

A Comparative Study about Improving Accuracy in Big Data

K.Krishnaveni, Dr.S.Siva Ranjani

Assistant Professor¹, Professor² Department of Computer Science and Engineering

Sethu Institute of Technology, Kariappati, Virudhunagar – 626 106

Corresponding author's email id: whittykrish@gmail.com, sivaranjani222@gmail.com

Abstract

Big data is one of the most popular and most significant used terms to describe the epidemic growth and availability of data in the modern age. It describes any multitudinous amount of both structured and unstructured data that has the potential to be mined for information. Due to the voluminous amount of data, it becomes more complex to perform in effective manner. Nowadays, it is not provide accurate result. If data collection (or) analysis may not have enough accuracy, then it will critically affect the decision based on an analytics. This paper provides a survey of recent work on accuracy approaches. In particular, the paper considers Classification techniques, Filtering techniques, Prediction method, Mining method and Estimation Method. Our results discussed the comparison of improving accuracy and different level of accuracy.

Keywords: Big Data Analytics, Accuracy, K- Means++, CSO, Ranking Approach, Load Balancing Technique;

INTRODUCTION

Definition and Features of Big Data

Big data is an intellectual concept. Apart from masses of data, it also has some other features, which determine the difference between itself and “massive data” or “very

big data.” At present, the importance of big data has been recognized in general, people still have different opinions on its definition. In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by

traditional IT and software/hardware tools within a tolerable time. For the definition of the Big Data, there are various different explanations of 3Vs. In the “3Vs” model, Volume means, with the generation and collection of masses of data, data scale becomes increasingly big; Velocity means the timeliness of big data, specifically, data collection and analysis etc., must be rapidly and timely conducted, so as to maximally utilize the commercial value of big data; Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data. More generally, a data set can be called Big Data if it is formidable to perform capture, duration, analysis and visualization on it at the current technologies.

Importance of Big Data

Big data is unlike from the data being stored in traditional warehouses. The data stored there first needs to be cleansed, documented and even trusted. Moreover it should fit the basic structure of that warehouse to be stored but this is not the case with Big data it not only handles the data being stored in traditional warehouses but also the data not suitable to be stored in those warehouses. Thus there comes the point of access to preserve the data and

better business strategies and decisions as analysis of more data is always better.

A. Log Storage in IT Industries

IT industries store large amount of data as logs to deal with the problems which seem to be occurring hardly in order to solve them. But the storage of this data is done for few weeks or so though these logs need to be stored for longer duration because of their value. The Traditional Systems are not able to handle these logs because of their volume, raw and semi structured nature. Moreover these logs go on changing with the s/w and h/w updates occurring. Big data analytics not only does analysis on the entire large data available to determine the point of failures but also would increase the durability of the log storage.

B. Sensor Data

Massive amount of sensor data is also a big challenge for big data. All the industries at present dealing with this large amount of data make use of small portion of it for analysis because of the lack of the storage infrastructure and the analysis techniques. Moreover sensor data is categorized by ysis techniques. Moreover sensor data is categorized by both data in motion and data at rest. Thus safety, profit and efficiency all require large amount of

data to be analyzed for better business insights.

C. Risk Analysis

It becomes important for financial institutions to model data in order to calculate the risk so that it falls under their acceptable thresholds. A lot amount of data is potentially underutilized and should be integrated within the model to determine the risk patterns more accurately.

D. Social Media

The most use of big data is for the social media and customer sentiments. Keeping an eye on what the customers are saying about their products helps business organizations to get a kind of customer feedback. This feedback is then used to modify decisions and get more value out of their business.

BIG DATA CHALLENGES

Big data may hold a lot of potential, but it can still be held back if the data being analyzed is inaccurate. Due to limitations on technology and other business deliberations, the analyses companies are getting back may not reflect what is really happening. If businesses want to ensure their big data intuitions get the desired results, they need to improve the accuracy

in their analytics efforts. In a perfect world, organizations would gather a vast amount of data, analyze it, and generate solutions to the problems they're facing. The truth is, as most know, we do not live in a perfect world. Insights from big data often have to be derived in a petite amount of time. The technology of a business has on hand might not be advanced enough to process so much information quickly. The Fig.1 [1] shows the continuously increasing Big Data. These restrictions lead many companies to performing big data analytics using sampling. In other words, they don't look at all of the data, but slightly analyze only smaller subsamples of information. Whereas this might be a go-to strategy for many businesses, the results have a greater chance of being inaccurate. Since it is vital for organizations to build accurate big data models, only looking at part of the data could lead to businesses forming the wrong conclusions. The key challenges are listed as follows: -

Data representation: Many datasets have definite levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility. Data representation aims to make the data more meaningful for computer analysis and user interpretation. However, an improper data

representation will reduce the value of the original data and may even impede effective data analysis. Efficient data representation shall reflect data structure, class, and type, as well as integrated technologies, so as to enable efficient operations on different datasets. –

Redundancy reduction and data compression: generally, there is a high level of redundancy in datasets. Redundancy reduction and data compression is effective to reduce the indirect cost of the entire system on the premise that the potential values of the data are not affected

Data life cycle management: compared with the relatively slow advances of storage systems, pervasive sensing and computing are generating data at unprecedented rates and scales. We are confronted with a lot of pressing challenges, one of which is that the current storage system could not support such massive data. In general, values hidden in big data depend on data freshness. Therefore, a data importance principle related to the analytical value should be developed to decide which data shall be stored and which data shall be precluded.

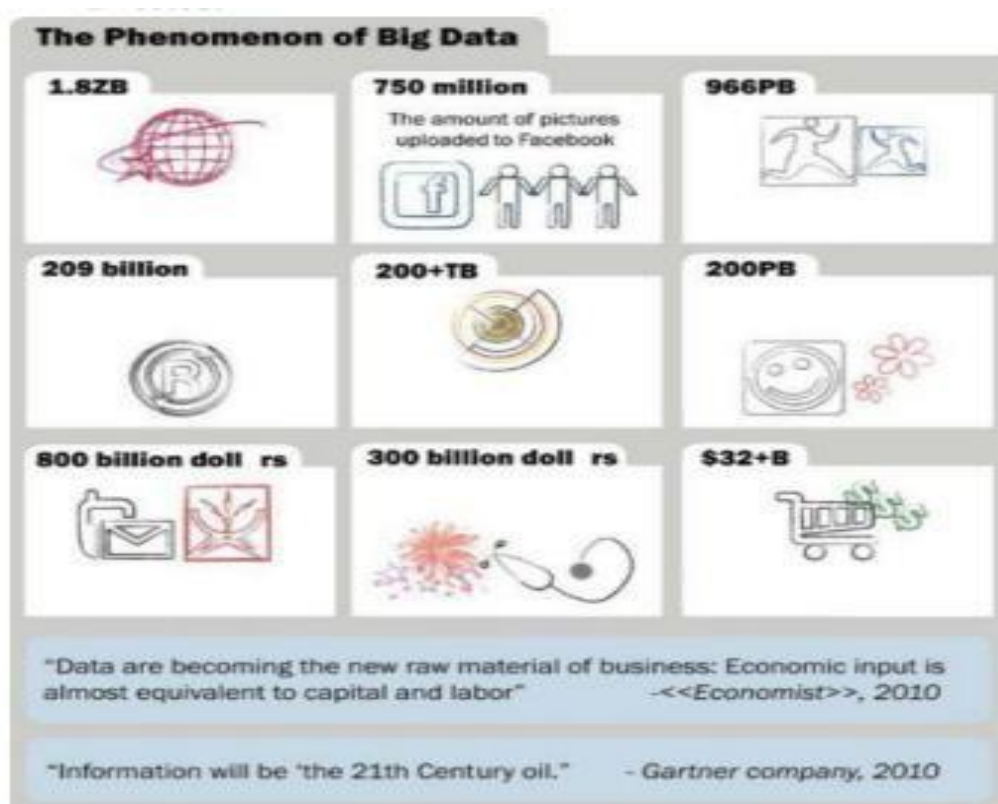


Fig.1 – The Continuously Increasing Big Data

Analytical mechanism

The analytical system of big data shall process masses of heterogeneous data within a limited time. However, traditional RDBMSs are strictly designed with a privation of scalability and expandability, which could not meet the performance requirements. Nonrelational databases have shown their unique advantages in the processing of unstructured data and started to become main stream in big data analysis. Even so, there are still some problems of non-relational databases in their performance and particular applications.

Data confidentiality

Most big data service providers or owners at present could not effectively preserve and analyze such immense datasets because of their inadequate capacity. They must depend on professionals or tools to analyze such data, which increase the potential safety risks. For example, the transactional dataset generally includes a set of widespread operating data to drive key business processes. Such data contains details of the lowest granularity and some sensitive information such as credit card numbers. Therefore, analysis of big data may be delivered to a third party for processing only when appropriate

preventive procedures are taken to protect such sensitive data, to ensure its safety.

Expendability and scalability

The analytical system of big data must sustenance present and future datasets. The analytical algorithm must be able to process progressively increasing and more complex datasets. – Cooperation: Analysis of big data is an interdisciplinary research, which requires experts in different fields cooperate to produce the potential of big data. A comprehensive big data network architecture must be recognized to help scientists and engineers in various fields access different kinds of data and fully exploit their proficiency, so as to collaborate to complete the analytical objectives.

BIG DATA APPLICATIONS

A. Public health:

1) analyzing disease patterns and tracking disease eruptions and transmission to improve public health investigation and speed response; 2) faster development of more accurately targeted inoculations, e.g., choosing the annual infection strains; and, 3) Turning large amounts of data into tortious information that can be used to identify needs, provide services, predict and prevent crises, especially for the benefit of populations.

B. Social Media

Social media can provide immediate news faster than traditional news channels or sources and can be a great wealth of information, but there is also an ever increasing need to verify and determine accuracy of this information. Nowadays, Social Media Application is the best display place for understanding the real-time customer choice or objectives and sentimentalities. By using such data, Social media business advertising and product marketing becomes easier.

C. IOT

IoT is not only an essential source of big data, but also one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT also develop interminably. Logistic enterprises may have greatly experienced with the application of IoT big data. The IoT is a significant concept of new technology; it has great potential to maintenance an ageing society, to improve the energy efficiency and to optimize all kinds of mobility and transport.

COMPARATIVE STUDY

METHOD I: *Distributed cluster analysis technique*

The incredible growth in data volumes has created a need for new tools and

algorithms to hastily analyze large datasets. Cluster analysis techniques, such as K-Means can be distributed across numerous machines. The accuracy of KMeans depends on the selection of seed centroids during initialization. K-Means++ improves on the K-Means seeder, but grieves from problems when it is applied to large datasets. Esteves[2], describes a new algorithm for Map Reduce implementation, developed for addresses of these problems and compared the performance with three existing algorithms and found that the algorithm improves cluster analysis accuracy and decreases variance.

K-Means

K-Means is a divider type of cluster analysis algorithm that attempts to solve the following clustering problem. Given an integer k , a distance measure d_m and a set of n data points in a d -dimensional space, the goal is to choose k centroids so as to minimize a cost function usually defined as the total distance between each point in the dataset and the closest centroid to that point.

El Agha and Ashour [3] discussed the K-Means algorithm is simple and forthright: first, it randomly selects k points from the whole dataset. These points represent the

initial centroids (or seeds). Each remaining point in the dataset is assigned to a cluster whose centroid is closest to that point. The coordinates of the centroid are then recalculated. The new coordinates of a specific centroid corresponds to the average of all points allotted to the respective cluster. This process iterates until a cost function converges to an optimum without a guarantee that it is the global one. Therefore, the initial selection of centroids during the initialization process to select the best possible set of centroids is essential. The accuracy of large dataset cluster analyses using K- Means depends on the accuracy of centroid initialization methods that are adapted to datasets distributed across several machines.

Pavan et al. [4] proposed K-Means++ is a stochastic algorithm, which means that the results produced by KMeans++ were extensively different across several analysis runs using the same initial conditions. K-Means++ was much better than K-Means when using a random selection of initial centroids. Second, K- Means++ is an integrally serial algorithm that is time-consuming for large datasets.

K-Means is a fast clustering method. However, initial seeding profoundly

impacts the cluster quality. A new strategy to parallelize K-Means++ that improves the speed and accuracy of cluster analysis for large datasets. Our new CK-Means is highly scalable and benefits from the use of Hadoop and Map Reduce. The new algorithm scales with the dimension of the dataset. The running time is more penetrating to variations in the number of data points and of clusters than to variations in the number of dimensions and addressed the problem of finding a good initial seeding in less time. Thus performing accurate cluster analysis over large datasets can now be done by using our new CK-Means approach.

Method II: *Feature Selection technique*

Feature selection is a type of optimization problem, generally achieved by combining an optimization algorithm with a classifier. Genetic algorithms and particle swarm optimization (PSO) are two frequently used optimal algorithms. Recently, cat swarm optimization (CSO) has been proposed and established to outperform PSO. However, CSO is limited by long computation times. Kuan [5], modify CSO to present an improved algorithm, ICSO. Then apply the ICSO algorithm to select features in a text classification experiment for big data. Results show that the proposed ICSO outperforms traditional

CSO. For big data classification, the results show that using term frequency- inverse document frequency (TF-IDF) with ICSO for feature selection is more accurate than using TF-IDF alone.

Cat swarm optimization algorithm (CSO)

CSO is an optimization algorithm that was proposed in 2007 Tsai PW [6]. It copycats the behavior of cats, which are excellent pursuers but which also display high levels of alertness even when at rest. This behavior can be described by two modes: seeking and tracing.

In the process of CSO, Hua Yin (2007) [7] discussed a number of parameters are required. N is the number of initial achievable solutions. Mixture Ratio (MR) is the ratio of cats in tracing mode. Seeking Memory Pool (SMP) is the number of candidate solutions in seeking mode. Self-Position Consideration (SPC) is a Boolean value; when SPC is true, a candidate solution will not be changed in seeking mode. Seeking Range of the selected Dimension (SRD) is the seeking range of selected position in seeking mode. Count of Dimension to Change (CDC) is the ratio or a count of dimensions to be changed in seeking mode. D is the number of dimensions of dataset features. C is a constant value in tracing mode.

Both the CSO and ICSO algorithms can reduce the large number of selected features for all datasets. Although, ICSO selected an equal or greater number of features, the ICSO algorithm generated higher classification accuracy than the CSO. Usually, the search algorithms find the same feature subset, the classification accuracy rates are the same. And tested the average accuracy of the two algorithms using Friedman two-way analysis of variance.

The confidence interval was set at 95%. And introduced two methods to improve the traditional CSO algorithm. Demonstrated that applying either of the methods individually can improve CSO and that ICSO overtakes CSO in terms of classification accuracy when applied to UCI datasets. However, ICSO also results in a greater number of selected features. In addition, ICSO to a real-world example of text classification to further demonstrate its effectiveness. Huang [8] applies ICSO to feature selection, it is likely that the proposed algorithm could benefit other optimization problems such as the Travelling Salesman Problem (TSP). In future work intend to develop methods which can expand the tracing mode of traditional CSO.

Method III: *Prediction Method in Smart Grids*

Smart grids are becoming popular with the initiation of sophisticated smart meters. They allow benefits to enhance energy consumption during peak hours by applying various demand response technique including voluntary curtailment, direct control and price incentives. To endure the curtailment over long periods of time of up to several hours utilities need to make fast and accurate ingestion predictions on a large set of customers based on a continuous flow of real time data and huge historical data sets. Given the numerous consumption patterns customers exhibit, different prediction methods need to be used to reduce the prediction error. The straightforward approach of testing each customer against every method is unfeasible in this large volume and high velocity environment.

To this aim, Marc [9] propose a neural network based approach for automatically selecting the best prediction method per customer by relying only on a small subset of customers and also introduce two historical averaging methods for consumption prediction that take advantage of the variability of the data and continuously update the results based on a sliding window technique. Its show that

once trained, the proposed neural network does not require frequent retraining, ensuring its applicability in online scenarios such as the defensible demand response.

Alfares [10], presented an efficient ANN based method for selecting the best consumption method in smart grids. This method is apposite for such environments where no single prediction method can be used due to different patterns in the time series. In the smart grid context, this is due to various consumption patterns among customers being detected. The proposed ANN, including training and running the Oracle on the train dataset is more than three times faster than the Oracle, which even though can result in better predictions is computationally expensive.

To further reduce prediction error induced by traditional methods in use by utilities and introduced an online historical averaging method with different weighting schemes. It analyzed its efficiency with respect to the number of times it outperformed other methods in the task of energy consumption forecasting. In future work plan to investigate the impact of customer sample selection as well as training dataset size on prediction error and conclude the plan to explore how to

incrementally update the ANN without retraining it from scrape when required.

Method IV: *Ranking-Based techniques*

Recommender systems typically try to predict the ratings of unknown items for each user, frequently using other user's ratings, and recommend top N items with the uppermost predicted ratings. Accordingly, there have been many studies on developing new algorithms that can improve the predictive accuracy of recommendations. However, the quality of recommendations can be evaluated along a number of dimensions, and depend on the accuracy of recommendations alone may not be enough to find the most relevant items for each user.

Recommendation Techniques for Rating Prediction:

Recommender systems normally perform the following two tasks in order to provide recommendations to each user. First, the ratings of unrated items are estimated based on the available information (typically using known user ratings and perhaps also information about item content or user demographics) using some recommendation algorithm. And second, the system finds items that maximize the user's utility based on the predicted ratings, and recommends them to the user.

Gediminas [11] approaches Ranking system are intended to improve the recommendation variety in the second task of finding the best items for each user.

Accuracy of Recommendations Numerous recommendation techniques have been developed over the last few years, and various measures have been employed for assessing the accuracy of recommendations, including statistical accuracy metrics and decision-support measures.

Standard Ranking Approach

Typical recommender systems predict unknown ratings based on known ratings, using any traditional recommendation technique such as neighborhood-based or matrix factorization CF techniques. Then, the predicted ratings are used to support the user's decision making.

Item-Popularity-Based Ranking Item

Popularity- based ranking approach ranks items directly based on their popularity, from lowest to highest, where popularity is represented by the number of known ratings that each item has. Here, despite the major diversity gain, such a significant accuracy loss would not be acceptable in most reallife personalization applications. Therefore, next introduce a general

technique to parameterize recommendation ranking approaches, which allows achieving significant diversity gains while controlling accuracy harms.

Recommender systems have made significant progress in recent years and various techniques have been proposed to improve the recommendation quality. However, in most cases, new techniques are considered to improve the accuracy of recommendations, whereas the recommendation diversity has repeatedly been overlooked.

In particular, we showed that, while ranking recommendations according to the predicted rating values (which is a de facto ranking standard in recommender systems) provides good predictive accuracy, it inclines to perform poorly with respect to recommendation diversity. Therefore, B. Sarwar [12], proposed a number of recommendation ranking techniques that can provide significant improvements in recommendation diversity with only a small amount of accuracy loss.

Method V: Estimation in Map Reduce Frameworks based on Map Reduce are used for processing large data-sets using a cluster of machines. Current systems, however, are not able to provide accurate

estimate of incremental results and are mostly suited for batch processing. Nikolay [13] presents EARL which is a simple framework for estimating results and errors for mining algorithms.

Modern parallel computing systems, such as MapReduce, often assume data values are uniformly distributed. However, in the real world, data is often highly crooked, which may cause workload imbalance among parallel running tasks. Wei Yan [14] addressed the reduce-phase skew problem in Map Reduce, where reduce tasks are often assigned imbalance load.

A sketch-based data structure for capturing Map Reduce key group size statistics and present an optimal packing algorithm which assigns the key groups to the reducers in a load balancing manner. An empirical evaluation with several real and synthetic datasets over two distinct types of applications Cormode[15] shows that load balancing algorithm can strongly diminish the reduce-phase skew.

Sketch-Based Load Balancing Algorithm

The design of the partition function, which maps the intermediate key to a reducer index based on the key group size information as concise in the global sketch. To address this objective, present

an optimal sketch packing algorithm. This algorithm works directly on the global sketch without the need for knowledge of the intermediate keys. As a result, the algorithm is scalable to a large number of key groups. The performance of the optimal sketch packing algorithm can be assessed using the reduce-phase load imbalance ratio σ , which is the ratio between the maximum reducer load and the average reducer load.

Haas, P.J et al [16], determined a scalable and robust load balancing solution for MapReduce framework based on a novel sketch-based data structure. Finally, Implemented and integrated the solution within Hadoop.

Experimental studies using Page Rank and Inverted Indexing over real and synthetic datasets show that the solution outperforms both Hadoop default implementation and the current state-of-the-art solution. A block-level sampling technique is introduced to reduce the extra running overhead brought by the profiling job. In the profiling job, each mapper processes the sampled input blocks.

CONCLUSION

Big Data encompasses a wide range of applications. Foremost, the goal of Health

care Data Application is to implement more effective healthcare delivery that allows limiting expenditure and at same time, it should help to increase quality of care settings. Social Media Application is the best display place for understanding the real-time customer choice or intentions and sentiments. By using such data, Social media business advertising and product marketing becomes easier. Commercial Applications uses the big data analysis for detecting fraud/risk, generating consumer's profile, reducing cost, changing travelling and consumption patterns. Agricultural applications are essentially focused by the desire to improve productivity and profitability of organizations along the agriculture-food supply chain. Societal-scale Sensor network can greatly improve the efficiency of energyprovision chain. The IoT is a substantial concept of new technology; it has great potential to support an ageing society, to improve the energy efficiency and to optimize all kinds of mobility and transport. Accuracy is the foundation on which all data-based decision rest, if accuracy is not given the consideration it deserves, the entire analytical and data- driven process becomes suspicious. Accuracy of the result is straight propositional to the Input data and processing algorithm. Data Quality is

primarily demonstrated in its accuracy. Big Data also defies the principles of Accuracy. So if feasibly improve the accuracy of input and algorithm, it will result in optimized solution.

REFERENCES

1. Min Chen · Shiwen Mao · Yunhao Liu, "Big Data: A Survey" @ Springer
2. Rui Máximo Esteves, Thomas Hacker and Chunming Rong "A new approach for accurate distributed cluster analysis for Big Data: competitive K-Means", Int. J. Big Data Intelligence, Vol. 1, Nos. 1/2, 2014
3. El Agha, M. and Ashour, W.M. (2012) „Efficient and fast initialization algorithm for K- Means clustering“, International Journal of Intelligent Systems and Applications (IJISA), Vol., No. 1, p.21.
4. Pavan, KK., Rao, A.A., Rao, A.V.D. and Sridhar, G.R. (2010) „Single pass seed selection algorithm for K-Means“, Journal of Computer Science, Vol. 6, No. 1, pp.60–66,doi:10.3844/jcssp.2010.60.66.
5. Kuan-Cheng Lin1 · Kai-Yuan Zhang1 · Yi-Hung Huang2 Jason Hung3 · Neil Yen4, "Feature selection based on an improved cat swarm optimization algorithm for big data classification", © Springer Science+Business Media New York 2016.
6. Chu SC, Tsai PW (2007) "Computational intelligence based on the behavior of cats". Int J Innov Comput Inf Control 3(1):163–173
7. Hua Yin1*, Keke Gai2, Zhijian Wang3, "A Classification Algorithm Based on Ensemble Feature Selections for Imbalanced- Class Dataset", 2016 IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing.
8. LinK-C,HuangY-H, Hung JC, LinY-T (2015) Feature selection and parameter optimization of support vector machines based on modified cat swarm optimization. Int J Distrib Sens Netw 2015:9. Article ID 365869.

9. Marc Frincu, Charalampos Chelmiss, Muhammad Usman Noor and Viktor Prasanna, "Accurate and Efficient Selection of the Best Consumption Prediction Method in Smart Grids", 2014 IEEE International Conference on Big Data.
10. H. K. Alfares and M. Nazeeruddin, "Electric load forecasting: Literature survey and classification of methods," International Journal of Systems Science, vol. 33, no. 1, 2002.
11. Gediminas Adomavicius, Member, IEEE, and YoungOk Kwon "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques" IEEE transactions on knowledge and data engineering, vol. 24, no. 5, May 2012.
12. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "ItemBased Collaborative Filtering Recommendation Algorithms," Proc. 10th Int'l Conf. World Wide Web (WWW), 2001.
13. Nikolay Laptev¹, Kai Zeng², Carlo Zaniolo³, University of California, Los Angeles, "Very Fast Estimation for Result and Accuracy of Big Data Analytics: the EARL System"
14. Wei Yan, Yuan Xue, Bradley Malin, "Scalable and Robust Key Group Size Estimation For Reducer Load Balancing in MapReduce", 2013 IEEE International Conference on Big Data.
15. G. Cormode and S. Muthukrishnan. An improved data stream summary:the count-min sketch and its applications. Journal of Algorithms,55(1):58–75, 2005.
16. Haas, P.J. et al. 1996. Selectivity and Cost Estimation for Joins Based on Random Sampling. Journal of Computer and System Sciences. 52, 3 (Jun. 1996), 550-569.