Original Article *Available online at www.bpasjournals.com*

Big Data Analytics Research – A Scientometric View

D. Samuvel Raja[1], Dr. P. Sivaraman[2]

[1]Research Scholar
Department of Library and Information Science
Annamalai University, Tamilnadu, India
samuelraja2024@gmail.com
[2]Professor and University Librarian i/c
Department of Library and Information Science
Annamalai University
Tamilnadu, India
psraman.p@gmail.com

How to cite this article: D. Samuvel Raja, P. Sivaraman (2024) Big Data Analytics Research – A Scientometric View, 44(3), 867-876.

ABSTRACT
In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. Big data is a combination of structured, semi-structured and unstructured data that organizations collect, analyze and mine for information and insights. It's used in machine learning projects, predictive modelling and other advanced analytics applications.Applications of big data and data science include information science, uncertainty modelling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Hence, much research is being done in developing and implementing tools for big data analytics. This investigation is an attempt to analyse the trend of research in the field of big data analytics. The results of the study show that India takes the lead in the research productivity. Though India has produced more number of papers, the high prolific authors are from Italy and Norway. The data do not confine with Bradford's law of scattering and Lotka's law of author productivity.

KEYWORDS: Big Data Analytics Scientometrics, Time Series Analysis, Author productivity, Bradford Distribution.

INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets.Big data is a combination of structured, semi-structured and unstructured data that organizations collect, analyze and mine for information and insights. It's used in machine learning projects, predictive modeling and other advanced analytics applications. Big data is characterised by the five V's: volume, velocity, variety, variability, and value. It is complex, so making sense of all the data in the business requires innovative technologies and analytical skills. Companies use big data in their systems to improve operational efficiency, provide better customer service, create personalized marketing campaigns and take other actions that can increase revenue and profits. Businesses that use big data effectively hold a potential competitive advantage over those that don't because they're able to make faster and more informed business decision.

A number of emerging technologies are likely to affect how big data is collected and used. Some of them are

- AI and machine learning analysis. Large data sets are getting larger and thereby less efficiently analyzed by human eyes. AI and machine learning algorithms are becoming key to performing large-scale analyses and even preliminary tasks, such as data set cleansing and pre-processing. Automated machine learning tools are likely to be helpful in this area.

- ➢ Improved storage with increased capacity. Cloud storage capabilities are continually improving. Data lakes and warehouses, which can be either on-premises or in the cloud, are attractive options for storing big data.
- ➢ Emphasis on governance. Data governance and regulations will become more comprehensive and commonplace as the amount of data in use increases, requiring more effort to safeguard and regulate it.
- ➢ Quantum computing. Although less known than AI, quantum computing can also expedite big data analyses with improved processing power. It's in its early stages of development and only available to large enterprises with access to extensive resources.

In recent years big data has been applied to many domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Big data analytics provides new opportunities in the knowledge processing tasks for the upcoming researchers. Now-a-days, big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modelling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Hence, much research is being done in developing and implementing tools for big data analytics. This investigation is an attempt to analyse the trend of research in the field of big data analytics.

Previous studies

RadhaRaghuramapatruni (2024) conducted a bibliometric analysis to find out the development of the field of big data science. It was found that countries like the US, China, Thailand, Australia, and India have a lot of publications in this area. The highest number of publications on big data was in China followed by US, India, UK and South Korea. The analysis showed that computer science, Decision Science, and math are the main areas where more research has been done on big data and data science.

Lars Lundberg(2023) analysed the researchliterature on Big Data for the time period 2012 to 2022 using data downloaded from the Scopus database. Among the four geographic regions (North America, European Union, China, and The Rest of the World), North America was the most active region during the first part of the time period. During the last years China is the most active region.

Justin Zuopeng Zhang, Praveen Ranjan Srivastava, DheerajSharma andPrajwalEachempati (2021) undertook a bibliometric study to analyze the contributions of major authors, universities/organizations, and countries in terms of productivity, citations, and bibliographic coupling. A sample of 2160 articles from the Scopus for the period 2006–2020 is the basis of the study. The publications are grouped into five clusters, of which Cluster 1 is consistently dominant in the information systems publication landscape. Cluster 2 includes published studies on the Internet of Things, security, and cloud computing, which have also been widely researched. Cluster 3, the third-largest cluster, has attempted to investigate social media analytics. Cluster 4 aims to look into the impact of classification and predictive, which is found to have sustained research interest. Topics with scant coverage in terms of papers are primarily in Cluster 5, indicating saturation in the area and the need for conducting inter-disciplinary studies.

According to Batistič and der Laken (2019) research in the domain of big data analytics has gained significant traction in recent years. Acharjya, D P and Kauser Ahmed P (2016) surveyed the various research issues, challenges, and tools used to analyze these big data. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing.

Methods

Data has been downloaded from Scopus database for a period from 2011-2023 using the author assigned keyword (AUKEY = Big data). The downloaded data was converted into MS access database and tables are generated using queries.

Results and Discussion

In 2012, Obama administration announced Big data research and development initiative which consisted of 84 big data programmes to address the problems faced by the government with the growing influx of data. As shown in **Table 1**, this is reflected in the current investigation since the growth rate is steep in 2012 (3.00) and 2013 (6.25). Though there are pits and rises throughout the study period, the average growth rate of research literature on big data is 1.15.

Table 1 Year-wise distribution of research literature on Big data analytics

Year	Publications	Percent	Growth rate
2011	1	0.02	
2012	4	0.10	3.00
2013	29	0.69	6.25
2014	66	1.58	1.28
2015	171	4.09	1.59
2016	269	6.44	0.57
2017	332	7.95	0.23
2018	419	10.03	0.26
2019	552	13.22	0.32
2020	496	11.87	-0.10
2021	588	14.08	0.19
2022	571	13.67	-0.03
2023	679	16.26	0.19
Total	**4177**	**100**	

Table 2 Authorship pattern

No of authors	Publications	Percent	Journal Articles	
			Publications	Percent
1	371	8.88	157	8.36
2	1090	26.10	447	23.81
3	1005	24.06	432	23.02
4	749	17.93	356	18.97
5	431	10.32	220	11.72
6	262	6.27	132	7.03
7	132	3.16	66	3.52
8	46	1.10	27	1.44
9	36	0.86	18	0.96
10	11	0.26	5	0.27
Anonymous	8	0.19	1	0.05
More than 10	36	0.86	16	0.85
	4177	**100**	**1877**	**100**

Big data research publications are produced by single authors to a team of more than 10 members. An analysis of overall literature shows that joint authored publications is highest with 26.10 per cent followed by three authored publications (24.06). The same pattern holds good for journal articles also.

Table 3 Leading countries in big data research

COUNTRY/TERRITORY	Publications	Percent
India	929	22.24

United States	740	17.72
China	478	11.44
United Kingdom	355	8.50
Italy	216	5.17
Australia	207	4.96
Malaysia	202	4.84
Canada	162	3.88
France	158	3.78
Germany	152	3.64
Saudi Arabia	133	3.18
Pakistan	120	2.87
South Korea	108	2.59
Taiwan	105	2.51
Norway	87	2.08

Literature on big data analytics is published by scholars from 87 countries of the world of which India has the highest productivity of 22.24 per cent followed by United States (17.72%) and China (11.44%). Here it is to be noted that India being the highest producing country takes more than 20 per cent of the productivity from the entire world.

Table 4: Prediction of Big data analytics – Time Series Analysis

Year	**Publications (Y)**	**X**	**X2**	**XY**
2011	1	-6	36	-6
2012	4	-5	25	-20
2013	29	-4	16	-116
2014	66	-3	9	-198
2015	171	-2	4	-342
2016	269	-1	1	-269
2017	332	0	0	0
2018	419	1	1	419
2019	552	2	4	1104
2020	496	3	9	1488
2021	588	4	16	2352
2022	571	5	25	2855
2023	679	6	36	4074
	4177	**0**	**182**	**11341**

Straight Line equation $Y_c = a + bX$

Since $\sum x = 0$

$a = \sum Y/N = 4177/13 = 321.31$ $b = \sum XY/\sum x^2 = 11341/182 = 62.31$

Estimated literature in 2030 is when X = **2030**– 2017= 13

= 321.31 + 62.31 *13 = 321.31 + 810.03 = 1131.34

The research productivity in Big data analytics will almost double after 7 years and it will double in 2033.

Table 5 Leading institutions in big data research

Rank	AFFILIATION	Publications	Percent
1	NorgesTeknisk-NaturvitenskapeligeUniversitet	52	1.24
2	Amity University	41	0.98
3	King Abdulaziz University	29	0.69
4	The Hong Kong Polytechnic University	28	0.67
4	UniversitÃ della Calabria	28	0.67
5	University of Technology Sydney	27	0.65
6	UniversitiSains Malaysia	26	0.62
6	K L Deemed to be University	26	0.62
7	University of Wollongong	25	0.60
7	National University of Sciences and Technology	25	0.60
8	Indian Institute of Technology Delhi	24	0.57
9	Chinese Academy of Sciences	23	0.55
10	Symbiosis International Deemed University	23	0.55
10	Vellore Institute of Technology	22	0.53
11	Bina Nusantara University	21	0.50

Among the institutions which are involved in big data research, NorgesTeknisk-NaturvitenskapeligeUniversitetfrom Norway ranks first followed by Amity University, Noida, India and King Abdulaziz University, Saudi Arabia. Indian Institute of Technology, Delhi, India is in the 8th place while Vellore Institute of Technology, Vellore, India is in the10th place.

Table 6 High Productive Journals – Core Journals

Journal	Publications
IEEE Access	48
Sustainability (Switzerland)	44
Technological Forecasting and Social Change	31
Journal of Big Data	31
Journal of Business Research	23
Big Data	23
International Journal of Information Management	21
Computers and Industrial Engineering	18
Annals of Operations Research	17
International Journal of Recent Technology and Engineering	16
Journal of Enterprise Information Management	14
Future Generation Computer Systems	14
Information and Management	13
International Journal of Advanced Computer Science and Applications	13
Industrial Management and Data Systems	12

IEEE Access ranks first with 48 publications. The other leading journals are Sustainability (Switzerland), Technological Forecasting and Social change and Journal of big data.

Table 7 Validation of Bradford's Law

Zones	No of Journals	Publications
Zone 1	47	623
Zone 2	207	625
Zone 3	579	629
Total	**833**	**1877**

An application of Bradfords law of scattering shows that the ratio of the three zones is 47:207:579 :: 1 :4.40 : 12.32 which is not in the form 1:n:n^2deviating Bradford's law. The reason may be that since Big data is a budding subject area, more number of journals in this subject have not emerged.

High prolific authors

In order to rank the authors according to their publications, three methods are employed, namely

- Direct count method - Calculating the total publications whether it is solo or collaborative research
- Equal Share method - Providing equal weightage to each author in a collaborative publication and summing the total weightage
- Positional share method – Providing weightage to an author according to their relative position in the named list of authors in a publication.

According to Dr.S.R.Ranganathan's cataloguing theory, the potency of an author is concentrated on the term in the first position and hence by applying this principle, the authors in a collaborative publication are given credits according to their position in the name list. The following procedure and formula proposed by J.P.S. Kumaravel is adopted.For example, if there are n authors for a publication, the weightage (w) of an author in p^{th} position ($p \leq n$) for that publication can be calculated as

$W = (n – p +1) / n\sum$ where $\sum = 1$ to nand $W \leq 1$

For example, the potency of each author in a work by 4 authors, can be calculated as

- 1^{st} Position = $(4 -1 +1) / 4\sum = 4 / (1+2+3+4) = 4 / 10$
- 2^{nd} position = $(4 -2 +1)/4\sum = 3/10$
- 3^{rd} position = $(4 -3 +1)/4\sum = 2/10$
- 4^{th} position = $(4 -4 +1)/4\sum = 1/10$

Table 8 High Prolific Authors – Direct count method

Author Name	Total Publication Count	Rank
Cuzzocrea, Alfredo	37	Italy
Bibri, Simon Elias	24	Norway
Mikalef, Patrick	18	Norway
Krogstie, John	18	Norway
Akter, Shahriar	16	Australia
Sun, Zhaohao	15	USA
Wamba, Samuel Fosso	13	USA
Chang, Victor	12	USA
Xia, Dawen	12	China
Mehmood, Rashid	11	Saudi Arabia
Li, Yantao	11	China
Kar, Arpan Kumar	11	India
Leung, Carson K.	11	Italy
Li, Huaqing	10	China
AL-Khatib, Ayman wael	10	Jordon
Wang, Lidong	10	USA

Table 9 High Prolific Authors – Equal Share count method

Author	Publications	Rank	Share
Cuzzocrea, Alfredo	37	1	22.69

Bibri, Simon Elias	24	2	21.25
AL-Khatib, Ayman wael	10	14	9.00
Krogstie, John	18	4	6.62
Sun, Zhaohao	15	6	6.00
Mikalef, Patrick	18	3	5.62
Wang, Lidong	10	16	4.83
Akter, Shahriar	16	5	4.77
Kar, Arpan Kumar	11	11	4.53
Wamba, Samuel Fosso	13	7	4.18
Ghasemaghaei, Maryam	5	85	3.83
Mehmood, Rashid	11	13	3.75
Singh, Vikram	7	42	3.70
Marine-Roig, Estela	6	45	3.58
Strang, Kenneth David	6	64	3.50
Chang, Victor	12	8	3.38

Table 10 High Prolific Authors – Positional share method

Author	Count	Rank	Positional value
Cuzzocrea, Alfredo	37	1	24.64
Bibri, Simon Elias	24	2	21.87
AL-Khatib, Ayman wael	10	14	9.33
Sun, Zhaohao	15	6	7.83
Mikalef, Patrick	18	3	7.40
Wang, Lidong	10	16	6.50
Akter, Shahriar	16	5	5.61
Kar, Arpan Kumar	11	11	4.43
Babar, Muhammad	9	18	4.25
Wamba, Samuel Fosso	13	7	4.22
Krogstie, John	18	4	4.20
Verma, Surabhi	6	66	3.83
Ghasemaghaei, Maryam	5	85	3.83
Marine-Roig, Estela	6	45	3.73
Dremel, Christian	8	23	3.73
Strang, Kenneth David	6	64	3.67
Raj, Pethuru	6	63	3.37
Xia, Dawen	12	9	3.36

If the authors are ranked according to their publication count (Table 8), Cuzzocrea, Alfredo from Italy ranks first with 37 papers andBibri, Simon Elias from Norway ranks second with 24 publications. The same sequence holds good for the other two methods also. In equal share method (Table 9), the entire ranking is changed. This is also true in case of positional share method (Table 10).

.

Table 11 Lotkas's Law of author productivity

No of Papers	No of Authors	Percentage of authors
1 Paper	10239	86.44

2 Paper	1095	9.24
3 Paper	275	2.32
4 Paper	120	1.01
5 Paper	49	0.41
6 Paper	25	0.21
7 Paper	14	0.12
8 Paper	8	0.07
9 Paper	4	0.03
10 Paper	3	0.03
11 Paper	4	0.03
12 Paper	2	0.02
13 Paper	1	0.01
14 Paper	1	0.01
15 Paper	1	0.01

Lotka's law states that ". . . the number (of authors) making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent". This means that out of all the authors in a given field, 60 percent will have just one publication, and 15 percent will have two publications ($1/2^2$ times . 60), 7 percent of authors will have three publications ($1/3^2$ times . 60), and so on. The total number of authors who have produced big data research literature is 11845 of which the number of authors who have contributed one paper each is 86.44 deviating Lotks's law.

Table 12 Subject Mapping

SUBJECT AREA	Publications	Percent
Computer Science	2855	68.35
Engineering	1491	35.70
Decision Sciences	836	20.01
Business, Management and Accounting	813	19.46
Mathematics	623	14.92
Social Sciences	536	12.83
Energy	247	5.91
Medicine	236	5.65
Environmental Science	190	4.55
Economics, Econometrics and Finance	177	4.24
Physics and Astronomy	161	3.85
Materials Science	130	3.11
Biochemistry, Genetics and Molecular Biology	81	1.94
Arts and Humanities	65	1.56
Chemical Engineering	61	1.46

Big data research literature falls into many subject areas, of which Computer Science takes a major share or 68.35 per cent followed by Engineering and Decision sciences having 35.70 per cent and 20.01 per cent respectively.

Table 13 Document type

Number of citations	Publications	Percent
Article	1877	44.94
Conference paper	1683	40.29
Book chapter	358	8.57
Review	190	4.55

Book	29	0.69
Editorial	22	0.53
Retracted	7	0.17
Note	5	0.12
Not assigned	3	0.07
Short survey, letters, Data papers	3	0.02
Total	**4177**	**100**

Research literature on Big data is available in various formats of which journal articles takes a major share of 44.94 per cent followed by conference papers forming 40.29 per cent.

Table 14 Number of citations

Number of citations	**Publications**	**Percent**
0	724	17.33
1	482	11.54
2	328	7.85
3	240	5.75
4	182	4.36
5	169	4.05
6	129	3.09
7	110	2.63
8	100	2.39
9	102	2.44
10	75	1.80
11	79	1.89
12	73	1.75
13	62	1.48
14	61	1.46
15	49	1.17
16	53	1.27
17	50	1.20
18	42	1.01
19	45	1.08
20	36	0.86
More than 20	986	23.61
Total	**4177**	**100**

Citations are acknowledgements given by the scholars to the previous research. The more the paper is cited, the more will be the versatility of the paper. In this research, 17.33 per cent of the papers do not have citation at all and 11.54 per cent of the papers have only one citation. As the number of citations increases, the number of papers decreases. The paper with highest number of citations (n=4173) is “Business intelligence and analytics: From big data to big impact by Chen H., Chiang R.H.L. and Storey V.C.in 2014 followed by “Internet of things in industries: A survey by Xu L.D, He W and Li S in 2014.

Conclusion

With the big data hype all around, it is the fuel of the 21st century. Big data is among the most talked about subjects in the tech world as it promises to predict the future based on analysis of massive amounts of data. Big data might be the hot topic around the business right now and hence the roots of big data run deep.

References

Acharjya, D P and Kauser Ahmed P (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. International Journal of Advanced Computer Science and Applications.7(2), 511-518.

Batistič, S et al(2019). History, evolution and future of big data and analytics: A bibliometric analysis of its relationship to performance in organizations. British Journal of Management.

Gupta D, Rani R. (2019). A study of big data evolution and research challenges. J Inform Sci.45(3):322–40.

Gupta V, et al. (2019). A quantitative and text-based characterization of big data research. J Intell Fuzzy Syst, 36(5):4659–75.https://www.projectpro.io/article/big-data-timeline-series-of-big-data-evolution/160

Justin Zuopeng Zhang, Praveen Ranjan Srivastava, DheerajSharma andPrajwalEachempati (2021). Big data analytics and machine learning: A retrospective overview and bibliometric analysis. Expert Systems with Applications. 184.

Kumaravel, J. P. S., et al. (2012).Dr. SR Ranganathan's Canon of Prepotence applied to Bibliometrics leading to a new indicator-Prepotency index (PI). 8th International Conference on Webometrics, Informetrics and Scientometrics (WIS) and 13th COLLNET Meeting in Seoul, Korea.

Lars Lundberg(2023). Bibliometric mining of research directions and trends for big data.Journal of Big Data.10, 122.

Liu X, et al.(2020). The research landscape of big data: a bibliometric analysis. Library Hi Tech. 38(2), 367–84.

Lohr S. (2013). The Origins of 'Big Data': An Etymological Detective Story. The New York Times.

Lotka, A.J. (1926). The Frequency Distribution of Scientific Productivity.Journal of Washington Academy of Sciences, 16, 317-323.

Lundberg L, Grahn H.(2022). Research Trends, Enabling Technologies and Application Areas for Big Data.Algorithms.15(8), 280.

RadhaRaghuramapatruni (2024). Trends and Pattern in Big Data: A Bibliometric Study. Intelligent Systems and Applications in Engineering. 12(13), 322–333.