Data Product: Analysis, Visualization and Prediction

Tejal Agrawal
Student
Department of Computer Engineering
Atharva College of Engineering, Mumbai, India

Akshen Doke
Student
Department of Computer Engineering
Atharva College of Engineering, Mumbai, India

Ashish Gaikwad
Student
Department of Computer Engineering
Atharva College of Engineering, Mumbai, India

Prof. Mahendra Patil
Head of the Department
Department of Computer Engineering
Atharva College of Engineering, Mumbai, India

Abstract

Today information technology is developing rapidly and adopting application of IT has created a significant revolution in many fields including business. One of the emerging and promising technology that provide efficient means to access various types of data and information all over the world is Data mining which also helps in decision making. The Objective of this paper is to provide a cross platform for data cleaning, analysis and visualization. This tool helps data scientist to come up with a solution to big data related issues by analyzing the output and predicting the trends. This tool will run on all major platforms which will enable the user to analyze data at one's ease.

Keywords- Data Mining; Data cleaning; data analysis; data visualization; acquisition; serialization

I. INTRODUCTION

Data is the foundation of digital age. In the last few decades there has been a tremendous amount of data which is stored in electronic format. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world doubles every 20 months[10]. With such an increasing quantity of data it is becoming more important for business to find a way to analyze it. One of the important phase in knowledge discovery and includes application of discovery, analytical methods on data to produce specific models across data is Data mining. It can be used to predict the future as well. Due to the widespread availability of huge, complex, information-rich data sets, the ability to extract useful knowledge hidden in these data and to act on that knowledge has become increasingly important in today's competitive world. Thus data mining is analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to data owner. Briefly, data mining is an approach to research and analysis. It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Sometime, data may be in different formats as it comes from different sources, irrelevant attributes and missing data. Therefore, data needs to be prepared before applying any kind of data mining. Data mining is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing. Many researchers and practitioners use data mining as a synonym for knowledge discovery but data mining is also just one step of the knowledge discovery process. All the techniques follow an automated process of knowledge discovery (KDD) i.e., data cleaning, data integration, data selection, data transformation, data mining and knowledge representation.

II. RELATED WORK

Traditionally two widely used tools have been presented for the purpose of Data analyzing and visualization: OLAP tool and WEKA tool.. These approaches perform well for analyzing and visualizing but still have some drawbacks.

A. OLAP (Online Analysis and Processing)

At Present OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions [6].

At the core of any OLAP system is an OLAP cube (also called a 'multidimensional cube' or a hypercube). It consists of numeric facts called measures which are categorized by dimensions. The measures are placed at the intersections of the hypercube, which is spanned by the dimensions as a vector space. The usual interface to manipulate an OLAP cube is a matrix interface, like

Pivot tables in a spreadsheet program, which performs projection operations along the dimensions, such as aggregation or averaging.

The cube metadata is typically created from a star schema or snowflake schema or fact constellation of tables in a relational database. Measures are derived from the records in the fact table and dimensions are derived from the dimension tables.

Each measure can be thought of as having a set of labels, or metadata associated with it. A dimension is what describes these labels; it provides information about the measure.

But the existing system has some drawbacks

– Pre-modeling as a must
– Great dependence on IT
– Poor computation capability
– Short of Interactive analysis ability
– Slow in reacting
– Great potential risk

These shortcomings can easily incur the failure of OLAP project, and bring about unrecoverable loss to the enterprise.

B. WEKA (Waikato Environment for Knowledge Analysis)

WEKA is a data mining/machine learning application developed by Department of Computer Science, University of Waikato, New Zealand. It is open source software in JAVA issued under the GNU General Public License. It is a collection tools for data pre-processing, classification, regression, clustering, association, and visualization. It is a collection of machine learning algorithms for data mining tasks . WEKA is well-suited for developing new machine learning schemes[9].

1) Disadvantages of WEKA

– Sequence modeling is not covered by the algorithms included in the Weka distribution
– Not capable of multi-relational data mining

2) Memory Bound

The main Drawback of WEKA tool is it can handle only small Datasets. Whenever a set is bigger than a few megabytes an Out Of Memory error occurs[8].

III. IMPLEMENTED SYSTEM

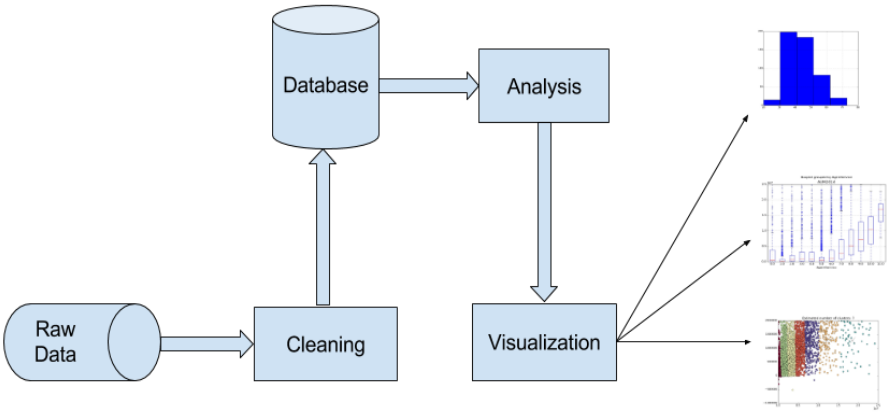

Fig. 1: System Block Diagram

To resolve the problem of setting up software's for data handling, we propose a new cloud based application which will be available to anyone at anytime from anywhere. It will have a client interface on all major platforms including mobile. It will be scalable as you grow, kind of a system where user don't have to do any upfront investment in software's.

Our application will have three main sections:

1) Data Cleaning
2) Data Analysis
3) Data Visualization

In data cleaning, users can perform various operations to convert raw or dirty data into easy to process data. For example, replace missing values, replace non numeric entry, string to float.

In data analysis, user can aggregate, disaggregate, summarize, check relationship between numbers such as ratios and normalize numbers etc.

In data visualization, we are primarily focusing on box plot, histogram, and cluster plotting and time series. Data scientist and also data enthusiast can then use these visualizations for further processing or prediction.

Data storage and computation will be done on cloud or powerful servers. Clients for accessing and interacting with data will be available on all major platforms such as Web, Desktop, Android and iOS.

IV. METHODOLOGY

As per lean methodology, we worked in iterations. In each iteration, we focused on feedback loop.

A. First Iteration (Prototype)

We started with writing python scripts for loading data in memory and storing it into CSV. After initial success in reading and storing data, we measured script's performance and figured out bottleneck. It was disk IO as storing to disk is expensive task in terms of time.

B. Second Iteration (Web Framework)

We designed basic architecture of the system with REST API. We tested many web frameworks in python such as Django, Flask, and Pyramid. Finally due to flexibility and simplicity of Pyramid, we choose it for our web.

C. Third Iteration (API & Frontend Framework)

In third iteration we started implementing our REST API as per design made during previous step. We successfully implemented and tested file upload and storage via API. We also setup web frontend MVC using AngularJS framework. As our team have previous experience in working with this framework, we choose it to speed up our development process.

D. Fourth Iteration (Auth & API changes)

In fourth iteration, we added user account feature with authentication through AJAX. After adding authentication we had to change few things about our API design such as making all API calls authentication aware.

E. Fifth Iteration (Dashboard)

This was very stressed out iteration in terms of time. We prototyped and tested many different dashboard designs with AngularJS. Finally settled down on simple and clean dashboard design.

F. Sixth Iteration (Data Processing Engine)

It was tricky to implement data processing engine in asynchronous manar. We implemented it using AJAX call which auto populates/updates dashboard once API call is completed. Data processing engine is set of scripts we developed during first iteration. Of course we have to modify those scripts to make them work with our web app.

G. Seventh Iteration (Visualization Prototype)

We prototyped data visualization through python scripts. We tried out histogram, box plot, cluster, time series etc. It took us some time to understand how data cleaning affects output of visualization. There we learn to deal with outliers, null values, wrong values, negative numbers etc.

H. Eight Iteration (Rework on architecture)

By this time our data processing engine was huge and now we have to integrate visualization in it. So we took time out to re work on the existing design of the system and separated out data cleaning and visualization. For example, we made different modules for cluster and box plotting. After doing re work out codes were very much manageable and simple to understand.

I. Ninth Iteration (Testing with large data)

We tested our web app with large data. We got it from open source data providing websites. Also we requested to a startup founder to let us use their data for testing our software. Initially we had few performance issues and failure of the entire web app due to such large data. We solved those issues by updating our python modules.

J. Tenth Iteration (Hybrid App)

We started implementing platform specific interfaces (clients) which can serve as alternative to our web app and make it easy for end users to access and process on their data from any platform. We completed hybrid mobile app using Ionic framework. We choose Ionic framework as it was based on AngularJS so our learning curve was very low to pick up mobile app development. We tested it on Android devices.

K. Eleventh Iteration (Desktop App)

This was also cross platform compatible i.e. it can work on Linux, OS X and Windows without re compiling. We used tried out different frameworks of python for this such as PyQt, WxPython. We also tested our hybrid mobile app on iOS platform.

V. RESULT AND ANALYSIS

1) LOGIN FORM in the application, first the Login Form is opened. Enter valid Username and Password, click on Submit as shown in Fig.2. The appropriate page will be opened.

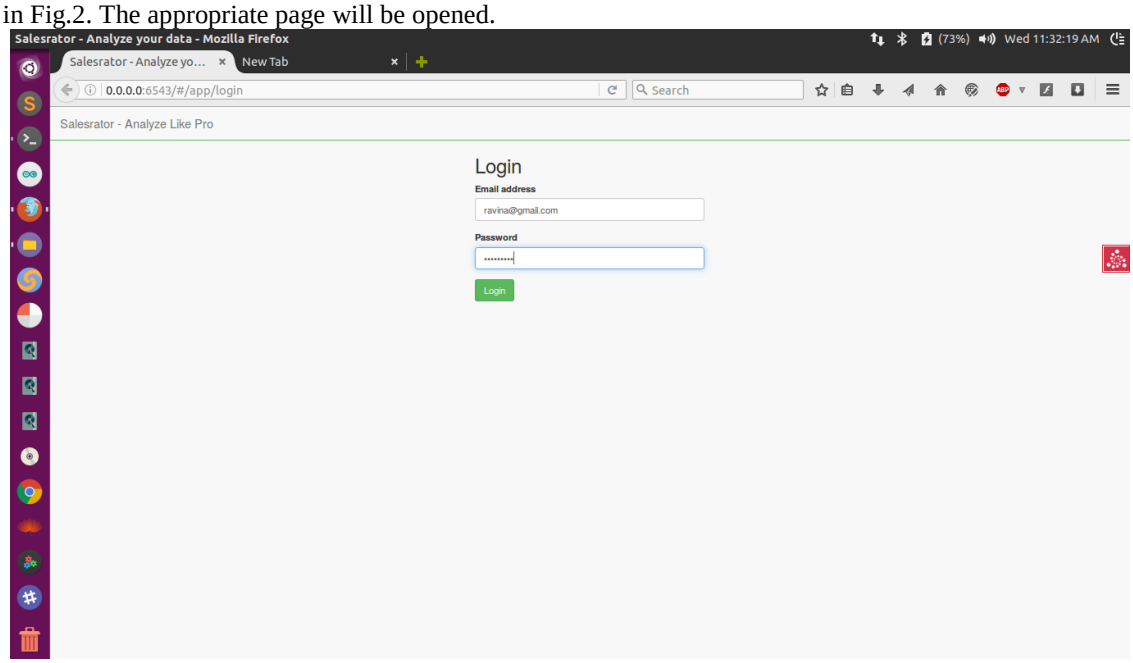

Fig. 2: Login Page

2) As the next page is loaded user selects the file in CSV format and uploads it.

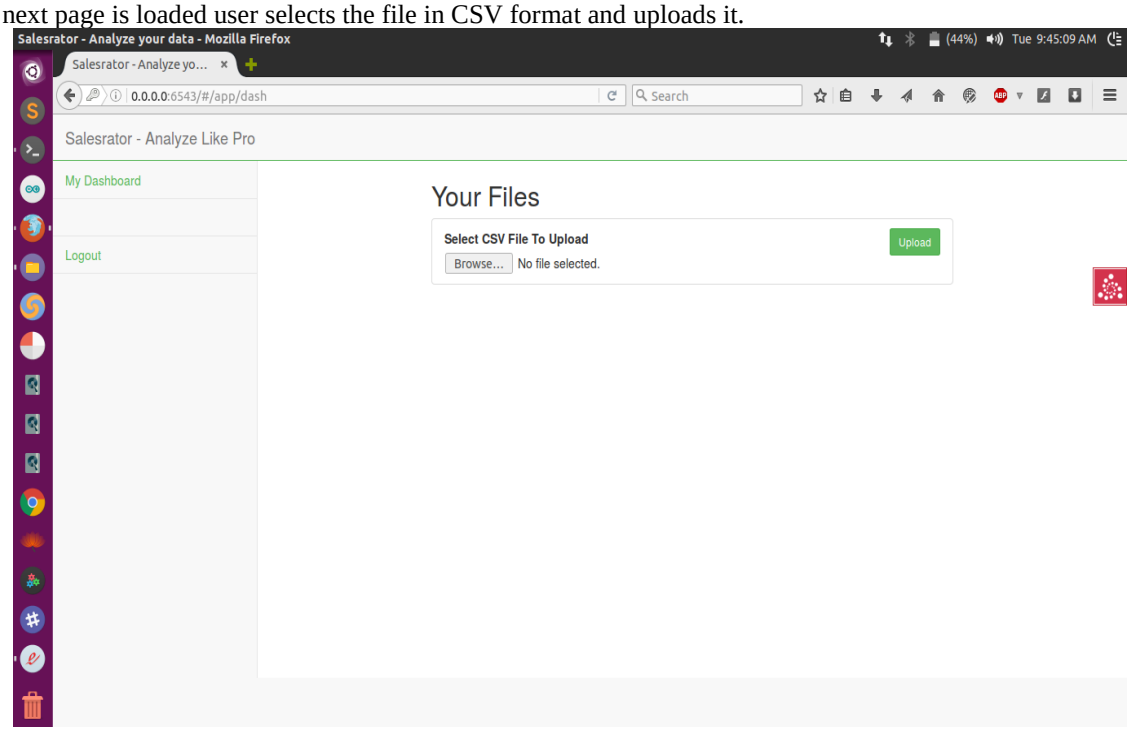

Fig. 3: uploading of CSV file

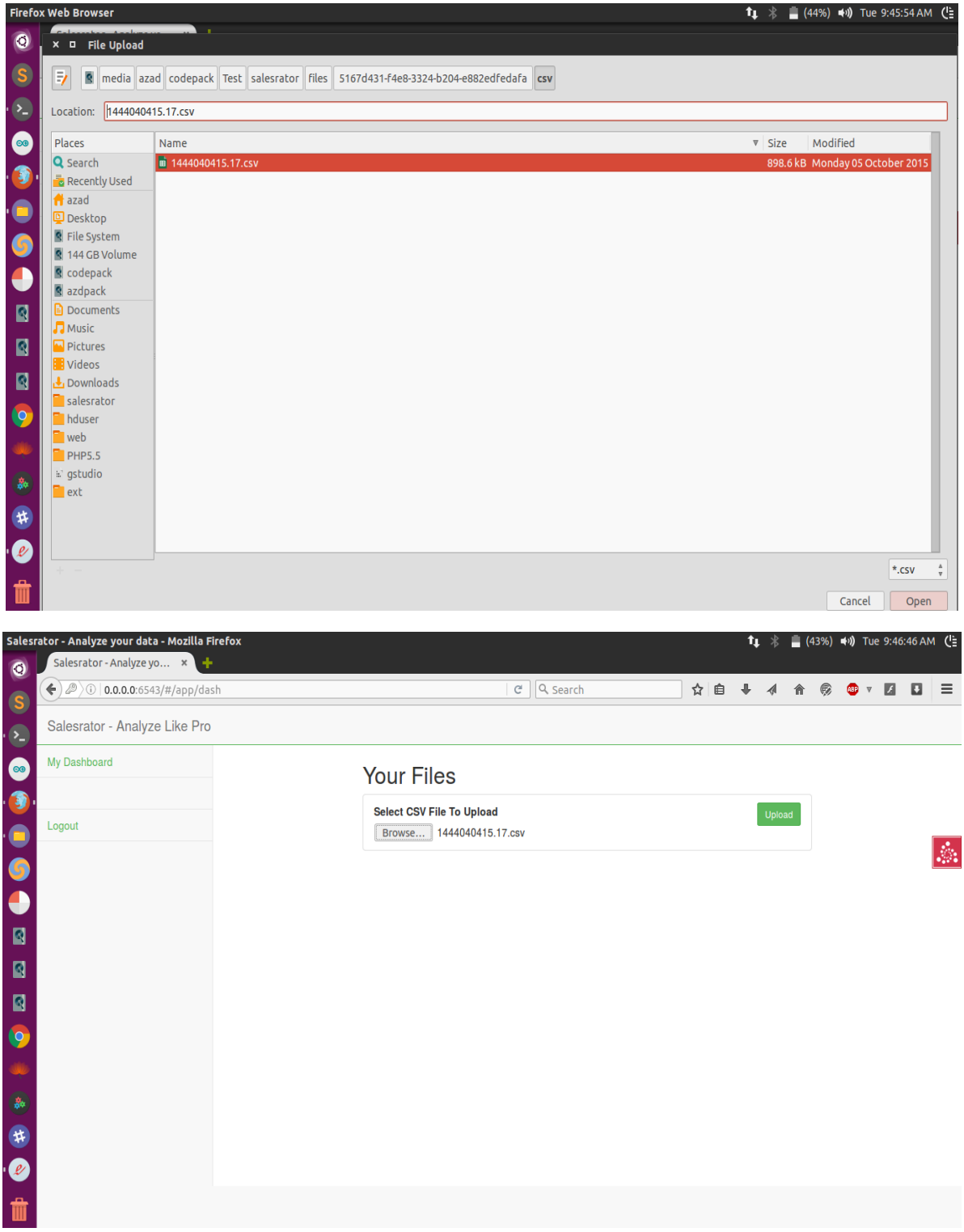

Fig. 4: upload successful

3) After uploading the file user selects the operation to be performed. It includes features like Cleaning and visualizations.

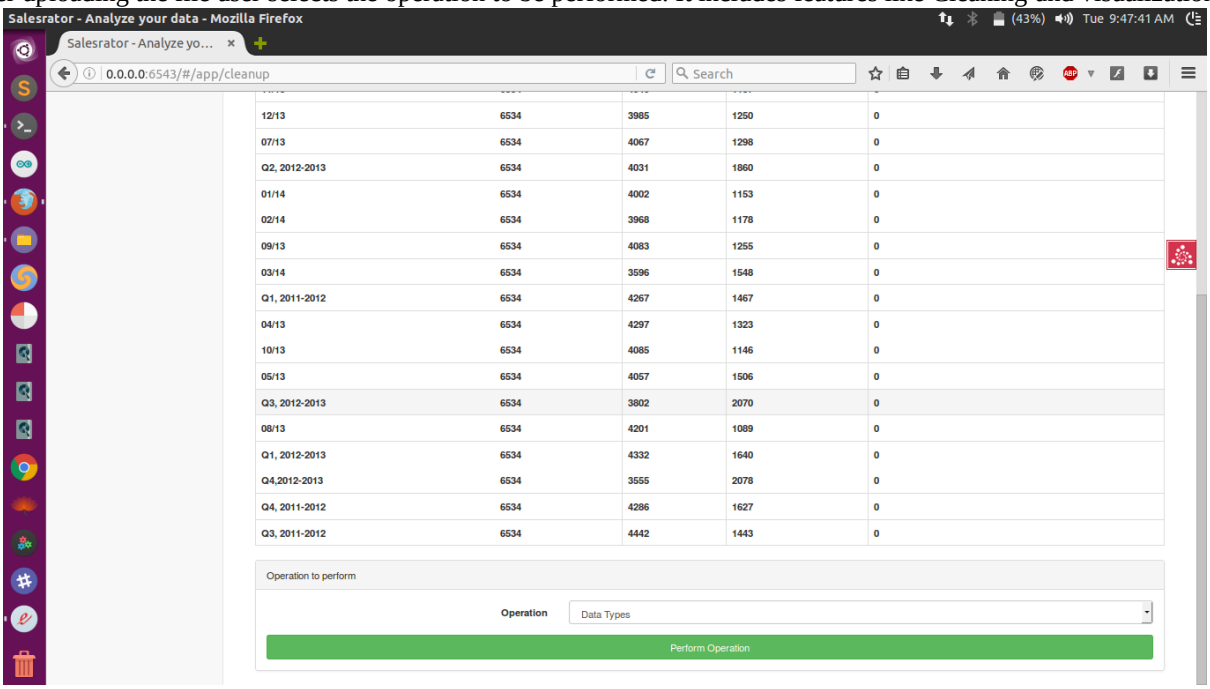

Fig. 5: File uploaded

4) Selection of type of visualization

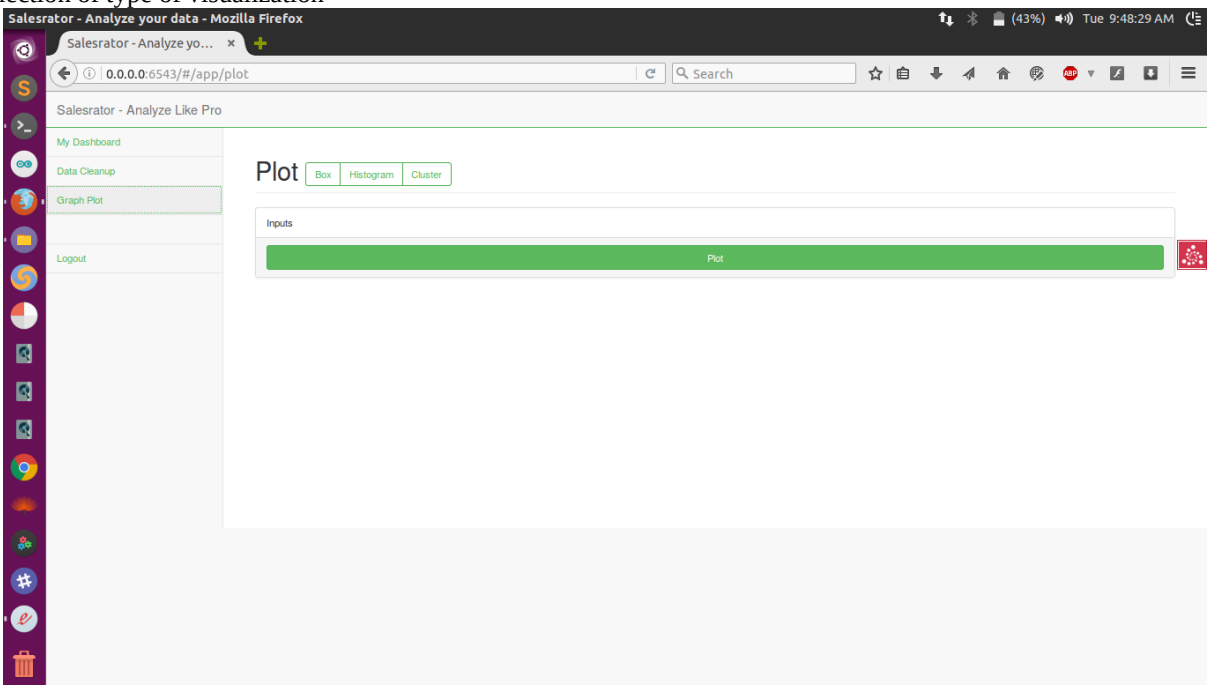

Fig. 6: Select type of visualization

5) Box Plot

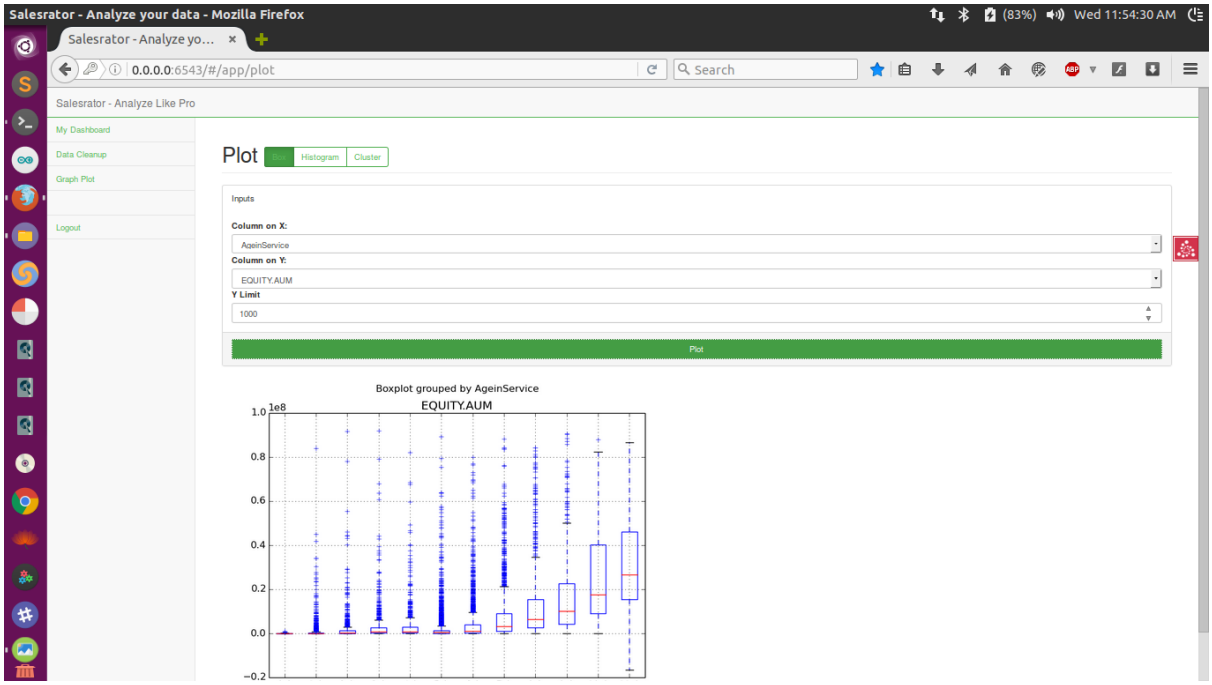

Fig. 7: Box plot

A. Step 4.2: Histogram

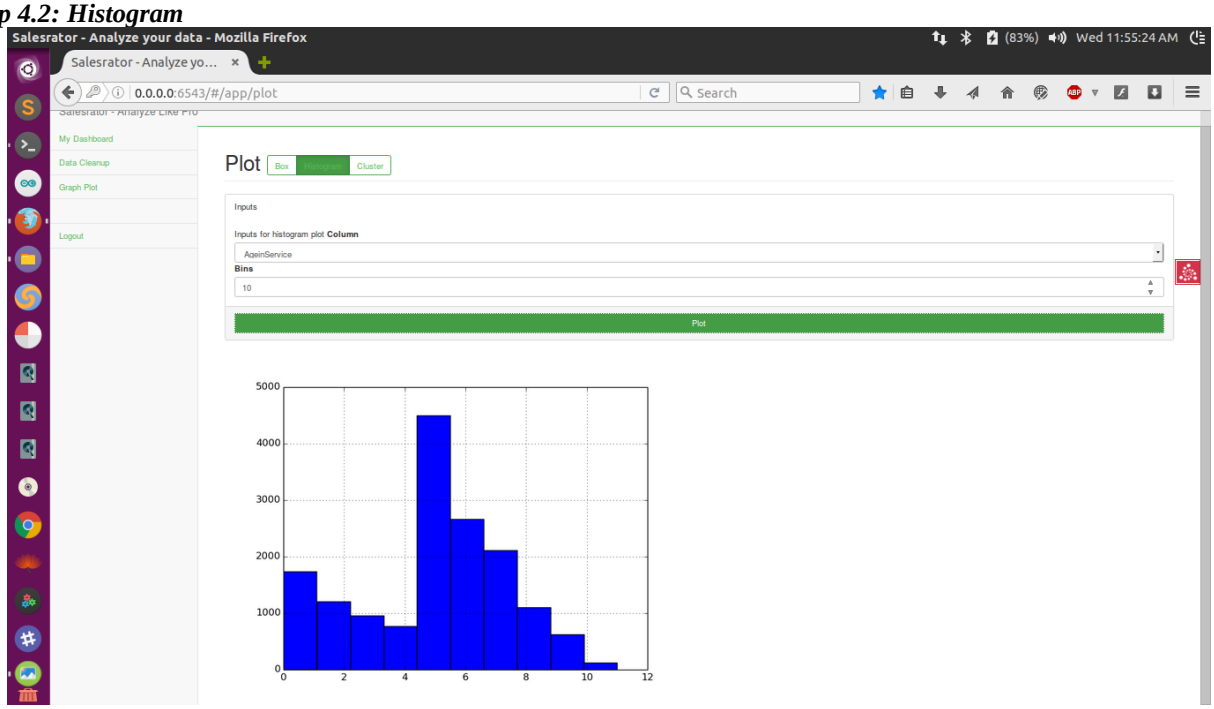

Fig. 8: Histogram

B. Step 4.3 Cluster

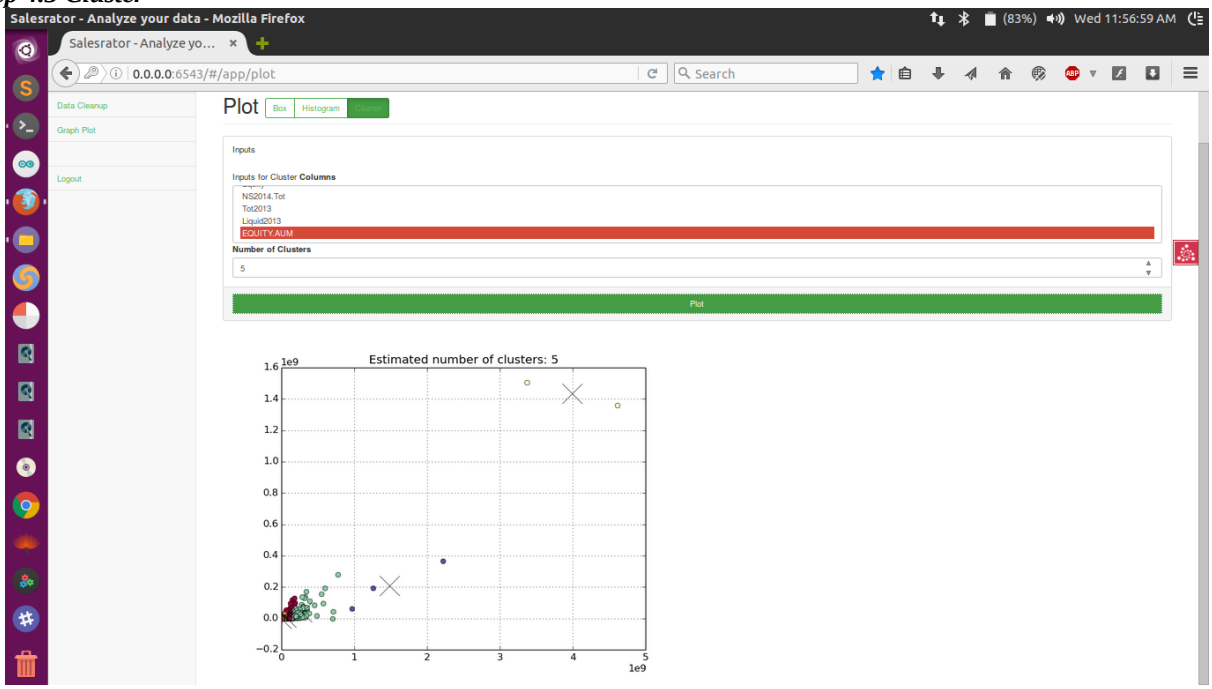

Fig. 9: Cluster

VI. CONCLUSION

Raw data is growing day by day at a rapid rate making it difficult for data analysts to analyze such huge data and visualize it securely and conveniently. The main purpose of our project is to develop a tool which can analyze and visualize this raw data and also help data scientist to predict future trends from analyzed data on all major platforms. Using this tool, any small-medium sized organization can also find out meaningful insights for their company without spending on system setup or paying to any third party for it.

REFERENCES

[1] Analytical Review of Data Visualization Methods in Application to Big Data, Evgeniy Yur'evich Gorodov and Vasiliy Vasil'evich Gubarev, Journal of Electrical and Computer EngineeringVolume 2013 (2013), Article ID 969458
[2] Data visualization, Zhao Kaidi School of Computing, National University of Singapore, Matrix Number:HT006177E
[3] O'Neil, Cathy and, Schutt, Rachel (2014). Doing Data Science. O'Reilly.ISBN 978-1-449-35865-5.
[4] http://www.businessdictionary.com/definition/dataanalysis.html
[5] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19
[6] Surajit Chaudhuri and Umeshwar Dayal (1997). "An overview of data warehousing and OLAP technology". SIGMOD Rec. (ACM) 26 (1): 65.
[7] www.sas.com/en_us/insights/big-data/data visualization.html
[8] Rahm, E., & Hai Do, H. University of Leipzig, Germany, (n.d.). Data cleaning: Problems and current approaches.
[9] www.bigdata4analytics.com/uploads/2/1/9/2/21928796/ data_visualisation.pdf
[10] http://www.ijarcsse.com/docs/papers/Volume_4/6_June2014/V4I6-0145.pdf