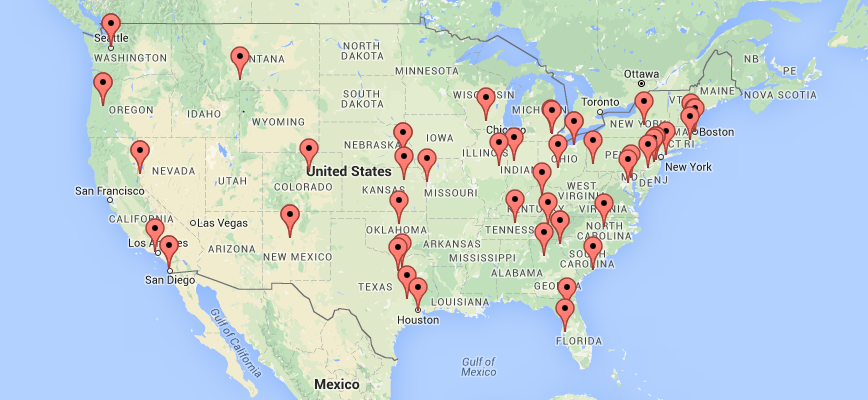
**PROJECT FINAL PHASE**

**Data Analytics Using Network Measurements  
Team: The Coding Ninjas**

**Objective:** Determine the geographical locations based on the Network measurements. **Introduction:** In this project we are exchanging data from different network nodes and calculating the delay in the data downloads. Based on the delay values obtained we are geographically determining the location of the existing network nodes using statistical methods. Error correction techniques are applied on the obtained delay values and the corresponding graphs are plotted.

**Logical Architecture:**

Geographic location of Selected nodes



Node N

Node 1

Round Trip Delay Time Round Trip Delay Time

Node 2

**Distance calculation between nodes:**

The time between starting the transmission of a data packet and receiving the corresponding immediate acknowledgement is referred to as Remote delay (dremote).  The duration of receiving one data packet and sending out the immediate acknowledgement is referred to as local delay (dlocal.

The overall propagation time is then estimated by subtracting the local from the remote delay.

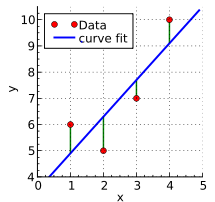
C = where c ≈ 3 · 108 m/ s being the speed of light.

**Progress until phase 1:**

* Access to the planet-lab (an online network services platform) and got access to the network nodes present in different location.
* Generated key and checked authentication for the different network nodes added.
* Installed Java 1.7 on all the added nodes and sorted few access issues.
* Developed a simple java Client Server program for testing the communication between the nodes.
* Added the sample file exchange feature to the simple java Client Server program and fixed the bugs received while development.
* Added functions to generate the Time Stamps and Delays.
* Learnt a brief about the different useful statistical methods for data correction.

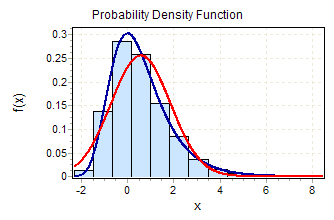
**Error correction techniques:**

Least Mean Square method:  
The LMSQ method is a kind of filter that estimates or mimics the right value by finding the points in the graph that produce the least mean square of the error which is the difference between the desired and actual value of the signal.



Probability Density Estimation:

In probability density Function also called as PDF or density of a continuous random variable, is a function that predicts the likelihood for the random variable to take on a given value. The density of the graph is higher at the point where most variables lie and the probability is more at this point. The probability density function is positive or zero an it’s integral is one over the entire domain.



**Progress until phase 2:**

* Added the file write functionality to the existing Client Server program to log the delay logs to it for each download.
* Fixed the file overwrite bugs.
* Altered the program implementing arrays for 60 network nodes and 5 different file sizes.
* Fixed few bugs in the program with the array loops.
* Ran the parallel Client Server program and recorded the logs in the text file.
* Collecting the log data for the data analysis.

**Final Phase:**

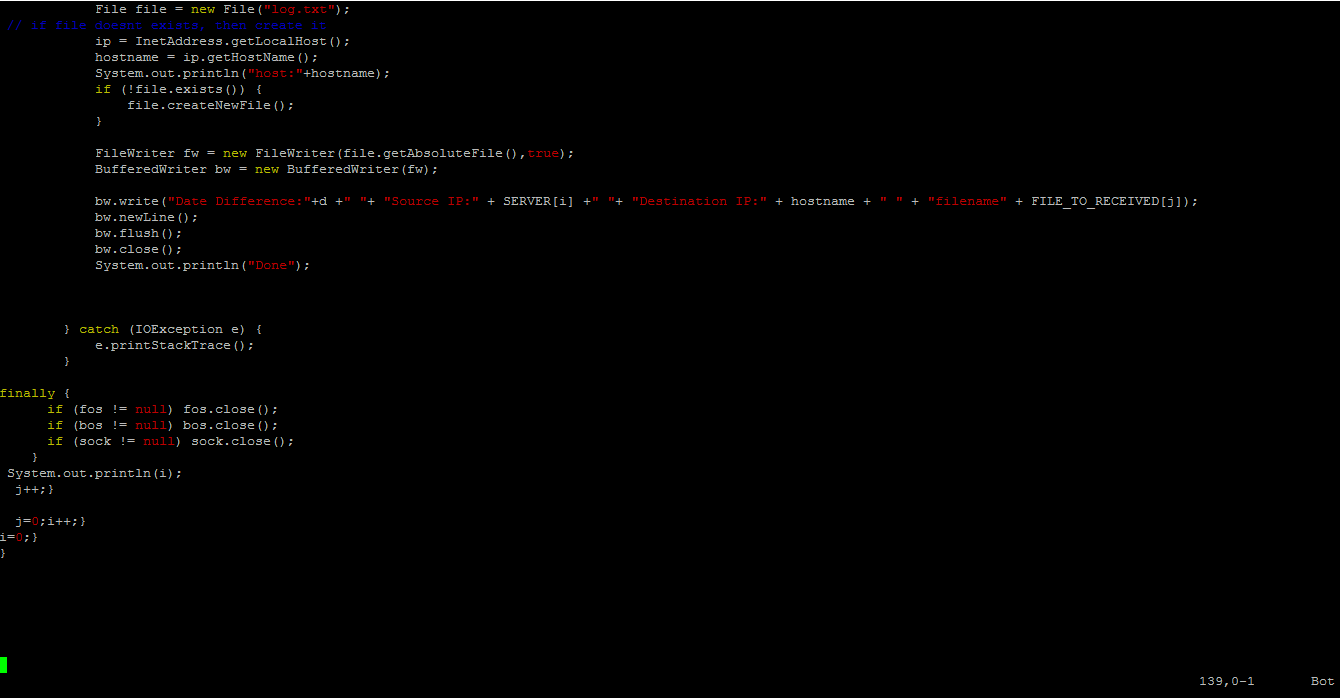
* Fixed the few bugs in the client-server final program.
* Automated the task of running all the network nodes in parallel through shell scripting.
* Created a database to maintain the log files and store the values of delay, source and destination IP’s, and corresponding file size downloaded.
* Calculated the DDR ratio and predicted values from the obtained delay, distance values.
* Applied error correction techniques to find the approximate delay between 2 network nodes for different file sizes.
* Plotted the cumulative distribution function curves for the obtained delay and distance values.

**Client Program:**



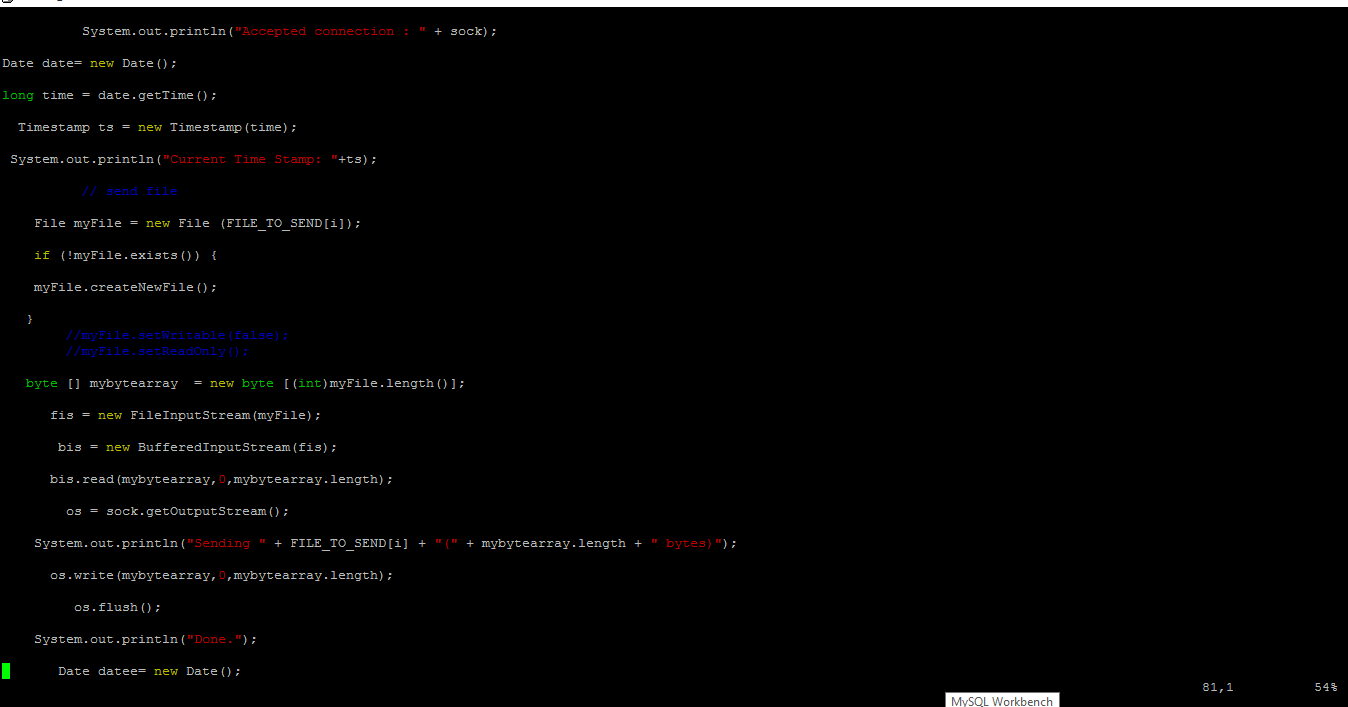






**Server Program:**

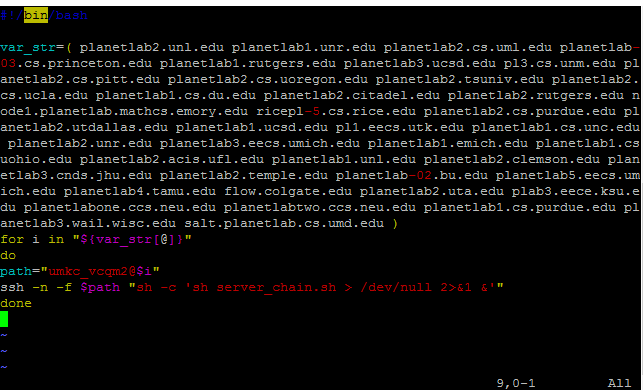


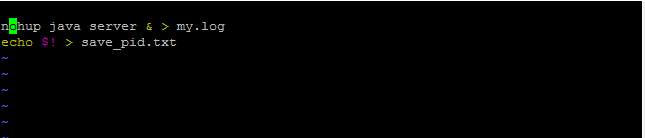




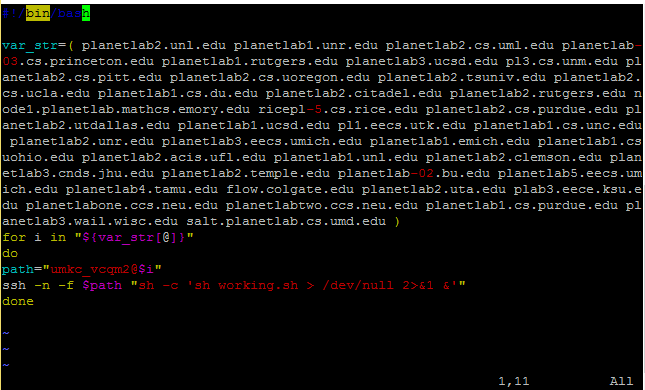
**Shell Scripting to automate the process:**

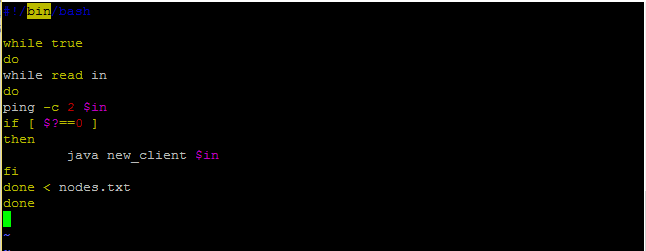
**To Start Server**



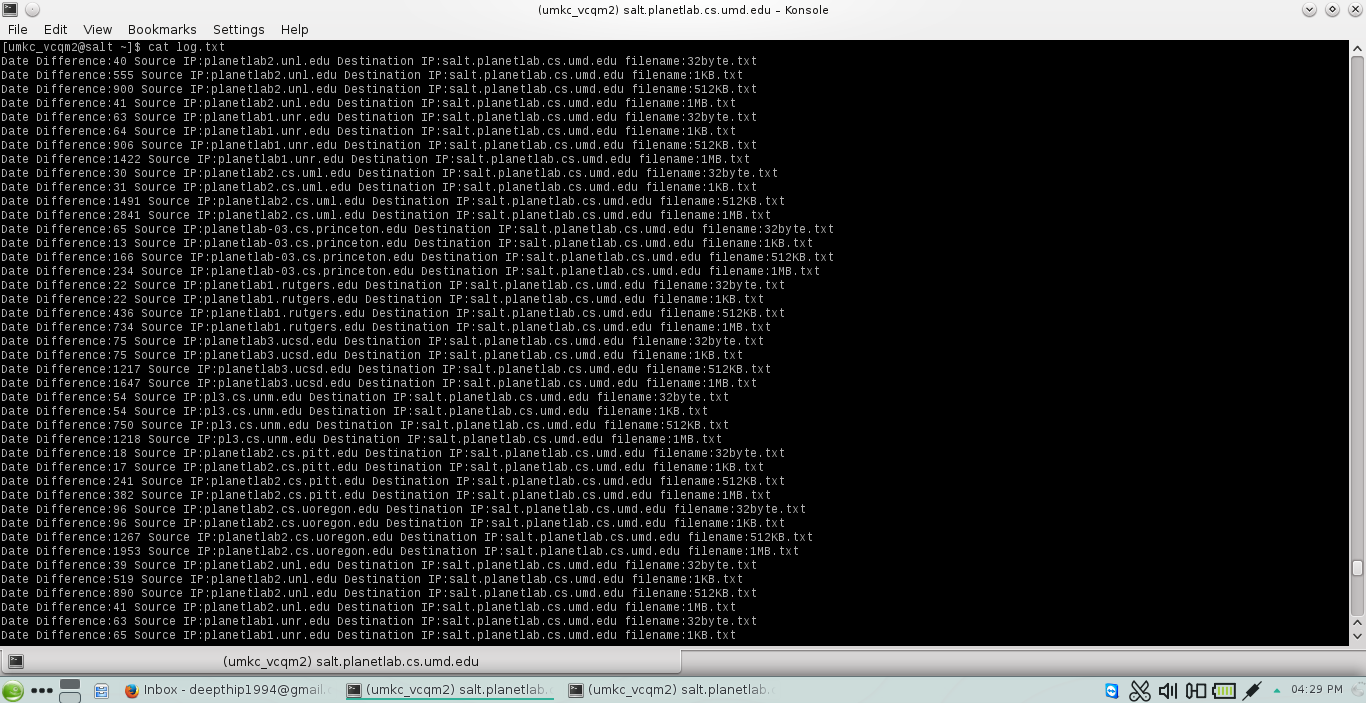


**To Start Client:**



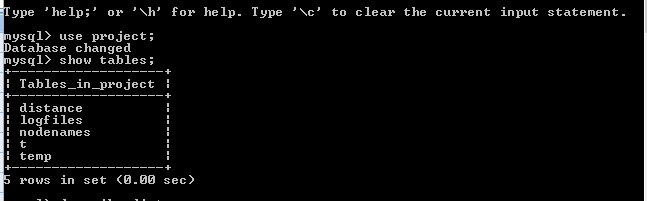


**Sample Log File Generated:**

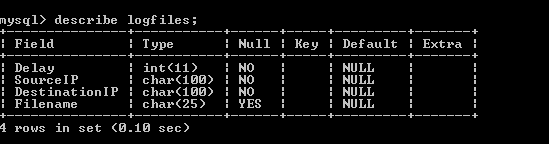
****

**Database Schema:**

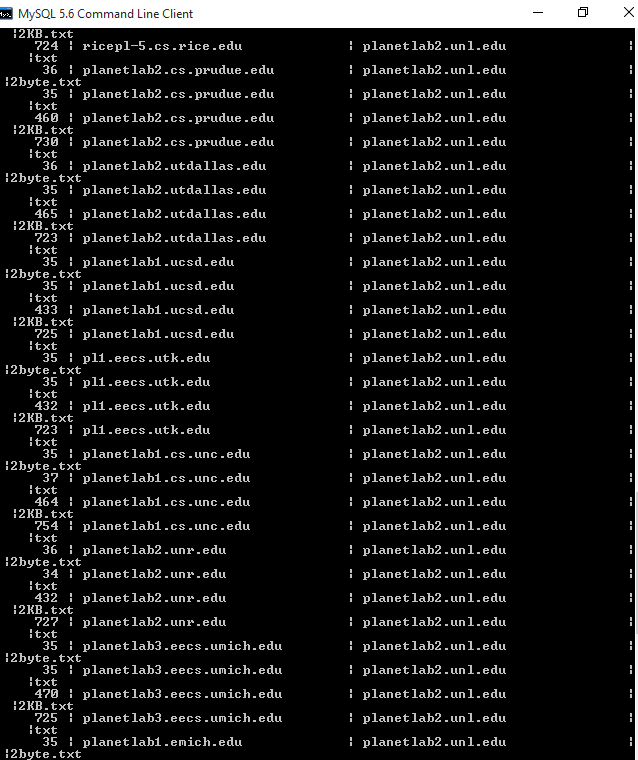
**Tables Created in database “ Project”:**

****

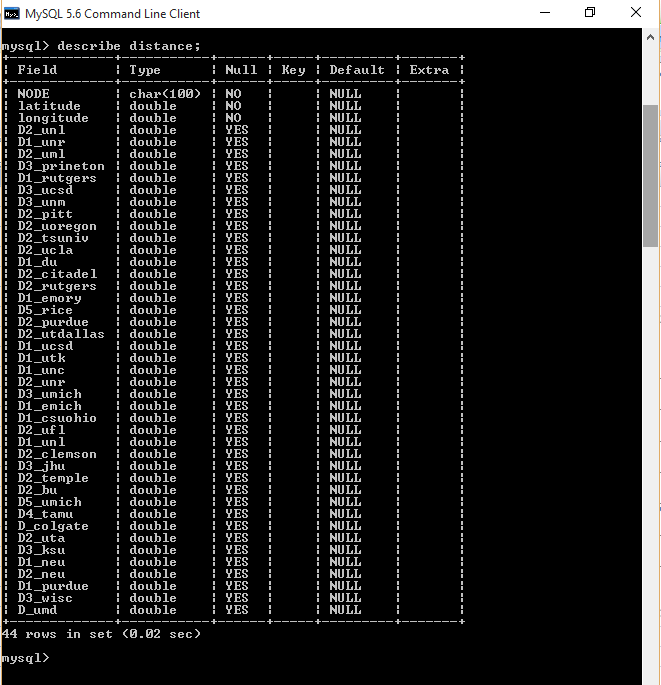
**Table “ logfiles”:**

****

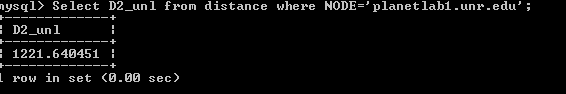
**Sample Result of the query on table logfiles:**

****

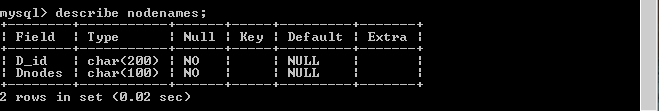
**Table “distance”:**

****

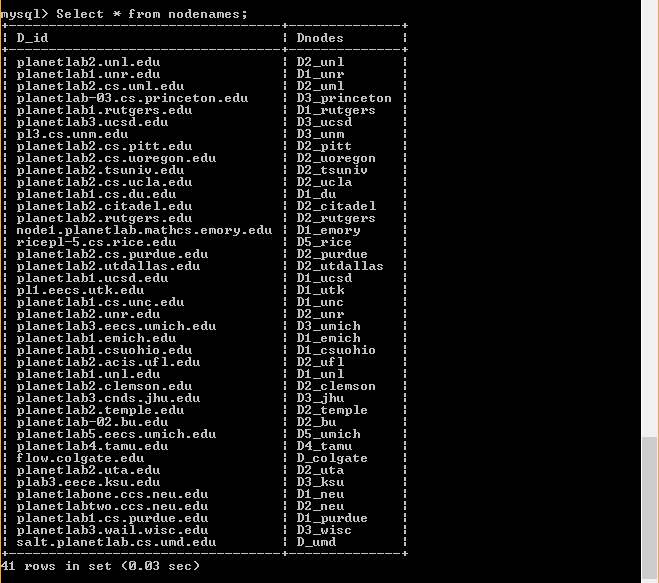
**Result of the query on table distance:**

****

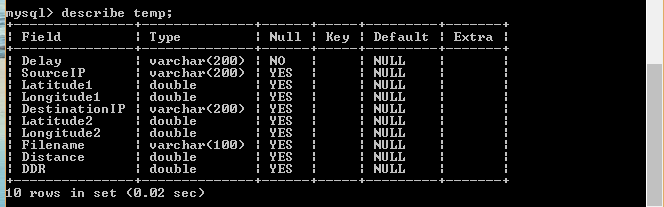
**Table “nodenames”:**

****

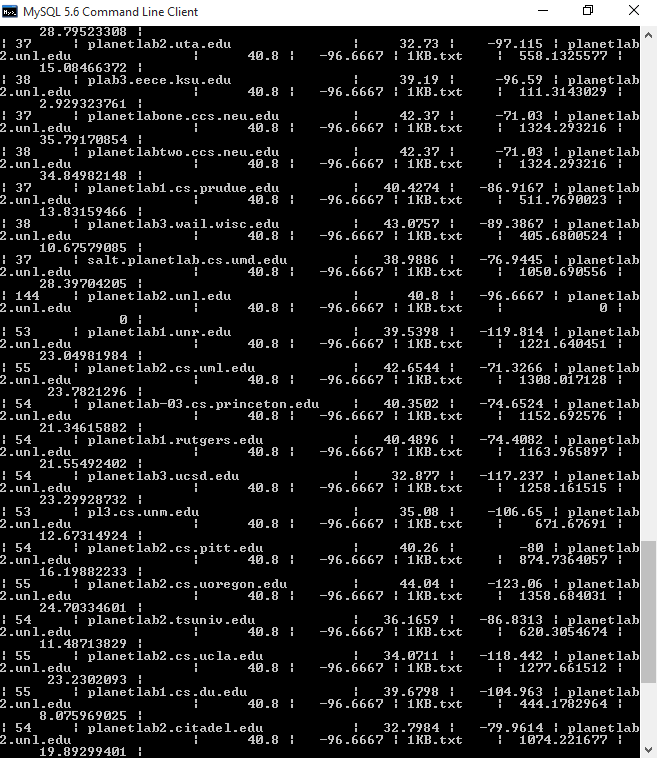
**Result of the query on table nodenames:**

****

**Table “temp”:**

****

**Result of the query on table temp:**

****

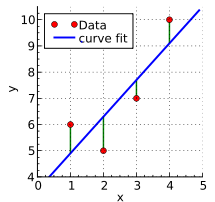
**Error Correction Techniques:**

**Linear Regression:**

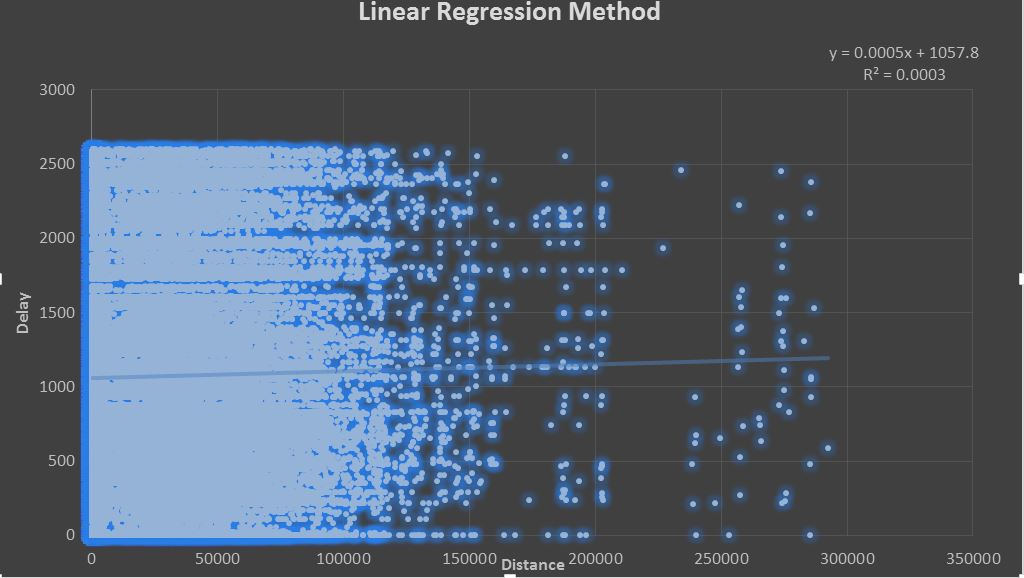
Linear Regression is an approach in which relationship between a dependent and independent variable is modeled based on predictor functions which are obtained from the estimated data. Here we would take a certain set of delay and distance value and plot it to a graph to predict the distance function. This linear function would help us correct the errors and predict a distance value close to the actual value.

This would in-turn reduce a lot of outliers by calculating the mean based on the distribution of the data.

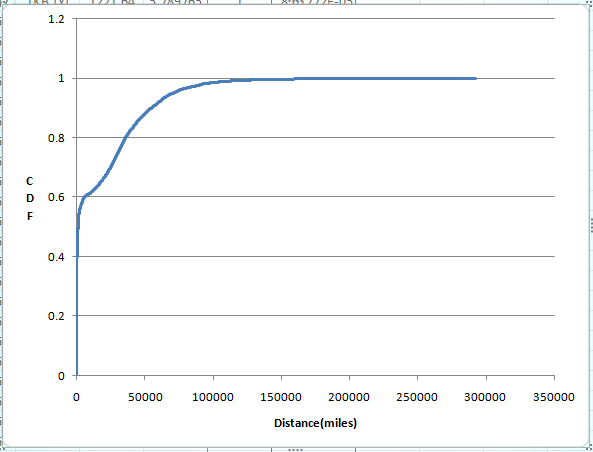
Sample linear regression using squares method.



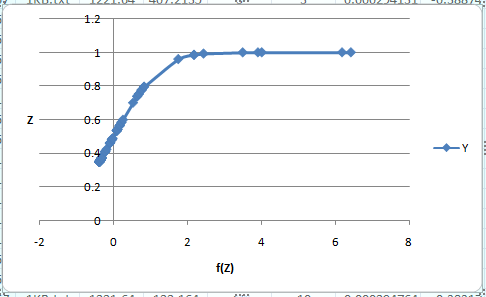
**Linear regression for the entire data set obtained:**



**Cumulative distributive function for the data set:**



**Normal Distribution function for the data set:**



**Challenges faced and their solutions**:

**Installation Phase:**

* Adding the Nodes from Planet lab and checking the status of the active nodes timely.

Sol: Keep adding the nodes that are active and include a ping function in your source code.

* Installation of JDK on each node and few of the nodes were not able to install the latest versions.

Sol: We had to replace those nodes and check for the nodes that could install the JDK 1.7.

* Few of the nodes gave authentication problems, not accepting the private key.

Sol: We have re-added those nodes.

**Development Phase:**

* Understanding how JAVA Socket Programming worked based on TCP.

Sol: We had to go through few online courses to understand the basis on which the data exchange works using TCP.

* After developing the Java Socket program that could exchange a file over two nodes, we were having few bind exception errors on the client side program.

Sol: We had figured out that a demon program that was running on the background that prevented the execution of the main client program.

* After successfully completing the data exchange, we had to log the details into a log file that was getting overridden.

Sol: We had to change the function used in the java source code to BufferedReader and remove logging functionality to avoid the complexity of the code and also alter the schema of the log.

* After developing the code, we had problems in execution as we included the nodes in an array within the java code which in turn started giving bind exception errors and port already in use errors due to the demon processes.

Sol: We had to break the array and store the nodes in a separate text file and wrote a shell script to execute the program sequentially.

* Once the program had successfully worked we had problems in data collection as the task had become manual and it needed execution on 40 different nodes.

Sol: We had to learn shell scripting in detail and know how to automate the starting of the client and server program.

**Data dumping into database and Data Analysis**:

* There was a lot of data with huge delay values due to inefficient nodes.

Sol: We had to use linear regression technique to remove the outliers and predict the distance value reducing the errors.

* We used MS-Excel for plotting the probability function and the cumulative distribution function as well as the linear regression graphs.

Sol: Plotting in excel was a bit tedious as the data set was huge, however we were suggested to use R- programming for data analytics as it being a very efficient language.

**Deficiencies and future scope:**

* We could improve the delay times by fine tuning the nodes.
* Data analytics part could have been improved using the R-programming.
* We can extend this data set to nodes across the world which could provide more relevant data for data analytics accounting for the losses due to propagation delay.
* We could use polynomial regression methods for error correction by including different delay factors for packet transmissions.

**Team Member Role:**

Deepthi: Nodes maintenance and Client-Server Programming.

Divya: Client-Server Programming and Database Design.

Varun: Shell Scripting and Data analytics.

Breno : Helped in fine tuning the code.

**Reference:**

**Determining the Geographic Location of Internet Hosts : Venkata N. Padmanabhan and Lakshminarayanan Subramanian**

<http://research.microsoft.com/pubs/69829/tr-2000-110.pdf>

Java Socket programming:

http://cs.lmu.edu/~ray/notes/javanetexamples/