ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

Approach for Big Data Mining in Hadoop Framework: A Survey

Saras Jarkad, Prof. J. E. Nalavade

Dept. of Computer Engineering, SIT Lonavala, Pune, India

ABSTRACT: Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies as well as research focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Frequent Itemset Mining is one of the classical data mining problems in most of the data mining applications. It requires very large computations and I/O traffic capacity. Also resources like single processor's memory and CPU are very limited, which degrades the performance of algorithm. In this paper we have proposed one such distributed algorithm which will run on Hadoop – one of the recent most popular distributed frameworks which mainly focus on mapreduce paradigm. The proposed approach takes into account inherent characteristics of the Apriori algorithm related to the frequent itemset generation and through a block-based partitioning uses a dynamic workload management. The algorithm greatly enhances the performance and achieves high scalability compared to the existing distributed Apriori based approaches. Proposed algorithm is implemented and tested on large scale datasets distributed over a cluster.

I. INTRODUCTION

Successive sets assume a key part in numerous Information Mining assignments that attempt to discover intriguing examples from databases, for example, affiliation rules, relationships, groupings, scenes, classifiers and bunch. The recognizable proof of sets of things, items, manifestations and qualities, which regularly happen together in the given database, can be seen as a standout amongst the most essential undertakings in Information Mining. The first inspiration for looking incessant sets originated from the need to investigate purported grocery store exchange information, that is, to analyze client conduct as far as the bought items. Incessant arrangements of items depict how regularly things are bought together. The current framework has issue of tradeoff in the middle of utility and protection in planning a differentially private FIM calculation. The current framework does not manage the high utility value-based itemsets. Existing strategies has expansive time multifaceted nature. Existing framework gives relatively substantial size yield blend. To take care of this issue, this task builds up a period productive differentially private FIM calculation. With correspondence, information stockpiling innovation, a gigantic measure of data is being gathered and put away in the Web. Information mining, with its guarantee to productively discover profitable, non-evident data from tremendous databases, is especially powerless against abuse. The circumstance may turn out to be more terrible when the database contains bunches of long exchanges or long high utility itemsets. To understand this, we propose a proficient calculation, in particular utilized hadoop, for parallel handling on high utility thing sets. Successive itemset mining (FIM) is a standout amongst the most fundamental issues in information mining. We introduce a structure for mining affiliation rules from exchanges comprising of diverse things where the datahas been randomized to save security of individual exchanges [2]. We proceed with the examination of the information mining are underneath.

- categorical information rather than numerical information, and
- Association guideline mining rather than grouping.

It will concentrate on the undertaking of discovering incessant itemsets in affiliation guideline mining. In the mining stage, to balance the data misfortune brought about by exchange part, It devise a run-time discovering technique to locate the real backing of itemsets in the first database. Here, we look the appropriateness of FIM procedures on the MapReduce stage. It is a parallel disseminated programming system presented in [4, 6], which can prepare a lot of

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

information in a hugely parallel manner utilizing straightforward merchandise machines. We utilize MapReduce to execute the parallelization of calculation, in this manner enhancing the general execution of successive itemsets mining.

II. RESEARCH BACKGROUND

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations. A MapReduce program is composed of a Map() procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The model is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once. As such, a single-threaded implementation of MapReduce will usually not be faster than a traditional (non-MapReduce) implementation; any gains are usually only seen with multi-threaded implementations. The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm. MapReduce libraries have been written in many programming languages, with different levels of optimization. A popular open-source implementation that has support for distributed shuffles is part of Apache Hadoop. The name MapReduce originally referred to the proprietary Google technology, but has since been genericized. By 2014, Google was no longer using MapReduce as their primary Big Data processing model, and development on Apache Mahout had moved on to more capable and less disk-oriented mechanisms that incorporated full map and reduce capabilities

III. LITERATURE SURVEY

Frequent item set mining used in wide range of application areas such as decision support, web usage mining, bioinformatics, etc. If the data is sensitive (e.g. web browsing history, medical records), releasing the discovered frequent item sets might pose considerable threats to individual privacy. Frequent itemset mining (FIM) is one of the most fundamental problems in data mining. Differential privacy [1] is one of the best solution to address such problem. Through adding right amount of noise, differential privacy assures that the output of a computation is insensitive to changes in any one personal record, and thus restricting privacy leaks from results. There are variety of algorithms have been proposed for mining frequent itemsets. The Apriori [4] and FP-growth [5] are the most well known. Compared with Apriori, FP-growth is depth first search algorithm which requires no candidate set generation. FP-growth only performs two database scan. Because of the appealing features of FP-growth proposed FIM algorithm is based on FP-growth algorithm. In this paper, we propose designing a differentially private FIM algorithm which can not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. Proposed differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. The PFP-growth algorithm consists of a preprocessing phase and a mining phase.

In preprocessing phase, we transform the database by limiting the length of transaction. To enforce such a constraint long transaction should be split rather than truncated. A proposed transaction splitting algorithm is used for splitting the transaction to transform the database. In the mining phase, we privately discover frequent itemsets. Even if the potential advantages of transaction splitting, might bring some information loss. So, we propose a run time estimation method to offset such a loss. Dynamic reduction method, dynamically estimate number of support computations, so that we can gradually reduce the amount of noise required by differential privacy. The tradeoff can be improved by our novel

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

transaction splitting techniques. By leveraging the downward closure property, a dynamic reduction method is proposed to dynamically reduce the amount of noise added to guarantee privacy during the mining process. Through formal privacy analysis we can say that PFP-growth algorithm is differentially private.

N. Li, W. Qardaji, D. Su, and J. Cao [1], in this paper, they sought the issue of how to perform incessant itemset mining on exchange databases while fulfilling no of protection. They propose a methodology, called PrivBasis, which influences a novel idea called premise sets. A θ-premise set has the property that any itemset with recurrence most noteworthy than θ is a subset of a few premise. They spoke to calculations for secretly building all premise set and after that utilizing it to locate the most incessant itemsets. Trials demonstrate that our methodology incredibly outflanks the cutting edge.

Maurizio Atzori, F. Bonchi, F. Giannotti [2], in this paper creator demonstrate that this conviction is poorly established. By idea of k-secrecy from the source information to the extricated designs, they formally portray the thought of a danger to namelessness in the connection of example, and gives a philosophy to proficiently and viably demonstrate all such conceivable dangers that emerge from the exposure of the arrangement of examples. On this premise, they pick up a formal thought of security assurance that permits the divulgence of the removed information while ensuring the obscurity of the people in the source database. Maybe with a specific end goal to handle the situations where the dangers to namelessness can't be kept away from, they concentrate how to dispose of such dangers by method for example contortion performed in a dataset.

Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke [3], Creator display a work for mining affiliation rules from exchange comprising of all out things where the information has been randomized to keep up protection of individual exchanges. While it is conceivable to recoup affiliation standards and save security utilizing a forward „„uniform"" randomization, the sought guidelines can tragically be abused to pick up protection. They examine the way of security and propose a class of administrators that are a great deal more viable than uniform randomization in restricting the breaks. They demonstrate formulae for a fair bolster estimator and its difference, which permit us to get back item set backings from randomized database, and demonstrate to join these formulae into mining calculations. Finally, they exhibit trial investigation that approves the calculation by applying it on genuine datasets.

W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis [4], they discovered continuous thing sets is the most unreasonable errand in affiliation principle mining. This undertaking to an administration supplier conveys a few advantages to the information proprietor, for example, cost help and a less commitment to capacity and computational assets. Mining results can be misfortune if the administration supplier is straightforward however makes blunder in the mining procedure, or is lethargic and decreases expensive calculation, returning deficient results, or is noxious and contaminated the mining results. They demonstrate the trustworthiness issue in the outsourcing procedure, i.e., how the information proprietor.

FP-Development: The FP-Development calculation skirts the competitor itemset era process by utilizing a reduced tree structure to store itemset recurrence data. FP-Development works in a separation and overcomes way. It requires two sweeps on the database. FP-Development first figures a rundown of continuous things sorted by recurrence in slipping request (F-Rundown) amid its first database check [5].

C. Dwork [6] the creator gives a general inconceivable possibility result demonstrating that a formalization of Dalenius" objective along the lines of semantic security can't be accomplished. In spite of instinct, a variation of the outcome undermines the protection even of somebody not in the database. This situation proposes another measure, differential protection, which, instinctively, catches the expanded danger to one's security brought about by partaking in a database. The methods created in a grouping of papers, coming full circle in those portrayed in, can accomplish any craved level of protection under this measure. By and large, amazingly precise data about the database can be given while at the same time guaranteeing abnormal amounts of security.

In recent years, business intelligence systems are playing pivotal roles in fine-tuning business goals such as improving customer retention, market penetration, profitability and efficiency. In most cases, these insights are driven by analyses of historic data. Now the issue is, if the historic data can help us make better decisions, how real-time data can improve the decision making process [7]. Frequent pattern mining for large databases of business data, such as transaction records, is of great interest in data mining and knowledge discovery [8], since its inception in 1993, by Agrawal et al. In this paper, we assume that the reader knows the basic assumptions and terminologies of mining all frequent patterns.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

IV. CONCLUSION

This paper introduction on how the Hadoop framework can be used for large data storage and analytics purpose through this paper. Large amount of source data from social media, web logs or third party stores is stored on Hadoop to enhance analytic models that drives research and discovery. Data can be stored on Hadoop clusters in cost effective manner and can be retrieved easily when needed. Operational cost of whole data analytics and data processing can be lowered by use of Apache Hadoop. Its MapReduce on HDFS provides a scalable, fault tolerant platform for processing large amount of heterogeneous data.The paper summarizes the current issues in data mining algorithms migration towards Hadoop platform. We have identified the current gaps and open research areas. Our future research will focus on these open problems and propose effective solutions for the same.

REFERENCES

[1] Ninghui Li, WahbehQardaji, Dong Su, Jianneng Cao,"PrivBasis: Frequent Itemset Mining with Differential Privacy.", in VLDB, 2012.
[2] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity preserving pattern discovery," VLDB Journal, 2008.
[3] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD, 2002.
[4] W. K.Wong, D.W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "An audit environment for outsourcing of frequent itemset mining," in VLDB, 2009.
[5] Revealing Information while Preserving Privacy Irit Dinur Kobbi Nissim
[6] C. Dwork, "Differential privacy," in ICALP, 2006.
[7] M. L. Gonzales, "Unearth BI in Real-time," vol. 2004: Teradata, 2004.
[8] B. Goethals, "Memory Issues in Frequent Pattern Mining," in Proceedings of SAC'04. Nicosia, Cyprus: ACM, 2004.
[9] Minal Shahakar and Rupesh Mahajan proposed Redistribution Of Task Using Load Balancing Heuristics In Heterogeneous Environment International Journal of Research in Computer Engineering & Electronics Volume 3 Issue 4 in sep 2014.