iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** I **Month of publication:** January 2020

DOI: http://doi.org/10.22214/ijraset.2020.1075

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue I, Jan 2020- Available at www.ijraset.com

Review in Data Stream Mining in Big Data

Padma Priya. R[1], Jothi. P[2]
[1, 2]*Associate Professor, Dept.of Bachelaor of Computer Application, Rathnavel Subaramaniam College of Arts And Science, Coimbatore, Tamilnadu, India*

Abstract: ***Data stream mining in big data is a process in which large streams of real-time data are processed with the sole aim of extracting insights and useful trends out of it. Streaming data study in real time is fetching the efficient and fastest way to obtain constructive knowledge from what is occurrence now, allowing concern to react quickly when problems emerge to detect new trends helping to recover their performance. In this paper, we have a tendency to reward the theoretical foundations of data stream in big data analysis and establish potential directions of future analysis. Mining data stream and big data techniques are being reviewed.***

Keywords*: Data Stream mining, Data mining, Big data, Big data techniques, Big data challenges.*

I. INTRODUCTION

Data mining defined to extract useful information from huge volumes of data [1]. In the same way, mining Data Streams consigns to extracting information from stable and constant river flow stream of data. Recently a lot of reports in the media promoter the hype of Big Data that are encountered in three problematic issues. They are three(3)volume of wheel challenges are known as; velocity problem that gives rise to a huge amount of data to be handled at growing high speed; variety of problem that construct data processing and integration difficult because the data come from various sources and they are formatted differently; and volume problem that builds the data in processing, storing and investigation over them both computational and archiving challenging. In review of these three volume challenges, the traditional data stream mining approaches which are based on the full batch form learning may run short in meeting the insist of analytic effectiveness.

II. BIG DATA

Big Data is a new emerging term used to identify the datasets that due to their large size, we cannot supervise them with the typical data mining software tools. Instead of defines a big data as datasets of a actual large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be handled without using recent algorithms or technologies. The McKinsey Global Institute (MGI) published a report on Big Data [14] that portrays the business opportunities that big data opens a potential value of $300 billion in the US health care, $149 billion in European government administration or improving the operating margin of retailer companies by 60 percent.

Big Data analytics in data stream is becoming an important tool to improve efficiency and quality in organizations, and its importance is going to increase in the next years. There are two main strategies for dealing with big data they are sampling and using distributed systems. Sampling is based in the fact that if the dataset is too large and we cannot use all the examples, we can obtain an approximate solution using a subset of the examples. A good sampling method will try to select the best instances, to have a high-quality performance using a small quantity of memory and time.

The foremost opportunities in Big Data tenders to developing countries in their White paper as 'Big Data for Development: Challenges & Opportunities'[15],describes the premature warning as expand fast response in time of disaster, detecting variance in the usage of digital media. Real time awareness is designed programs and policies with a more fine grained representation of reality. Real time feedback as check what policies and programs fails, monitoring it in real time, and using this feedback make the needed changes. The Big Data stream mining revolution is not restricted to the industrialized world, as mobiles are spreading in developing countries as well. It is estimated then there are over five billion mobile phones, and that 80% are located in developing countries.

III.DATA STREAM MINING IN BIG DATA

Data stream mining in big data has turn out to be a mainstream field now. Since the usual and traditional methods cannot resolve the data stream issues there are various challenges to solve them some of them [4] are frequently changing dynamic nature, huge volume and speed with which data is created, memory constraints, managing these continuous flow of data create a bigger problem and challenge for the researchers working on streaming data. In traditional data sets we could store the data and analyse it many times but this cannot be done with data streams due to huge volumes of data. Many new techniques keep evolving to contract with these issues, the bottom line being that the algorithms must frequently update their models to hold the inconsistencies in the data.

©IJRASET: All Rights are Reserved

The current review establishes the need for the distributed data stream in the world of technology today. The review will further highlight the advantages of real-time scalability of the distributed data streaming in increasing its advantages and applicability in the big data environment. Additionally, the present review also works towards identifying the importance of data streaming across the various frameworks of big data to stress upon the fault tolerant and high availability features of the data streaming system in big data. The review will further highlight the strengths and shortcomings of the big data stream system and analyze the architecture and systems being followed by various companies across industries utilizing the big data streams.

IV. DATA STREAM MINING IN BIG DATA –CHARACTERISTICS AND CHALLENGES

Major characteristics faced by a system designer while designing big data stream systems are reviewed by:

1) *High Availability:* Data stream in big data are distributed systems ensure availability, i.e. even if a component of a system fails, the system doesn't fail, and there is another component that replaces it and keeps it running.
2) *Scalability:* To ensure that it doesn't fail, a distributed system is flexible so that you can easily increase the number of machines when the workload increases.
3) *Transparency:* Data stream in big data are distributed systems provide a global view of all the underlying machines as a single machine and hide their internal workings from end users.
4) *Flexibility:* Due to advancements in the field of microprocessors, networking, and storage,distributed systems have made it possible to make changes to both hardware and software components, without affecting performance.

Major challenges faced by a system designer while designing big data stream systems are reviewed by:

a) *Transparency:* The challenge here is to decide up to what extent the distributed big data stream system should appear as a single system to the user. The thumb rule here is to hide the complexity of a distributed system as much as possible.
b) *Prone To Failures:* Distributed big data stream systems are prone to failures. Hence, the system designer should implement methods for quick detection of failures and also provide workarounds to address the failures.
c) *Data Security And Privacy:* The data stored in distributed systems is highly sensitive, hence it's mandatory for the system to have strong security and privacy techniques for safeguarding the data.
d) *Concurrency:* Big data distributed systems are often required to serve a lot of end users in parallel. As system resources such as user data, computing hardware etc. are shareable, this leads to problems of concurrent access to shared resources and multi-tenancy. For a resource to stay safe during parallel accesses, various synchronisation techniques such as isolation, locks etc. are implemented. Implementing these synchronisation techniques in a distributed environment is not straightforward and is highly challenging.

V. DATA STREAM MINING IN BIG DATA –LITERAURE REVIEW

1) Zhang P., Li J., Wang P., et al[3] discussed about the semi-supervised learning strategy is used in document to reduce the influence of dimension on classifiers by performing low dimensional subspace mapping on the high dimensional data streams. A classifier for each unlike data stream is built. The most individual classifiers are established using the integrated learning – thinking, thought to establish an integrated multi-data stream classification.
2) Jing Liu, Guo-sheng Xu, Da Xiao, Li-ze Gu, Xin-xin Niu [4] discussed about the Data stream classification based on incremental learning In document [5], the traditional learning vector quantization (LVQ) algorithm is improved by using the incremental learning approach. With the incremental learning concept not only the learned knowledge in the traditional algorithm (LVQ) is retained in the update process, new knowledge is also learned. It greatly improves the accuracy of the model classification.
3) Apache Spark Apache Spark [6] here in this paper discussed about a powerful processing framework that provides an ease of use tool for efficient analytics of heterogeneous data. It was originally developed at UC Berkeley in 2009 [7] .
4) Spark has several advantages compared to other big data frameworks like Hadoop MapReduce [8] and Storm [9] here in this paper discussed about a key concept of Spark is Resilient Distributed Datasets (RDDs). RDD is basically an irreversible collection of objects spread across a spark cluster. In spark there are two types of procedure on RDD they are transformations and actions. Transformations consist in the creation of new RDDs from existing ones using functions like map, filter, union and join. Actions consist of final result of RDD computations. Spark Streaming is a Spark library that enables scalable and high-throughput stream processing of live data streams.

©IJRASET: All Rights are Reserved

5) 2.2. Apache Storm [9] here in this paper discussed about an open source framework for processing large structured and unstructured data in real time. Storm is a fault tolerant framework that is suitable for real time data analysis, machine learning, sequential and iterative computation. A Storm program is represented by a directed acyclic graphs (DAG). The edges of the program DAG symbolize data transfer process.The nodes of the DAG are divided into two types they are spouts and bolts. The spouts or way in points of a storm program symbolize the data sources. The bolts symbolize the functions to be performed on the data. Storm is based on two daemons called Nimbus (in master node) and a supervisor for each slave node. Nimbus supervises the slave nodes and assigns tasks to them. If it detects a node failure in the cluster, it reassigns LADaS 2018 - Latin America Data Science Workshop 18 the task to another node. Each supervisor controls the execution of its tasks (affected by the nimbus). It can stop or start the spots following the instructions of Nimbus. Each topology submitted to Storm cluster is divided into several tasks.

6) Apache Flink Flink [10] here in this paper discussed about is an open source framework for processing data in both real time mode and batch mode. It provides several benefits such as fault-tolerant and large scale computation. The programming model of Flink is similar to MapReduce [8] here in this paper discussed about By contrast to MapReduce, Flink offers additional high level functions such as join, filter and aggregation.Flink allows iterative processing and real time computation on stream data collected by different tools such as Flume [11] and Kafka [7]. It offers several APIs on a more abstract level allowing the user to launch distributed computation in a transparent and easy way.

7) Apache Samza Apache Samza [12] here in this paper discussed about an open source distributed processing framework created by Linkedin to solve various kinds of stream processing requirements such as tracking data, service logging of data, and data ingestion pipelines for real time services. Since then, it was adopted and deployed in several projects. Samza is designed to handle large messages and to provide file system persistence for them. It uses Apache Kafka as a distributed broker for messaging, and Hadoop YARN for distributed resource allocation and scheduling.

VI. DIFFERENT FRAME WORKS USED IN DATA STREAM MINING IN BIG DATA

Comparison of popular stream processing frameworks [13] is discussed below in fig-1.

	Spark	**Storm**	**Flink**	**Samza**
Data format	DStream	Tuples	DataStream	Message
Data sources	HDFS, DBMS, and Kafka	Spoots	HDFS, DBMS, and Kafka	kafka
Programming model	Transformation and action	Bolts	Actions functions (map,groupby...)	Mapreduce Job
Programming languages	Java, Scala and Python	Java	Java	java
Cluster manager	Hadoop YARN, Apache Mesos	Zookeeper	Hadoop YARN, Apache Mesos	YARN
Latency	Few seconds	Sub-second	Sub-second	Sub-second
Messaging	Exactly once	At least once	Exactly once	Exactly once
Machine learning compatibility	SparkMLLIB	Compatible with SAMOA API	FlinkML	Compatible with SAMOA API
Elasticity	Yes	Yes	No	No
Sliding win-dows/Windowing	time based	time based and count based	time based	time based and count based
Auto-parallelization	On demand	Pipelined processing	Pipelined processing	On demand
Streaming query	SparkSQL	No	No	Yes (Samza-SQL API)
Data Partitioning	Yes	No	No	Yes
API	Declaratif	Copositionnel	Declartaif	Copositionnel
Data transport	RPC	RPC	RPC	Kafka

Fig -1: Comparison of popular stream processing frameworks [13]

©IJRASET: All Rights are Reserved

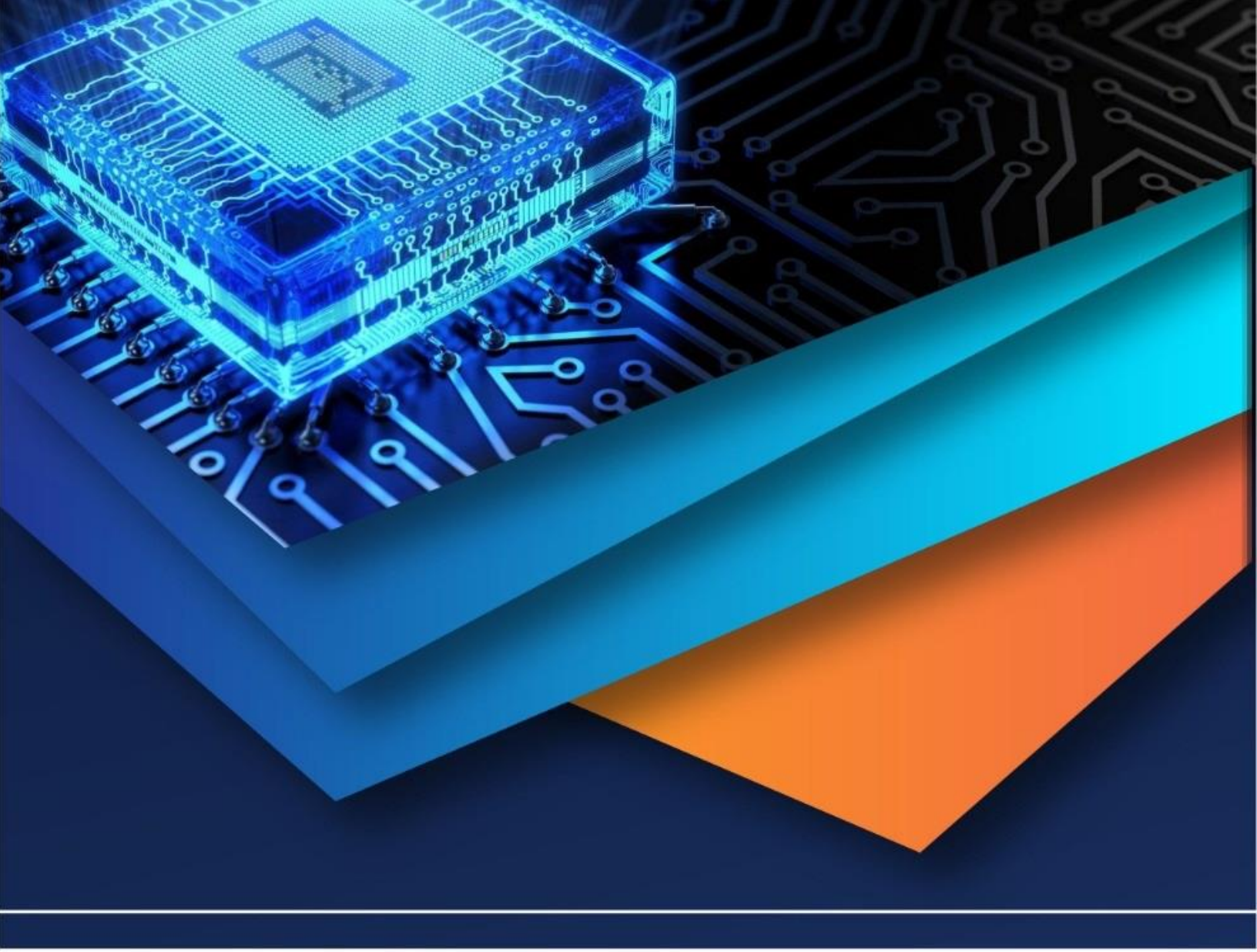
VII. CONCLUSIONS

Data streams are active ordered, fast changing and gigantic, immeasurable and infinite sequence of data objects. Big Data grows continually with fresh data and are being generated at all times; hence it requires an incremental computation approach which is able to monitor large scale of data dynamically. In the field of stream data mining for big data and many problems remain to be solved, the application prospect of data mining in cyberspace is very broad, and the research work in this region is presented with high practical value and academic potential. In this paper, we have a tendency to reward the theoretical foundations of data stream in big data analysis and establish potential directions of future analysis. Mining data stream and big data techniques are being reviewed.

REFERENCES

[1] Aggarwal.C, Ed., "Data Streams – Models and Algorithms", Springer, 2007.
[2] Aggarwal.C.C, J. Han, J. Wang, and P. S. Yu,"A framework for clustering evolving data streams," In Proc. of VLDB, pages 81-92, 2003.
[3] Zhang P., Li J., Wang P., et al. Enabling fast prediction for ensemble models on data streams[C]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, San Diego, CA, USA, Aug 21-24, 2011: 177-185.
[4] Jing Liu, Guo-sheng Xu, Da Xiao, Li-ze Gu, Xin-xin Niu. A Semi-supervised Ensemble Approach for Mining Data Streams[J]. Journal Of Computers, 2013,8(11):2873-2879.
[5] Xiao J. , Xie L., He C. Z. ,et al. Dynamic classifier ensemble model for customer classification with imbalanced class distribution[J]. Expert Systems with Applications, 2012,39(3):3668-3675.
[6] Apache Spark. Apache spark: Lightning-fast cluster computing, 2015.
[7] Nishant Garg. Apache Kafka. Packt Publishing Ltd, 2013.
[8] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.
[9] Ankit Toshniwal, Siddarth Taneja, Amit Shukla, Karthik Ramasamy, Jignesh M Patel, Sanjeev Kulkarni, Jason Jackson, Krishna Gade, Maosong Fu, Jake Donham, et al. Storm@ twitter. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages 147–156. ACM, 2014.
[10] Apache Flink. Scalable batch and stream data processing, 2016
[11] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R Henry, Robert Bradshaw, and Nathan Weizenbaum. Flumejava: easy, efficient data-parallel pipelines. In ACM Sigplan Notices, volume 45, pages 363–375. ACM, 2010.
[12] Apache Samza. Linkedin's real-time stream processing framework, by riccomini, c, 2014.
[13] A Comparative Study on Streaming Frameworks for Big Data Wissem Inoubli 1 , Sabeur Aridhi2 , Haithem Mezni3 , Mondher Maddouri4 , Engelbert Mephu Nguifo5 1University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH, Tunis, Tunisia 2University of Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France 3University of Jendouba, SMART Lab, Jendouba, Tunisia 4College Of Buisness, University of Jeddah, P.O.Box 80327, Jeddah 21589 KSA 5University of Clermont Auvergne, LIMOS, Clermont-Ferrand, France.
[14] The McKinsey Global Institute (MGI) ,may 2011,copy right McKinsey company.
[15] Big Data for Development: Challenges & Opportunities ,un globel pulse ,May 2012.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)