SYNOPSIS

**On**

## ***Data Analysis of IT Sector in India using Big Data***

(Minor Project)

Tanya Gupta - 04214803114

Archit Gupta - 04114803114

Group: I-123

Dated: 16th Aug`17

Submitted to:-

Mr. Nitesh Wadhera

## **Statement about the Problem:-**

## The IT industry is continuously growing in India but there hasn’t been any tool yet which can analyze the growth of this sector with such large dataset with immediate results. Such problem can be addressed using a tool which can fetch any analysis related query on huge datasets and can give immediate results.

## **Why is the particular topic chosen?**

This topic finds its relevance in analysis of growth of IT industry of India to judge the increase in the number of IT companies in various states and at central level too.

This tool would be able to handle huge sized dataset of companies which normally are found to be difficult to access in a fast manner to fetch relevant results

## **Objective and scope of the project**

Using dataset of companies to:

1. Observe IT growth in India since past few decades in terms of various factors such as Sate wise growth to understand development needed in the same.
2. Understand private and public sector growth of industries in India.
3. Understand the capital investment involved in various sectors of industry and many more.

## **Methodology/Process description:-**

Dataset of companies which is huge in size will be first accessed through Cloudera software using Hadoop technology.

Using this technology, various queries would be coded down to use the dataset to give back all the results needed in minimal time.

Those results would be then converted into graphical representation to study the growth.

## **Required Resources** :

## Software-

* 1. Cloudera
  2. Eclipse

## **What contribution would the project make?**

* This will help in studying IT structure of India.
* Various parameters needed to decide future steps to be taken for improvement in various states can be figured out using this analysis.
* Analyzing growth pattern of various industries in India.
* Ultimately creating a tool which would be able to handle any big size of industry data and would give much faster statistical results than normal processors.

The Schedule of the project

* Identify Statistics needed: (2 days)
* Data Acquisition: (5 days)
* Process/Clean Data: (1 week)
* Exploratory Analysis: (1 week)
* Designing Queries: (5 days)
* Creating code: (5 days)
* Implementing Code & Validation: (1 week)
* Debugging code: (5 days)
* Running code and fetching results: (1 week)
* Graphical Conversion of results: (5 days)
* Visualize Results: (5 days)

References/Bibliography

Data is taken from:

[http://](http://simplystatistics.tumblr.com/post/15182715327/list-of-citiesstates-with-open-data-help-me)data.gov.in