IMPLEMENTATION OF APPLICATION FOR HUGE DATA FILE TRANSFER

Taner Arsan, Fatih Günay and Elif Kaya

Department of Computer Engineering, Kadir Has University, Istanbul, Turkey

ABSTRACT

Nowadays big data transfers make people's life difficult. During the big data transfer, people waste so much time. Big data pool grows everyday by sharing data. People prefer to keep their backups at the cloud systems rather than their computers. Furthermore considering the safety of cloud systems, people prefer to keep their data at the cloud systems instead of their computers. When backups getting too much size, their data transfer becomes nearly impossible. It is obligated to transfer data with various algorithms for moving data from one place to another. These algorithms constituted for transferring data faster and safer. In this Project, an application has been developed to transfer of the huge files. Test results show its efficiency and success.

KEYWORDS

Network Protocols, Resource Management in Networks, Internet and Web Applications, Network Based Applications.

1. INTRODUCTION

Data means raw fact collection from which conclusions may be drawn. Letters of handwritten, printed books, photographs of the families, movie on video tape, printed and signed copies of mortgage papers, bank's ledgers and account holder's passbook are examples of the data.
In the past, people use the limited forms to save or share their data for example; paper or film. Then, computer invented and they can convert the same data into variety forms by the computer such as an e-mail message, text documents, images, software programs or etc. Then, these all datum are stored in strings of 0s and 1s in a computer. Since, all computer data is binary format and this forms data is called digital data as shown in Figure 1.

Data is classified of two types which are structured and unstructured as shown in Figure 2. Structured data uses row and columns format to definite and order the information. This data is stored in database management system (DBMS). Unstructured data is not stored in row and column. Data is stored in variety forms such as e-mail, images, pdf, or etc. Unstructured data is more than structured data because of do not require the storage space [1].

From past to present, information increased even it came to present exponentially. Result of this there is one expression occurred is called "Information Trash". A lot of software companies worked about this topic and Big Data expression formed. Big data is a form that data which is collected from community media sharing, network dailies, blogs, photographs, videos, log files and formed meaningful manner. According to some belief which is destroyed now data which is not structural was worthless, but big data showed us that there would be occurred how important, useful data from the manner is called Information Trash. Big Data is consist of web server logs,

DOI : 10.5121/ijwmn.2014.6403

statistics of internet, social media publications, blogs, microblogs, climate sensor and information which is coming from similar sensors, Call Records from GSM operators.
Big Data allows users and companies to manage risks better and making innovation if it used with true analysis methods on the right way.

Most companies are still making decisions through result of their conventional data store and data mining. But to be able to foresee consumer wishes, there should be big data analysis and to be able to proceed according to these analysis. Big data analysis, flows, store is difficult to deal with using old traditional database tools and algorithms. Computers and data storing unit is not enough for big data and not enough have capacity to use. By 2012 data is produced 2.5 quintillion bytes daily in the today's world. Handling, transferring this big data or issues about it is called Big Data.

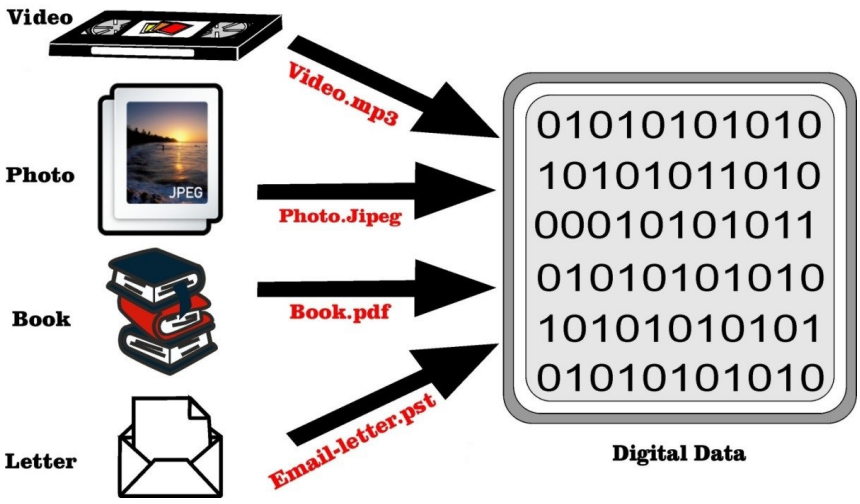

Figure 1. Digital Data

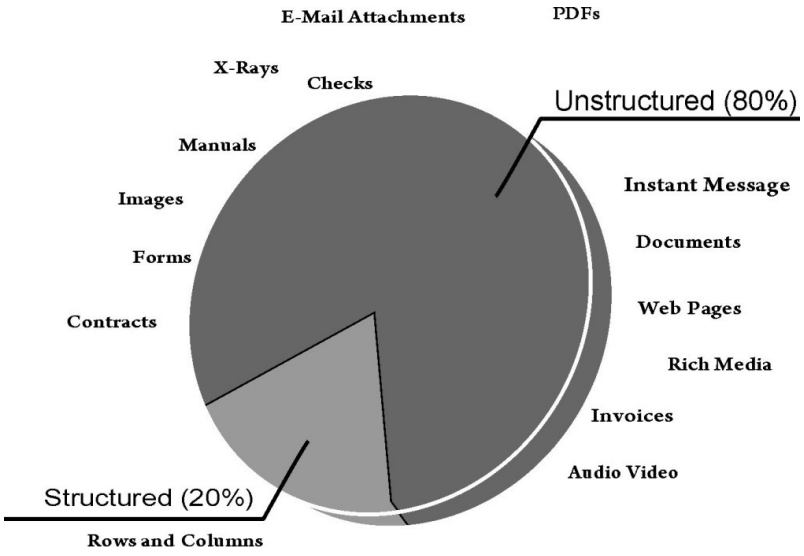

Figure 2. Data Types: Structured and Unstructured.

Databases are not enough to hold this growing data today. When relational database can hold at gigabyte level, with big data it could store at petabyte level. But big data is only proper for Batch calculations. There is no property which is critical in developed database like Transactions. Because of the reading, writing and updating databases could make through transactions, these processes accepted as atomic process and changing from different processes and making it inconsistent can be blocked. Big data should be used when it is written in once and read many times. Because data is parallel processed more than one place. This big data is produced in many sectors like RFID sensors, social media, hospitals etc. Notably data which are come from DNA sequence analysis, weather forecast sensor as a big data is need for us in many areas like data process areas.

There are 5 components in big data platform which are variety, velocity, volume, verification and value. Generally it is called 5V.

Variety: Data which is produced not structure %80 and every new technology can produce data in different formats. Its needed to deal with data type in variety comes from phones, tablets, integrated circuit. They are also in different languages, could be Non-Unicode, they need to be translated and formed in same format.

Velocity: Producing time big data is too fast and it has been growing. Process number and variety of related to the data is increasing too.

Volume: According to IDC Statistics data amount will be 44 times bigger than 2009's data amount in 2020. The fact that we called it huge and big systems which we use will be bigger 44 times than now. This fiction has to made that how to deal with data store, proceed, integration technologies with this much data volume. In 2010's total amount of informatics expense was increased by the ratio of % 5 but, amount of data producing was increased %40.

Verification: There is also another component that must be exist is security in data density. During the flow, it must be seen by true people, controlling with enough security level, and must be latency.

Value: The most important component is creating value. Big Data has to create positive value for company after the data producing and processing as describes as at the above. For example governmental institution which makes decisions and strategies about health must see illness, cure, and doctor distribution in district, city, and county etc. details. Air Forces must see every inventory's location and situations and must watch histories about them. A bank should see the person's information including eating, vacation habits and even social media activities for giving the credit to that person.

Our daily life needs and applications for them on the Internet, after sales records for the customer delight and storing data in the companies for this reason, it caused problem about lack of free space. It was the beginning of the new search about this area. Recently companies have to keep their customer's data with new services as individual and personal.

Hospitals; keeping data of their patients in a numerical environment for serve them individually and efficiently.

Governments; have to keep data of their citizens and have to work on it, for example according to rules of governmental authorities, media records have to be stored for last 1 year.

Because of the producing and consuming data on the internet affected internet service providers about ensuring big data forming meaning and reuse again.

Banks became more efficient about internet banking, more updated about daily records of the customers, availability for 7 days 24 hours online.

Power supply companies have to record of individual data by using smart system and counter.
Pharmaceutical Industries have to be available always for researches and big genomic databases such as cancer researches.

Additionally GPS, GSM and cameras which can produce high resolution pictures and sounds data decreases efficiency of data storage areas. Social media platforms like Facebook and Twitter push entrepreneurs through the big data because of the data which have to be stored.
Entrepreneurs, investors and media-consulting companies have opportunities about Big Data. Clouding technology has spread around and became cheaper so that its usage increased and also it changed balance of data forming economy. After the evolution of most important technology market on Big Data, its expecting that market will exceed 50 billion dollar in 5 years. Annual increase of data in the world now is 59% and it's expecting to increase more. Traditional and new data sources are laying on this growth. IDC digital records will be 1.2M zeta bytes at the end of the year and it will be 44 times more than now in this following 10 years. Essential source of growth is on structured data. Myth about 80% of unstructured data is worthless is now destroyed because of the result of the success for e-commerce companies and search motors. Main necessity is storing of structured and unstructured data, and analysing them and data mining.

In 1980s companies' main aim was the produce main object and provide to access to customer when production was more important. Essential purpose for the developing of ERP systems is collecting customer, distributing centres, supplier and production in one platform. People started to ask this question "who is the right customer for me?" when this system has enough satisfaction.

Born of CRM systems also started with the same question, CRM focus one point which is "providing right production to the right customer with right price by right way on right time and in right place." So now it's not customer for the production, its production for the customer. This methodology keep increase its importance in last 10 years.

2. BIG DATA COMPUTING

Thanks to the developments of communications, digital sensors, computation and storage have generate very big collections of data, and catching information of prize for the business, science, government and society. Most popular search engine companies such as Google, Yahoo! and Microsoft have developed a new sector and business making the information freely accessible on the World Wide Web and support to the users in functional ways. These companies gather trillions of bytes of information every day and they are developing their services such as satellite images, driving directions, and image retrieval. These services are advantages of sociality such as collecting people's information, gathering them together, knowing how to use people's data and their value are measureless.

Some of the big data computing forms will change the companies' actions, scientific researches, practitioner of medical, security intelligence and nation's defence like how search engines have changed accessing of the data for people.

In the modern medicine world, patient's information is collected via some technological equipment. Such as CAT scans, MRI, DNA microarrays, and the many other of equipment. Due to collection of big data sets of patients and applied them to data mining, many developments are possible for the medical world.

The collection of data on the World Wide Web is showing up to be a principal and can be mined and ran in various ways. For instance language translation programs can be used by statistical language models which are created by result of analysis of the billions of documents in the languages which we want to translate from and the source, along multilingual documents. Advanced web browsers spare documents according to the reader levels of English from the beginners to adults. Carnegie Mellon University has a research of information storage on the human brains and in which way human brains are storing the information. The study depends on word combinations which are existing in the web documents.

The big data technology has the importance has growing because of the many technologies.
Sensors: Many different kind of sources creating digital data such as digital imagers, chemical and biological sensors, and foundations and people. Such as digital cameras, MRI machines, microarrays, environmental monitors.

Computer Networks: Localized sensor networks such as Internet can gather data together from the various sources.

Data Storage: Thanks to the magnetic disk technology development, cost of storing data has decreased. For instance cost of storing one trillion bytes of data in a one terabyte disk drive is about $100.Acoording to the reference of the one estimation which is converting the all books in the Library of Congress to the digital form can be done it with around 20 terabytes .

High Speed Networking: In spite of the fact that one terabyte data can be stored on a disk for $100, it's transferring with a group is requires one hour or more and approximately a day by "high speed" Internet connection. Out of the ordinary the most common way to transfer mass data between sites is to guide disk drive through Federal Express. Bandwidth limitations expanding the difficulty of efficiency usage of computing and storing resources in a group. The power of connecting a cluster and end user which are geographically distributed also limited by them. The difference between amounts of data which is using for storing and which is using for communication will increase. For the combination of increasing bandwidth and decreasing cost, there should be "Moore's Law" technology.

Cluster Computer Programming: Programming for the large scale of the data in distributed computer systems is an existing problem since running large sets of data become required in an optimum time. The software should divide the data and make computations between nodes in a data set, and whenever there is an error occurs in the hardware or software, software must investigate and fix the error. Great inventions such as MapReduce programming framework presented by Google, these inventions are built for the organization and programming the systems. To understand the power of the big data computing, more effective and powerful techniques must be improved.

Expanding the access of cloud computing: AWS has a technological limitations such as bandwidth, and it's not appropriate for the large data and its computations, however Amazon making good profit with AWS. Additionally, limited bandwidth causes time consuming and expenses for taking data inside and sending it out. The cloud systems should be separated from each other because the fact that decreasing the sensibility for the disasters. But data mobility and interoperability is needed in more developed levels. There is a project called The Open Cirrus which is directing attention to this point by creating international environment to make experiments on connected cluster systems. Organizations must tune to new costing model on their managerial part. For example, governments do not take money from the universities for the capital costs such as purchasing new machines but they do for operating costs. We can imagine that ecology of cloud systems which are supporting general capacity of computing and right on target special services or storing expert data sets.

Security and Privacy: Data sets include sensitive data and tools which can be used for take-off information and produce opportunities for access without authorization and use. Many of privacy protections in our community are based on current uselessness. For instance people have been watching by video cameras from many locations such as markets, ATMs, airport security. The possibility of abuse becomes important when these sources gathered together and advanced technology of computing let it associate and analyzing the data strems. For harmful agents, cloud facilities became a cost effective platform for the start a botnet or implement huge parallelism for breaking the cryptosystem. We must generate assurances to prevent misuse together with evolving technology to enable beneficial properties.

Remarkably companies which have available environment for the Internet service comes first in the industry. These companies also providing billions of dollars for their computing systems and overshadow the conventional ones. Companies like Google, Yahoo!, and Amazon are the pioneers of the cluster computing systems' installation and programming. Other companies receiving news of the business benefits and the operating efficiency which are discovered by these companies.

University researchers do not have enough access to large scale cluster computing and new sight's appreciation. This situation is quickly changing because of their colleagues' success in the industry, providing access and training. Google, IBM, Yahoo! and Amazon ensure access of their computing resources for the students and researchers. It is enough to make satisfy the some potential needs but clearly it's not enough for providing all potential needs for spreading the application of data intensive computing. Some of the large scale scientific projects are making plans for managing and providing computing capacity for their stored data. There is a spirited debate between new approximation of data management and traditional one. Weakness of research funding is an important block for attendance and encouraging of university researches.
The development of big data computing is one of the best innovations in recently. The importance of its properties such as collecting, organizing and processing of the data has been recognized newly. It's simply to provide the development of it by the federal government's financial support.

3. CONCEPT OF DATA TRANSFER

Data is transferred by some applications such as electronic mail, file transfer, web documents, so bandwidth and timing are important things for data transfer. Figure 3 refers data transfer;

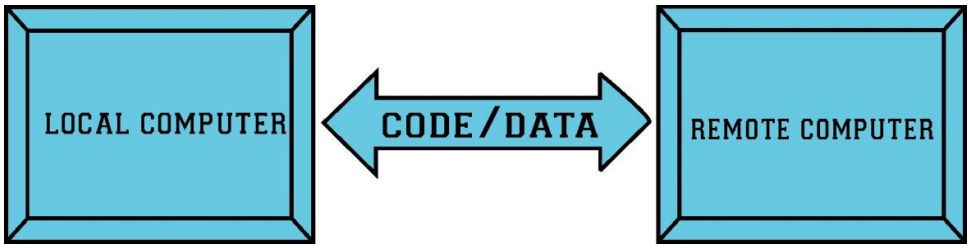

Figure 3. Data Transfer

If you want to transmit small data, you need small rate bandwidth such as the application of internet telephony encodes voice at 32 kbps. However, if you have huge files and want to transmit them, you need more bandwidth. This is more advantages than small rate bandwidth.
Timing is important when you transmit the data. Applications should provide quick data transferring to save time. For example, real-time applications of internet telephony, virtual environments, multiplayer games or etc.

3.1. Internet Transport Protocols

Two transport protocols are used to applications which are UDP and TCP [2]. Each of these protocols provide different service model for applications, so you should choose one of them when you create a network applications for the Internet [3].

3.1.1. Transmission Control Protocol (TCP)

Transmission Control Protocol (TCP) guarantees the reliable data transfer. TCP service models include connection-oriented service as shown in Figure 4. The connection is a full-duplex between two hosts. TCP exchanges data between applications as a stream of bytes. Also, TCP implements highly congestion-control mechanism. TCP has high quality connection between initiator and receiver as shown in Figure 5.

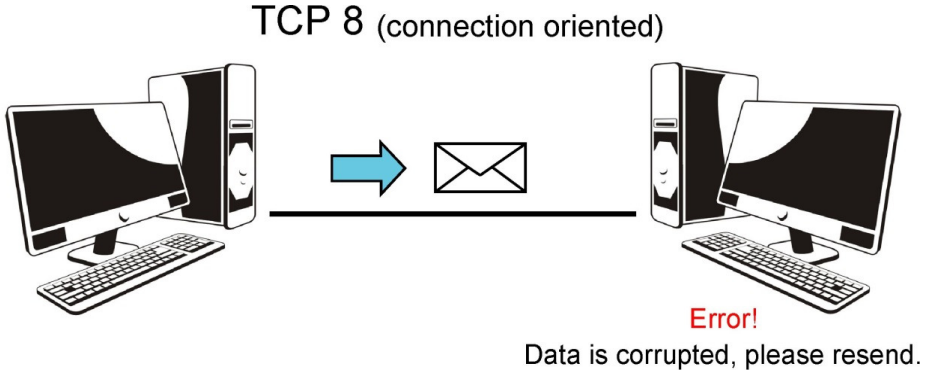

Figure 4. TCP (connection-oriented)

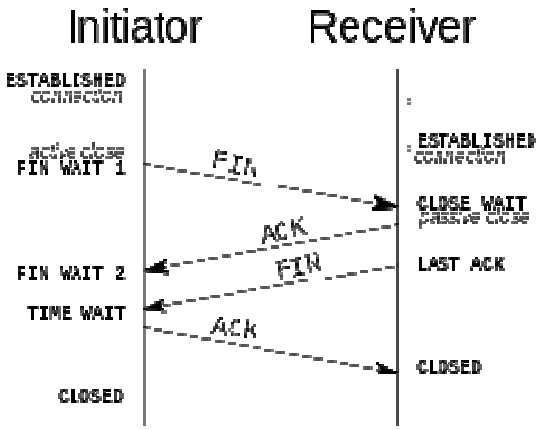

Figure 5. TCP connection

3.1.2. User Datagram Protocol (UDP)

User Datagram Protocol (UDP) is connectionless, so handshaking is not used in this service before two process start to communicate as shown in Figure 6. Also, UDP provides unreliable and unordered data transfer. There is no guarantee the messages sending to any loss. It has small

segment header. Also, UDP doesn't have congestion control mechanism, so it is very fast, so useful for some applications.

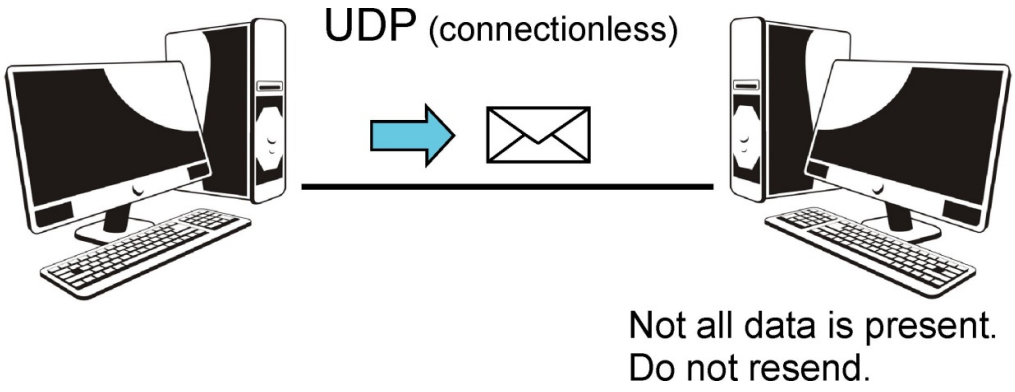

Figure 6. UDP (connectionless)

UDP and TCP service models have many different features as shown in Table 1.

Table 1. TCP and UDP features

TCP	UDP
Sequenced	Unsequenced
Reliable	Unreliable
Connection-oriented	Connectionless
Virtual circuit	Low overhead
Acknowledgments	No acknowledgment
Windowing flow control	No windowing or flow control

4. METHODS FOR FILE TRANSFER

4.1. The File Transfer Protocol (FTP)

The File Transfer Protocol was appeared by Abhay Bhushan and brought out as RFC 114 on 16 April 1971.It was changed by a TCP/IP version in June 1980 as RFC 765, and changed in October 1985 as RFC 959. RFC 959 is the current version and it's expanded by the some standards which are RFC 2228, RFC 2428. RFC 2228 is for security expansion and RFC 2428 is for providing support for IPv6 and establishes new passive mode.

The File Transfer Protocol is a standard network protocol which allows transferring computer files from one host to another host through a TCP based network [4]. It was came out to allow allocation of files and support to enhance remote computer's usage. FTP based on client server architecture and manages various control and data connections among client and the server as shown in Figure 7. Clear text sign in protocols are used by clients, which allows user to have

password and username also connection without them (anonymously). SSL, TLS (FTPS) and SFTP secure the FTP.

Commands of FTP;

USER username: Used to send the user identification to the server.
PASS password: Used to send the user password to the server.
LIST: Used to ask the server to send back a list of all the files in the current remote directory. The list of files is sent over a (new and nonpersistent) data connection rather than the control TCP connection.
RETR filename: Used to retrieve (that is, get) a file from the current directory of the remote host. Triggers the remote host to initiate a data connection and to send the request files over the data connection.

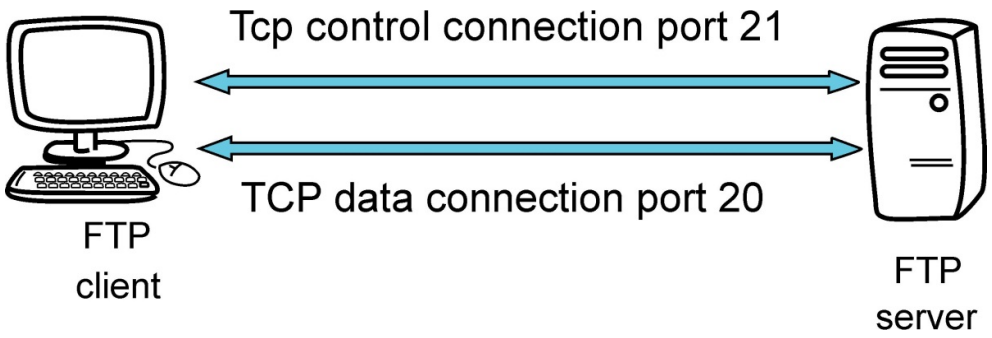

Figure 7. Control and data connection

FTP replies;

331 Username OK, password required

125 Data connection already open; transfer starting

425 can't open data connection

452 Error writing file

4.1.1. FTP Communication and Data Transfer

There are two modes for constituting data connection such as active and passive, FTP uses both modes. TCP control connection from a non-specific port N to the FTP server command port 21 will be created by client as shown in Figure 8. In active modes, incoming data connections on port N+1 from the server will be received by client. Port N+1 is used for report the server when FTP command is sent by client. In passive mode there is client behind a firewall and also not available for incoming TCP connections. In this mode PASV command is sent to the server by client through the control connection. Afterwards server sends a server IP address and server port number. In this way client can use it for opening a data connection from random client port to the server IP address and server port number [5]. Active mode and passive mode were updated to support IPv6 in 1998.

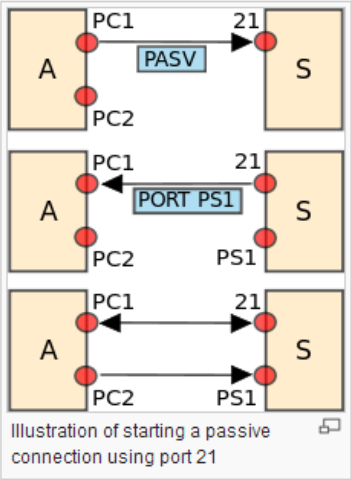

Figure 8. Illustration of starting a passive connection using port 21

The server reply with three digit status codes in ASCII upon control connection. These messages consist 100 Series, 200 Series, 300 Series, 400 Series, 500 Series etc. For example 200 OK means last command was successful. The numbers describe the code for the response and optional text describes explanation for humans. The current transfer may interrupt while data transferring. There are 4 representations can be used:

ASCII Mode: Data is converted. From sender's character representation to 8 bit ASCII before the transmission and to the receiver's character representation if there is any need. If files have data differently from plain text it will be unsuitable.

Image Mode: (Binary) Sender sends file byte for byte and receiver stores byte stream.

EBCDIC Mode: It uses EBCDIC characters among hosts for plain text.

Local Mode: It allows two computers to send data without convert them ASCII in a proprietary format.

For text files, there are various format control and record structure options. For the files which are containing Telnet or ASA, there are features to promote them.

Stream Mode : In this mode data is sent as persistent stream which is reliving FTP and making FTP trim from any processing.

Block Mode : In this mode data will break into different blocks. And will be passed on to TCP.
Compressed Mode : In this mode data is compacted using a single algorithm.

4.1.2 Basic Operations of FTP

FTP is founded on a client-server architecture which clients are transferring files to a server and receiving files from a server as shown in Figure 9. FTP period consists two connections which are transmitting standard FTP commands, responses and transferring the actual data.

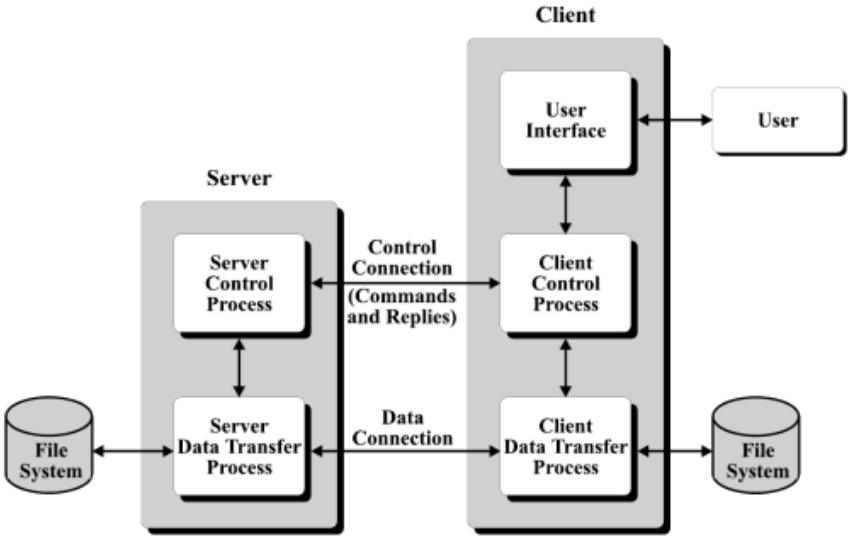

Figure 9. FTP Client-Server Application

The Client Control Process starts the Control Connection and in this process FTP commands and responses are transmit as shown in Figure 10. So that data connection will begin and will be used as needed. The Control Connection should be stay open mode during the data transferring time [6]. If any collision occurs during the data transfer FTP Data connection will be closed and session will be failed. There are parameters such as transfer mode, file structure, data representations during the data transfer and they are sent by FTP commands. When operation and parameters are transmitted, client will look up predefined TCP port and server.

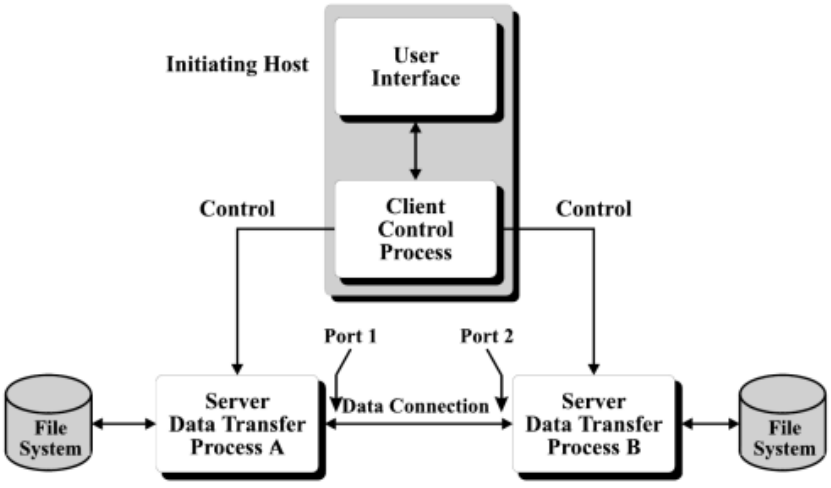

Figure 10. FTP Client Server Process

There is no Data connection need for between server and host in FTP. A FTP user process must guarantee that one of the third party machines starts on a designated port and other starts the Data Connection. In this module user creates a control connection and organize for the founding the Data connection among servers. At the figure below explains this scenario.

There is an alternative way which allows to users to start data connections and which also allows to greater network administrator. This is called FTP Passive Open.

4.1.2.1 Active FTP step by step

Client connect to FTP server with command port number 21.

FTP sends message and username query to the user

Client enters connection information as needed.
Server controls the information of sender and replies user.
If the information is true, FTP command line will showed to the client.
Client creates port bigger than 1024 and announce this port to the FTP server.

FTP server connects with this port number and transferring starts.
Client sends confirmation message.

4.1.2.2 Passive FTP step by step

Client connects to FTP server with command port number 21.

FTP server sends message and username query to the user.

Client enters information of connection as needed.
Server controls the information of sender and replies user.
If the information is true, FTP client wait for the additional port from the server through PSAV command.

FTP client connects this port and starts to data transfer. Client sends a confirmation message.

4.1.2.3 Active FTP and Firewall

Even FTP user creates a port on its side and if there is a firewall, there would be a problem because that firewall will not give permission for this port as shown in Figure 11.

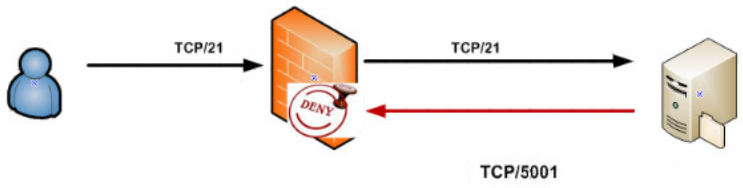

Figure 11. Active FTP and Firewall

4.1.2.4 Passive FTP and Firewall

In this module, even if FTP server opens additional port and if there is firewall, permission is needed from the firewall to access this port as shown in Figure 12.

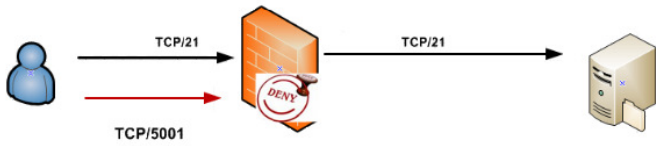

Figure 12. Passive FTP and Firewall

4.2. Secure File Transfer Protocol (SFTP)

There are methods for securing FTP such as FTPS and SFTP. Explicit FTPS is a FTP standard which allows users to demand encrypted FTP session. This session can be done with AUTH TLS command. Connections which are requesting TLS can be allow or deny because of server has these both options. There is a recommended standard called RFC 4217. Implicit FTPS is unconfirmed standard for FTP that needed the usage of SSL or TLS connection. It is defined for using different ports than plain FTP. Secure File transfer protocol which is also called as secure FTP is a protocol that transmits files and set of command for users, but its rest on different software technology as shown in Figure 13. SFTP uses the Secure Shell protocol to transmit files. It encrypts commands and data in terms of avoiding passwords and information which are susceptible. It cannot run with FTP software.

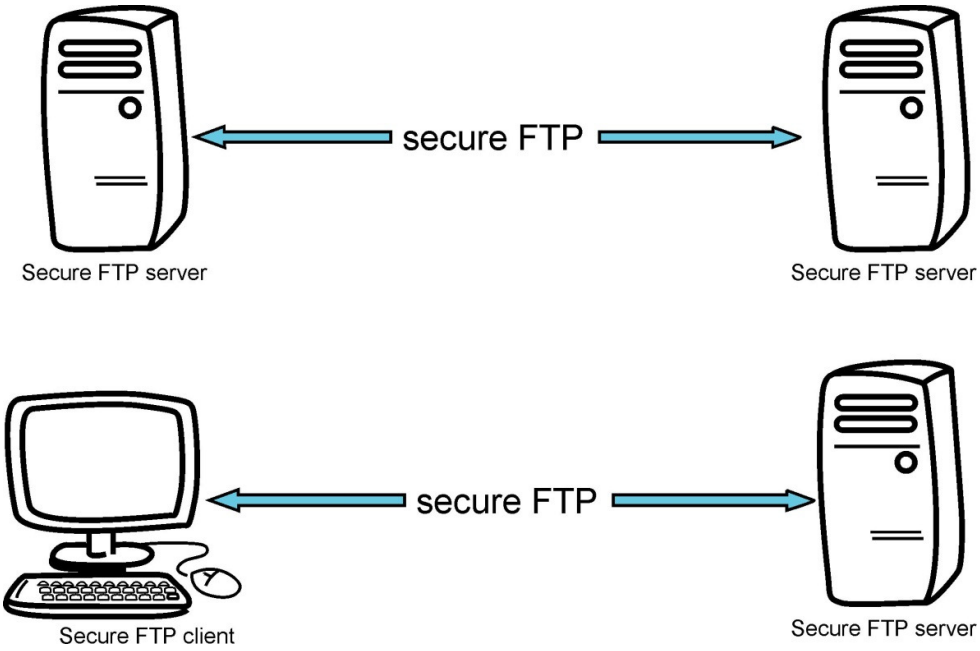

Figure 13. Secure FTP

5. METHODOLOGY

Before start working on this project, first searching programs like zip, gzip pack big data with size of scale is an issue. Inputs are analysed and various tests are run on data. Speed of data transfer are tried to control with Duplicati program which is preferred nowadays. Duplicati is an open

source software because of that we will analyse codes and will learn what kind of algorithm is processing. Our aim is writing this program with C language which is the closest language of machine. Because we want to benefit from machine's hardware when we need it. We think that C language is most efficient language when it is used with efficient way, our gain will be a lot.

5.1. Scenario

For example, we want to send the file through FTP Server, we need to select FTP Button, and then we should continue with entering hostname, username and password. Then select the file we want to transfer. After that choose the backup path destination from the button of backup path. Click the send button. Our timer will start to work .Algorithm which is running at the background, controls the format of the file if it's compressed file or not. For example .mp3 and .mp4 is the compressed file format. If it's not compressed format, algorithm make the compression of the file. If it's already in the compressed format, then format will stay at the same format. After the process is completed, we are starting to send the file. Sending algorithm take two streams, one from server side, the other one is from file side. Then pieces of 1024 bytes of data are written on the server using streams. After the completing the sending file if it's a successful process, timer, streams and connection from server are closed. You can display the file on the server.

For example, we want to send the file through SFTP Server, we need to select SFTP Button, and then we should enter hostname, username and password. File is needed to select by user which we want to send. Using the backup path destination button, we will choose the destination of compressed file. Click the send button. Our timer will start to running. Algorithm which is running at the background controls the format of the file if it's compressed file or not. For example .mp3 and .mp4 is the compressed file format. If its not compressed format, algorithm make the compression of the file. If it's already in the compressed format, then format will not change. After the process is completed, sending process will begin. Sending algorithm take two streams, one from the server side and the other one is from file side. Then 1024 of bytes of data pieces is written on the server with usage of two streams. After the finishing the sending file, if it's a successful process, timer, streams and connection from server will closed. You can display the file on the server.

This Figure 14 is our project use cases diagram. It shows how application works. First of all, our main menu includes two file transfer protocols which are FTP and SFTP. User 1 chooses protocol one of them. Then, He or she continues to enter hostname, username and password such as hostname: 78.185.128.224, username: Ahmet FTP and password: 10. Moreover, the user chooses the sending file and send it. If, there is not any problem, user 1 can send the file successfully and other user 2 takes the sending file. However, if there is any problem such as loss connection, power supply problem or etc., he or she has to go back to enter the hostname, username and password menu and continue to send the file again.

Project main menu is shown in Figure 15. We have two choices which are File Transfer Protocol (FTP) and Secure File Transfer Protocol (SFTP). Then, we choose one of them to send the file and click the "Next" button. Otherwise, we click the "Exit" button and close the application as shown in Figure 15.

Selection of FTP or SFTP file transfer is shown in Figure 15, If the user wants to send file to FTP Server, the user should use FTP button. If the user wants to send file to SFTP Server, the user should use SFTP button.

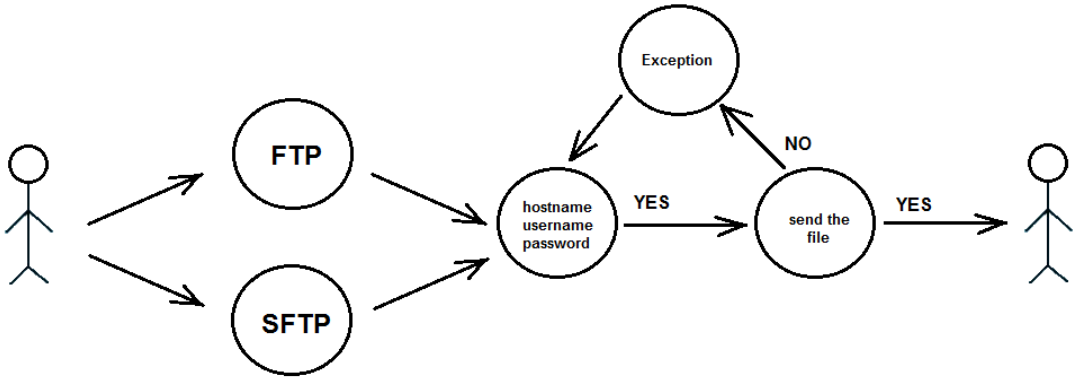

Figure 14. Use case Diagram for Huge Data File Transfer.

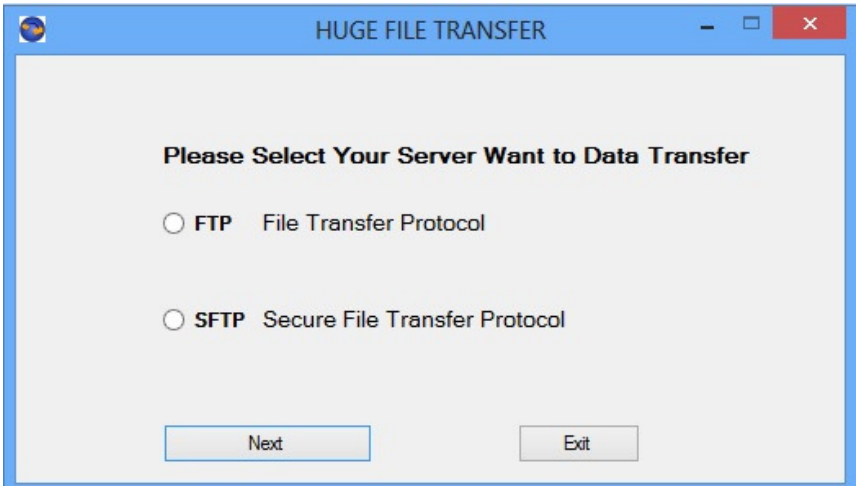

Figure 15. Select of SFTP or FTP

In Figure 16, our project flow chart diagram shows our application working steps. When, we start the application, we select FTP or SFTP. Next, we enter host, username and password. Then, if there is not a problem sending the file is successful. Otherwise, we enter the host, username and password again and continue to send file.

As shown in Figure 17, FTP authorization can be established by entering the necessary information.

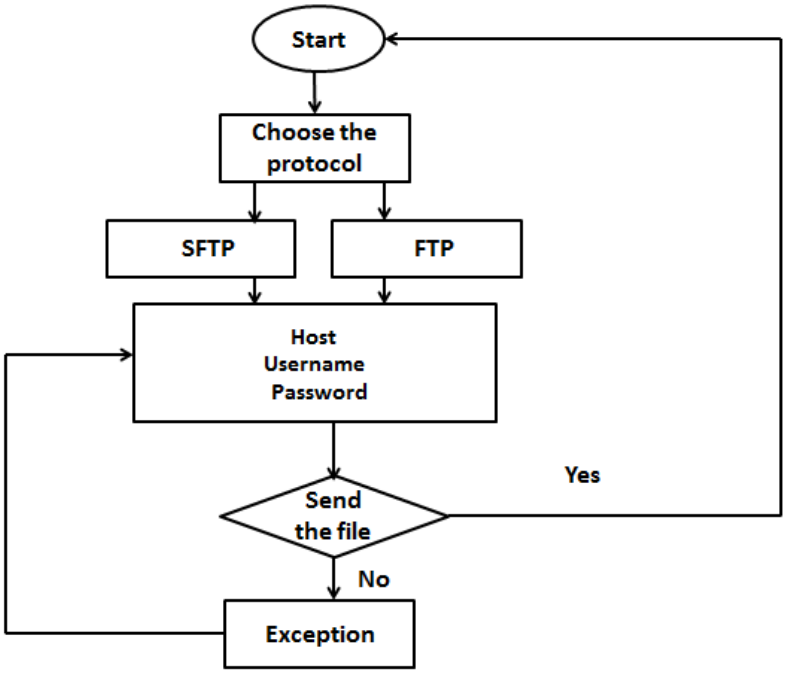

Figure 16. Flow Chart

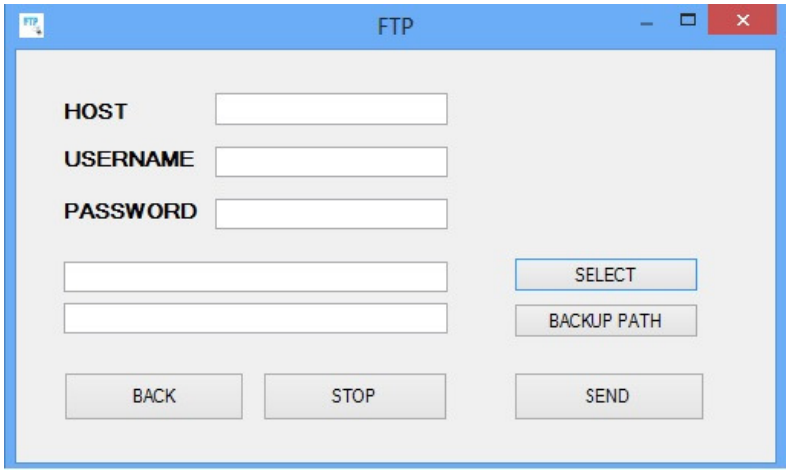

Figure 17. FTP authorization

This is FTP menu includes host, username, password. User enters the host, username and password. Then, he or she clicks The "SELECT" button and select which file will send. Also, user clicks the "BACKUP PATH" button to put the copy of this file zip format. If he or she wants to go back, user clicks the "BACK" button. In addition, if user wants to stop the sending file, he or she clicks "STOP" button. Then, user, send the file to click "SEND" button as shown in Figure 17. Finally, user can see how much time file sends.

Hostname: The IP of FTP Server
Username: Name of user for login to server.
Password: Password of user account.
Select button: The File which we want to send.

Backup Path Button: Path of compressed file's destination.
Send Button: Button for sending file to the server.
Stop Button: Button for stoping file transfer.
Back Button: Turn the selection page of choosing FTP or SFTP.

When user chooses the FTP selection, we use 21 port and program converts the file zip format. However, if file is zip, mp3, mp4 format, program doesn't convert this file because these are compressed file. Then, file divides 1024 bytes packets and sends it and other user takes the file zip format.

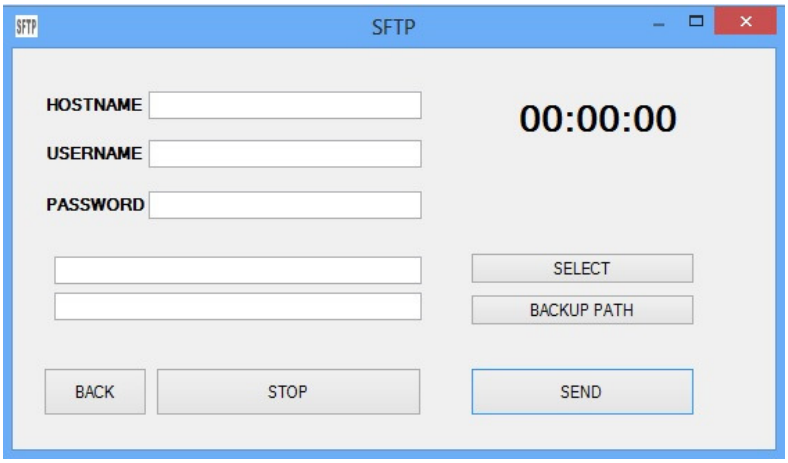

Figure 18. SFTP authorization

This is another selection to send the file which is SFTP as shown in figure 18. User enters hostname, username and password. The "SELECT" button and select which file will send. Also, user clicks the "BACKUP PATH" button to put the copy of this file zip format. If he or she wants to go back, user clicks the "BACK" button. In addition, if user wants to stop the sending file, he or she clicks "STOP" button. Finally, user, send the file to click "SEND" button. Finally, user can see how much time file sends.

If user selects the SFTP protocol, we use 22 ports. Firstly, file is converted the zip format exception of the zip, mp3 and mp4 format files. In addition, user sends the file and another user takes the file. This selection sending time is longer than FTP. Because, this is securely.

Hostname: The IP of SFTP Server
Username: Name of user for login to server.
Password: Password of user account.
Select button: The File which we want to send.
Backup Path Button: Path of compressed file's destination.
Send Button: Button for sending file to the server.
Stop Button: Button for stoping file transfer.
Back Button: Turn the selection page of choosing FTP or SFTP.

6. PERFORMANCE COMPARISON OF PROPOSED SYSTEM AND DUPLICATI

In Figure 19, we showed the program time speed between Duplicati for 100 Mb file in FTP protocol. Our program is faster than Duplicati.

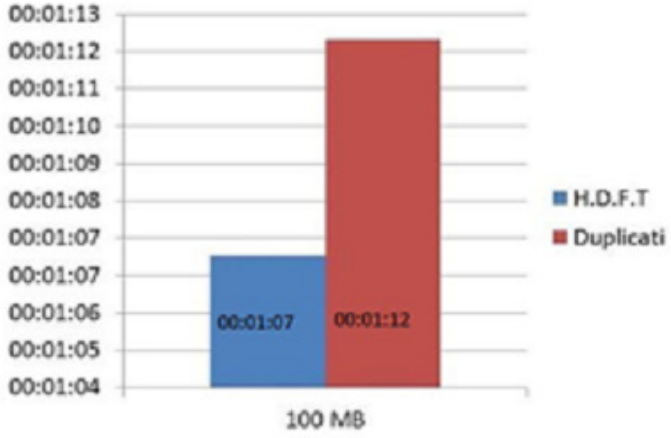

Figure 19. 100 Mb for FTP

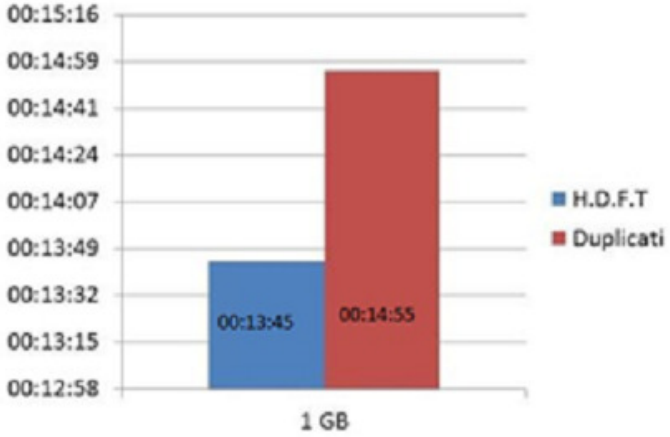

Figure 20. 1GB for FTP

In Figure 20, we tested the file sending to use FTP for 1 GB between our program and Duplicati. Our program sends the file shorter time than Duplicati.

We compared the sending file time our program to use FTP with Duplicati for 10 GB as shown in Figure 21. Our program sends the file faster than other.

We sent the 100 mb file both our program and Duplicati in SFTP protocol as shown in Figure 22. Our program sends the file very short time.

We sent the 1 GB file between Duplicati and our program in SFTP as shown in Figure 23. We sent the file faster than Duplicati.

Finally, we sent the 10 GB file to use our program and Duplicati in with SFTP protocol as shown in Figure 24. Our program takes very short time.

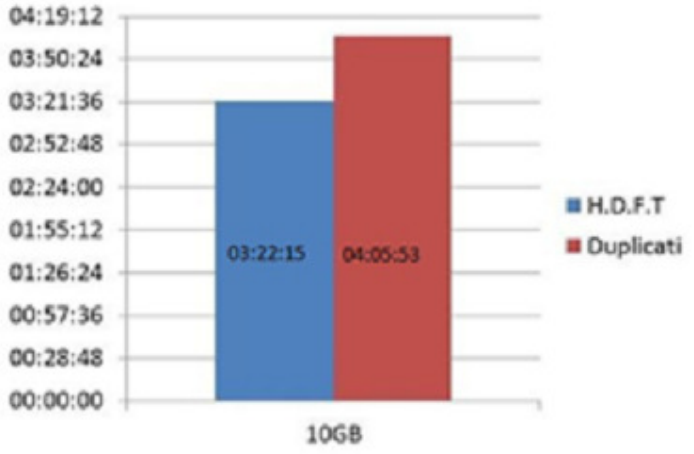

Figure 21. 10 GB for FTP

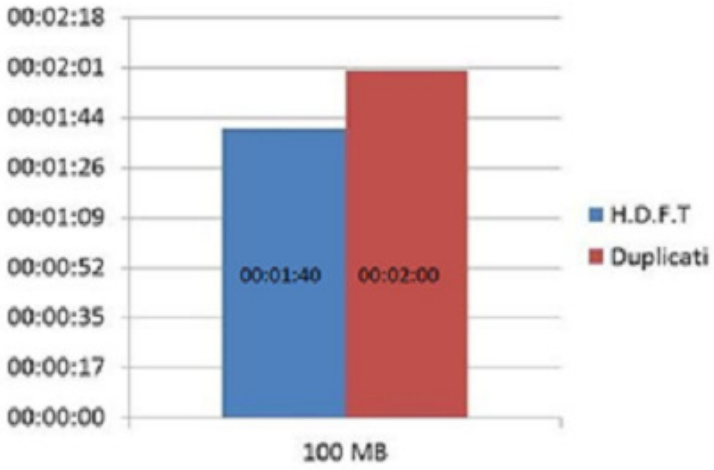

Figure 22. 100 Mb for SFTP

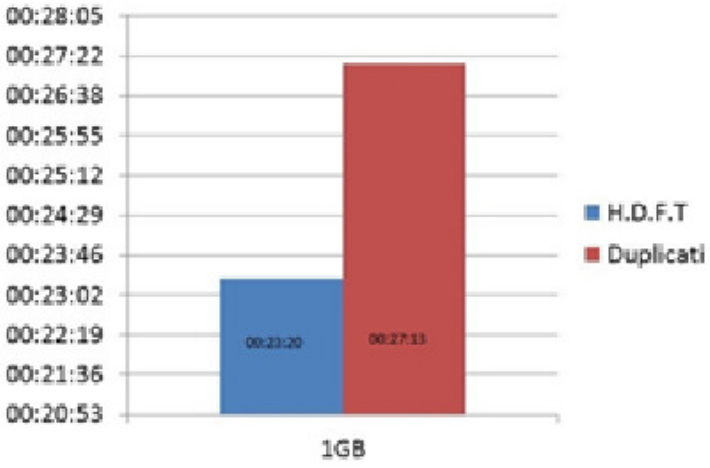

Figure 23. 1 GB for SFTP

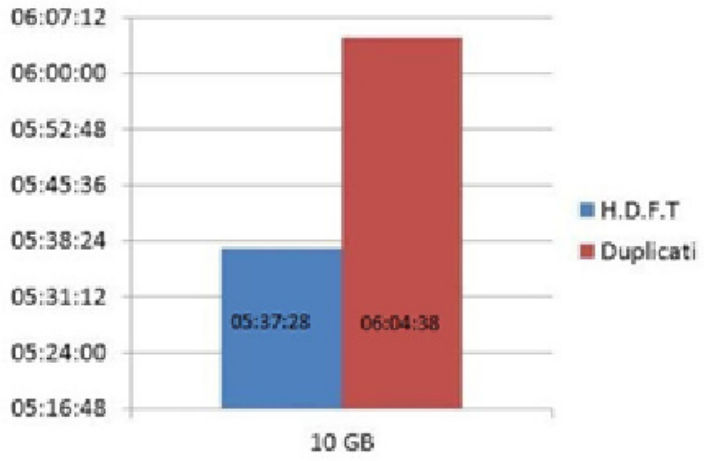

Figure 24. 10 GB for SFTP

3. CONCLUSIONS

The first aim of this paper is to provide a better and more efficient performance and to develop a faster application than Duplicati. First of all, the algorithm of Duplicati is examined. Then, we develop and create our own algorithms for sending huge data file. After obtaining the test result, we assure huge data file transfer application to complete the same process faster than Duplicati less than two times. So, huge data file transfer becomes more effective and faster.

The algorithm, which is implemented in this paper, consists compressing the file with the format of .zip and splitting the file into the pieces. It can send the file successfully, efficiently and faster. This is the main contribution of this paper.

REFERENCES

[1] Information Storage and Management Storing, Managing, and Protecting Digital Information Edited by G. Somasundaram Alok Shrivastava, EMC Education Services, 27-29.

[2] The Embedded Internet TCP/IP basics, implementation and applications Edited by Sergio Scaglia , 2007,225-228

[3] Computer Networking A Top-Down Approach Featuring the Internet third edition Edited by James F. Kurose, Keith W. Ross 96-98.

[4] Data Transfer Linda Woodard Consultant Cornell CAC Workshop: Parallel Computing on Stampede: June 18, 2013, 2.

[5] Managed File Transfer Solutions using DataPower and WebSphere MQ File Transfer Edition Edited by IBM,4-5.

[6] Computing Community Consortium Version 8 : December 22 , 2008