IJRRECS/November 2015/Volume-3/Issue-11/4782-4791 **ISSN 2321-5461**

INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE

BIGDATA PROCESSING - BUILDING COHERENT SCHEME FOR RECOGNIZING QUALITY INFORMATION

Dr.B.Vijayakumar

Professor, Computer Science & Engineering

Malla Reddy Engineering College for Women, Maisammaguda, Secunderabad

drvijaybel@gmail.com

ABSTRACT:

In the modern times, Big Data processing relies on models of parallel programming and offers public, a platform of cloud computing concerning Big Data services. As a consequence of dynamic features of application data concerned in a distributed setting, one of most momentous characteristics relating to Big Data is to perform computing by means of complex computing procedure. Big Data originate with huge-volume independent sources with dispersed control, and seeks to discover complex relationships among data. These characteristics make it a remarkable challenge for finding out practical knowledge from them. To carry out mining of Big Data, holding an effectual data access method is extremely important, for users who utilizes a third party to practice their data. Mining of Big Data in general provides benefits to go ahead of recognized relational databases to depend on less structured data which are mined for constructive information. This paper uses HACE theorem that explains the features of Big Data revolution, and present a model of Big Data processing, from the viewpoint of Bigdata mining through regression algorithms.

Keywords: Big Data, Cloud computing, HACE theorem, Data mining, Database, regression algorithms.

1. INTRODUCTION:

To handle the challenges concerning Big Data and grab the opportunities provided by novel data focused resolution, US National Science Foundation has announced Big Data solicitation. This activity has resulted in quite a lot of appealing projects to examine the Big Data management for extensive scientific literatures and other works [2]. These projects tried to discover methods, and frameworks, that permit us to convey huge amounts of data down towards a human manageable as well as interpretable scale. Because of enormous and active features of application data concerned in a distributed setting, one of most significant characteristics concerning Big Data is to perform computing by means of complex computing process. Thus, utilizing a parallel computing infrastructure and software models resourcefully extracting the dispersed data are the critical goals for processing of Big Data to modify from quantity towards quality. Generally algorithms of data mining typically need to scan all the way through training data for finding the statistics to solve model parameters [7]. For applications connecting Big Data and incredible data volumes, it is general the case that data are physically dispersed at several locations, which means that users no more hold their data storage. To perform mining of Big Data, having an effectual data access mechanism is very important, in particular for users who aim to employ a third party to practice their data. A Public key-based mechanism is employed to allow third-party auditing so users can securely permit a third party to analyze their information without breaching the security settings. As Big Data are stored at numerous locations and data volumes may constantly increase, an effective computing proposal should get hold of distributed data storage for computing [6]. This paper uses HACE theorem [1] that describe the features of Big Data revolution, and present a model of Big Data processing, from the viewpoint of data mining.

2. BIG DATA AND HACE THEOREM

For most of the applications of Big Data, privacy concerns spotlight on excluding the third party from directly accessing original information. General solutions are to depend on a number of privacy-preserving mechanisms to defend the data. To adjust to enormous, active Big Data, researchers have prolonged existing methods of data mining in numerous ways, comprise efficiency

enhancement of single-source methods of knowledge discovery scheming a data mining method from a multisource viewpoint and examination of stream data. The most important motivation for finding out knowledge from huge data is improving efficiency of single-source mining methods. On regular enhancement of computer hardware functions, researchers carry on exploring ways to get better the effectiveness of knowledge discovery algorithms to make them improved for huge data. As massive data are in general collected from several data sources, knowledge discovery of enormous data must be performed by means of a multisource mining mechanism. Thus, enormous, various and instantaneous features of multisource data offer necessary differences among single-source knowledge discovery as well as multisource data mining [4]. In numerous situations, knowledge extraction procedure has to be extremely efficient and secure to real-time as storing observed data is almost infeasible. Big Data initiates with huge-volume, independent sources with dispersed control, and seeks to discover complex relationships between data. These features make it a tremendous challenge for learning constructive knowledge from Big Data. One of basic features of Big Data is enormous volume of data represented by varied as well as diverse dimensionalities. Since different data collectors choose their own protocols for recording, nature of several applications moreover results in varied data representations. Independent data sources data with distributed controls are a major trait concerning applications of Big Data. Being self-sufficient, every data source is able to produce and gather information without concerning any centralized control [5]. While volume of Big Data enhances, so do complexity and relationships underneath data.

The HACE theorem[8] present that important features of Big Data are enormous with heterogeneous and different data sources; independent with distributed control; complex and progress in data associations. These features imply that Big Data need a big mind to strengthen data for utmost values. HACE theorem is theorem to replica of characteristics Big Data. It starts with Heterogeneous & large-volume, distributed Autonomous resources with and de-centralized control, and seeks to consider Complex and Evolving associations among data. An example of this theorem is demonstrated in Figure 1. The goal of each

blind man is to extract conclusion based on the part of information he collects by touching the elephant. Each blind man concludes independently that the elephant "feels like" a wall, a snake, a mat, a tree or a rope depending on the part of their touch limited to. To mark the problem even more complex, we may accept that the elephant is increasing quickly and its posture varies continually, and each blind man may have his own information sources that tell him about subjective knowledge about the elephant, e.g., one blind man may converse his feeling about the elephant with another blind inherently subjected. Exploring the Big Data in this scenario is equivalent to form various information from different sources (blind men) to help to draw a best possible design to uncover the actual sign of the elephant in a definite way. HACE theorem proposes the following key characteristics: i) Huge with various and mixed data sources. ii) Autonomous sources with circulated & separate Control. iii) Complex and Evolving associations.

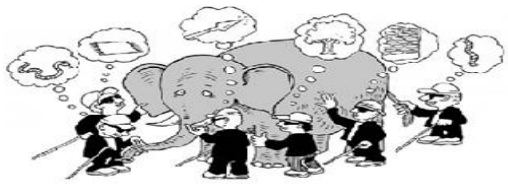

Fig. 1: Five blind men and huge elephant: The limited view of each blind man leads to biased conclusion.

3. BIG DATA PROCESSING STRUCTURE

In recent times, processing of Big Data mostly depends on models of parallel programming and offers a cloud computing proposal of Big Data services for public. For a system of intelligent learning database to hold Big Data, the necessary key is to increase to exceptionally huge volume of data and make available treatments for the characteristics. The overview of system of Big Data processing including three tiers was shown in fig1.3. They are data accessing represents Tier I, data privacy as well as domains represents Tier II, as well as regression algorithms of data mining as shown in Tier III. This paper uses regression algorithms in Tier III.

The task of regression commence with a data set in that the target values are known. For instance, a regression model that predicts house-values might be confined

based on observed data for many houses over a time period. Additionally, the data might track the life time of the house, square-footage, number of rooms, nearness of market and shopping centers, taxes, and so on. Here, House-value is the target, the other attributes are predictors, and the data for each house will compose a special case. In the training process, the value of the target is estimated by the regression algorithm as a function of predictors for all the cases in the build data. This is actual model build process. The relationships among are summarized predictors and target in a model, that can then be applied to a different data set in which the target values are unknown. Regression models can be tested by computing various statistics that compute the variation between the predicted values and the expected values. Regression modeling may be applied in many areas i.e. in business planning, trend analysis, marketing, financial forecasting, time series prediction, environmental modelling, biomedical and drug response modeling, and so on.

4. PROPOSED ALGORITHMS FOR BIG DATA MINING AND METHODOLOGY

Regression analysis is used to determine the values of parameters for a function that cause the function to best-fit a set of data remarks that we provide. The following equation (1) state these relationships in symbols. It shows that regression is the process of estimating the value of a continuous target, **Y**, as a function **f** of one or more predictors (x1, x2 ... xn), a set of parameters (θ1, θ2, ..., θn), and a count of error is **e**.

$$Y = f(x, \theta) + e \quad \text{.....(1)}$$

The predictors are independent variables and the target is a dependent variable. The error is residual that is the variation between the expected and predicted value of the dependent variable. The regression parameters are also known as regression coefficients. The training process of regression model involves deciding the parameter values that decrease the error. There are various regression functions and ways to measure the error.

Proposed techniques are Linear, Multivariate-Linear and Non-linear regressions.

A linear regression method is used if the link between the predictors and the target can be estimated with a straight line. Linear regression with a single predictor is

expressed through the following equation(2).

$Y = \theta 2.x + \theta 1 + e$ .....(2)

Here, θ2 (the slope of the line) is the angle between a data point and the regression line. And θ1 (Y intercept) is the point where x crosses the y axis (x = 0).

The Multivariate-linear regression consign to linear regression through two or more predictors (x1, x2, … xn). If many predictors are used, then the regression line cannot be seen in 2-dimensional space. Though, the line can be calculated simply by escalating the equation for single-predictor linear regression to comprise the factors per the predictor.

$Y = \theta 1 + \theta 2x1 + \theta 3x2 + ... \theta n\ xn-1 + e$ ...(3)

In Multivariate-linear regression, the parameters are repeatedly consigned to as coefficients. To model multivariate-linier regression the algorithm computes a parameter for each of the predictors used by the model. This parameter is a gauge of the impact of the predictor **x** on the target **Y**. One can analyze the regression parameters to estimate how fine data fits the regression line.

Frequently the relationship between **x** and **Y** can't be estimated with a straight line. For this, a nonlinear regression method may be used. Otherwise, the data could be pre-processed to construct the linear relationship.

In Nonlinear regression model, **Y** is a function of **x**, that is more complex than the linear regression equation. The figure1.1 shows Linear Regression with a Single Predictor and the figure1.2 shows and y have a nonlinear relationship.

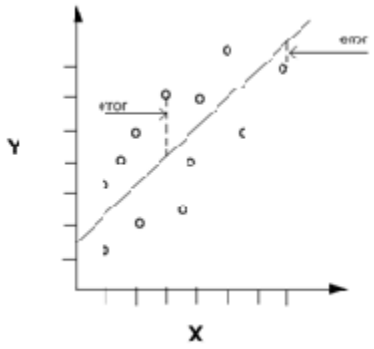

Figure 1.1 Linear regression with a Single Predictor

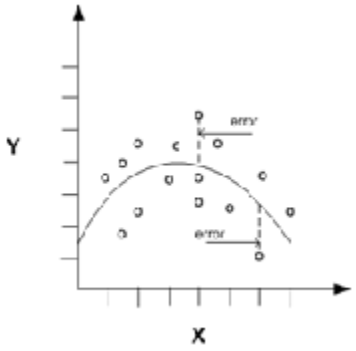

Figure 1.2 Non-linear regression with a Single Predictor

Testing with quality parameters

To test a regression model, apply the test data with known target-values and compare with predicted-values with the known

values. The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model. Test metrics are used to assess how accurately the model predicts these known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

The RMSE (Root-Mean-Squared-Error) and the MAE (Mean-Absolute-Error) are used as quality parameters for estimating the overall quality of a regression model.

The value of RMSE is the square root of the average squared distance of a data point from the fitted line. Following SQL expression and equation (3) computes the RMSE value.

SQRT (AVG ((predicted_value - actual_value) * (predicted_value - actual_value)))

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad \text{.....(4)}$$

In this equation (4) **Σ** represents summation; **j** represents the current predictor, and **n** represents the number of predictors.

The value of MAE is the average of the absolute value of the residuals. The MAE is very similar to the RMSE but is less sensitive to large errors.

Following SQL expression and equation(4) calculates the MAE.

AVG(ABS(predicted_value – actual_value))

$$\mathbf{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \quad \text{.....(5)}$$

In this equation(5) **Σ** represents summation; **j** represents the current predictor, and **n** represents the number of predictors.

This paper proposes GLM(Generalized Linear Models) algorithm and SVM(Support Vector Machines) algorithm for Big Data mining with data sets that have very high dimensionality (many attributes), including transactional and unstructured data.

GLM-Generalized Linear Model:

GLM is a popular arithmetical technique for linear modelling to apply in Big Data Mining for regression and also for binary classification. GLM provides broad coefficient-info and model statistics, and row diagnostics. It also chains confidence bounds.

SVM-Support Vector Machines:

SVM is a powerful technique for linear and nonlinear regression to apply in Big Data Mining for regression,

classification, and anomaly detection. SVM regression uses two kernels i.e. the linear kernel for linear regression and Gaussian kernel for non-linear regression.

In distinctive systems of data mining, the mining procedures necessitate demanding computing units for data analysis. For Big Data mining, since data scale is far away from capacity that a single personal computer can hold, representative Big Data processing structure will depend on cluster computers by means of computing scheme, with data mining task being organized by running a number of parallel programming tools. Big Data mining provides benefits to go ahead of established relational databases to depend on less structured data which are mined for constructive information. The challenges focused at Big Data mining platform spotlights on computing methods of data accessing. Since Big Data are regularly stored at several locations and data volumes might always grow, an effectual computing proposal will have to obtain dispersed important data storage for computing [6]. The challenges at Big Data semantics as well as application knowledge for several applications of Big Data provides extra benefits to the procedure of mining, in addition to technical barriers to Big Data access as well as mining algorithms. Challenges of data mining focus on Regression algorithms for Big Data mining at Tier III and managing of difficulties that are raised by volumes of Algorithms of Big Data mining typically need to scan all the way through training data for finding the statistics to solve model parameters. Semantics as well as application knowledge in Big Data refer to several aspects connected to policies, user knowledge, as well as domain information. It includes sharing of privacy and data; application knowledge. While applications of Big Data are marked with independent sources, combining of distributed data sources towards a centralized site for mining process is steadily prohibitive due to possible transmission cost as well as privacy concerns. Mining system of Big Data has to facilitate an information exchange to make sure that the entire distributed sites can act as a team to attain a global optimization objective.

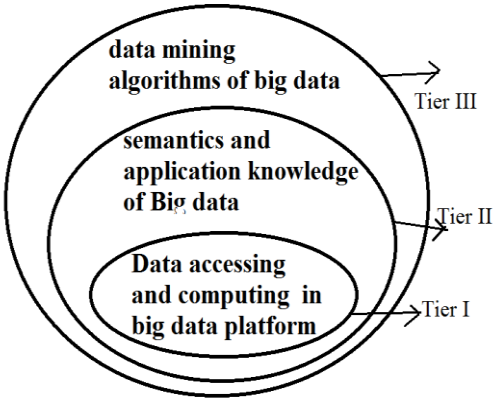

Fig 1.3: An overview of system of Big Data processing.

4. CONCLUSION

Fundamental features of Big Data are enormous volume of data represented by varied as well as diverse capacity. Autonomous data sources with distributed controls are the most important characteristics concerning applications of Big Data. Being independent, each data source produces and gathers data without concerning any centralized control. The challenges that are focussed at Big Data mining platform spotlight on computing methods of data accessing. While Big Data are repeatedly stored at quite a lot of locations and data volumes might always grow, an effectual computing proposal will have to obtain dispersed important data storage for computing. A HACE theorem is used to describe the features of Big Data revolution, and put forward a model of Big Data processing, from the perspective of Big Data mining. For the proposal of intelligent learning database to hold Big Data, the necessary key is to increase to exceptionally huge volume of data and make available treatments for the characteristics. In the model of proposed system, data accessing represents Tier I, data privacy as well as domains represents Tier II, as well as regression algorithms for Big Data mining as shown in Tier III. The introduced HACE theorem presents important features of Big Data such as: enormous with heterogeneous and different data sources; independent with distributed control; complex and progress in data associations which mean that Big Data need a big mind to strengthen data for extreme values.

REFERENCES

[1] Wu, X., Kumar, V., Quinlan J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14, 1, pp. 1-37.

[2] R. Chen, K. Sivakumar, H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data",

Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[3] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array", Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[4] G. Cormode, D. Srivastava, "Anonymized Data: Generation, Models, Usage", Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[5] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, "Ricardo: Integrating R and Hadoop", Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998, 2010.

[6] C.T. Chu et al., "Map-Reduce for Machine Learning on Multicore", Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.

[7] Y.C.Chen, W.C.Peng, and S.Y.Lee, "Efficient Algorithms for Influence Maximization in Social Networks", Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[8] Deepak S. Tamhane, Sultana N. Sayyad, "Big Data Analysis Using Hace Theorem", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 1, January 2015.

AUTHOR:

Dr.B.Vijayakumar, working as Professor in Computer Science & Engineering in Malla Reddy Group of Institutions. He is having 22 years of total experience including both teaching and industry. His areas of interest in research are: Digital Image Processing, Big Data, Data Mining and Parallel Computing, Steganography, Network Security and App. Development.