DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4105

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Big Data Analysis And Deterministic Encryption Challenges

Aasim Shafi
Student, CSE (Computer Science Engineering)
SSM College of Engineering & Technology
J&k, India

Sahila Fareed Shah
Assistant Prof. , CSE (Computer Science Engineering)
SSM College of Engineering & Technology
J&k, India

Abstract: with the rapid change in the technology and innovation in the last decade, big data analysis has gone for an exponential and tremendous growth and will most probably continue to perform spectacular developments due to the emergence of innovative trends and new interactive multimedia applications and the use of highly integrated systems driven by the rapid growth in information services and microelectronic devices. Up to yet, large no. of the existing mobile systems is mainly targeted to voice communications with low transmission rates. Big-Data has always been a part of our lives knowingly or unknowingly [7]. This is a particular review on known big-data systems that contain a set of tools and technique to load, extract, and process and improve dissimilar datasets while leveraging the immensely and most parallel processing power to perform the idiosyncratic transformations and analysis. Big-Data" technology faces a list of technical challenges.

Keywords: Big-Data, Structured data, Un-Structured data, Random Encryption, Deterministic Encryption

I. INTRODUCTION

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guess work or on painstakingly constructed models of reality can now are made based on the data itself. Such Big Data analysis now drives nearly almost every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences and physical sciences [2].

While the potential benefits of Big-Data are real and significant and some initial success have already been achieved, there are some technical challenges that must be addressed to fully realize this potential.

Data becomes 'Big-Data' when the individual datum stops mattering, and only aggregates or derived analysis matter. For example, sensors generate lots of data, but we are only interested in the detecting and recording their irregularities. This data which is collected by the sensors are then analysed by the Big-Data algorithms to show the relationship between various datasets. Big-Data is a buzz-word, or catch-phrase, used to describe a massive volume of both structured and un-structured data that is so large that it is difficult to process using traditional database and software techniques. "Big-Data" system faces a series of technical challenges, including: First, due to the large variety of different data sources and the huge volume, it is too difficult to collect, integrate and analysis of "Big Data" with scalability from scattered locations. Second "Big Data" systems need to manage, store and integrate the gathered large and varied verity of datasets, while provide function and performance assurance, in terms of fast retrieval, scalability and secrecy protection.

Third "Big Data" analytics must effectively excavation large datasets at different levels in real time or near real time - including modelling, visualization, prediction and optimization - such that inherent potentials can be revealed to improve decision making and acquire further advantages. To address these challenges, the researcher IT industry and community has given various solutions for "Big Data" science systems in an ad-hoc manner. Cloud computing can be called as the substructure layer for "Big Data" systems to meet certain substructure requirements, such as cost-effectiveness, resistance, and the ability to scale up or down. Distributedfile systems and No SQL databases are suitable for persistent storage and the management of massive scheme free datasets. Map Reduce, R is a programming framework, has achieved tremendous potential in manipulating "Big Data" group-aggregation tasks, such as web application ranking.

Hadoop integrates data storage, data processing, system management, and other modules to form a powerful system-level solution, which is becoming the mainstay in handling "Big Data" challenges. We can build various "Big Data" application system based on these innovative technologies and platforms. In light of the of big-data technologies, a systematic frame work should be in order to capture the fast evolution of big-data research [2].

II. BIG DATA CHALLENGES

Each and every technology comes with some challenges that it had to overcome, so is the case with the Big-Data Analysis .These are some of the challenges before Big-Data below-

A. How to keep data private & allowing querying over it [2].

Private data problem or confidential data leaks is the main concern of the Big-Data .As most of the Big-Data companies involves the use of cloud based architecture, there are lots of concerns about the safety and confidentiality of the data. In 2014 hackers compromised 70 million credit cards numbers and pins from target.

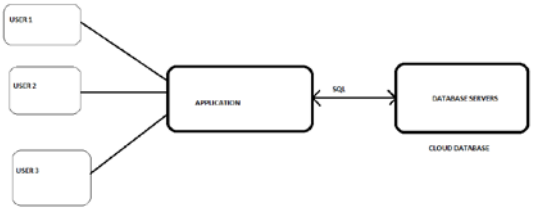

Fig. 1 A Typical cloud database server

These databases are storing sensitive data on the cloud servers and we have to prevent this data from getting leaked to the hackers. The intruders can be malicious system administrators or hackers. The data confidentiality can be protected as shown below-

1) Encrypt data –

Even if the malicious attackers get the access to the data, they won't be able to interpret it .By encryption we mean changing the original form of the data to some another random symbols which cannot have any specified meaning as such. This can be done by using MONOMI / CRYPTDB

1) Process sql queries on encrypted data. This involves hiding database from system administrator, outsource database to the cloud.

2) Modest Overhead.

3) No changes to DBMS (e.g., Postgres, Mysql) and no changes to applications.

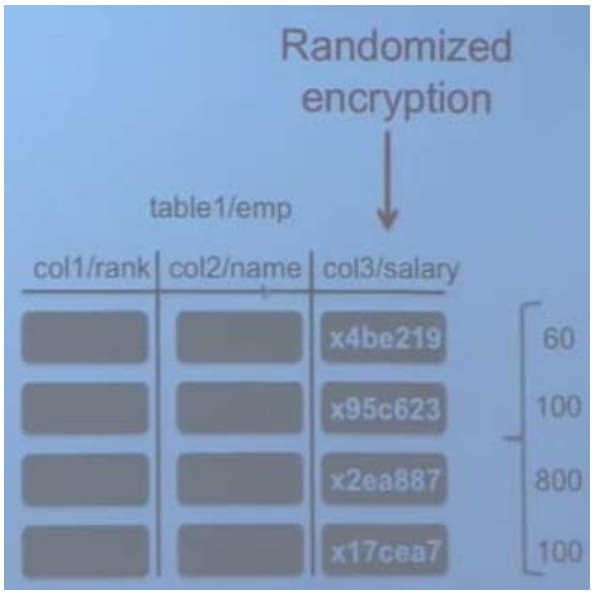

Fig. 2 sql queries on encrypted data

The standardized random encryption maps every value to random string from the point of view of somebody completely outside. But somebody who has access to the key, can look at one of those random values and determine what the actual value is. It is very hard to break randomized encrypted data, if one does not have the access to the key. It uses complex mathematical approach.

Each occurrence of the same value (say 100) maps to the different encrypted data. But the problem with this scheme is that we could not do the sql queries like

Select * from emp
Where salary =100

Because every occurence of 100 in the table is mapped to a different encrypted data in the database.so we could not get the columns.So,we have reduced the encryption to Deterministic encryption technique in which the same value (say 100) could be pointed to the same encrypted data. In this way, we can get the data of our choice . This can be extended in the same way to all operations of sql queries and likewise the data is retrieved.

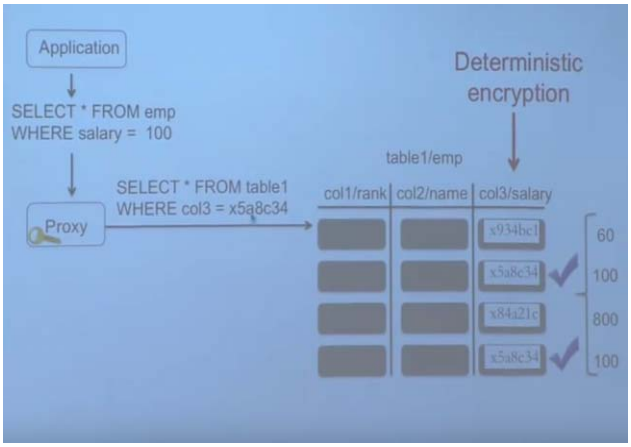

Fig. 3 sql query on Deterministic Encryption

B. Data Access and Sharing of Information [3].

If the data in the company's information systems is to be used to make accurate decisions in time, it becomes necessary that it should be available in accurate, complete and timely manner.

C. Analytical challenges .

Big-Data brings with it some huge analytical challenges. The type of analysis is to be done on this huge amount of data which can be unstructured, semi-structured or structured requires a large number of advance skills. Moreover the type of analysis which is needed to be done on the data depends highly on the results to be obtained i.e., decision making. This can be done by using the two techniques either incorporates massive data volumes in analysis or determine upfront which Big-Data is relevant.

D. Interoperability.

Integration of Big-Data technologies with existing enterprise solutions .Data ingestion, Data modelling, Data visualization using existing tools. Most companies should standardize the tools and should keep with the technology until fruitful results are not obtained because it is a time consuming process .

E. Work on Reliable Data.

With the burst in the volume of the present data, the challenge is how to bifurcate the signal of data and valuable information. But a lot of companies are facing difficulties in identifying the correct data and determining how they can optimally use it. Results produced by the Big-Data should be consistent because they are based on the predictive models of analysis and not on explanation. Sources of data are reliable whether it may belong to the social network, banking and sensor related industry .

III. BIG DATA ANALYSIS

The most crucial part of the phases in the "Big Data" value chain is to do the analytics part i.e., data analysis, the objective of it is to get beneficial data, suggest optimal plans for industry and to support decision-making system of the company to stay intact with growing competition in the market.

Descriptive Analytics: uses the previous data to explain what occurred in past years of analytics. For instance, a mathematical regression analysis technique may be used to find the simple trends in the datasheets, visualization patterns presents information in a meaningful and easy way, and information modelling is used to collect, store, process and cut the data in most efficient way. Descriptive analysis is particularly related with business intelligence or visibility systems.

Predictive Analytics: it concentrates on the inference of future known probabilities and trends. For example, predictive modelling uses statistical techniques such as line static and mathematical logistic regression to know trends and calculate coming times outcomes, and data mining techniques extracts patterns to provide careful insight and efficient forecasts. Prescriptive Analytics: is another tool which addresses the sound decision making capability to the industry and efficiency. For example, simulation tool is used to analyse intricate systems to have views into system latency and earmark issues and optimization tool are used to find best solutions under given constraints [7].

IV. BIG DATA CLASSIFICATION ALGORITHM

1) Decision Tree
2) Random Forest
3) Support Vector Machine
4) Chandelier Decision Tree

Decision tree guiding uses a rule based tree as a predictive tool that points observations about an item to achieve the conclusions about the item's target value. It is one of the predictive analytical approaches which are employed in the fields of statistics, data extraction and robot learning. Tree models in which the target variable can have a discrete set of values are called classification trees. In these tree structures, leaf nodes represent class labels and branch nodes represent conjunctions of characteristics that lead to those class labels. Decision trees where the goal variable can take analogue values are called regression trees. In decision analysis, a decision tree can be used to visually and elaborately represent decisions and decision making. In data mining, a plan chart tree explains information flow but not decisions; rather the verifying classification tree can be a record for decision making [7].

Random forest is a supervised learning technique and, as the name suggests, it forms forest-like structures with decision trees that are generated using the random sampling with replacement. These decision trees may either be the classification trees or the regression trees; therefore, the random forest can be applied to both classification problems and regression problems. The advantage of the random forest is that it provides multiple trained decision tree classifiers for the testing phase. This property of the random forest supervised learning technique makes the random forest a preferred technique over regular decision tree learning [6].

Support vector machine provides a classification learning model and an algorithm rather than a regression model and an algorithm. It uses the simple mathematical model $y = wx+ \gamma$, and manipulates it to allow linear domain division. The support vector machine can be divided into linear and nonlinear models. It is called linear support vector machine if the data domain can be divided linearly to separate the classes in the original domain. If the data domain cannot be divided linearly, and if it can be transformed to a space called the feature space where the data domain can be divided linearly to separate the classes, then it is called nonlinear support vector machine. Therefore, the steps in the linear support vector machine are the mapping of the data domain into a response set and the dividing of the data domain. The steps in the nonlinear support vector machines are: the mapping of the data domain to a feature space using a kernel function, the mapping of the feature space domain into the response set, and then the dividing of the data domain. Hence, mathematically, we can say that the modelling of a linear support vector machine adopts the linear equation $y=wx+ \gamma$, and the modelling of a nonlinear support vector machine adopts the nonlinear equation $y =w\varphi(x) +\gamma$. The kernel function makes it non-linear. The classification technique using a support vector machine includes the parameterization and optimization objectives. These objectives mainly depend on the topological class structure on the data domain. That is, the classes may be linearly separable or linearly non separable. However, linearly separable classes may be separable. Therefore, the parameterization and optimization objectives that focus on the data domain must take these class properties into consideration.

Nonlinear Support Vector Machine is that in which the scatter plots play a major role in classification by facilitating the domain divisions. In the scatter plots, the dimension (i.e., each axis) is defined by a feature, and the space defined by the feature is called the vector space. The scatter plot describes the relationship between the features, and thus the correlated and uncorrelated data points can be identified in the vector space. The classification (in other words, the domain division) may be carried out either in a vector space or in a feature space, where the vector space is defined as the space that contains the scatter plot of the original features, and the feature space is defined as the space that contains the scatter plot of the transformed features using kernel functions [4].

Chandelier Decision Tree The unit circle algorithm (UCA) proposed in a recent paper is one of the important contributors to the chandelier decision tree and random chandelier techniques. The main concept of the UCA is the transformation of a given data domain to a circular data domain (or hypersphere) and the execution of domain division on that circular domain for classification. To achieve these objectives, the data have been represented in unit circles with two regions called the inner circle and the outer circle and then used to classify two classes. The chandelier decision tree is the machine learning technique that integrates the URM approach and the concepts used in the decision tree. It divides the circular data domain into a tree of circular sub domains, where a node has the ring with data that gives two branches (sub domains) that have maximum information gain [1].

V. EXAMPLES AND ADVANTAGES

There are large no. of fields in which the Big-Data is used. These applications are in the IT-Industry as well as in the other aspects of human information patterns.

- Credit Card Transactions – 10000/sec in the banking

Sector can be used in the aggregate to predict the change in the financial exchanges.

- Wal-Mart which is the largest discount store in America uses the Big-Data technologies to predict the buying behaviour of their potential customers and by this it is estimated around 1 million customer transactions/ second.
- RFID Systems.
- Big-Data are used by the companies to analyse the temperature in the industrial plants by means of the sensors and the data collected is send to the cloud databases where the Big-Data algorithms is then applied on them to get the predictions about various plans to be followed by companies . For example in the oil plants.
- It is used by the video streaming sites like Netflix to get the watching patterns of their customers .In this way they produce the successful series with the perfect combination of actors ,directors etc.
- Big-Data of traffic is being analysed to develop a car that can drive completely by itself accident free.
- In the Future we can get the Big-Data of DNA , in this way curing the genetic diseases like cancer would become much easier[2].

ADVANTAGES OF BIG-DATA

- Making our cities smarter.
- Oslo, Norway reduced street light consumption by 62%.
- Memphis police department, USA reduced serious crime by 30% since 2006.
- Portland, USA optimized timing of traffic signals & eliminated CO2 emissions.

VI. Conclusions

Training in Big-Data helped us to know what the crazy trend in IT-Industries and how technology is becoming more fruitful to human development. Big-Data is a future. Currently a lot of research is going on in this industry. As data is expanding at an alarming rate thus there is a immense need of such tools and technology which can handle it. Hadoop is the most emerging framework used by most of big-firms like facebook, Microsoft, Amazon and many more. The Big-Data is the technology on which most of the multinational companies are relying for their activities. To gain the prominence in the world of innovation and technology, every industry should reap the benefits of Big-Data. It not only helps in growing the industry but also increase in profits.

. VII. REFERENCES

[1] Shan Suthaharan, Machine Learning Models and Algorithms for Big Data Classification , Integrated Series in Information Systems 36 , Series Editors: Ramesh Sharda · Stefan Voß ,Springer.

[2] J. Breckling,MIT Center for Transportation and logistics.

[3] Du Zhang," Inconsistencies in Big Data", IEEE 978-1- 4799-0783-0/13, PP 61-67.

[4] Swag tam Das, Ajith Abraham, Senior Member, IEEE, and Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE 2008, PP 218-237.

[5] Carson Kai-Sang Leung, Richard Kyle MacKinnon, Fan Jiang, "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data", IEEE 2014, PP 315-322.

[6] Vrushali Y Kulkarni," Random Forest Classifiers: Survey and Future Research Directions", International Journal of Advanced Computing, ISSN: 2051-0845, Vol.36, Issue.1, and April 2013.

[7] http://ijcsmc.com/docs/papers/june2017/v616201710.pdf Archives review papers.