BIG DATA: Explored using HPCC Systems

Dr.Parul Agarwal [1]
Department of Computer Science,
Jamia Hamdard
New Delhi-INDIA

Mohd Rashid [2]
Department of Computer Science,
Jamia Hamdard
New Delhi-INDIA

Corresponding author : rash93mls@gmail.com

Abstract:

In current computer age, handling data that is too big, difficult to store and manage and hard to be perceived is "Big Data". In this paper, the general concept of big data has been reviewed besides presenting an overview on connected techniques and tools such as the commonly referred to Hadoop. The technical challenges in using the above principal is described. The paper emphasizes and demonstrates in a procedural manner the method for using HPCC (High Performance Computing Cluster) systems. This includes step by step procedure explained extensively for a child dataset and then using it as per the need for inference of relevant knowledge. It is then concluded with a mention of future directions associated with it.

Keywords:

Big data, Hadoop, Map Reduce technique, child data sets, HPCC systems.

I. INTRODUCTION

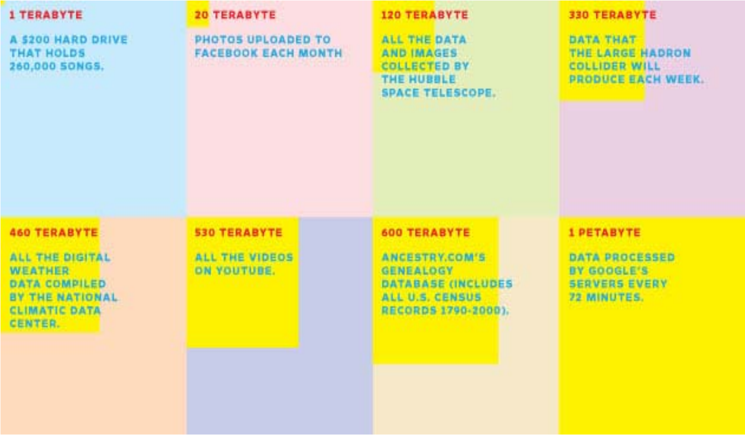
Over the past, data has exponentially grown and is estimated to grow at a faster speed in the years to come. The term big data has earned importance in few years for the fact that it differs from huge or the so called massive data. Several questions like "What is Big Data", "What is its source of origin", "how do we store and then use it", "How can it be used for data mining", etc need to be answered. We start off by answering each of them in this paper. Big data[5] is data which is massive but additionally the one that needs real time analysis. It also introduces with it challenges in terms of storage and inference. Owing to its reality of origin, i.e. websites generating tens and hundreds of Petabytes of data, several definitions of it have been suggested [19,20].

A. Challenges of Big Data

There are many challenges associated with the analysis of big data [1], [2], as the present research is still in early stage. To improve the storage, efficiency of display and analysis of big data a considerable research efforts are needed [4]. The supporting technology which includes cloud computing, grid computing and big data architectures have to be fully explored. The hardware and software requirements [8] as on today are high and thus restrict its access for most of the users. A minimal requirement framework has to be identified that makes the storage and maintenance of big data computing easy. The source of big data [21] is heterogeneous in nature. Thus a method is required which would build these unstructured data into a format which is compatible with each other.

For the above stated heterogeneous nature of data, big data oriented databases have to be developed that can be configured on minimal requirements. Searching of relevant real time big data[15] is a challenge which has to be overcome. Big data is huge in size. Thus is susceptible to redundancy .The redundant data is an overhead for processing which needs to be removed in such a manner that the values in the data are not lost. In addition, good compression techniques might be useful for the storage of this type of data. Once, we overcome these challenges, its applications in sciences, data mining and many real life application areas can be fully exploited. Fig 1[7] below contains a description of how fast this data is increasing.

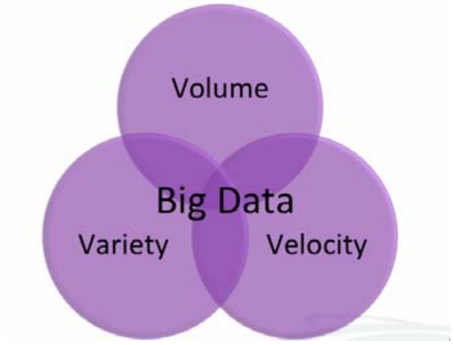
Figure 1 The continuously increasing big data


1 TERABYTE
A $200 HARD DRIVE THAT HOLDS 260,000 SONGS.
20 TERABYTE
PHOTOS UPLOADED TO FACEBOOK EACH MONTH
120 TERABYTE
ALL THE DATA AND IMAGES COLLECTED BY THE HUBBLE SPACE TELESCOPE.
330 TERABYTE
DATA THAT THE LARGE HADRON COLLIDER WILL PRODUCE EACH WEEK.
460 TERABYTE
ALL THE DIGITAL WEATHER DATA COMPILED BY THE NATIONAL CLIMATIC DATA CENTER.
530 TERABYTE
ALL THE VIDEOS ON YOUTUBE.
600 TERABYTE
ANCESTRY.COM'S GENEALOGY DATABASE (INCLUDES ALL U.S. CENSUS RECORDS 1790-2000).
1 PETABYTE
DATA PROCESSED BY GOOGLE'S SERVERS EVERY 72 MINUTES.

B. Importance Of 3V’s

In [3, 10, 21], the author has mentioned the challenges and opportunities faced because of increasing data in a 3V’s Model where the V’s denote Volume, Velocity and variety[11]. Fig. 2 describes the 3V’s Model[20] by employing a diagrammatic approach. Volume denotes the amount of data, Velocity signifies that collection and analysis of big data is to be done in a timely manner so that full exploitation of its value can be done and variety denotes heterogeneity of data[18] i.e. data is coming from various sources and has no common format of its representation like audio, video, webpage, and text, as well as unstructured data. Though recently, another parameter has gained importance in connection with big data i.e. hidden values in the data. It deals with identifying values from data having an enormous scale and deals with their generation. In Section II, we discuss the latest work as literature review. Section III focuses on various tools and technologies supporting big data.

Figure 2. 3 V’s Model



II. LITERATURE REVIEW

In [9], the author discusses the scope, methods, advantages and challenges of Data. They described that as the data increases it becomes more complex. The main challenges is that how to extract useful data from big data. In [22] the author show the how to extract the useful data from the largest amount of data. To implement Google’s Mapreduce Model, an open source platform “Hadoop” is used. In [14], this paper addresses the problem that occur in mapreduce technique of hadoop. Hadoop Distributed File System (HDFS) is explained for storage and Map reduce programming framework for parallel processing to process large data sets. In [13], the author discusses the benefits of Grid computing center i.e. how high processing power can be harnessed for big data and how its storage capability can be high.

III. TOOLS AND TECHNIQUES

To handle the processing of BIG DATA, several supporting related technologies exist[3] like Data Centers, Cloud Computing, IoT, Hadoop, HPCC systems to name a few. Its worth mentioning that the development in big data is leading to enhancements in these technologies. Thus, big data and these tools are supplementing each other’s growth.

Data Centers: Data Center is a platform that stores massive data, in addition to managing, gathering and organizing data. The need of the hour is to develop an efficient data center which would capacitate a large number of nodes and build a high speed network whose transmission capacity would be high.

For smooth functioning of big data, Cloud computing is another solution. Cloud Computing is an ideal solution where huge storage resources are needed. It would indeed accelerate the computing of big data significantly.

In this section, we discuss the widely used Hadoop framework and discuss how HPCC systems can be explored to support big data.

A. HADOOP

Hadoop[14] is a programming framework which is used for supporting the processing of large data sets in a distributed computing environment. Google’s Mapreduce has developed Hadoop [14], and it is a software framework for breaking an application into various parts. Hadoop caters to the needs of two categories of users. The first time users for whom a single node setup is defined and a distributed cluster setup for those who have had an experience in handling it.

The present Apache Hadoop ecosystem includes Hadoop Kernel, Mapreduce, HDFS and a numbers of various components like Apache Hive, Base and Zookeeper. Hadoop is already in use in industry in the area of big data applications (for example yahoo, Facebooks, twitter etc).

1. MAPREDUCE

MapReduce[14] is a programming model and an associated implementation to process and generates large data sets. Two functions (Map and Reduce) are expressed by Map Reduce library for computation. Map, written by the user, after taking an input pair, it produces a set of intermediate key/value pairs. Then the reduce function, merges together these values to form a possibly smaller set of values.

2. Architecture of MAPREDUCE

MapReduce is a software framework meant for processing multi tera bytes of data for parallel processing on large clusters i.e. thousands of nodes in a fault tolerant manner. A Map Reduce job usually works in the following manner: Split the input data into independent chunks. These chunks are then processed by the map tasks in a completely parallel manner. The output of the maps in then sorted, which become input to the reduce tasks. The storage of the input and the output is a file system.

Thus the Map Reduce framework and the file systems referred as HDFS execute on the same set of nodes. The Map Reduce framework has one master job tracker and one slave task tracker for each cluster node. The function of master job tracker is to schedule the job’s component responsibilities/ task on the slaves. It has a monitoring role also besides re-executing the failed task in case of failures. The slaves on the other hand execute the tasks as directed by the master.

An HDFS [14] installation has *Namenode* i.e. it is a master server which functions to manage the file system namespace besides regulating access of files by clients.

The responsibility of *Datanodes* includes serving read and write requests from file system clients. Besides the above functions, it has the function of performing: to create block, to delete, and to replicate upon instruction from the *Namenode*.

3. Limitations of Hadoop

- ➢ Unable to control the order in which the maps or reductions are executed.
- ➢ Maps and Reduces are dependent on data generated in the same MapReduce job (i.e. stateless)
- ➢ Executing a map reduce task is time consuming
- ➢ The reduce operations do not take place until all Maps are complete.
- ➢ Most of the beginners find Hadoop framework too complex to initiate and end up in a hole.

Thus another easy to understand and use technology has been suggested in the next subsection HPCC Systems.

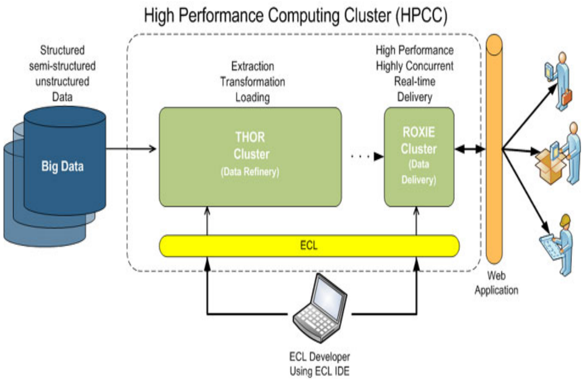
B. HPCC SYSTEMS

The HPCC Systems[6] architecture incorporates the Thor and Roxie clusters as well as common middleware components, an external communications layer, client interfaces which provide both end-user services and system management tools, and auxiliary components to support monitoring and to facilitate loading and storing of filesystem data from external sources. An HPCC environment can include only Thor clusters, or both Thor and Roxie clusters. Each of these cluster types is described in more detail in the following sections.

The diagram[6] below illustrates a high level overview of the platform architecture and how the components all work together as a powerful solution for managing Big Data. These components are described in detail below.

Figure 3. An Overview of HPCC systems

High-Level HPCC Architecture



Thor[6] , as the name implies is a data refinery cluster that consumes vast amount of data, transforming, linking and indexing that data. It plays a very important role in big data field such as it performs parallel processing power across several nodes. It’s a single threaded ECL[6].

Roxie[6] is another important concept associated with HPCC. It means Rapid online XML Inquiry Engine and is a data delivery engine of HPCC. It is a multi Threaded ECL which helps to support thousands of requests per node per second.

ECL[6](Enterprise Control Language) is the powerful programming language that is ideally suited for the manipulation of Big Data. Its features are multi fold which include transparency, modularity, reusability, Easily extensible using C++ libraries.

ECL IDE [6] is a modern IDE used to code, debug and monitor ECL programs and has access to shared source code repositories. It provides a complete development, debugging and testing environment for developing ECL dataflow programs.

We now discuss the various parameters and steps required for its installation, and usage.

System Requirements

Running HPCC in a virtual machine[16] requires (at minimum):

- A personal computer running Windows XP, Vista, Windows 7 (either 32- or 64-bit)
- A minimum of 2 GB ram, with at least 1.5 GB of free memory available.
- Intel Pentium D (or better) or AMD Ahlon64/Opteron/

Phenom processor

- A virtualization software package: VMware Player or Server (version 4.0 or later) or Oracle VM VirtualBox (version 4.0 or later).
- Internet Explorer 7 or 8, Google Chrome 10, or Firefox™ 3.0 (or later).

Getting the Tools and the VM Image

To run the virtual machine version of the HPCC System, you need virtualization software. These packages allow you to run virtual images inside a single host. For our experiments, the VMware player(a virtualization software used to run the HPCC virtual machine) was used.

In the following sections, we discuss the following:

- Download and install the VMware Player
- Download the HPCC virtual machine image from HPCC Systems.
- Open and import the image in the VMware Player

Once you have completed these steps, you can evaluate the HPCC Platform and learn how to use it.

We now discuss the first step i.e. Downloading and Installing the VMware Player.

1.Go to the VMware site: http://www.vmware.com/products/player/.

2. Click on download link, then follow the instructions to download the VMware Player for 32-bit and 64-bit Windows..Registration is required, but the player is free.

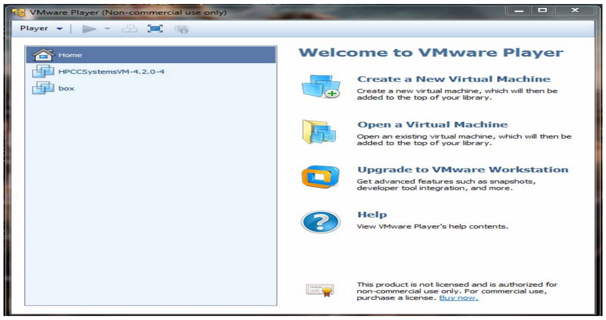
3. Download the VMware Player (save to a folder on your machine).

4. Follow VMware's on-screen instructions and install the VMware Player.

These are the following steps to fetch the data.

Step 1:

Fig. 4 Opening of the VMware workstation.



Welcome to VMware Player

Create a New Virtual Machine
Create a new virtual machine, which will then be added to the top of your library.

Open a Virtual Machine
Open an existing virtual machine, which will then be added to the top of your library.

Upgrade to VMware Workstation
Get advanced features such as snapshots, developer tool integration, and more.

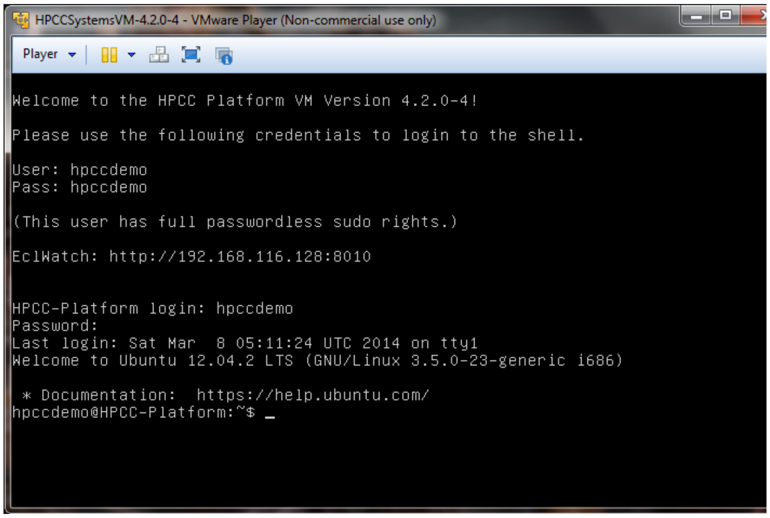
Help
View VMware Player's help contents.

This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.

Step 2:

The user now needs to login to the HPCC systems. A window would appear where this process takes place.

Fig 5. Login on the hpcc system.

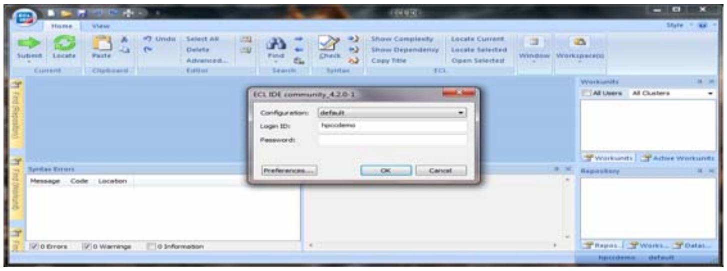

HPCCSystemsVM-4.2.0-4 - VMware Player (Non-commercial use only)
Player

```
Welcome to the HPCC Platform VM Version 4.2.0-4!

Please use the following credentials to login to the shell.

User: hpccdemo
Pass: hpccdemo

(This user has full passwordless sudo rights.)

EclWatch: http://192.168.116.128:8010


HPCC-Platform login: hpccdemo
Password:
Last login: Sat Mar  8 05:11:24 UTC 2014 on tty1
Welcome to Ubuntu 12.04.2 LTS (GNU/Linux 3.5.0-23-generic i686)

 * Documentation:  https://help.ubuntu.com/
hpccdemo@HPCC-Platform:~$ _
```

Step 3:

The next step is opening the ECL IDE which is shown in Fig. 6.

Fig. 6 Opening the ECL IDE

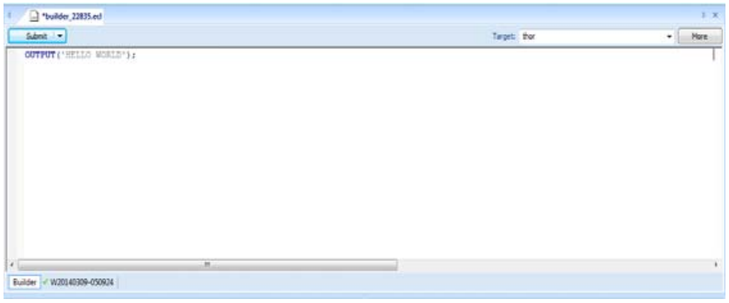


ECL IDE community_4.2.0-1

Configuration: default
Login ID: hpccdemo
Password:

Step 4:Once a connectivity has been established, a new builder has to opened whwere the user would type output "HELLO WORLD".

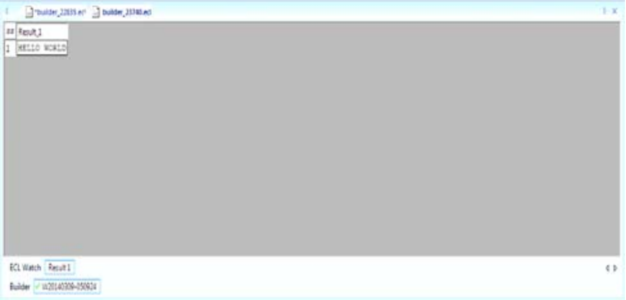
Fig 7 HELLO WORLD in ECL IDE.



```
OUTPUT('HELLO WORLD');
```

Step 5:

Fig 8 Output of the HELLO WORLD program.



```
1  HELLO WORLD
```

Step 6:

We now discuss how a dataset[17] shown in Fig 9 which can be passed using the following syntax of HPCC systems.

*DefinitionName***(DATASET(** *recstruct* **)** *AliasName* **)**
:= *expression***;**
The required *recstruct* defines the layout of fields in the passed DATASET parameter. The required *AliasName* names the dataset for use in the function and is used in the Definition's *expression* to indicate where in the operation the passed parameter is to be used.

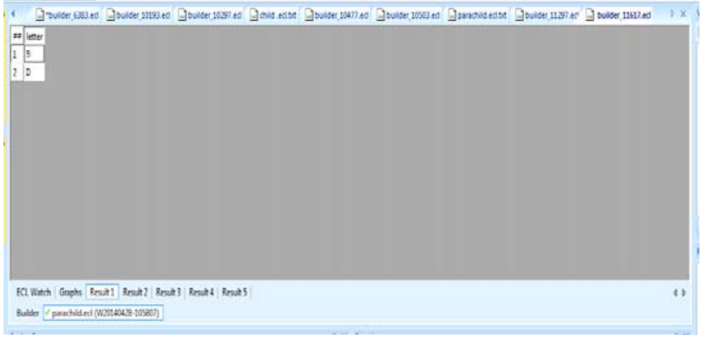
Fig 9 A Child dataset fetched using HPCC systems

```
MyRec := {STRING1 Letter};


SomeFile :=
DATASET([{'A'},{'B'},{'C'},{'D'},{'E'}],MyRec);


FilteredDS(DATASET(MyRec) ds) := ds(Letter NOT IN
['A','C','E']);

        //passed dataset referenced as "ds" in expression
```

```
OUTPUT(FilteredDS(SomeFile));
```

Step 7: Fig 10 contains the output. The screenshot clearly shows that B,D are displayed as output when the dataset was executed.

Fig 10 output of Child Dataset



##	letter
1	B
2	D

Conclusion:

We have entered an era of data explosion i.e. Big Data. Various tools are available as on now which support big data. Each having its advantages and disadvantages in terms of amount of support that they provide and the domain area. This paper discusses two major technologies which are sure to make waves and bring a revolution so as to meet the challenges of big data namely Acquisition, storage, Processing, Security and result interpretation. A lot has been done and even more work needs to be done in this domain. In future, we would be proposing a new clustering technique that produces good results for big data.

REFERENCES

[1]“Big data analytics: Turning insight into action”,[www.teradata.com]

[white paper08.13 eB 6839]

[2] Jagadish, H. V., et al. Univ. of Michigan (Coordinator)], “Challenges and Opportunities with Big Data”, 2012.

[3] Chen, M., Mao, S., Liu, Y. “Big Data: A Survey” published online in Springer Science + business media, New York, 2014.

[4] Shilpa, Kaur, M. “BIG Data and Methodology-A review” in International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, 2013.

[5]http://strata.oreilly.com/2012/01/what-is-big-data.html

[6] http://hpccsystems.com/Why-HPCC/How-it-works

[7] Dissertation Thesis by Gianmarco De Francisci Morales , 2012, “Big Data and theWeb: Algorithms for Data Intensive Scalable Computing” in IMT Institute for Advanced Studies, Lucca, Italy.

[8] Bakshi, K., ” Considerations for big data: Architecture and approach”, Aerospace conference,2012 IEEE(3-10 MAY,12).

[9] Sagiroglu, S.; Sinanc, D. 2013 ,”Big Data: A Review”, International conferences on collaboration technologies and system (CTS),2013, pp 42-47.

[10] Katal, A Wazid, M. ;Goudar, R.H.,.” Big data: Issues, Challenges, Tools and Good practices”in Contemporary Computing ,2013 Sixth International Conferences on

[11]http://dashburst.com/infographic/big-data-volume-variety-velocity/

[12] http://www-01.ibm.com/software/in/data/bigdata/

[13] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G. ;”A Big Data implementation based on Grid Computing”, 2013.

[14] Aditya B. Patel, Manashvi Birla, Ushma Nair, “Addressing big data problem using Hadoop and Map Reduce”, (NUiCONE), Proceedings published in Nirma University International Conferences on Engineering, 2012.

[15] Vaswani, G., Bhatia, A. “A Real Time Approach with Big Data-A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, volume 3, Issue 9, 2013.

[16] www. vmware.com

[17] hpccsystems.com/download/docs/ecl-language-reference/html/DATASET_as_a_parameter_type.html

[18] http://www.nature.com/news/specials/bigdata/index.html

[19] http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/

[20]O. R. Team (2011) Big data now: Current Perspectives from OReilly Radar. OReilly Media

[21] Grobelnik M (2012) Big data tutorial. http://videolectures.net/eswc2012grobelnikbigdata/

[22] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., ” Shared Disk data Analytics with Apache Hadoop”, published in 19th International Conference on High Computing,, 2012.

[23] Laney, D., “ Data Management: Controlling data, Volume,Velocity and Variety, META Group Research Note, 2001.

AUTHORS PROFILE

1) **Dr. Parul Agarwal , Ph.D(Computer Science)**

Has 14 years of teaching experience at University level. Specialization includes Fuzzy Data Mining, Soft Computing. Has published several papers in Indexed, reputed International Journals. Have reviewed several papers published in International Conferences, Journals.

2) **Mohd Rashid,(Student)**

He completed his senior secondary certificate from Allahabad and currently pursuing B.Tech in Computer Science Engineering from Jamia Hamdard University. He has presented various papers in national conferences .