International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

"Big Data Analytics in Cyber Physical Systems"

Ashish Jadhao, Swapnaja Hiray
Department of Computer Engineering, Sinhgad College of Engineering, University of Pune, India.

Associate Professor, Department of Computer Engineering, Sinhgad College of Engineering, University of Pune, India

Abstract

The cyber physical world has computation and communication power which grows in vast manner. Thus as because of this it produces large volume of data to handle this processes. There are four main challenges related to big data and they are volume, variety, velocity, veracity. Volume and variety are managed by some store data processing system like Hadoop. But the velocity and veracity of such large amount of data is too much complex process.

In this paper we are going to implement such system which can handle high speed and different pattern of data with its large volume. We are going to implement correlation analytics and mining on the data stream to extract meaningful information. The system should provide real time data processing so that it used Event processing engine as Esper which has different language queries to generate different events. To catch real time data and for simple filtering of that data stream Storm is used which used topology. Correlation and mining takes place by two different algorithm Apriori and FP-Growth algorithm.

Keywords: *Analytics, Apriori, Big data, CEP, Cyber Physical System, Esper, FP-growth, Mining, Sliding window, Storm.*

I. INTRODUCTION

Big data sizes are a continuously achieving its peak, as at present it is changing its range from gigabytes to number of terabytes in a single data storage. Much of this data explosion is the result of a dramatic increase in devices located at the periphery of the network including embedded sensors, smart phones, and tablet computers. So now a day's cyber physical system has face big data handling property and obtain the core data from that is also a main challenge with fast growing resources. There are large numbers of challenges with big data its veracity, size, accuracy, hardness, and security. This challenges rise right away during data collection, where this data tsunami force us to take such decision, which are not practically meaningful about which data is stored and which is discarded, and how that data is going to store according to pattern of that data. The format of data today is not properly structured; for example, in which tweets and blogs are loosely structured parts of text, while images and video arrangement provides storage and display, and not used in search and semantic content: and converting such content into structured is the biggest challenge.

Cyber-Physical System is integrations of computation and physical processes. In this system the design of communication infrastructure is of key importance since it conveys information from sensors to controllers. As the use of cyber physical system increases in vast manner the data related to that system also gathered in very big amount. The offline operation on cyber physical data is not too much difficult because we already know the pattern of data stream at each time. But the biggest problem rise when there is manipulation of real time data.

II. RELATED WORK

Big data subject to where number of amount of data transaction of large amount of high speed data with different variety of it takes place or compute capacity for accurate and timely decision making. In this system the design of communication infrastructure is of key importance since it conveys information from sensors to controllers. The row data coming from various sources are stored in storage where already has terabyte of data is stored so to handle such data and get useful data from that is main problem for this it has to implement data stream analytics and mining for cyber physical system. There are many important cyber physical systems in practice such as smart grid and unmanned aerial vehicle networks.

Data and Information or Knowledge has a significant role on human activities. Data mining is the process where of evolution of knowledge by studying the large volumes of data from various perspectives and summarizing it into useful information. And Analytics are the methods of decomposing concepts or substances into smaller pieces, to understand their workings. In past the data arrive from various sources are stored in big data storage. Here all the data stream analytics and mining process takes place. As the big data contains terabytes of data already taking analytics and mining in storage is hectic and also

prone to various challenges. This challenges are not simple to eliminate because it already knows that the data comes from large sources is in large volume with very high velocity. To make analytics and mining of such data lead to complex process. And after storing data in storage and then implement all this operation on that data increases cost very high.In initial days, data mining algorithms work best for numerical data collected from a single storage of data, and this techniques of data mining have developed for continuous files, and also where data is stored in table format. In past days of data mining most of the algorithms employed only statistical techniques.

There are two challenges the process has to face they are designing fast mining methods for data streams and need to detect promptly changing concepts and data distribution as real time because of highly dynamic nature of data streams. Memory management is a main challenge in stream processing because many real data streams have irregular arrival rate and variation of data arrival rate over time. In many applications like sensor networks, stream mining algorithms with high memory cost is not applicable. Therefore, it is necessary to develop summarizing techniques for collecting valuable information from data streams. By considering the size of memory and the huge amount of data stream that continuously arrive to the system, it is essential to have a perfect data structure for storage, frequent improvement and to access the stored information [2].If such structure is not present, quality of using of mining algorithm will sharply decrease. Some traditional used for mining are slow and not so efficient for online data processing. The algorithm which are used for analytics and mining have to consider all the factors which are effecting the value of data and importance of that system. Also number of data sources and fluctuations of data properties processing the data online is also a problem.

Sometime streaming data coming from different sources lead to various errors in which one is missing the tuple, out of ordering of it during sending the data to storage. Sometime wrong values are send to the operations which will gives wrong result. Also to avoid unwanted data for saving further CPU, energy cost is important. In all of this the big challenge is to implement stream correlation and rule mining on the same system so that at the same time it will work on the online data stream with rule mining and emerge new data mining rule which store the same system so that they can be useful for future.

III. MAIN METHOD

3.1 System Architecture

System architecture of data stream analytics goes through three parts they are data stream management system, complex event processing engine and at last business process management and visualization.

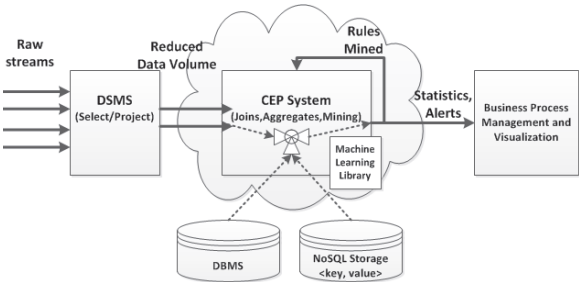

Fig 3.1: Data stream analytics and mining architecture

The row material or row data is first subject to the data stream management. Thus data stream management system is a pipeline structure where basic main working is to eliminate unwanted data from the row data so that in future it will not affect the stream analytics and mining process and also it will avoid CPU utilization and storage and memory cost. Data stream management system also provide scheduling and proper maintaining of data stream so that there will not be traffic of collision of data tuples with each other. Thus data stream management system provides all pre-processing that will require before the actual complex processing. Also it reduces the data size to row wise and column wise. The next part include in it is complex event processing engine where all core filtering takes place and which is used offline or online according the system need. Here it also contains one important part that is data base management system with No SQL queries which are used for the statistical analysis. In complex event processing engine the basic filter data is correlated with the standard data which already stored in the DBMS system by attacking various No SQL queries so that the knowledgeable data that required for some specific application are drawn out. No SQL queries are nearly similar like structured query language but it will provide some extra opportunities. By combining this two structures DSMS and complex event processing tool it lead to main filtering tool and it will handle as below.

3.2 Stream Analytics

Stream analytics is the process of making correlation of raw data with the standard data that values are already store in the database system. Thus system used Pearson product moment correlation over the

row data stream. Stream analytics analyze big data to find patterns and relationships, make informed predictions, deliver actionable intelligence, and gain business insight from this steady influx of information. Organizations in every industry are trying to make sense of the massive influx of big data, as well as to develop analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information. If big data is properly handle and manage it can provide priceless information related to market related problem, damage of equipment, buying patterns, maintenance cycles and many other business issues, decrement in costs, and able to make more unique business decisions. To obtain value from big data, you need a cohesive set of solutions for capturing, processing, and analyzing the data, from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems.

Actually correlation is the covariance of two variables divided by their standard deviation. Thus it means obtain the input of the data as variable of considering example or application and taking covariance means current values compare with each other and then divided by standard value which is stored already by studying all the patterns and conditions related to the stream of data. By taking correlation it get the result of statistical analytics which will be in range of (-1 to1).Lets consider if there is high positive correlation then it will provide +1 value, if there is no relation then its value is 0 and for highly negative correlation its result is -1. The application aim is to study the daily routing of bus transportation. This application provides the actual position of bus on daily path and compares its parameters with itself and other vehicle. For statistical correlation of bus application standard data about the bus transport has to be already store in database management system so that correlation should be maintain with proper parameter. Thus for example it finds the correlation of two buses running on different following queries is used to find their analytics. This analytics can be carried out by the use of sliding window. If there is small window size then there are less result are available to compare and it will raise alarm for very short fault or errors. So large window will provide large result to compare so that small fault are neglected and this is useful for bus application as bus will catch its fault in some time further. To reduce the amount of processing and output produced, we could use tumbling windows which only publish results at the end of a time or count period. For tumbling windows the delay increases with window size, because all the data that is collected until the end of a time interval is processed at once.

3.3 Rule Mining

The structure of Data streams are frequently changing, arranged for some specific time, vast and potentially with no limit in real time system. Because of high volume and speed of input data, it is needed to use semi-automatic interactional techniques to extract embedded knowledge from data. The main algorithm is association rule mining under which two different algorithms implemented and they are Apriori and FPGrowth algorithms which are used for rule mining in various applications. In an association rule denoted by ***XY (S,C)***, X and Y refer to the frequent item sets, while S is support which is the percentage of record that contain item set either X or Y or both. C is the confidence which is is the percentage of record which contains both X and Y.

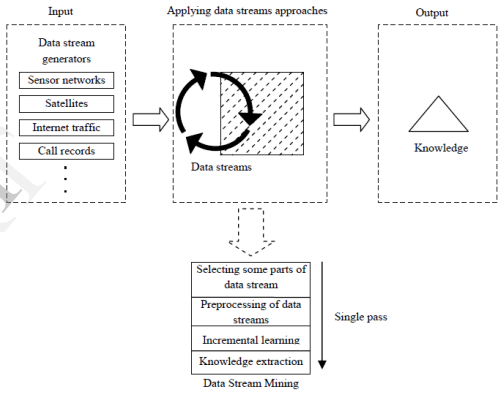

Fig 3.2: General Process of data stream mining

As it is already knows that there is very large amount of data stored in data set before as there is continuous, unlimited, and very high speed fluctuated data streams in both offline and online condition and because of that scanning the data again and again is not efficient by using traditional data mining algorithms. So it best to use algorithm like Apriori which counts frequent item sets, generates candidate item sets using the minimum support value, prunes the infrequent ones, calculates confidence on all permutations of the frequent item sets and selects those above the given Confidence threshold. Next is FP-Growth algorithm. FP-Growth algorithm does a first pass over the transactions creating a frequency-sorted database of items, omits the in frequent items, and finally creates an FP-tree Compared with Apriori based algorithms; it achieves higher performance by avoiding iterative candidate generations.

Thus total process goes through Data Stream, which are the simple tuples with fixed number of field comes from various sources is taken into Apache

Kafka as messaging structure which send it to storm which has elements like spouts and bolts. Data Streams are first taken into SPOUTS (data emitters), which retrieve the streams and put them into the storm clusters. This data inserted into BOLTS (data processor) which will perform some primary processing task and then emits those data into one or more streams. The data streams comes from storm is taken into Esper where Complex event processing is done using CEP engine, where complex event are process by continuously firing the NoSQL queries over the data to filter it and then applying the Apriori & FP-Growth for the stream mining, for that some kind of threshold data is with the system for association.

IV. CONCLUSIONS

The basic aim is to achieve or to develop particular system that will provide proper data stream analytics and mining on the real time data stream. It provides the past technique for analytics and mining. On the basis of past technique it compares its working and show that how this system overcome nearly all the challenges of past technique. The main thing is that it provides the analytics and mining structure on the same system so that it possible to perform this operation online as it provides application of sliding type window. The main conclusion of this system is that it can handle the data stream by using various tools like Esper, Data stream management system tool like Storm. There are number of algorithm studied under this implementation they are association rule mining, Apriori and FPGrowth.

REFERENCES

[1]B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom, Models and issues in data stream systems, *ACM PODS, 2002, June, pp 1-16.*

[2]C. Borgelt, An Implementation of the FP-growth Algorithm, *ACM Workshop of Open Source Data Mining Software, (OSDM), pages 1-5, 2005.*

[3]M. M. Gaber, A. Zaslavsky, S. Krishnaswamy; Mining Data Streams:A Review; *ACM SIGMOD Record Vol. 34, No. 2; June 2005.*

[4] N. Jiang, L. Gruenwald, Research issues in data stream association rule mining, *In SIGMOD Record, Vol 35, No 1, March 2006.*

[5]What is Big Data? http://www-01.ibm.com/software/data/bigdata/

[6]Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten: The Weka data mining software: An update, *In SIGKDD Explorations, Volume 11, Issue 1, page 10-18, 2010*

[7]MahnooshKholghi, MohahmmadrezaKeyvanpour: An Analytical Framework For Data Stream Mining Techniques Based on Challenges And Requirements, *In IJEST, Vol. 3 No. 3, page no.2507-2513, Mar 2011*

[8]Hua-Fu Li, Suh-Yin Lee, Man-Kwan Shan: Online Mining (Recently) Maximal Frequent Item sets over Data Streams, *(RIDE-SDMA'05) 1097-8585/05*

[9]R. Manickam, D. Boominath, V. Bhuvaneswari: An analysis of data mining: past, present and future, *(IJCET), Volume 3 Issue 1, January-June (2012), pp. 01-09.*

[10] Ismail Ari, Erdi Olmezogullari, Ömer Faruk Çelebi†: Data Stream Analytics and Mining in the Cloud, *2012 IEEE 4th International Conference on Cloud Computing Technology and Science*.