Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

REAL TIME DATA PROCESSING USING STORM

Prof.Minal Nerkar, Aravind Iyer, Swapnil Joshi, Mehulkumar Shilu, Ashit Singh

Computer Science, AISSMS IOIT

Abstract *—As the world is developing, internet has become important part of our life. Every Person is accessing the Internet through Smart phones, Laptops, IPad, and Desktop. It is very difficult to store and process the real-time data by Internet Service providers and Network Administrator. The Streaming source or continuous source can be structured, unstructured, Semi-structured. The given data is scattered in such a way which will take great efforts to find the relevance. Performing such operations manually is infeasible hence we can introduce machine learning to fasten up the process. Hence we propose a new architecture, designed to perform Real-time stream processing and analysis using Apache SAMOA and Apache Storm.*

Keywords*-STORM, SAMOA, Stream Data, Spout, Bolt, Vertical Hoeffding Tree Algorithm, Processor Item.*

I. INTRODUCTION

One of the most important issues for Internet Service Providers and Investigation department is to, process and to analyse for legal reasons, a great amount of internet data and real time data set. For example social media website like twitter, receives 10000 tweets per second and it is not feasible to apply data mining technique in this type of huge information on batch mode, but a scalable architecture to process all the data in real time manner is required. We need a system which is Horizontally Scalable, which is Robust and Fault Tolerant and has guaranteed message processing. Hence we propose a new architecture, designed to perform Real-time stream processing and analysis is done where:

1-Real Time Stream Analysis is performed by Scalable Advanced Massive Online Analysis (SAMOA). SAMOA is framework that allows to perform data analysis by distributed Streaming Machine Algorithms

2-Real Time Stream processing is ensured by Stream Processing Engine called Storm. Apache Storm is free and open source distributed real-time computations system designed to tolerant to fault of its own components.

The Goal is to perform real-time analysis of network traffic in order to recognize data traffic. For example Skype traffic analysis.

II. RELATED WORKS

Traditional database query processing systems are inappropriate when a huge amount of data must be processed in real-time. So, new stream processing engines have been designed to overcome the drawback of the traditional databases query processing. Traditional databases introduce high latency in analysing data when they are streamed continuously and also this operation is done via offline instead of that stream processing engine are designed to process data on Real time. And also the old databases are design in a way that to perform one time queries while stream engine are planned to perform queries continuously when data stream is received.

There are different tools, platforms and architecture for managing the huge amount of data on across cluster of server has been proposed. Map Reduced is developed by Google to perform parallel programming and distributed execution on large clusters. In the Map Reduced the data are divided into smaller chunks, stored, and safely replicated across all nodes of the cluster. Hadoop uses the Map Reduced model to setup of distributed, scalable and parallel applications by decomposing a massive job into smaller tasks and a massive data set into smaller partitions such that each task processes a different partition in parallel using HDFS (Hadoop distributed file system).there is another Hadoop related Project is Apache Mahout It has been presented as a framework implementing parallel machine learning algorithms to perform online analytics. It provides a set of machine learning algorithms that were parallelized and tuned for scalability and Map Reduce execution

A. LITERATURE SURVEY

1. A framework for Internet data real-time processing: a machine-learning approach[1], Mario Di Mauro, Cesario Di Sarnoy,IEEE 2014.In this paper objective is to perform processing and analysis of real time data using Apache Storm and Samoa framework by using the vertical hoeffding tree algorithm. Advantage is structured data can be processed and analyzed .

2. Amazon Kinesis and Apache Storm Building a Real-Time Sliding-Window Dashboard over Streaming Data[2] , Rahul Bhartia,oct 2014,aim of the paper is to do real time analytics of data stream using Amazon kinesis,ApacheStorm,node.js. Advantage of this paper is implementation of real-time visualization of sliding windows analysis over streaming data. Disadvantage of this paper is it requires a decoupled architecture for streaming, processing, storage, and delivery of data.

3. Decision Trees for Mining Data Streams Based on the McDiarmids Bound[3], LeszekRutkowski, Fellow, IEEE, Lena Pietruczuk, PiotrDuda, and MaciejJaworski,IEEE June 2014,objective of paper is to develop decision tress for mining data stream by using The MC diarmid tree algorithm with proper explaination.

4. Comparing data streams using Hamming norms (How to Zero In) [4], Graham Cormode, MayurDatar, PiotrIndyk, and S. Muthukrishnan, IEEEjune2013, Objective of this paper is to do real time data analytics by using powerful data stream comparing architecture like Hamming norm computation model and advantage of this paper is has Given powerful data stream comparing architecture for analysis purpose .Disadvantage of paper is the estimation of data is accurate to within a few percentage points.

5. On the processing time for detection of Skype traffic[5], P.M. Santiago del R o, J. Ramos, J.L. Garc a-Dorado, J. Aracil, A. Cuadra-Sanchez, M. Cutanda- Rodrguez., IEEE 2011,objective is to do real time traffic classification of SKYPE application using NUMA (Non Uniform Memory Access) architecture. Advantage of this paper is has given proper description of how Skype traffic is classified and disadvantage is performance and accuracy of traffic classification should be increased.

6. Data Stream Processing on Real-time Mobile Advertisement[6], Manuel Couceiro, David Suarez, David Manzano, Luis Lafuente, 2011 IEEE. Objective is to create a Real-time Mobile Advertisement System data Stream processing technology on a telecommunications network by using Multi Service Proxy, CEP engine, Data Stream Management System. Advantage is that flexibility in data source handling, disadvantage is that the Request Synchronizer limits the throughput.

7. Efficient Data Streams processing in the Real Time Data Warehouse [7], FiazMajeed, Muhammad SohaibMahmood, MujahidIqbal, IEEE 2010, aim is to do real time data stream processing for data warehouse, by using Token Bucket Technique, Data Streams Approximation techniques, Time Stamping. Advantage of this paper is that streams processor accepts value data contents from the data stream elements for achieving maximum accuracy and disadvantage is that it is necessary to work on data streams extraction and loading according to the requirements of data stream.

8. STORM @TWITTER [8], AnkitToshniwal, SiddarthTaneja, AmitShukla, aimis to use STORM with twitter application. In this paper Data models and execution architecture of STORM are used. In this paper detail description of storm working and main components is given.

9. Real-Time Query Processing for Data Streams[9], Yuan Wei,Objective is to do real-time data stream query model named PQuery.In this paper PQuery model, a real-time data stream management prototype system named Real Time Stream is used.

10. The 8 requirements of real time-stream processing[10], Michael Stonebraker, Ugur Cetintemel and StanZdonik, In this paper Stream Processing Engines, Database Management Systems technologies are used.They have also given the 8 major requirements to perform real time stream processing.

11. Real-time Business Intelligence System Architecture with Stream Mining [11], Yang Hang, Simon Fong, objective is to do real time stream mining by using rt-BI System Architecture.

III. THE PROPOSED SYSTEM

To perform real time data processing and analysis using apache storm and SAMOA framework using machine learning algorithm

Architectural Diagram:-

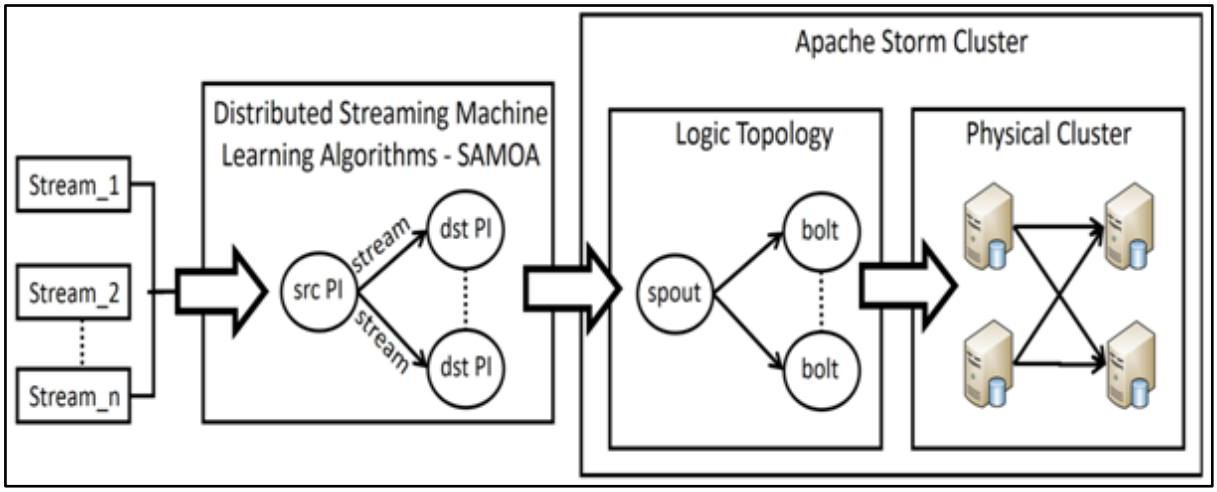

Figure 1: Architectural Diagram

As shown in figure we show the proposed architecture in charge to perform both real-time stream processing and analysis. Stream sources represent the entry points of the architecture in other words the information flow to be analysed. The first component of presented architecture is the framework SAMOA.SAMOA implements the distributed streaming version of widely adopted machine learning algorithms using both Supervised and Unsupervised approach e.g., Hoeffding Tree or k-means-based clustering. Hence it is suitable to perform Real-time data analysis. Three main SAMOA components are: Processing Item, Processor and Stream.

Where Processor implements the business logic of the machine learning algorithm. Processing Item wraps a Processor in order to perform some tasks of a specific machine learning algorithm using the available nodes provided by stream processing engine. The number of Processing Items running in SAMOA environment defines the degree of parallelism allowed. And the Stream describes a connection that allows to exchange data between multiple Processing Item(s).

The second component shown in proposed architecture is a stream processing engine called Apache Storm. A stream is defined as an unbounded sequence of tuples where each tuple can represent data types as integers, floats and so on. The component responsible for feeding messages into the topology for processing is called spout while the component in charge of processing any number of input streams is called bolt. Of course Storm allows to create multiple instances both of spout and bolt. The network generated by connecting the spout(s) with bolt(s) is called Storm topology.

Vertical Hoeffding Tree

Vertical Hoeffding Tree (VHT) is a distributed classifier that uses vertical parallelism on top of the Very Fast Decision Tree (VFDT). VHT is implemented using the SAMOA API. The diagram below shows the implementation:

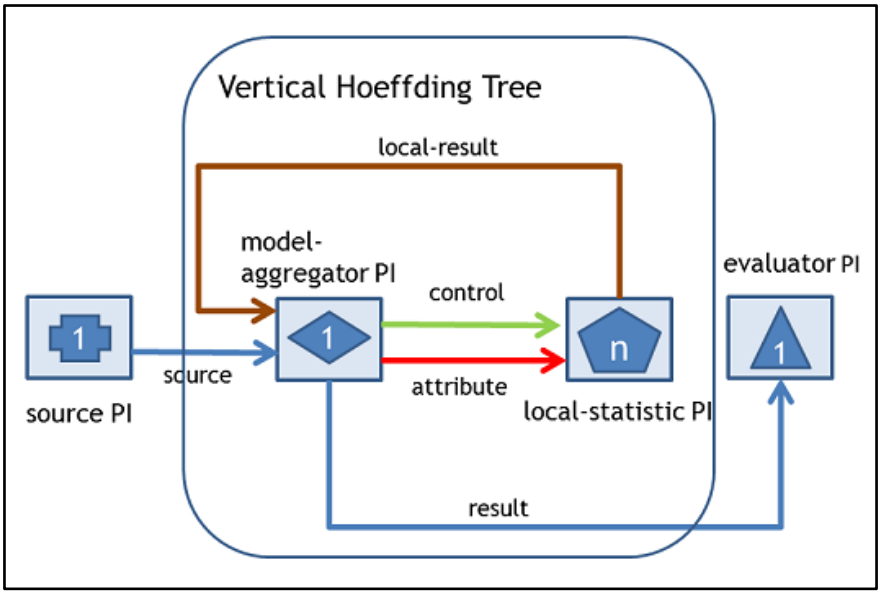

Figure 2: Vertical Hoeffding Tree Algorithm

The source Processor and the evaluator Processor are components of the prequential evaluation task in SAMOA. The model-aggregator Processor contains the decision tree model. It connects to local-statistic Processor via attribute stream and control stream. The model-aggregator Processor splits instances based on attribute and each local-statistic Processor contains local statistic for attributes that assigned to it. The model-aggregator Processor sends the split instances via attribute stream and it sends control messages to ask local-statistic Processor to perform computation via control stream. Users configure n, which is the parallelism level of the algorithm. The model-aggregator Processor sends the classification result via result stream to the evaluator Processor for the corresponding evaluation task or other destination Processor. Incoming instances to the model-aggregator Processor arrive via source stream.

Algorithm

Vertical Hoeffding tree induction algorithm

1. Let HT be a tree with single leaf(the root)
2. **for** all training examples **do**
3. Sort example into leaf l using HT
4. Update sufficient statistics in l
5. Increment n_1,the number of examples seen at l
6. **If** n_1 mod n_{min} = 0 and examples seen at l not all of same class **then**
7. compute $\bar{G}_1(Xi)$ for each attribute
8. Let X_a be attribute with highest $\bar{G}_1$
9. Let X_b be attribute with second highest $\bar{G}_1$
10. Compute Hoeffding bound € = $\sqrt{\frac{R^2}{2N}\ln(1/\delta)}$
11. If $X_a \# X_b$ **and** ($\bar{G}_1(X_a)$-$\bar{G}_2(X_b)$) > € or € < Y) then
12. Replace l with an internal node that splits on X_a
13. for all branches of the splits **do**
14. Add a new Leaf with initialized statistics
15. end for
16. end if
17. end if
18. End for

What are machine learning algorithms?
Machine learning algorithms are algorithms designed for machines so that machine can learn different consequences of events predict output and perform certain actions.

Types of Machine Learning:
i) Supervised Learning where input training data are labelled and accompanied with an expected output per item

ii) Unsupervised Learning where the input is unlabelled and no instance expected outputs exist to evaluate the learned functions**.**

iii) Reinforcement Learning related to the problem of teaching a decision agent on how to perform actions in a specific environment in order to meet a set of end goals according to a reward function. No input or output data is provided**.**

Goals and objectives

Objective:-

To perform real time data stream processing and analysis using storm.

Scope:-

1-**Input**- Input will be unbounded stream of real time data which is to be processed and analyzed.

2-**Output**- Analyzed and processed data.

IV. MATHEMATICAL MODEL

S= {I, O,Ds,Su,Fa,Functions}
I-Input- Unbounded stream of real time data network traffic.

O-Output- To recognize and classify real-time Skype traffic which is hidden within network traffic generated from Personal Computers which are equipped with VOIP application like Skype.

Ds-Identify data structures: Queue for real time data streams.

Su-Success case: - successfully processed and analyzed real time stream of data.

Fa-Failure case:- some problems occurred to process and analyze real time stream of data in some modules of proposed system.

Functions used:-Apache storm, Samoa framework, hoeffding tree algorithm.

V. CONCLUSION

Nowadays, the big data processing and analysis represents challenging task in different fields as telecommunications security, lawful interception and others. This is because traditional technologies for data storage as databases are not suitable to manage a huge amount of streaming data because of standard SQL-like queries work on an off-line environment. Also traditional data mining techniques are designed to perform data analysis in batch mode. This represents a strong limitation to perform the real-time analysis on streamed data.

In this paper we proposed a scalable architecture designed to process and analyze streaming data in real-time fashion. The architecture is based on Storm framework and SAMOA, a set of algorithms machine learning-based for distributed environments. Storm allows to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. In our study we used a specific machine learning algorithm called Vertical Hoeffding Tree i.e., the distributed version of Hoeffding Tree algorithm. In future we plan to analyze the same dataset with different machine learning techniques. In particular, it could be interesting to apply an unsupervised approach on a set of unlabelled data in order to obtain a partition in clusters.

VI. REFERENCES

[1] Mario Di Mauro, Cesario Di Sarnoy, "A framework for Internet data real-time processing: a machine-learning approach", IEEE 2014.

[2] Rahul Bhartia, "Amazon Kinesis and Apache Storm Building a Real-Time Sliding Window Dashboard Over Streaming Data", October 2014.

[3] LeszekRutkowski, Fellow, Lena Pietruczuk, PiotrDuda, and MaciejJaworski, "Decision Trees for Mining Data Streams Based on the McDiarmidsBound", IEEE ,Vol 25 ,No 6, June 2013.

[4] Graham Cormode, Mayur Datar, PiotrIndyk, and S. Muthukrishnan, "Comparing data streams using Hamming norms (How to Zero In)", IEEE June 2013.

[5] P.M. Santiago del R o, J. Ramos, J.L. Garc a-Dorado, J. Aracil, A. CuadraSanchez, M. Cutanda- Rodrguez, "On the processing time for detection of Skype traffic", IEEE 2011.

[6] M. Couceiro, D. Suarez, D. Manzano, and L. Lafuente, "Data stream processing on real-time mobile advertisement: Ericsson research approach," vol. 1, pp. 313–320, 2011.

[7] F. Majeed, M. Mahmood, and M. Iqbal, "Efficient data streams processing in the real time data warehouse," vol. 5, pp. 57–61, 2010.

[8] AnkitToshniwal, SiddarthTaneja, AmitShukla, "STORM @TWITTER".

[9] Yuan Wei, Sang H. Son, John A. Stankovic , "RTSTREAM: Real-Time Query Processing for Data Streams".

[10] Michael Stonebraker, Uuretintemel, Stan Zdonik, "The 8 requirements of real time-stream processing".

[11] Saeed Shahrivari, "Beyond Batch Processing: Towards Real-Time and Streaming Big Data", *Computers* 2014.

[12] Yang Hang, Simon Fong, "Real-time Business Intelligence System Architecture with Stream Mining".

[13] S. Babu and J. Widom, "Continuous queries over data streams",vol. 30, no. 3, pp. 109–120, 2001.

[14] Arinto Murdopo, Antonio Severien, Gianmarco De Francisci Morales,Albert Bifet, "*Samoa Developer's Guide"*.

[15] Bifet and R. Kirkby, "Data stream mining a practical approach", 2009.

[16] Zliobaite, A. Bifet, J. Read, B. Pfahringer, and G. Holmes, "Evaluation methods and decision theory for classification of streaming data with temporal dependence," 2014

