Investigating the Inclinations of Research and Practices in Hadoop: A Systematic Review

Megha Sharma
Amity University
U.P, India
megha.rsystems@gmail.com

Nitasha Hasteer
Amity University
U.P, India
nhaster@amity.edu

Anupriya Tuli
Amity University
U.P, India
anupriyatuli6@gmail.com

Abhay Bansal
Amity University
U.P, India
abansal1@amity.edu

Abstract— **Hadoop Technology is commonly being used to manage Big Data projects. The research done in this domain has increased over the past few years. In order to investigate whether Hadoop Technology is the next big thing we have conducted systematic literature review on the usage of Hadoop technology in Big Data projects. This work outlines the usage of Hadoop technology from year 2010 to 2013 using publications of conference proceedings, journals and magazines of IEEEXplore as primary studies. With the help of search strategy followed, we identified 690 research papers out of which 296 were identified as relevant papers. We observed that 166 studies by different authors were utilizing Hadoop Technology as tool in Big Data projects whereas only 130 studies were purely on Hadoop. This review will assist researchers to understand the present scenario of research in Big Data projects using Hadoop technology.**

Keywords— **Big Data, Hadoop technology, Systematic Literature Review, Map Reduce, HDFS**

I INTRODUCTION

Advancement in technology has contributed in generation of huge amount of data which is collectively known as 'Big Data'. It was in 1998, that the term Big Data was first used in a Silicon Graphics (SGI) slide deck, by John Mashey in his work, "Big Data and Next Wave of InfraStress [1]". Also, the first book on 'Big Data' by Weiss and Indrukya appeared in the same year [2].

Management of Big data forms the backbone of modern science and business as the huge amount of data is generated from emails, images, online transactions, videos, click stream, audios, search queries, posts, logs, health records, sensors, social networking interactions, mobile phones and science data. It becomes very difficult to store, capture, manage, analyze and share such a large amount of data via typical database software tools. Today many technologies are being used to manage and analyze Big Data.

One of the solutions to manage Big Data is Hadoop, which is an open-source distributed computing software developed by Apache Software foundation. Initially Hadoop had its source in Apache Nutch, one of the sub items of Apache. It was 2006, when Hadoop became independent software and got its name "Hadoop". Important components of Hadoop include MapReduce, HDFS and HBase [3].

This paper investigates the research direction in Big Data projects using Hadoop Technology since 2010. The systematic review is carried out by identification of research, selection of studies by various authors, deciding the inclusion & exclusion criteria and analyzing the amount of publications done in this domain over the time period of year 2010 to 2013. This paper limits its scope to publications done only in IEEE Digital Library (IEEE Xplore).

The next section of the paper includes literature review. Section III, discusses about the research methodology used to extract the relevant data for our systematic review. Section IV, comprises of result set, followed by conclusion in Section V.

II LITERATURE REVIEW

Nowadays, Hadoop technology is being widely used in various fields, Internet being one of them. H. Lu et.al. [3], have stated in their work that major companies like Yahoo, Facebook and Baidu are using Hadoop for web searching, analysis of data and web data mining. Besides Internet industry, public sector, telecom industry, finance industry and many other organizations have also started paying attention to applications based on Hadoop. Their work also includes the survey result of application trend of Hadoop, conducted by IT128 web site over the period of 2011-2012 which shows that Hadoop has become famous among computer hardware, computer software and network equipment manufacturers.

Kai Ren, et al. [5], performed analysis on Hadoop usage in the year 2012. Their work suggested that Hadoop usage for academic research was still in adolescence stage. Also, Hadoop features, tool and extensions were underused as many applications were responsive to non-Hadoop solutions then.

978-1-4799-4236-7/14/$31.00 ©2014 IEEE

III RESEARCH METHODOLOGY

A Systematic Literature Review (SLR) or Systematic Review (SR) is a form of secondary study. It is a method used for identification and evaluation of available research work done by various authors, relevant to a specific domain of interest or formulated research question. To carry out reported study, we have used Kitchenham and Charters guidelines [4]. In accordance with these guidelines, we have followed the following steps to conduct this research:

- *Selection of primary studies:* The individual work or studies by various authors which are used in a systematic review are called primary studies. SLR's goal is to find out innumerable primary studies related to selected domain. After collecting relevant data, there is need to find out the actual relevant data.
- *The assessment of quality:* It is important to evaluate the quality of collected data. This is carried out by determining the strength of each individual research papers collected (whether they play crucial role in the area of selected research domain or not) and also to give more detailed criteria of inclusion/exclusion.
- *Extraction of relevant data:* Data extraction should be conducted with the help of two or more expertise. Once the data has been extracted by individual researcher, either additional independent personnel or the same researchers should compare the extracted data in order to collect the relevant data.
- *Data synthesis*: It is carried out by keeping in mind the research questions. The extracted data from the primary studies is then tabulated in a consistent manner. It is very crucial to calculate whether the formulated results from the extracted data are consistent with each other or not.
- *Result Formulation:* In last step, formulation of the final result is carried out. This final result set provides assistance in answering the formulated research questions.

Aim of this work is to perform SLR to answer formulated questions, to seek the research direction in the big data projects. The formulated questions are:

RQ1. Is Hadoop technology the next big thing?

RQ2. How frequently Hadoop is being implemented as a tool in Big Data projects?

RQ3. What all are sectors/ areas in which Hadoop Technology has been successful in providing solution from 2010-2013?

3.1 Selection of Data Source & Primary Studies

We have limited our search to research papers that are available online and written in English only. Only Manual search was performed to collect primary studies across conference proceedings, journals and magazines of IEEEXplore (www.ieeexplore.ieee.org/Xplore/).

In stage 1, different keywords or search terms were used to search the digital database. These keywords are:

- Hadoop and Big Data
- Hadoop Map Reduce Framework
- HDFS
- Hadoop Distributed File System
- Big Data and Hadoop Technology

On performing manual search, we got 690 'hits' with only 451 non-redundant studies from 2010 to 2013. Out of 690 studies, 68 were found in 2010, 88 were found in 2011, 201 were found in 2012 and 333 were found in 2013. The process of how we reviewed and identified number of research papers or studies at particular stage is shown in fig. 1.

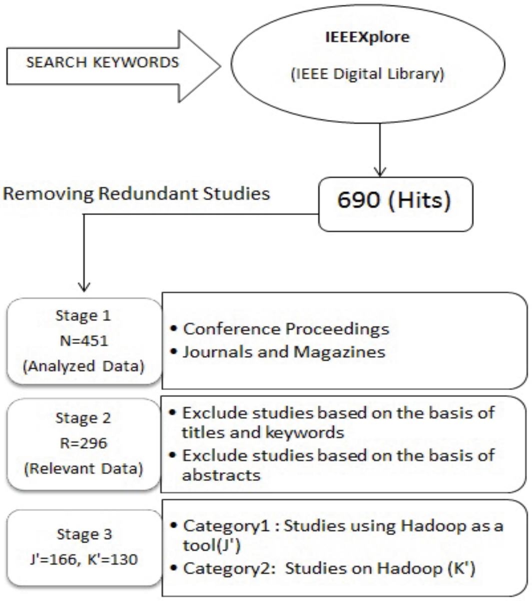
SEARCH KEYWORDS
IEEEXplore
(IEEE Digital Library)
690 (Hits)
Removing Redundant Studies
Stage 1
N=451
(Analyzed Data)
• Conference Proceedings
• Journals and Magazines
Stage 2
R=296
(Relevant Data)
• Exclude studies based on the basis of titles and keywords
• Exclude studies based on the basis of abstracts
Stage 3
J'=166, K'=130
• Category1 : Studies using Hadoop as a tool(J')
• Category2: Studies on Hadoop (K')

Fig. 1. Research Methodology

3.2 Extraction of Relevant Data using Inclusion/Exclusion Criteria

After completion of search from year 2010 to 2013 in stage 1, we made Notes (doc files) where we stored all the unduplicated citations relevant to the study (N=451). To decide on inclusion/exclusion of studies, we maintained Notes for each year (2010-2013) separately. These Notes included title and abstract of all the unduplicated studies (N). At stage 2, on the basis of their title and abstract, we excluded those studies from the Notes which were out of scope of our study. On reviewing all the Notes, we identified 296 relevant studies (R). All the relevant studies used for this review are presented in Appendix_738, available at: [https://drive.google.com/file/d/0B66VIdl-khXfYzIzRjRnbUdrblk/edit?usp=sharing]

3.3 Data Synthesis:

The following four graphs in fig. 2 depict data synthesis from the year 2010 to 2013 respectively. It can be clearly seen from these graphs that the number of relevant studies (R) has increased in past four years. We found 15 relevant studies in 2010, 44 in 2011, 89 in 2012 and 148 in 2013 respectively.

3.4 Data Classification:

At stage 3, we found that some studies were purely on Hadoop technology such as Hadoop architecture, Hadoop performance, Hadoop Distributed File System (HDFS) and more. Whereas others were using Hadoop technology as implementation tool, for developing new framework, for creating new algorithms and more. We classified the set of relevant studies (R) as:

- *Category 1:* It includes studies using Hadoop as a tool, which are defined in set J'.
- *Category 2:* It includes studies on Hadoop Technology, which are defined in set K'.

The Table 1 shows year vise categorization of relevant data into above mentioned two categories. The following are mathematical equations (1) and (2), used to represent the data in the table.

- Studies under Category 1= Set J' (J1, J2,…., Jm)

$$\text{Jm} = \sum_{m=1}^{166} \qquad (1)$$

- Studies under Category 2= Set K' (K1, K2,…., Kn)

$$\text{Kn} = \sum_{n=1}^{130} \qquad (2)$$

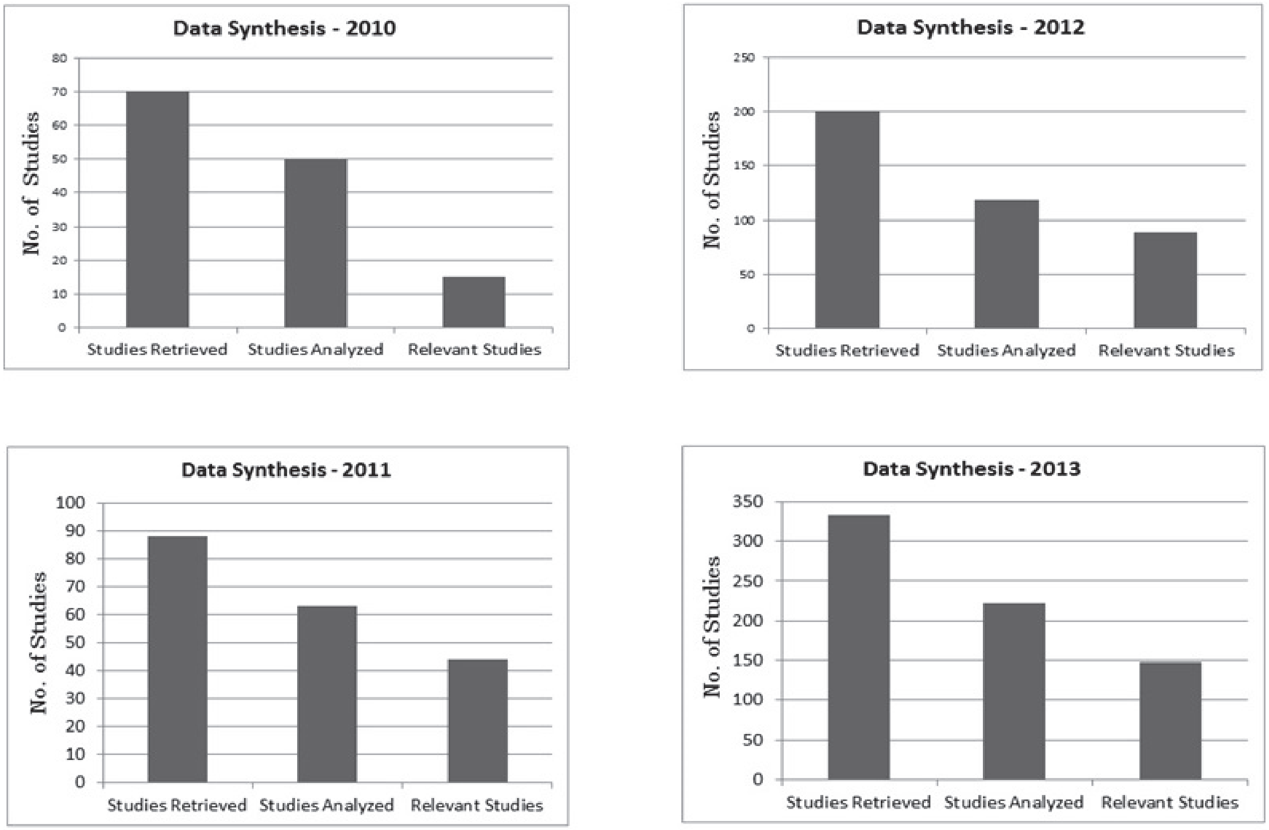

Fig. 2. Graphical representation of Data Synthesis from the year 2010 to 2013 respectively

TABLE 1. Classification of relevant data (R)

	J'		K'	
	Studies	Frequency	Studies	Frequency
2010	J1- J8	8	K1- K7	7
2011	J9 - J35	27	K8 - K24	17
2012	J36 - J83	48	K25 - K65	41
2013	J84 - J166	83	K66 - K130	65
Total	166		130	

IV RESULT

Our aim was to give a synthesized overview on the trend of the research publications of Hadoop technology. After a detailed analysis of the relevant studies, we have found the solutions to the formulated research questions.

RQ1. *Is Hadoop technology the next big thing?*

The analysis of extracted data shows that Hadoop Technology has drawn interest of various researchers in past four years. We can clearly see from fig. 3 that the number of publication of research papers in conference proceedings, journals and magazines of IEEEXplore has significantly increased from the year 2010 to 2013.

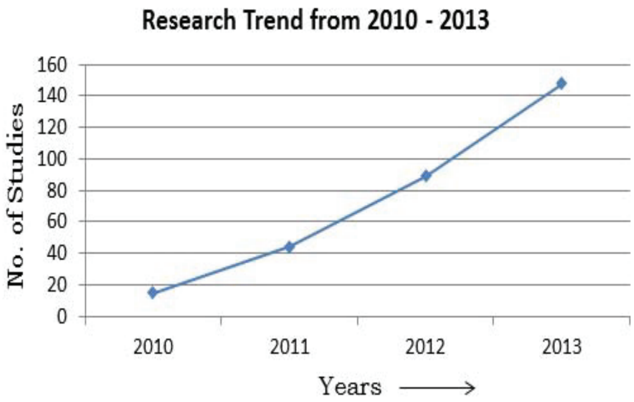

Fig. 3. Publication Trend (2010-2013) for Hadoop Technology

RQ2. *How frequently Hadoop is being implemented as a tool in Big Data projects?*

While further reviewing relevant studies (R=296), we found that only 43.77% of studies were purely on Hadoop Technology as shown in fig. 4. This outlines the fact that most of the Big Data projects are adapting Hadoop Technology as a tool to either manage Big Data or for proposing new models.

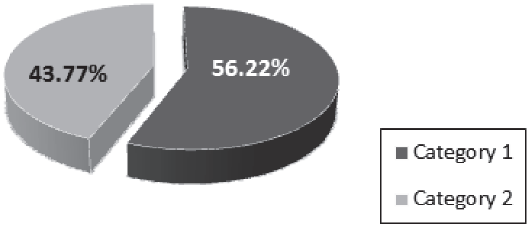

Fig. 4. Representation of classified data

RQ3. *What all are sectors/ areas in which Hadoop Technology has been successful in providing solution from 2010-2013?*

We found that Hadoop technology is becoming popular in the following areas:

1. Used as a tool in order to provide framework for cloud computing
2. In Big Data analysis (storage, biological data, road traffic, travel and tourism, enterprise data, citizen's info)
3. In implementing map reduce algorithms for providing solutions to various problems of handling large amount of data.
4. Internet data management (storage, load balancing).
5. In proposing new models by using HDFS.

V CONCLUSION

On performing the systematic review, we found that trend of research and practices in Hadoop has increased from past few years (2010-2013), making Hadoop the next big thing. Though there is increase in number of studies published in IEEEXplore, most of the studies talk about using Hadoop as a tool to either solve the Big Data problem at hand or to propose a new solution for Big Data projects. This work also classifies the areas where Hadoop technology was able to provide a successful solution in past four years.

The information provided by the result of this review would prove beneficial to the researchers and practitioners, to know the present scenario of Hadoop technology in Big Data projects. This work can be extended by considering more journals like Elsevier and Springer.

REFERENCES

[1] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012

[2] S. M. Weiss and N. Indurkhya. Predictive data mining:a practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998

[3] H. Lu et al., “Research on Hadoop Cloud Computing Model and its Applications”, in Networking and Distributed Computing (ICNDC), 2012, Third IEEE International Conference, doi: 10.1109/ICNDC.2012.22

[4] B. Kitchenham and S. Charters, “Guidelines for Performing Systematic Literature Reviews in Software Engineering” EBSE Technical Report, EBSE-2007-01, 2007

[5] Kai Ren, Garth Gibson, Yong Chul Kwon, Magdalena Balazinska and Bill Howe, “Abstract: Hadoop's Adolescence; A Comparative Workloads Analysis from Three Research Clusters”, High Performance Computing, Networking, Storage and Analysis (SCC), SC Companion, pp. 1452, 2012. doi:10.1109/SC.Companion. 2012.253