**5. BIG DATA ANALYSIS ON**

**MEDICAL INSURANCE DATASET**

1. **ABSTRACT**

Data analysis is important to businesses will be an understatement. In fact, no business can survive without analyzing available data. Merely analyzing data isn't sufficient from the point of view of making a decision. How does one interpret from the analyzed data is more important.

Unleashing the value of analytics in insurance scenario can transform how insurers do business. It can significantly improve the profit margins for the insurers along with closing or reducing frequency of a particular policy based on losses encountered for them. Traditionally actuaries used advanced math and financial theory to analyze and understand the cost of risks and other such cost matrix. Indeed, the analytics performed in such traditional ways are critically important to any insurance company. We propose to analyze the insurance data using more modern techniques such as Big Data analysis.

Traditional data analysis techniques focus mainly on numerical and experience based data. The results are presented in the form of tables and rules. We intend to use Database query techniques to mine the knowledge and present it in pictorial form such as pie charts bar graphs etc. Big Data analysis is both efficient and fast. It does not require deep domain related knowledge. Presenting results in the form of graphs and pie charts helps in understanding the findings more efficiently.

1. **INTRODUCTION**

We study here the problem of secure mining and analysis of relational databases. In that setting, there are several tools (or applications) that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities. The goal is to find valid knowledge or rules based on the data available in relational databases while protecting any personal or private information. The information that we would like to protect in this context is not only individual transactions in the different databases, but also more global information such as what association rules are supported locally in each of those databases.

1. **EXISTING SYSTEM**

**3.1. Big Data**

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn’t refer to any specific quantity, the term is often used when speaking about peta bytes and Exabyte’s of data.

Big data is used to describe a massive volume of data that is so large that it’s difficult to process. The data is too big that exceeds current processing capacity.

Big data can be characterized by 3Vs: the extreme volume of data, the wide variety in types of data and the velocity at which the data must be must processed.

**3.2. DISADVANTAGES OF EXISTING SYSTEM**

**3.2.1. SECURITY CONCERNS**

Managing a complex application such as Hadoop can be challenging.

**3.2.2. VULNERABLE BY NATURE**

The framework is written almost entirely in Java, one of the most widely used yet controversial programming languages in existence.[**Java has been heavily exploited by cybercriminals**](http://www.darkreading.com/attacks-and-breaches/java-under-attack-again-disable-now/d/d-id/1108136?)and as a result, implicated in numerous security breaches.

**3.2.3. NOT FIT FOR SMALL DATA**

Due to its high capacity design, the Hadoop Distributed File System or HDFS lacks the ability to efficiently support the random reading of small files.

**3.2.4. POTENTIAL STABILITY ISSUES**

Hadoop is an open source platform. That essentially means it is created by the contributions of the many developers who continue to work on the project. While improvements are constantly being made like all open source software, Hadoop has had its fair share of stability issues**.**

1. **PROPOSED SYSTEM:**

Since its inception, Hadoop has become one of the most talked about technologies. Why? One of the top reasons (and why it was invented) is its ability to handle huge amounts of data – any kind of data – quickly. With volumes and varieties of data growing each day, especially from social media and automated sensors, that’s a key consideration for most organizations.

1. **CONCLUSION**

In this project we successfully took a publically available dataset from a source and performed big data analysis on it. We removed any anomalies that were present in the dataset and after thoroughly cleaning the data we loaded it into hdfs (Hadoop distributed file system).

We finally used query language to query on the data and get rich knowledge from the database. This knowledge was then graphically and pictorially represented in the form of pi charts, bar graphs etc. The result so obtained will help to take valid decisions in terms of opening new hospitals, closing existing ones, enhancing over crowded hospitals etc.