Systematic Survey on Evolution of Machine Learning for Big Data

[1]**R.Swathi**
Research Scholor
Dept.of.CSE
S.V.University,Tirupati
Swathi.mani08@gmail.com

[2]**Dr. R. Seshadri**
Professor
Dept.of.CSE
S.V.University,Tirupati
ravalaseshadri@gmail.com

Abstract: Advanced data processing techniques with massive and high dimensional data, dramatically increased storage capability and complex data formats cause the Big data. In this realm, to solve the various issues of computational time to extract the valuable information without sensitive information loss, the Big data need modern advanced technologies and/or techniques. To overcome those problems, a novel and rapidly expanding research domain have been recently proposed: Machine Learning. Generally Machine learning algorithms have been considered to learn and find useful and valuable information from large volumes of data. The goal of this paper is to build the effective universal architecture which defines the quality and durability of a system software. The paper intends to add to the Systematic Literature Review (SLR) to help specialists who are endeavoring to contribute around there. The principle target of this audit is to deliberately recognize and dissect the as of late distributed research subjects identified with Machine learning in big data as to research action, utilized apparatuses and systems, proposed methodologies and spaces. The connected strategy in SLR depends on three chose electronic databases proposed by (Kitchenham and Charters, 2007).

Keywords: Machine Learning, Big Data, Software Architecture

I. INTRODUCTION

The expression "Big data" has turned into a popular expression and it for the most part alludes to information that is too huge or excessively mind boggling, making it impossible to handle on a solitary machine. As per International Data Corporation's yearly Digital Universe think about [1], the measure of information on our planet is set to reach 44 zettabytes by 2020 which would be more than ten times than it was in 2013. Numerous businesses are as yet creating information too vast to be in any way prepared effectively utilizing customary strategies. Ancestry.com, for instance, stores billions of records totaling around 10 petabytes of information [2].

As we move into an all the more innovatively propelled work process, the requirement for greater information stockpiling in an advantageous stage turned into the request of numerous huge organizations. Big Data is defined by using the following data characteristics: volume, velocity and variety appeared in 2001 when Gartner analyst Doug Laney used it to help identify key dimensions of big data. IBM and others added veracity. Then Viability, value, variability, and even visualization got included. It implies that some point in time, when the volume, variety and velocity of the information are expanded, the present strategies and advances will be unable to deal with capacity and preparing of the information. Regardless of whether the information is structured or unstructured information. Its about understanding what the information truly looks like and how you will utilize it. For IT organization, the most interesting thing about Big Data isn't that it's tremendous anyway it is colossal, and growing exponentially; IBM gauges that 90% of the information on the planet has been created over the most recent 2 years [3].

Web organizations have been advanced effectively with big data speculation ventures, because of the colossal measure of constant information that they are taking care of. While Google was ordering a million pages for a couple of million inquiries in 1998, it was ordering more than a trillion pages ten years after the fact, for more than 3.5 billion pursuit questions played out each day, or 1.2 trillion ventures a year, as indicated by the following site, Internetlivestats.com. Moreover, Facebook is dealing with around a billion substance data inquiries consistently, and Netflix has aggregated billions of watcher appraisals, with individuals seeking and including many things consistently [4].

While 2012 has been the time of Big information advances and the year 2013 is turning into the zone of Big Data Analytics. Assembling and keeping up vast accumulations of information is a certain something, yet extricating the helpful data from these accumulations is much all the more difficult. Big Data not simply changes the mechanical assemblies one can use for farsighted examination, it also changes our entire perspective about data extraction and interpretation.

As the cost of the capacity of information has tumbled

978-1-5386-2745-7/17/$31.00 ©2017 IEEE

down and the development of the superior of PCs have turned out to be all the more broadly and effortlessly open, the development of machine learning (ML) has seen into a large group of enterprises including account, business, diversion, medicinal services[5][6], and law implementation. As hypothetical research is getting a handle on into down to earth errands, the instruments of machine learning are progressively observed and incorporated to numerous business operations.The utilization of Machine Learning has spread quickly all through the Computer Science, Statistics and past. Machine Learning is utilized as a part of Spam Filters, Web seek, credit scoring, extortion recognition, stock exchanging, medicate outline and in numerous different applications [7]. "Advancing execution paradigm by utilizing illustration information and past experience" by E. Alpaydin [8]. Machine Learning is "Modifying a PC to enhance execution criteria by utilizing case information (or) past experience". Information is a main part in Machine Learning and the learning calculations are utilized to find the learning from the information. Quality and amount of the dataset will influence the learning and expectation execution. A current report from the Global Institute affirms that Machine Learning will be the driver of the following huge flood of advancements [9].

As indicated by big data setting, the term 'Machine learning' happens mostly in two parts:

(i) Machine learning as enabling technology: Due to high volume, velocity and variety of the data, traditional modelling or manual inspection becomes impossible. The methods of Machine learning together with their scientific foundation offer conceivable outcomes to extricate precise data from such information. So to deal with huge information(Big data), Machine learning techniques are required.[10].

(ii) Machine learning techniques for processing of Big data: Big Data strengths machine learning exploration to venture out of the established setting of equivalently limited learning undertakings and i.i.d[independent and identically distributed]. Information which are accessible before preparing. Out of ten famous machine learning strategies just a couple of Machine learning systems are readily applicable for big data according to this article [11]; Machine learning research needs to confront the difficulties forced by Big data investigation.

Lamentably, Big Data is, practically by definition, past the limit or persistence of any individual to physically examine and dissect utilizing ordinary apparatuses and questions. On the off chance that we need to interpret Big Data and stay aware of its exponential development, we have to instruct computers to think more like people. Specialists have been considering and executing machine learning algorithms for a considerable length of time that can accomplish human-like learning from information to perform prediction, classification and clustering in barely characterized conditions, and late advances in accessible computational power and additionally the algorithms themselves are making these strategies priceless for removing valuable information from piles of crude information[12]. Thus, the field of machine learning is ending up plainly firmly connected with Big Data.

Machine learning and Big data are becoming an active area of research. In the design and development phases of a software system , the Architecture plays a Vital role. As of now there is little contribution to systematic literature review for mapping software architectures and Machine Learning with Big data.

Research questions:

The main intention of this survey paper is, to find and interpret the published literature related to software architecture of Machine learning with big data environment. This is additionally itemized in the accompanying examination questions:

RQ1 How much activity was carried out recently?

RQ2 What inquire about topics are being tended in Machine learning and big data?

RQ3 What are the different tools, Frameworks and technologies that were used?

RQ4 What are the application domains?

II.RELATED WORK

from that point forward the quantity of the domain has increased and we focused on a systematic literature review from 2013 to 2016 from all repositories with different search strings.

Search Procedure

We performed our search on scientific electronic databases which includes high impact factor conferences, journals and articles. The search process follows the guidelines suggested by (Kitchenham and Charters, 2007). Refer to Table 1 for a list of selected electronic databases.

Electronic database	**URL**
IEEE	http://ieeexplore.ieee.org

Science direct	http://www.sciencedirect.com
Springer	http://www.springerlink.com

Table 1: Selected electronic databases

Search string

The Search string is utilized to catch all outcomes identified with Machine learning with big data. The inquiry string utilized as a part of all databases (Machine learning and big data).

In order to include relevant publications in our review, we defined selection criteria and based on that we performed inclusion and exclusion of published literature. We selected papers published in peer review conferences, journals from 2013 to 2016. We selected papers that are relevant to our research questions. We excluded papers that are not related to software architecture of Machine learning with big data.

The process of selection of conferences and Journals was conducted as follows:

1.Search is performed on every database and references are spared in list of sources document.
2. The Scholar peruses all titles and modified works of the distributions and noticed the substance according to the criteria.
3.The scholar classifies the conferences and journals according to type, topic, and domain.

The data analysis is represented as below:

1. The databases and number of query results
2. The number of important distributions every year regarding scenes
3. The graph that will show publication of journals and conferences, which are generated from the final results

4. A Detailed selection process performed on selected databases.

III. OUTCOMES

The conveyance of outcomes for every database is recorded in Table 2.

Database	Results (%)
IEEE	30%
Science direct	50%
Springer	20%

Table 2: Number of results per database

RQ1 How much activity was carried out recently?

We plotted graph for a number of relevant publications per databases in Figure 2, and per year in Figure 3. For past 4 years, there is no software architecture was founded on machine learning for big data. However, the research publications were focused mainly on deep learning, support vector machines(SVM), extreme learning machine, quality, multi-tenancy, frameworks, security and application domains. The papers mainly focused on horizontal research rather than a vertical approach. Figure 3 shows numbers of papers published from 2013 to 2016. In the year of 2016 publication were less, because the search date was taken up to march 2016.

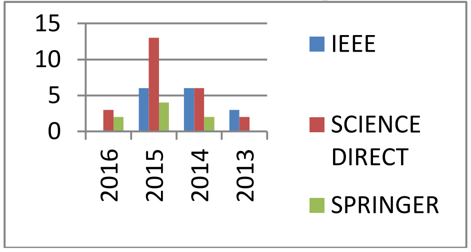

Fig:3 list of publications per year

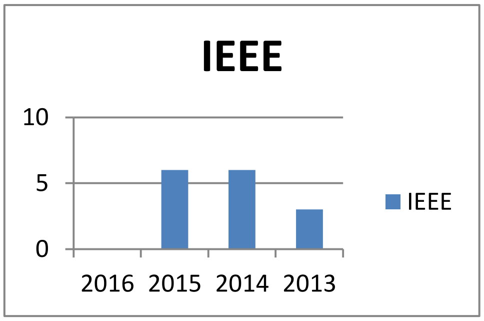

Fig:3a list of IEEE publication per year

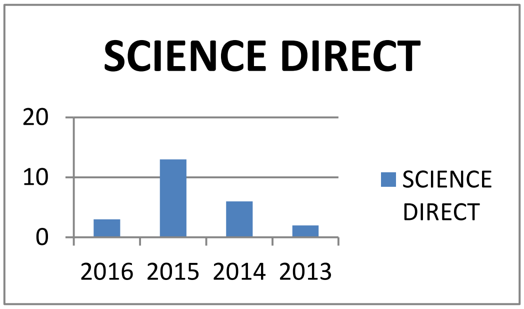

Fig:3b list of Science Direct publication per year

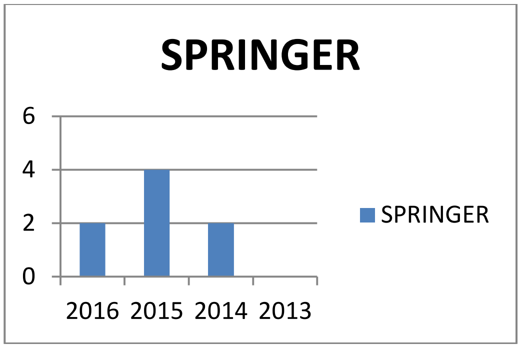

Fig:3c list of Springer publications per year

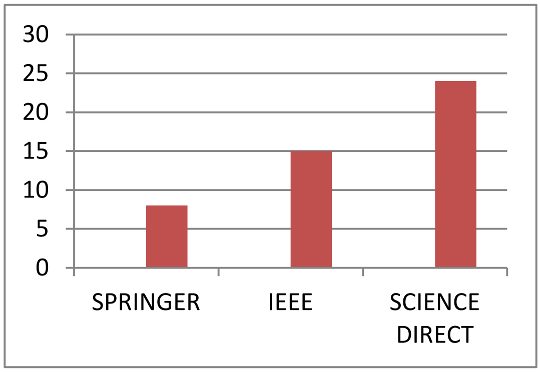

Fig:4 Total number of included publications per database

RQ2 What inquire about topics are being tended in machine learning and big data?

To identify the research topics that are focused in Machine Learning for Big Data architectures, we have created a weighted topic and taxonomy in Figure 5, which are derived from titles, keywords and topics are shown in Tables 4–7. The research topics are broadly classified into classification, clustering, regression, deep learning, neural networks, ensemble learning, extreme learning machine, security, Big Data tools and techniques, Machine Learning approaches for Big Data, Data streaming methodology, Different frameworks for machine learning and different application domains.

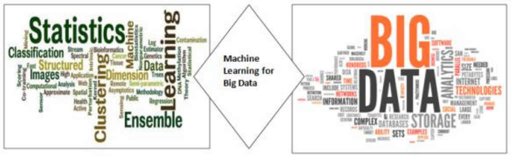

Fig: 5 Weighted research topics of Machine learning for Big data.

RQ3 What are the different tools, Frameworks and technologies that were used?

Mahout

Mahout is a one of the standout instrument for Machine Learning. Mahout is updated in the year of 2015. With this discharged adaptation , the attention is presently on a math domain called Samsara, which incorporates linear algebra, factual operations, and data structures. The calculations incorporated into Mahout fundamentally concentrates on order, grouping and community separating .Currently offered Classification algorithms in Mahout are Logistic Regression, Naïve Bayes, Random Forest, Hidden Markov Models, and Multilayer Perceptron. The Mahout's library utilizes the mainstream k-Means Clustering calculation, including the customary k-Means, and Streaming k-Means. Mahout is likely the best-known structure for collective sifting devices.

SPARK MLlib

MLlib is created as a major aspect of the Apache Spark extend. It consequently gets tried and refreshed with each Spark release. MLlib contains numerous algorithms and utilities. It covers the algorithms used in Mahout, and also adds regression models which are also not included in Mahout and also included the algorithms for topic modeling and frequent pattern mining. Extra tools incorporate dimensionality reduction, feature extraction and change, optimization, and essential measurements.

MLbase

In spite of the truth fact that it is not yet at present accessible, but rather there has been continuous innovative work at Berkeley's AMP lab on a stage called MLbase, which wraps MLlib, Spark, and different tasks to make machine learning on informational collections of all sizes accessible to a more extensive scope of clients.

H2O

Out of the majority of the apparatuses which are talked about in this paper, H2O is the main device that can be viewed as an item, instead of a venture. While they offer a wander variant with two levels of support, practically their offerings are available open source as well and can be used without the purchase of a permit.The most prominent components of this item are it gives a graphical UI (GUI), and various apparatuses for profound neural systems. There is another organization offering open source usage for profound learning,

SAMOA

SAMOA[Scalable Advanced Massive Online Analysis], a platform for machine learning from streaming data. It is an adaptable structure that can be run locally or on one of a few stream processing engines, like Storm, S4, and Samza.

Processing engines	Execution model	Supported language	Associated tools
Map reduce	Batch	Java	Mahout
Spark	Batch and Stream	Java, Python, R, Scala	MLlib, Mahout,H2O, Sim Sql,graphlib,giraph
Flink	Batch and Stream	Java, Scala	Flink-ML, SAMOA
Storm	Stream	Any	SAMOA
H2O	Batch	Java, Python, R, Scala	MLlib, Mahout,H2O

Table 3 : Machine Learning processing tools

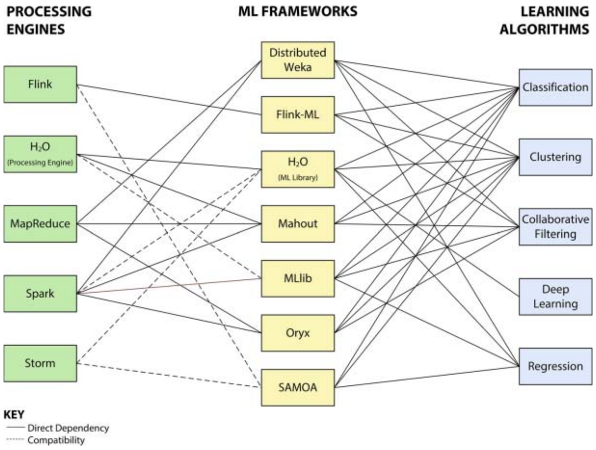

Fig:6 Machine Learning Frameworks and Learning Algorithms.

IV.DISCUSSION

This area gives a dialog of the outcomes and constraints for this review.

Conclusions

In the wake of incorporating information gathered through this SLR, we watched number of research patterns in Deep Learning, Extreme Learning Machine, Neural Networks, Different classification techniques and different approaches for Machine Learning with Big Data. The maturity of Extreme Learning Machines is still in its early stages. However, we can find a clear growth in maturity and researchers need to focus on a vertical approach. More case studies will improve the confidence of researchers and practitioners regarding the benefits of Machine Learning for Big Data Architectures.

Threats to legitimacy

In this paper the SLR gives a study of software architectures for Machine Learning for Big Data. Though the results of reviews are reliable, they have potential threats to validity. The main threats of this review are the bias in our selection of studies to be included, data extraction from different sources and data synthesis. In order to mitigate potential threats to legitimacy, we define a research protocol, which contains research questions, inclusion/exclusion criteria, research strategy and followed the guidelines of a systematic review (Kitchenham and Charters, 2007). We scanned for basic terms and consolidated them in our inquiry string, which decreases bias and increases search work. In order to mitigate reliability threat several researchers are involved in reviewing the included papers to achieve high validity of the study.

V. CONCLUSIONS

The objective of this study was to deeply understanding the existing research on Machine Learning for Big Data and associated topics that consider developing an assemblage of information. We considered 47 out of 300 reviewed publications significant with respect to research protocols, research question and categorized them according to the research area. On that premise, we gave scientific categorization to speaking to research territories, application space, tools and technologies. We identified unexplored areas by synthesizing collected data, making those available for future research. We observed vast interests towards resource management, service management and security areas.

We also observed a lack of tools and also lack of evidence for architectural adaption to develop common and secure architecture. This field is still in its initial stages and to develop, Machine Learning for Big Data researchers should come together by proposing a common research agenda.

REFERENCES

[1] International Data Corporation. Digital Universe Study. 2014. http://www.emc.com/leadership/digital-universe/index.htm. Accessed 1 Jun 2015.

[2] Ancestry.com Fact Sheet. http://corporate.ancestry.com/press/company-facts/. Accessed 1 Jun 2015.

[3]http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html.

[4] Amatriain X (2013) Beyond Data: from user information to business value through personalized recommendations and consumer science, CIKM'13. San Francisco, CA, USA

[5] Praveen Kumar Rajendran, A. Asbern, K. Manoj Kumar, M. Rajesh, R. Abhilash ,"Implementation and Analysis of MapReduce on Biomedical Big Data," Indian Journal of Science and Technology (IJST), ISSN/E-ISSN: 0974-6846 / 0974-5645, Vol. 9, Issue. 31, pp. 1-6, August 2016.

[6] K. Manoj Kumar, Tejasree S, S. Swarnalatha, "Effective implementation of data segregation & extraction using big data in E - health insurance as a service," 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 2016, pp. 1-5.

[7] Pedro Domingos, A few useful things to know about Machine Learning Communications of the ACM, Vol. 55 No. 10, Pages 78-87.

[8] E. Alpaydin, Introduction to Machine Learning.

[9] Big data: the next frontier for innovation, completion and productivity. Technical report, MCkinsey Global institute 2011.

[10] Barbara Hammer, Haibo He , and Thomas Martinetz, Learning and modelling big data, ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014.

[11] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. Knowl. Inf. Syst., 14(1):1–37, Dec. 2007.

[12] Sara Landset, Taghi M. Khoshgoftaar, aron N. Richter and Tawfiq Hasanin,| A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal of Big Data 2015, DOI: 10.1186/s40537-015-0032-1, Published: 5 November 2015.