Systematic Mapping Study of Big Data Analytics Tools and Techniques

Sheena Rewari¹, Dr. Williamjeet Singh²

¹Research scholar, Punjabi University, Patiala, Punjab, (India)
²Assistant Professor, UCOE Punjabi University, Patiala, Punjab, (India)

ABSTRACT

Big data is a huge dataset demonstrating the aspects of volume, velocity, variety, veracity, validity, value, variability and vagueness in an OR relationship. More than sufficient new insights are discovered while dealing with big data. There are many software and hardware solutions available in the technology prospect that facilitate grabbing, saving and consecutive analysis of big data. Hadoop and its correlated high-tech solutions are one of them.

Keywords: Big data analytics, tools, Hadoop, challenges, application areas.

I. INTRODUCTION

With all the devices available today to collect data such as RFID readers, microphones, cameras, sensors and so on, we are seeing an explosion in data being collected worldwide. Big data is a term that is used to describe these large number of datasets that may be unstructured and grow so quickly that it is impossible to manage with a regular database or statistical tools. [1]

1.1. Big Data Adoption

The big data adoption process goes through a number of phases are given in figure 1.

![Figure 1_big data adoption](image)

1.2. Big Data Analytics

Big data analytics is the process of analyzing and acquiring intelligence from data to gain meaningful patterns in big data. Big data analytics helps a business acknowledge the needs of a customer so that businesses can expand their customer base and maintain the existing ones with admissible offerings of their product. [2]

The three supreme types of analytics are shown in figure 2.
1.3. Motivations and Objectives

This study is carried through in order to evaluate the existing big data analytics tools and techniques which are more applicable. This analysis is necessary to make it possible to know which class of big data analytics tools, techniques and application areas have been shielded in past research and helps to identify gaps.

This study plans at periodically reviewing the big data analytics tools, techniques and application areas used in existing studies. The results may help the researchers to get an outline of the state of big data analytics and the feature the research gaps.

This paper is structured as follows. Section 2 describes the research methodology used in this study. Section 3 gives the classification of big data analytics research papers considering the following criteria: (1) application domains; (2) techniques used; and (3) platform/framework used. Section 4 presents the mapping of studies. Section 5 discusses the paper. Section 6 presents the conclusions and future directions.

II. RESEARCH METHODOLOGY

The research methodology is composed of two stages. The first stage involves the research of works related to big data analytics. The second stage is concerned with establishing a classification scheme described in Section 3.

1) RQ.1: What are the different categories of tools and techniques in the area of big data analytics? Section 3.2 of this paper answers this question.

2) RQ.2: What are the application areas in big data analytics research?
The section 3.1 describes the hot big data analytics application areas with their future directions.

3) RQ.3: What are the various platforms/framework used in big data analytics?
To answer this question table in section 3.3 describes the platform involved. The research is initialized with these queries and then follows the steps described.

2.1. Search Strategy and Screening

Sources of information

To increase the probability of relevant articles, a set of appropriate databases must be chosen. For this review, the major databases of electronic journals are searched. The digital libraries considered are:

- IEEE Explore (http://ieeexplore.ieee.org)
- ACM Digital Library (www.acm.org/dl)
- Science Direct (www.sciencedirect.com)
- Springer (www.springerlink.com) Additional sources

IBM’s Big Data University website: bigdatauniversity.com

Technical Reports.

2.2. Study Selection

Research papers published by journals, conference proceedings and workshops are thought to be worthy and reliable. Keyword based search is employed to select the most relevant works. The keywords used are big data analytics, big data analytics tools and techniques and big data analytics applications.

The Table 1 shows the defined search strategy and number of results obtained. From the returned studies, firstly irrelevant studies are excluded on the basis of title. Certain studies could not be estimated from the title, then their abstract is considered. If even abstract is not evident then after reading the full text of papers, irrelevant studies are excluded.

<table>
<thead>
<tr>
<th>S. no.</th>
<th>E-resource</th>
<th>Studies returned</th>
<th>Excluded</th>
<th>Search string</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>based on</td>
<td>based on</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>title</td>
<td>abstract</td>
</tr>
<tr>
<td>1.</td>
<td>ieeexplore.ieee.org</td>
<td>115</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td><a href="http://www.acm.org">www.acm.org</a></td>
<td>136</td>
<td>70</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td><a href="http://www.sciencedirect.com">www.sciencedirect.com</a></td>
<td>133</td>
<td>79</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td><a href="http://www.springerlink.com">www.springerlink.com</a></td>
<td>146</td>
<td>126</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3. Establishing a Classification Scheme

The selection process resulted in 359 papers selected from five different digital libraries as on 27th January 2017.
III. CLASSIFICATION METHOD

The research papers are classified by giving consideration to following criteria: (1) application areas of big data analytics, (2) techniques used, and (3) platform used.

3.1. Classification based on application areas

On the basis of research papers studied, the operative big data analytics research fields are explained in the table below:

<table>
<thead>
<tr>
<th>APPLICATION AREA</th>
<th>ROLE OF BIG DATA ANALYTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEALTHCARE</td>
<td>Provides a comprehensive view of treatment delivery for meeting future needs.</td>
</tr>
</tbody>
</table>
3.2. Classification based on techniques used

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSOCIATION RULE MINING [2]</td>
<td>It looks for relationship between variables or objects.</td>
</tr>
<tr>
<td></td>
<td>It is a popular and well researched method for discovering interesting relationships between</td>
</tr>
<tr>
<td></td>
<td>variables in large databases.</td>
</tr>
<tr>
<td>CLASSIFICATION [2]</td>
<td>Classification is a method of recognizing categories that an observation belong to, based on</td>
</tr>
<tr>
<td></td>
<td>its attributes.</td>
</tr>
<tr>
<td>GENETIC ALGORITHMS [21]</td>
<td>Genetic algorithm is a method for solving optimization problems that are based on natural</td>
</tr>
<tr>
<td></td>
<td>It provides the computers the ability to learn without being clearly programmed.</td>
</tr>
<tr>
<td>REGRESSION ANALYSIS [17]</td>
<td>Regression analysis involves examining independent variables to see their impact on dependent</td>
</tr>
<tr>
<td></td>
<td>variables.</td>
</tr>
</tbody>
</table>
### 3.3. Classification based on platform used

Platforms engaged are used as the third standard for classification of research papers.

**Table 5 Platforms used for Big Data Analytics**

<table>
<thead>
<tr>
<th>PLATFORM USED</th>
<th>DESCRIPTION</th>
<th>ADVANTAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>HADOOP [7]</td>
<td>Hadoop is an open source project of apache foundation. It uses map reduce technology as its foundation. Hadoop replicates its data across different computers so that if one goes down, the data is processed on one of the replicated computers.</td>
<td>Hadoop handles massive quantities of data using commodity hardware that is relatively inexpensive computers.</td>
</tr>
<tr>
<td>MONGODDB [19][3]</td>
<td>Mongo DB is an open source document and nosql database. Mongo DB uses sharding technique for splitting the data evenly across the cluster for</td>
<td>It handles large amounts of structured, unstructured and semi structured data with ease.</td>
</tr>
<tr>
<td>RAPID MINFR [25]</td>
<td>Rapid miner is a platform for data mining and machine learning. It uses operator tree modeling knowledge discovery process. Mostly database and excel files are smoothly</td>
<td>Data transformation, data integration, data modeling as well as visualization methods are incorporated by rapid miner. Rapid miner provides fully integrated platform for data analysis.</td>
</tr>
<tr>
<td>HIVE [9]</td>
<td>HIVE provides data warehousing tools to extract, transform and load data. This query data is stored in Hadoop files.</td>
<td>Using HIVE the data is portioned into tables to improve performance.</td>
</tr>
<tr>
<td>POSTGRE SOL [18]</td>
<td>Postgresql is an open source object-relational database svstem.</td>
<td>Postgresql runs stored procedures in more than 12 languages.</td>
</tr>
</tbody>
</table>

It is the process of making a group of objects into classes of similar objects. A cluster of data objects can be treated as one group. It is also pronounced as anomaly detection. It is the process of identifying and excluding the items or observations that do not match to an expected pattern.


IV. DISCUSSION

The systematic mapping study is derived from 30 publications. 5 papers are published in journals and rest of the papers is published in conferences and workshops. Any technique such as classification, clustering, association rule mining, regression analysis, genetic algorithms, and outlier analysis can be used for big data analytics. A large number of studies confirmed that application of big data analytics tool depends on the situation and objective. The result of same research may vary with the use of different tool, result depends on tool used.

V. CONCLUSION

Big data analytics is an active area of research. The result of this study may help new potential users in understanding the range of available big data analytics tools and techniques. This study presents big data analytics active application areas. One of its most well liked application area is crime analysis and governance for better law and order. Crime data analytics is major example that shows how big data analytics can be used by law enforcement communities to take full advantage to retain public support. Crime data analytics is used to provide whole crime statistics of the region that provides benefit to the society by striking the government that why the crime is increasing. The government can take better decisions for better living of the citizens that would naturally add up to lot of lives.

REFERENCES


