost analyzer which arguebly states that the use of Cloud Computing is better in terms of cost.

The evaluation is carried by analyzing the obtained logs using the MapReduce and the specification is designed according to the utilized resources. The cost is compared between the on premise and the cloud system. It is clearly presented that the total cost of on premise compared to cloud cost for the same resources is more than 70%. This analyzes that the Cloud Computing is better option for educational institutions for their purposes.This set up can also be used to analyze the utilization of each server in the cluster.

This work can be extended in:

- Providing the dynamic cost report for the usage
- Building a web application to compare the cost based on the utilization of the resources and provide the report to the users on demand based.

Bibliography

[1] Google app engine for education. http://www.google.com/enterprise/apps/education.

[2] Wholeschool-education service providers. http://www.wholeschool.ie.

[3] Kirsti Ala-Mutka, Christine Redecker, Yves Punie, Anusca Ferrari, Romina Cachia, and Clara Centeno. The future of learning: European teachers visions. In *Report on a foresight consultation at the 2010 eTwinning Conference, Seville*, pages 5–7, 2010.

[4] Mohssen M Alabbadi. Cloud computing for education and learning: Education and learning as a service (elaas). In *Interactive Collaborative Learning (ICL), 2011 14th International Conference*, pages 589–594. IEEE, 2011.

[5] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[6] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona 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.

[7] Deka Ganesh Chandra and Malaya Dutta Borah. Cost benefit analysis of cloud computing in education. In *Computing, Communication and Applications (ICCCA), 2012 International Conference*, pages 1–6. IEEE, 2012.

[8] Ganesh Ch Deka, Jasni Mohamad Zain, and Prabhat Mahanti. ICT's role in e-governance in india and malaysia: A review. *arXiv preprint arXiv:1206.0681*, 2012.

[9] Jing Dong, Jun Han, Jing Liu, and Xu Hongyun. The shallow analysis of the enlightenment of cloud computing to distance education. In *E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference*, volume 2, pages 301–303. IEEE, 2010.

[10] Tuncay Ercan. Effective use of cloud computing in educational institutions. *Procedia-Social and Behavioral Sciences*, 2(2):938–942, 2010.

[11] Dan R Herrick. Google this!: using google apps for collaboration and productivity. In *Proceedings of the 37th annual ACM SIGUCCS fall conference*, pages 55–64. ACM, 2009.

[12] Na Jeyanthi, Hena Shabeeb, and N Ch SN Iyengar. A study on security threats in cloud. *International Journal of Cloud Computing and Services Science (IJ-CLOSER)*, 1(3):84–88, 2012.

[13] Philip Koehler, Arun Anandasivam, MA Dan, and C Weinhardt. Cloud services from a consumer perspective. In *AMCIS*, page 329. Citeseer, 2010.

[14] Ramkumar Lakshminarayanan, Binod Kumar, and M Raju. Cloud computing benefits for educational institutions. *arXiv preprint arXiv:1305.2616*, 2013.

[15] Saju Mathew. Implementation of cloud computing in education-a revolution. *education*, 1(5):6, 2012.

[16] Peter Mell and Timothy Grance. The nist definition of cloud computing (draft). *NIST special publication*, 800:145, 2011.

[17] Marinela Mircea and Anca Ioana Andreescu. Using cloud computing in higher education: A strategy to improve agility in the current financial crisis. *Communications of the IBIMA*, 2011.

[18] Karen Myers and Patrice De Broucker. *Too many left behind: Canada's adult education and training system*. Canadian Policy Research Networks, Incorporated, 2006.

[19] Pankesh Patel, Ajith Ranabahu, and Amit Sheth. Service level agreement in cloud computing. In *Cloud OOPSLA*, 2009.

[20] Richard A Piña and Bharat Rao. The emergence and promise of cloud computing for under-developed societies. In *Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET'10:*, pages 1–10. IEEE, 2010.

[21] K Praveena and T Betsy. Application of cloud computing in academia. *IUP Journal of Systems Management*, 7(3):50–54, 2009.

[22] Dan Sanderson. *Programming Google app engine*. O'Reilly Media, 2010.

[23] Nabil Sultan. Cloud computing for education: A new dawn? *International Journal of Information Management*, 30(2):109–116, 2010.

[24] Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, and Guohui Zhou. SAM: A semantic-aware multi-tiered source de-duplication framework for cloud backup. In *Parallel Processing (ICPP), 2010 39th International Conference on*, pages 614–623. IEEE, 2010.

[25] Luis M Vaquero, Luis Rodero-Merino, Juan Caceres, and Maik Lindner. A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1):50–55, 2008.

[26] Bo Wang and HongYu Xing. The application of cloud computing in education informatization. In *Computer Science and Service System (CSSS), 2011 International Conference*, pages 2673–2676. IEEE, 2011.

[27] Yang Wang and Yuan-ting Yan. Design and implementation of cloud computing educational service platform based on p2p. *Computer*, 16:045, 2010.

[28] David C Wyld. *Moving to the cloud: An introduction to cloud computing in government*. IBM Center for the Business of Government, 2009.

[29] Xun Xu. From cloud computing to cloud manufacturing. *Robotics and computer-integrated manufacturing*, 28(1):75–86, 2012.

Appendix A

Apache JMeter

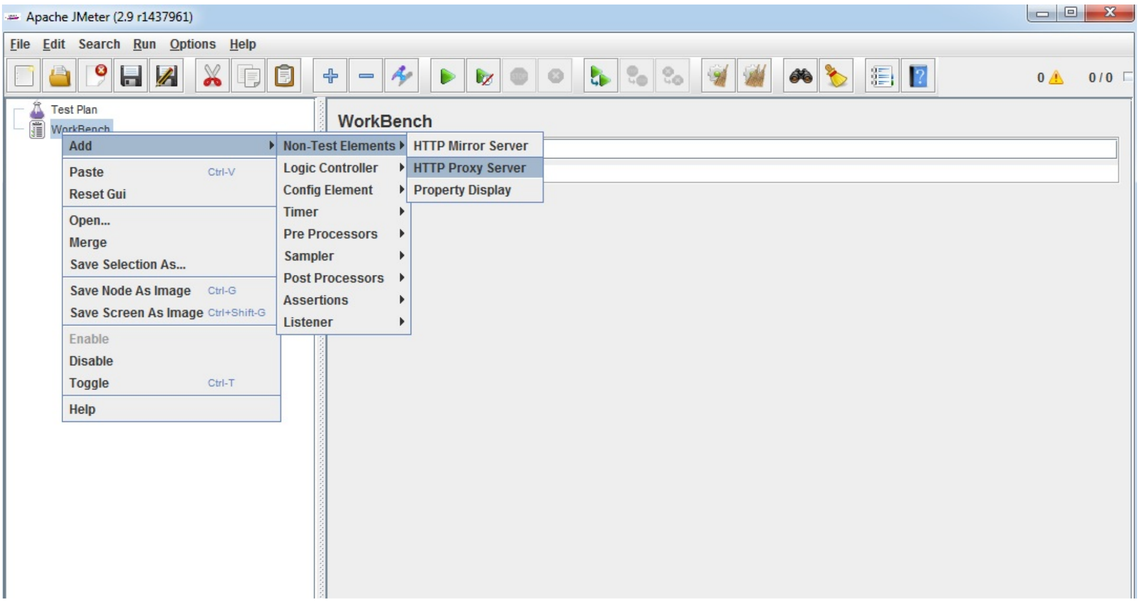


Figure A.1: Jmeter Proxy server initialization

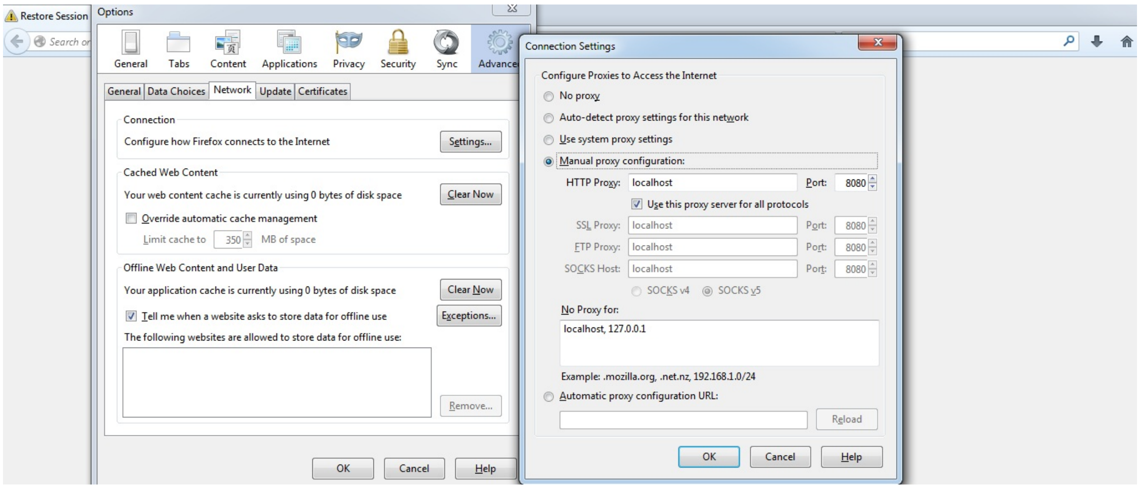


Figure A.2: FireFox network settings

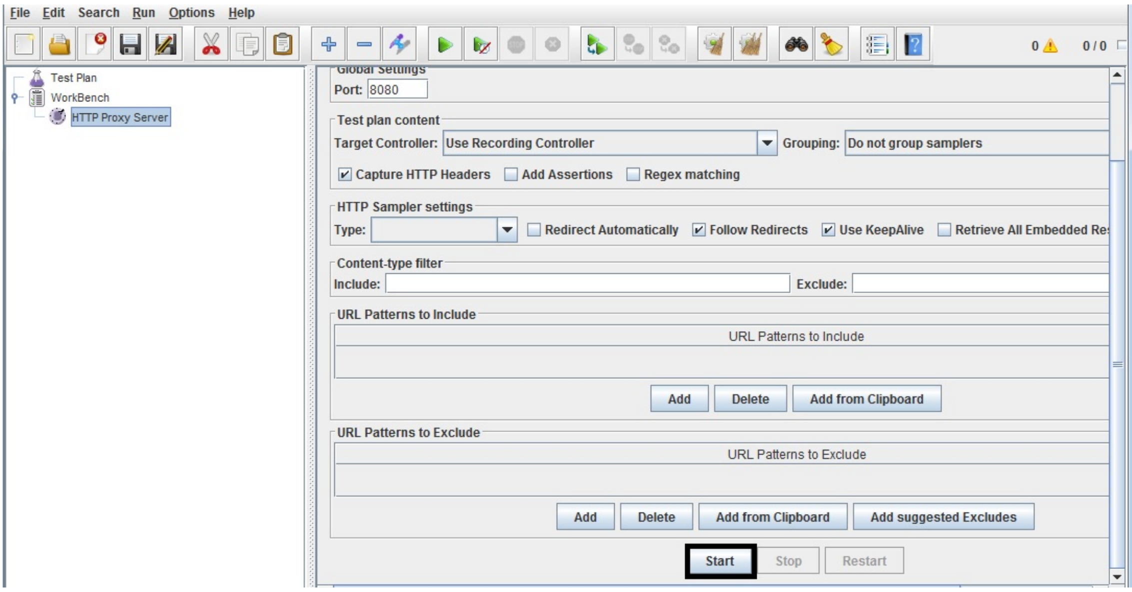


Figure A.3: JMeter Proxy server start

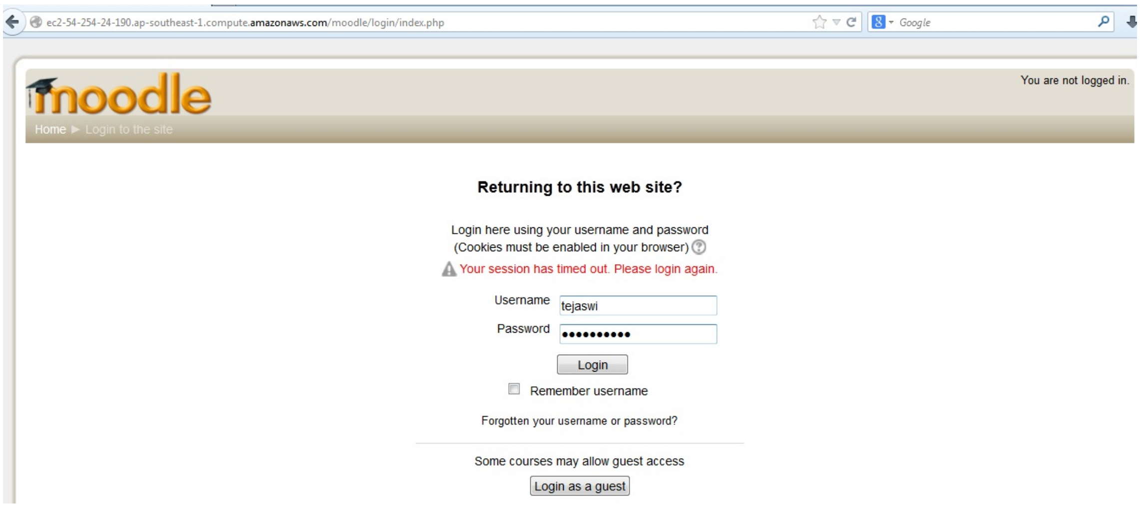


Figure A.4: Running the Moodle in the firefox to generate script

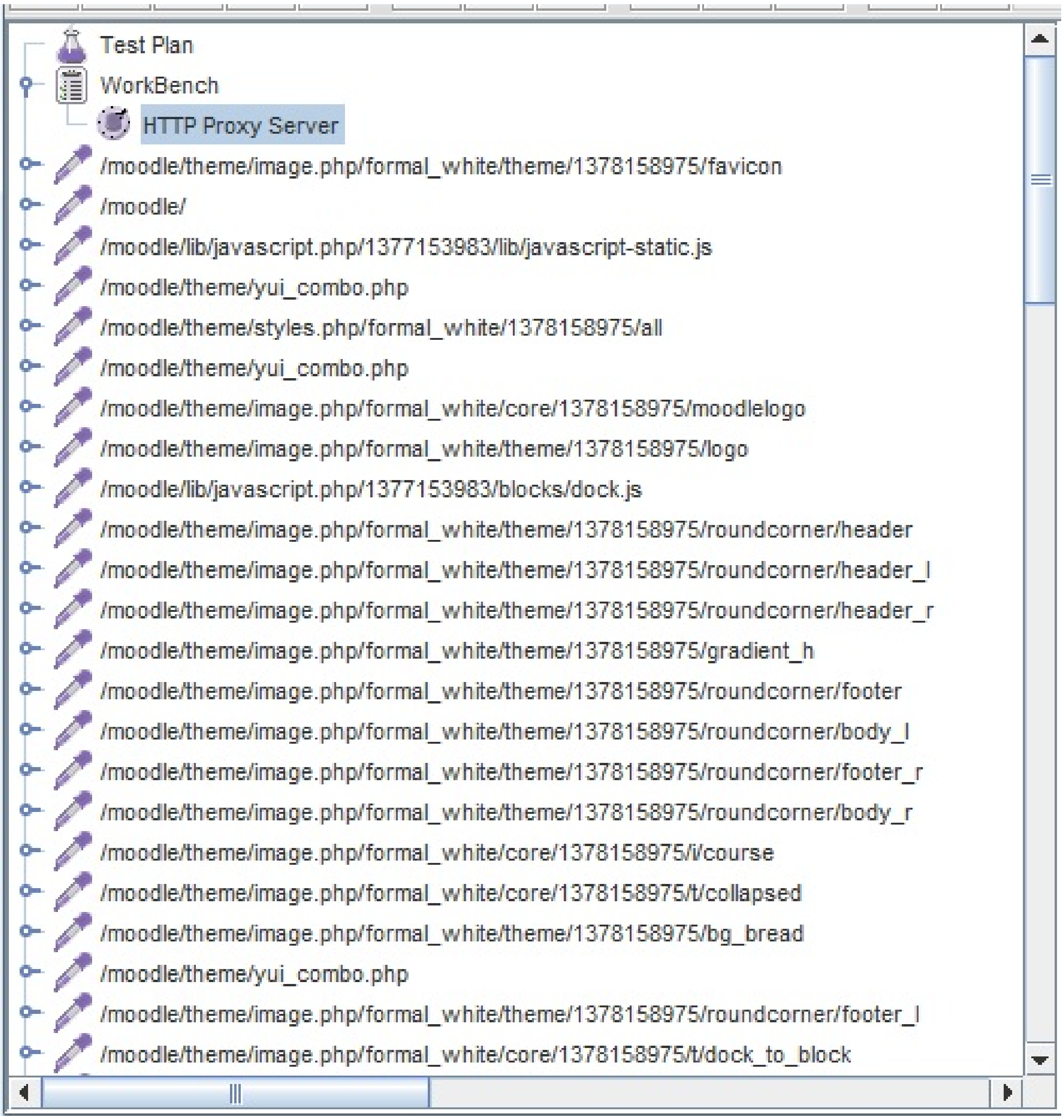


Figure A.5: Recorded script of the Proxy server

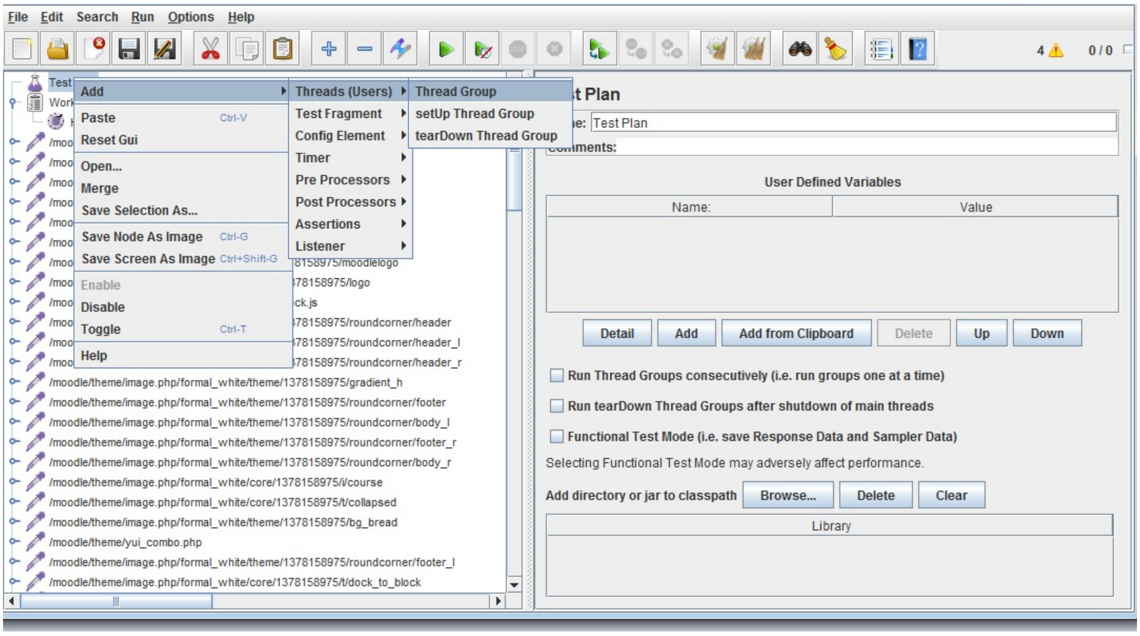


Figure A.6: Adding ThreadGroup

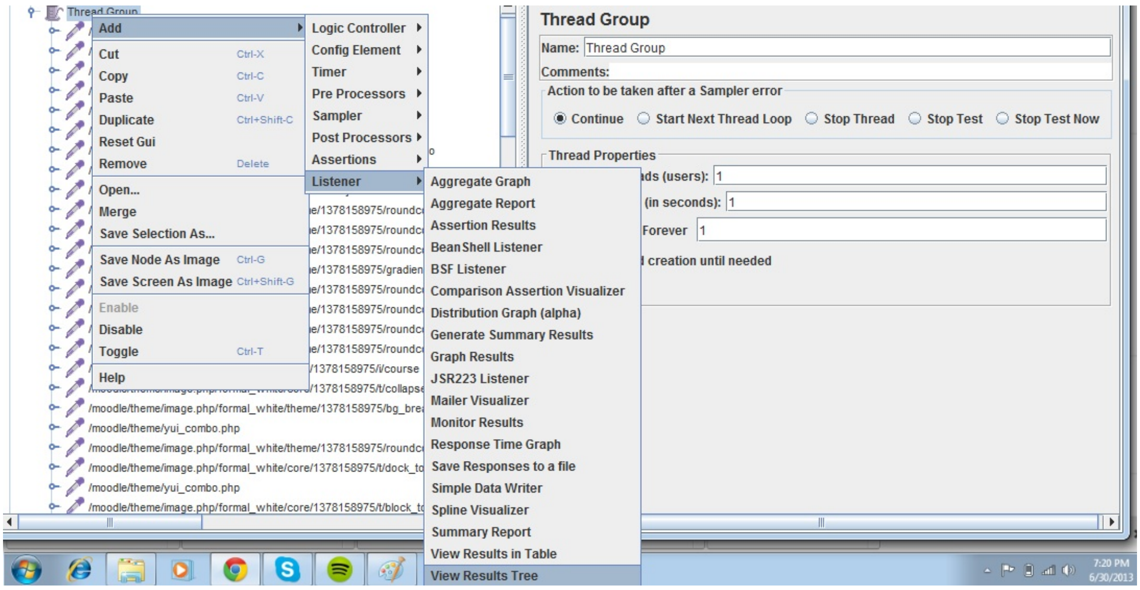


Figure A.7: Adding the view result tree to the ThreadGroup

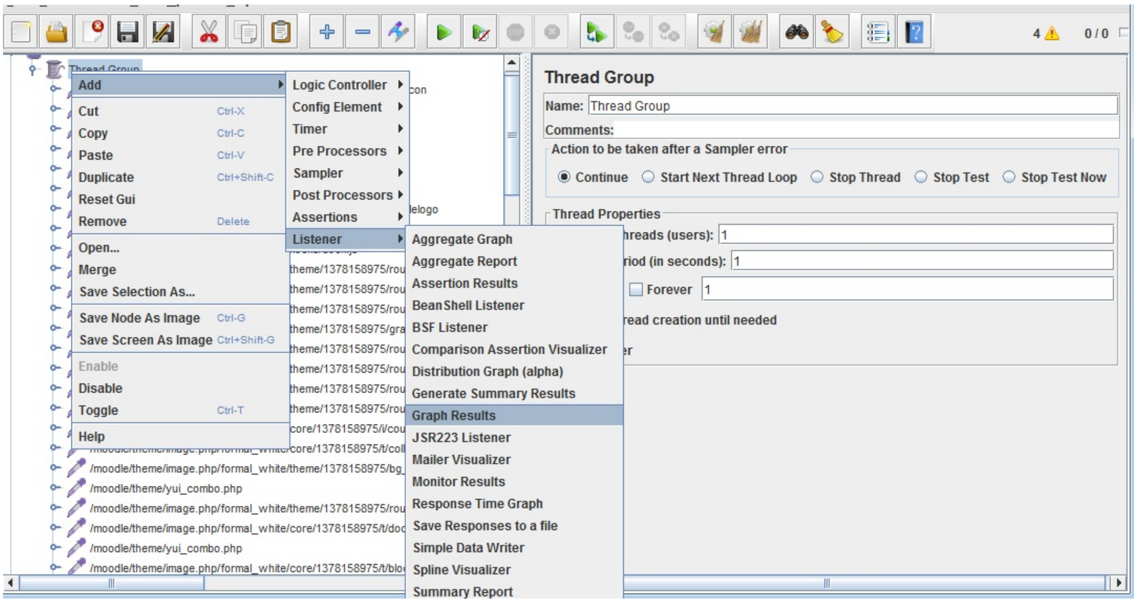


Figure A.8: Adding the result graph in the ThreadGroup

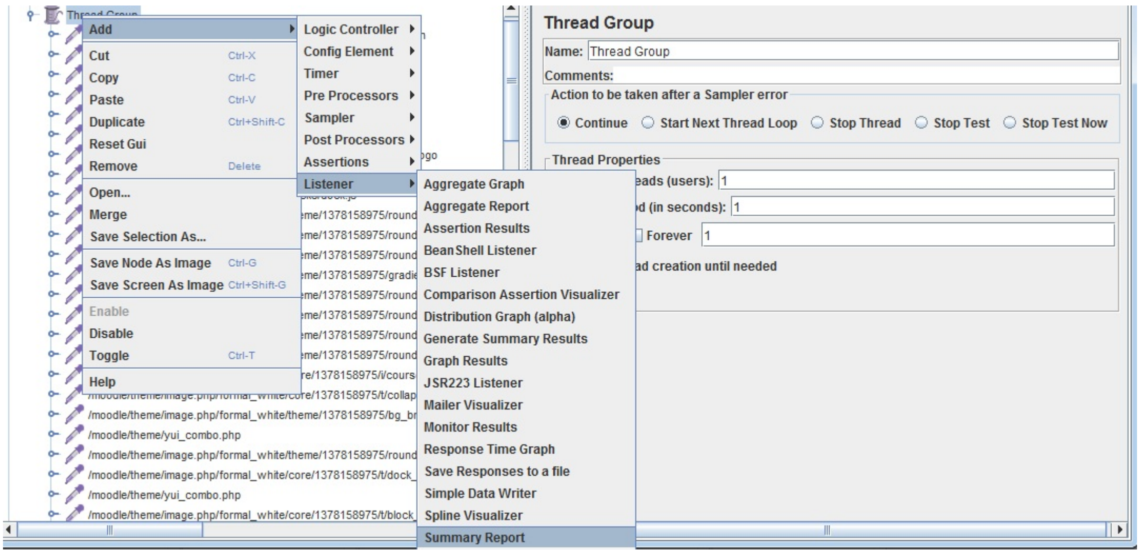


Figure A.9: Adding the summary report in the ThreadGroup

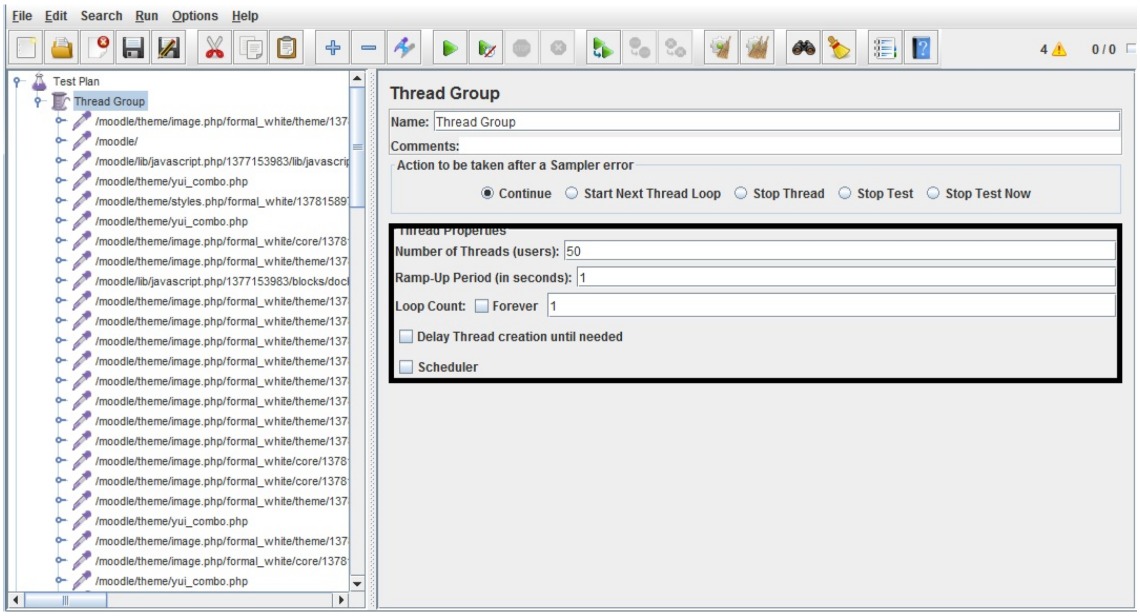


Figure A.10: Increasing the number of Threads

Appendix B

Hadoop Infrastructure

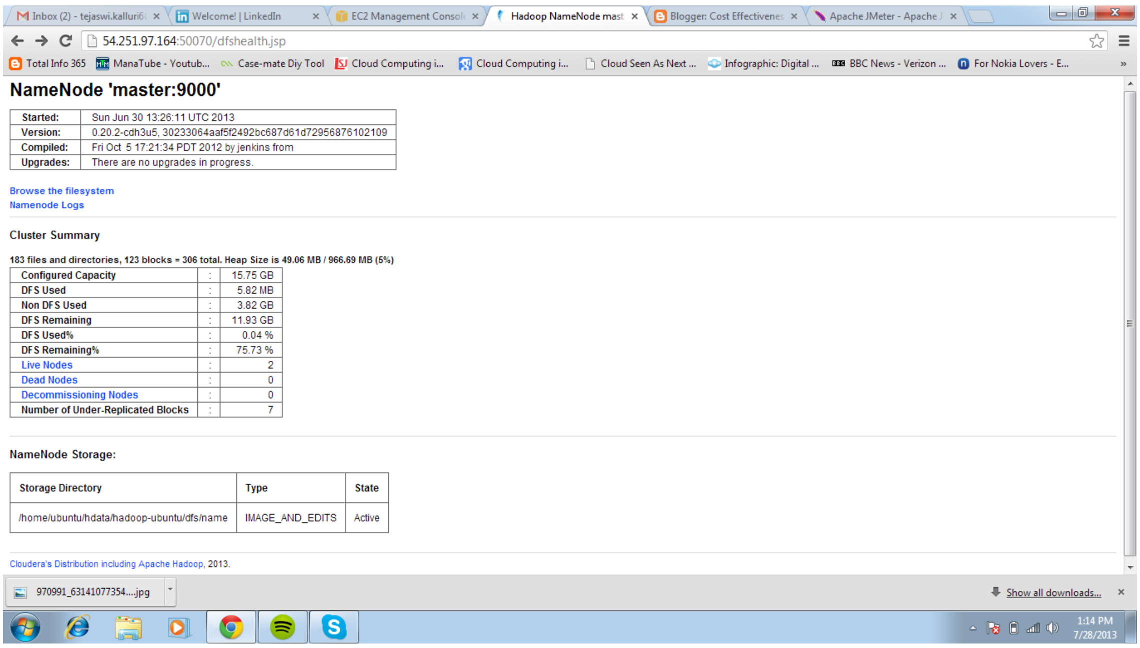


Figure B.1: Hadoop Master Node

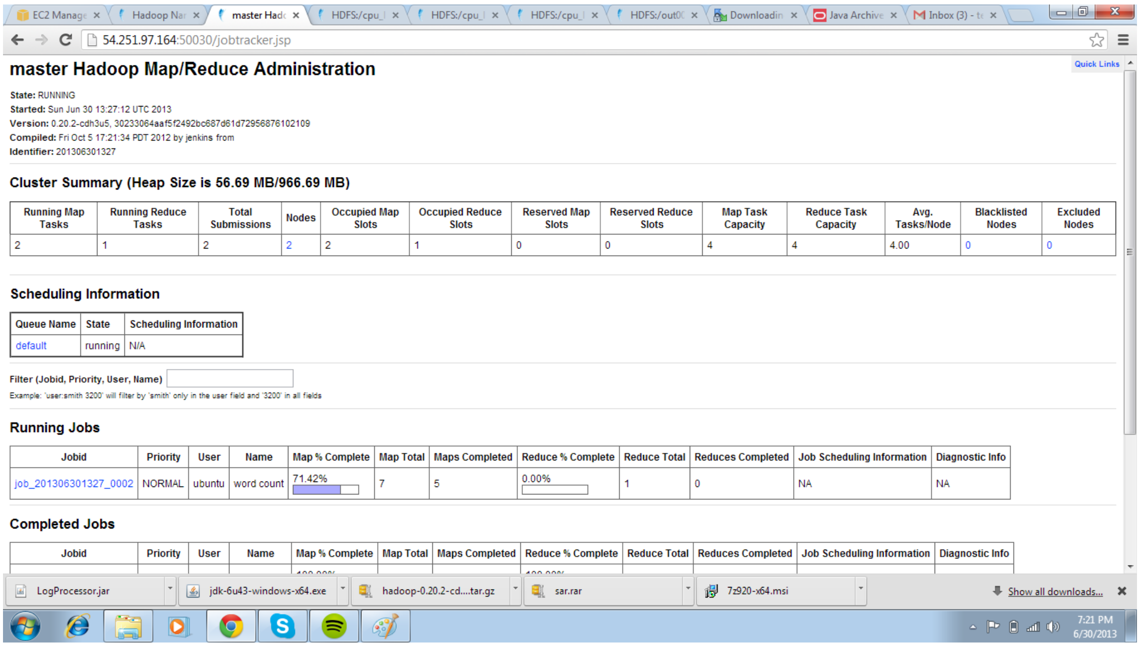


Figure B.2: JobTracker

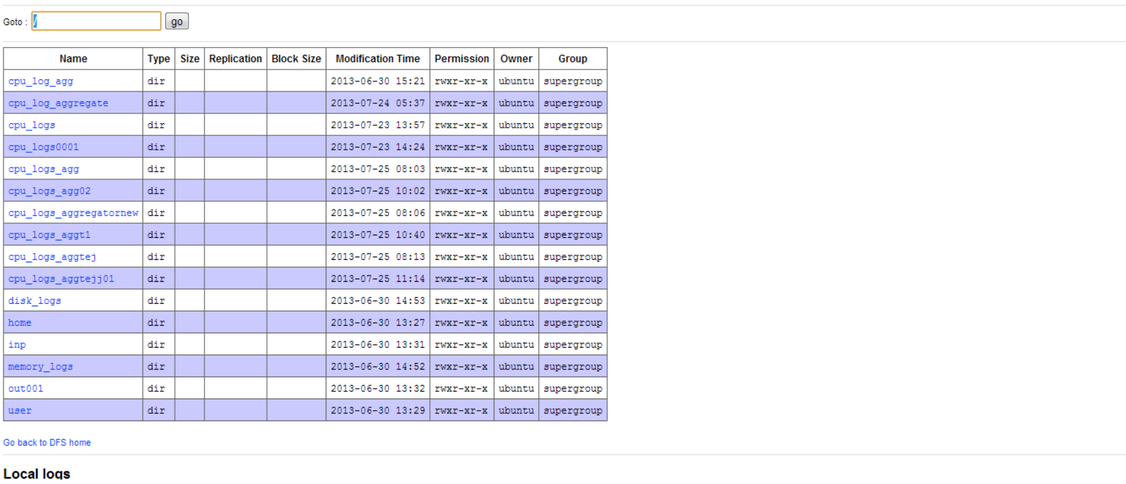


Figure B.3: Logs Directories collected from Flume

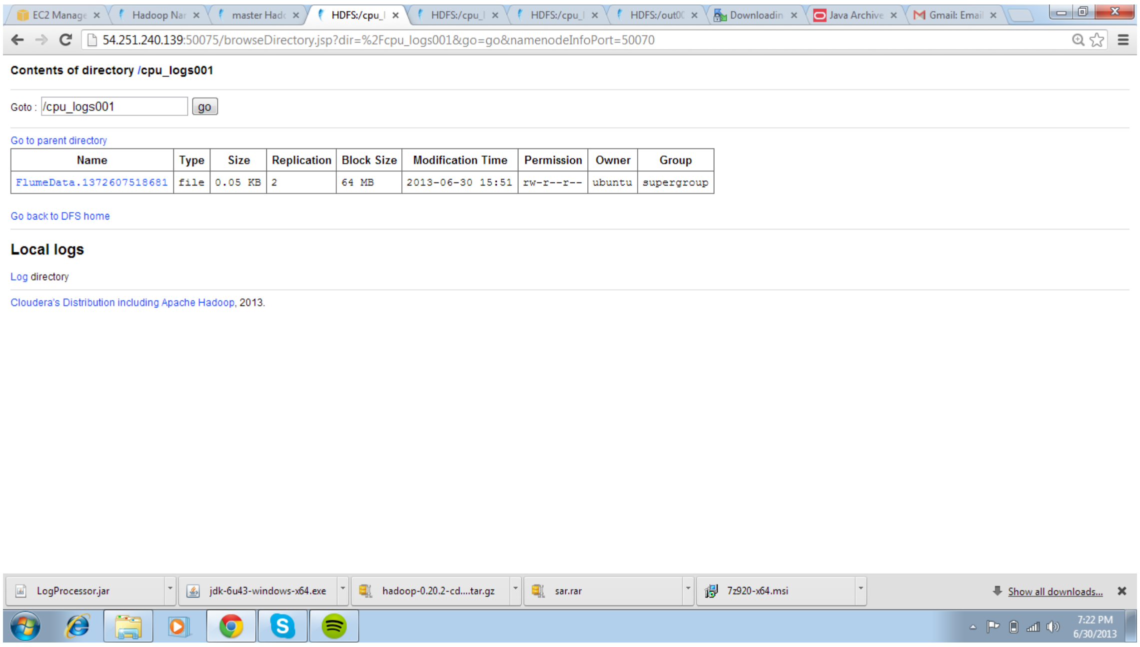


Figure B.4: Log Files

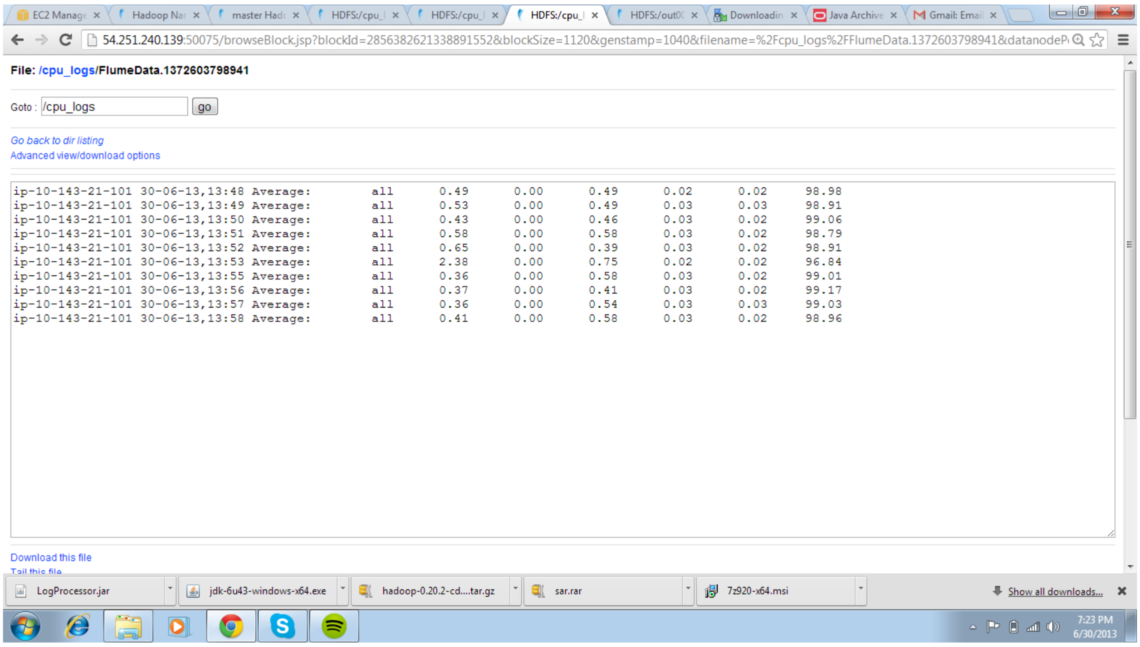


Figure B.5: Log Files