JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Insights into the NoSQL Databases: A Bibliometric Analysis

[1]Aditi Badhan, [2]Anita Ganpati
[1]Research Scholar, [2]Professor
Department of Computer Science Himachal Pradesh University, Shimla, India
Department of Computer Science Himachal Pradesh University, Shimla, India

Abstract

The study attempts to address the growing pace of technological advancement and digitization over the period. Digitization outbreak has resulted in producing a massive amount of data which has led to the introduction of the term Big Data. With the notion of Big Data, the existing data processing applications are inadequate to efficiently process, analyze and visualize the large quantity of data. To efficiently process a massive amount of data or Big Data the NoSQL approach has been introduced. It helps to extract the most fruitful information, to process a massive amount of data at the same velocity produced from the heterogeneous sources. "NoSQL is an acronym for "Not Only SQL", a non-relational distributed database. The study is accomplished using bibliometric analysis as it helps to evaluate the importance and impact of articles published with the title and evolution of the domain in the research area. Biblioshiny tool present in R-Studio used for the analysis of data. This study aims to provide a comprehensive review on the growth of NoSQL databases and elaborating the capabilities compared to relational databases. This paper provides insight into the advent and how database users are fascinated by the NoSQL databases.

Originality/Value: To the best of the author's knowledge no bibliometric analysis has been done on the nosql databases.

Keywords: NoSQL, Database System, Bibliometric Analysis, Big Data, Databases.

1. Introduction

The relational model was leading the market in late ‘90s, but swiftly with the increasing number of digital devices accumulated growth in data stocks were consist structured, semi-structured and unstructured data in large "Volume, Variety and Velocity" termed as Big Data [1]. The new concept of Big Data has affected the potential of a conventional way of data processing applications [2].In the age of information technology, Big Data is gaining popularity. The International Data Corporation(IDC) defines Big Data as: Big Data technologies represent a new generation of technologies; that promotes new architecture designed to derive valuable data at high velocity [3].Relational databases (RDB) uses the relational model to store data. The constraints of RDB like strict schema, no support for horizontal scalability, inadequate to process unstructured data, etc.

[4] does not cope with the current requirement of the database users. The existing data processing tools are inadequate to efficiently process and get insight into the data, fulfil the needs of database users, to extract the data from unstructured and semi-structured data: these are the trends that motivated and boosted the development of the NoSQL Databases. NoSQL or Not Only SQL is an open-source, non-relational distributed database used to process, analyze and visualize the massive amount of data produced from heterogeneous resources. NoSQL databases generally do not follow the approach of table/row/column as relational databases[5]. NoSQL databases, gaining attention for many enterprises and database users as it is not a replacement of relational models or databases but adopted as an alternative model chosen as the problem at hand [6]. The NoSQL database benefits the users with many efficient solutions over relational databases like expediting data processing, supports horizontal scalability, better performance, bare query language, no strict schema, etc. [7].

NoSQL databases support four kinds of data models like Key-value store, Column-oriented, Document-oriented and Graph databases. Each database has its pros, cons and use cases. Each model can be applied according to the use case scenario. To deal with relations, the graph database is helpful; when we have to work with semi-structured data, column-oriented databases are used. There is a dearth of research that represents the emergence of the Nosql databases and their significance to cope with the current requirements of the database's users. This study provides a glimpse of the Nosql database to the new researchers who are interested in the area of study. In the study, the bibliometric analysis has been performed using a Biblioshiny tool available in the R-package.

Recently numerous studies, literature reviews have been performed on NoSQL databases, but to the best of author's knowledge, no bibliometric analysis depicts the growth trend that justifies the need for NoSQL databases. The bibliometric analysis performed in the study helps to improve the quality of literature systematically. Many research papers have been published on the review of nosql databases [8] that has discussed the NoSQL databases. Other studies were performed on the comparison of various NOSQL and RDBMS databases, models of NoSQL databases etc.

Method for Bibliometric analysis

For the accomplishment of the study, the Bibliometric Analysis is performed using the steps suggested by [9] are followed. The steps used in the study are defined using the diagram.

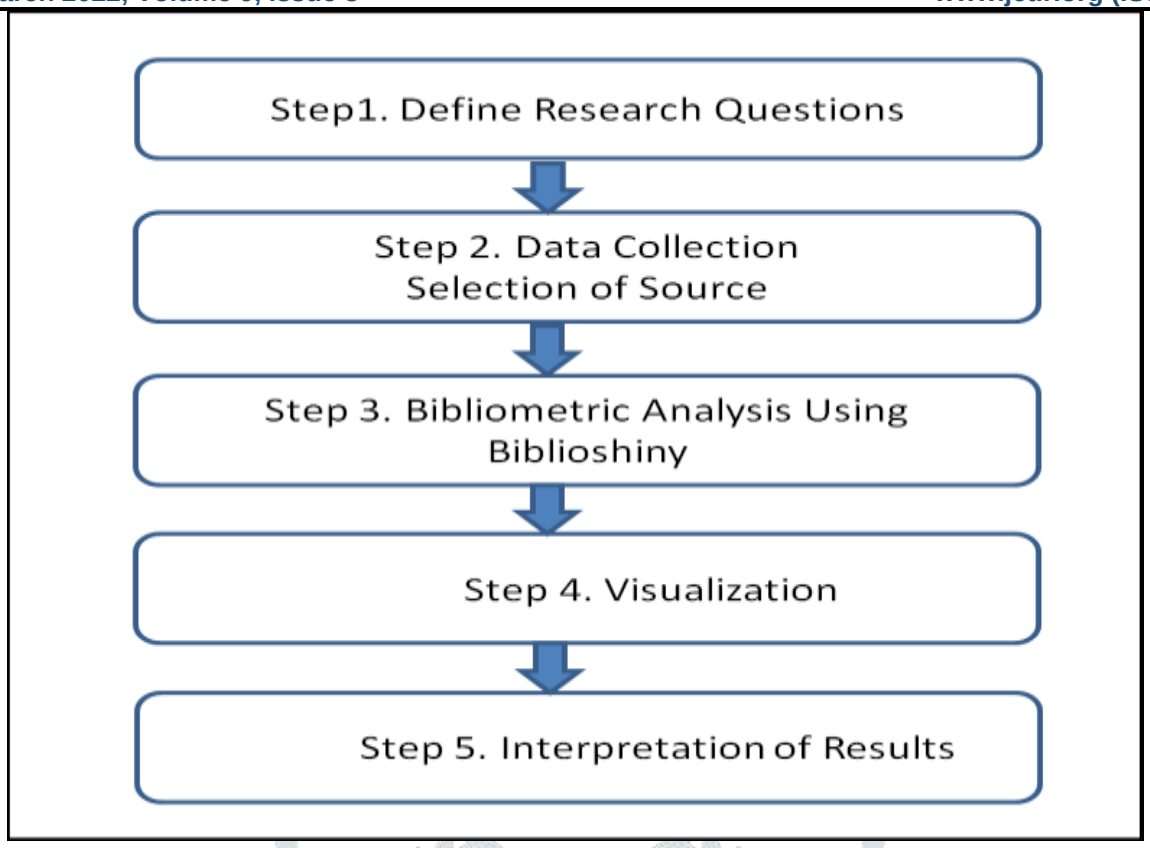

Figure1. Bibliometric Analysis Procedure (Nasir et al., 2020)**.**

Research Questions

The literature review is done by the author's attempts to address the following research questions:

1. To find the top authors, countries, publications and keywords in the domain?

2. To understand why the Non - relational databases are influencing the users of the database?

3. What is the state-of-the-art of NoSQL databases?

4. What is the research Trend in the field?

The remainder of this paper is assembled as follows. Section II describes the Research Methodology. Section III outlines the bibliometric analysis of 1168 papers. In Bibliometric Analysis, Conceptual structure, Intellectual structure, keyword analysis of the most frequent words in the title, abstract etc., and the top author impacts etc., addressed.

2. Research Methodology Numerous databases are available in the market**.** To address the research question," The Scopus Database" was chosen as a source for the accomplishment of the study. It contains more than 20,000 titles of Peer-Reviewed (Journal, Conference proceeding, etc.) [11]. The data is derived using a query statement. A total of 1386 articles were elicited, later inclusion and exclusion criteria were applied on the research articles, out of which 1168 research articles were taken into account. Several tools are available to conduct bibliometric analysis like VOS Viewer, Bib EXCEL, CiteSapce, SciMAT, Biblioshiny, etc. In this study, Biblioshiny, a tool present in "The R-package", has been selected for analyzing the data. To find the answer of research questions 1 and 2, the core areas of study, top authors, top countries and keywords in the literature are identified using the descriptive analysis. To find the answer of the question 3 and 4, the "Science Mapping", which reveals the growth of specific

domain and the association of Authors, etc. have been performed. Furthermore, the Thematic Evolution exhibits the evolution of a theme in the area of the study.

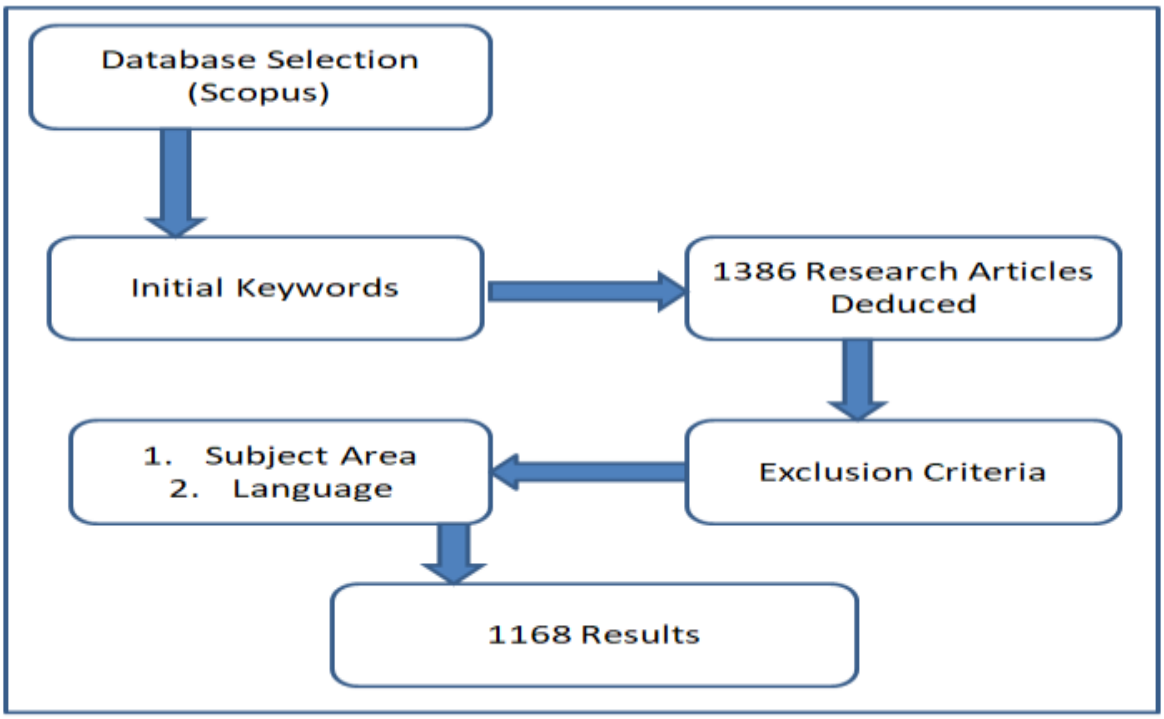

Figure 2. Design Methodology

Figure 2 demonstrates the research methodology used during the study. At the early stage, the database has been chosen for the study. Initially, the keyword "nosql" was investigated, 1386 research articles were derived from the database. To restrict the study the exclusion criteria is employed. The research articles published in the domain of Computer Science and Applications and the articles published only in English language are considered, Out of which 1168 articles are investigated for the accomplishment of the study.

3.Bibliometric Analysis

Bibliometric analysis is the scientific application of quantifying the quality and measuring the impact[12]. It describes the way the specific domain and research field evolved with time. The Biblioshiny tool of R - Package has been selected to perform the bibliometric analysis on the articles. A web-based graphic interface tool can be effortlessly learned and used by the non-coders. The Developer of a tool divided it into seven divisions like Sources, Authors, Documents, Clustering, Conceptual Structure, Intellectual Structure and Social Structure. The Biblioshiny tool for performing the bibliometric analysis, to extract the most pertinent information like Keyword analysis; Scientific Mapping reveals the relationship among Authors, Countries and Institutions include Intellectual Structure, Conceptual Structures and Social Structure and the development of the area over time.

Table I. Document Main Information

Description	Results
Time Span	**2010:2020**
Sources(Journals ,Books ,etc.)	646
Documents	1168
Article	228
Book Chapter	57
Conference Paper	874
Review	9
Document Contents	
Keyword Plus (ID)	4901
Authors Keyword	2475
Authors	2864
Authors Appearances	3860
Authors of Single-authored Document	71
Authors of Multi-authored Documents	2793
AUTHORS COLLABORATION	
Single-authored Documents	90
Documents Per Author	0.408
Authors Per Documents	2.45
Co-Authors Per document	3.3
Collaboration Index	2.59

Table 1 describes the research articles that are used for performing the bibliometric analysis. To accomplish the "bibliometric analysis", the data set has to undergo two stages. The first stage is "Data Collection", the source has been chosen from where the data is collected. Various databases are available like "The Scopus, The web of Science and Google Scholar" etc. "To meet the purpose of the study", the Scopus database has been chosen for the accumulation of data. The second stage is "formation of the query and Application of filters". The "search query" is formed, research

articles published in the domain of Computer Science and Applications have elected to collect data, and multiple filters are employed to search the data. The keyword "nosql databases" has been keyboarded 1386 articles extracted from the Scopus database. Next, to restrict the study, different constraints were employed—the research articles published from 2010 to 2020, Paper of the conference, book chapters, Articles and Papers published only in the English language considered for the study. The purpose of confining our study to one language is that it produces efficient bibliometric analysis and provides tools to compare keywords, articles and sources. In addition, the "articles" are investigated manually, and then from 1386, articles 1168, articles selected are considered for analysis.

Visualization of Bibliometric Analysis

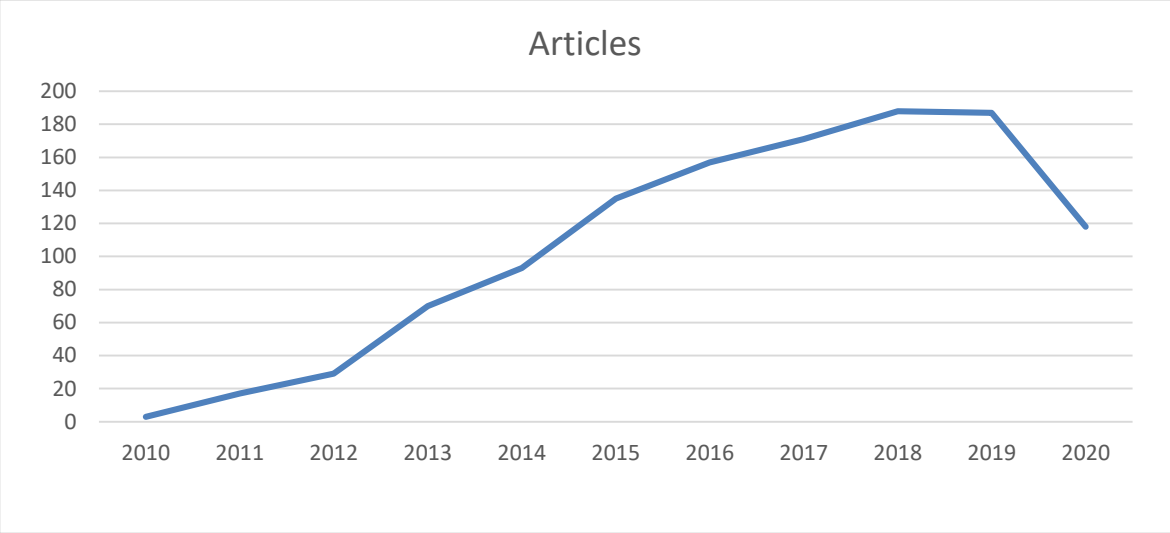

Figure 3. Annual Scientific Production

The Annual Production and Average Citation over the years of NoSQL databases are shown in Figure 3 and Figure 4. Initially, there is a limited production of articles in the area of study. The production of literature in the domain of Nosql databases increased with time and with the current requirements of the database users. The graph shows an increase in production from 2010 to 2018, but in 2019 and 2020, there is slight less growth that might be due to Covid -19.

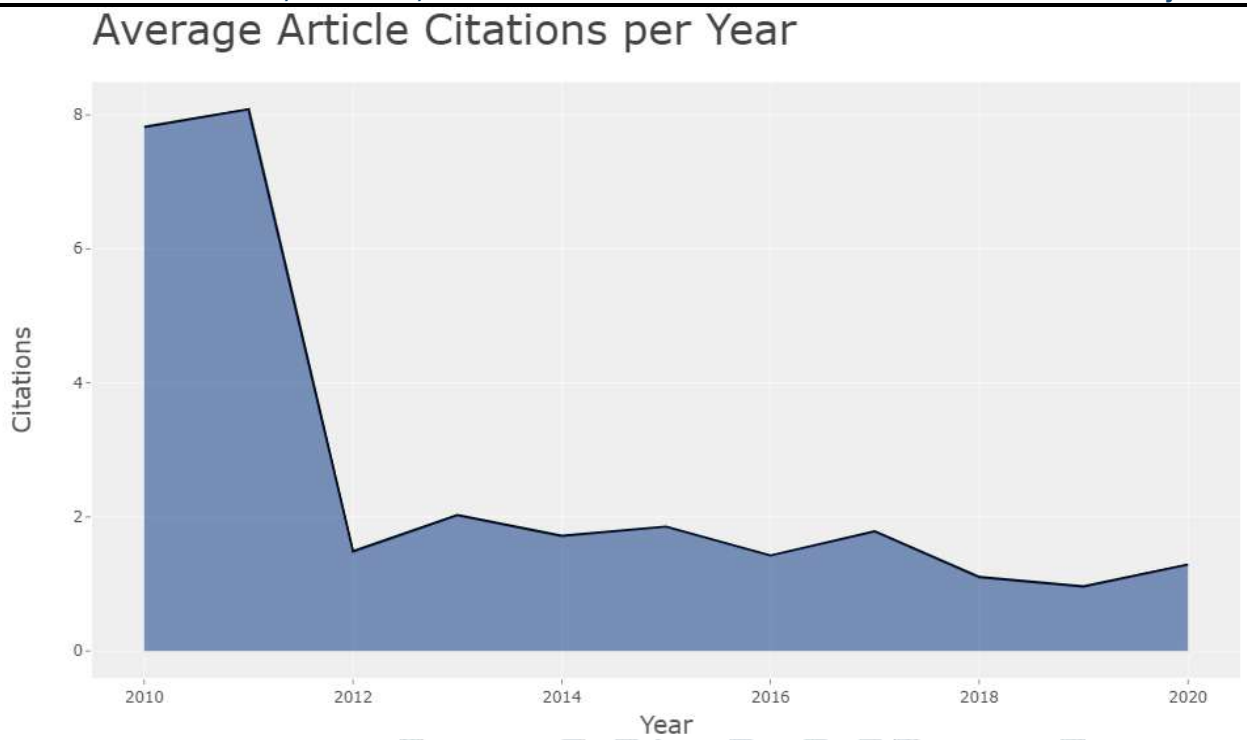

Figure 4. Average Article Citation per Year

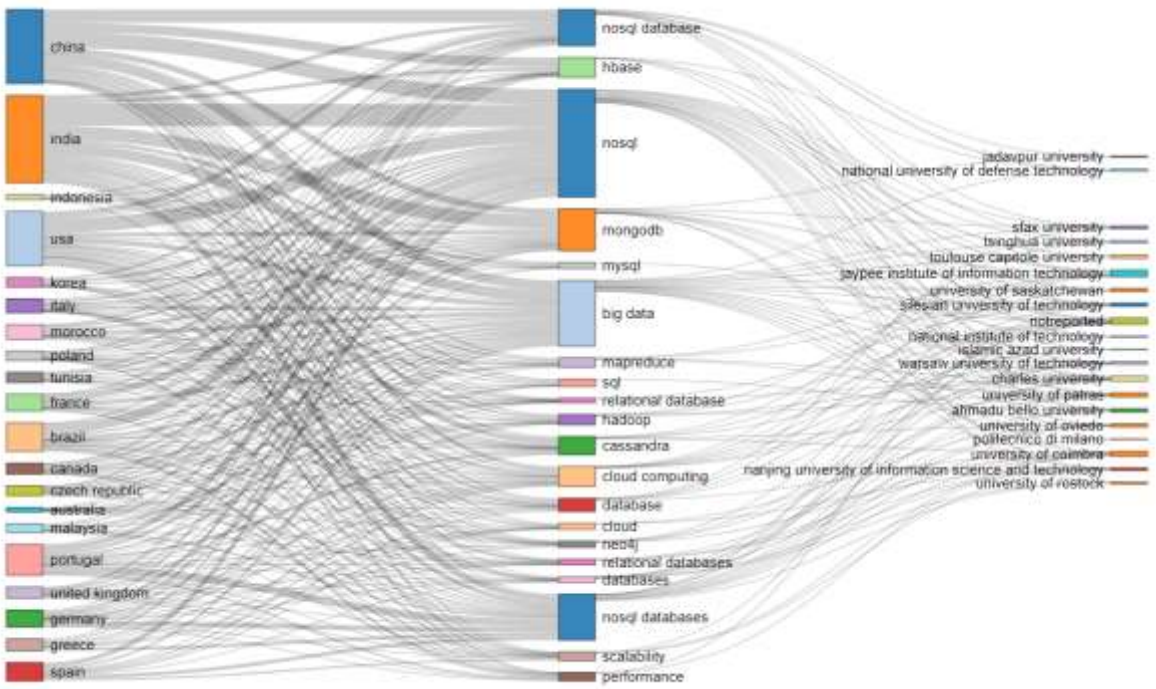

Figure 5. Three-fold Analysis of NoSQL databases

The study carried on the annual production and Average citation of articles. Considerably, there is a need to probe the "Relevant topics, Places "and Affiliations" of the field.

Figure 5. Present the "Three-fold-Analysis" of NoSQL Publications. In the Three-Fold- Analysis, a keyword takes on the left side, keyword plus in the middle and Countries on the right side of the Figure. The Figure depicts that the word "NoSQL", "big data", "MongoDB" are the

most prominent area of study, most of the countries are working in this area. China is the leading country predominantly working in the areas, following the USA and INDIA.

ANALYSIS OF CORE WORDS

Table 2 represents the keyword analysis of the most frequent words used in the literature of nosql databases. The investigation is performed on the most frequent word appear in the abstract, keyword in the title, keyword plus and authors keywords. Thus the table is formed into four parts. In all components, the word nosql is the most appearing word. Nosql, Big data and database system are the words that occur in all articles. It has been acknowledged that the term "nosql "is the most significant in all categories like the keyword plus, abstract, title and author's keyword.

Table II. MOST FREQUENT WORDS

Words	occurrences	Words	Occurrences
Keyword Plus		**Authors Keyword**	
nosql database	470	**nosql**	381
big data	361	**big data**	216
database system	361	**mongodb**	147
Nosql	312	**nosql databases**	146
digital storage	279	**nosql database**	117
query processing	237	**cloud computing**	65
information management	207	**cassandra**	59
relational database	155	**hbase**	54
Mongodb	140	**database**	40
data handling	139	**Sql**	36
Abstract		**Title**	
Words	**occurrences**	**words**	**occurrences**
Data	4871	**nosql**	482

Nosql	2468	**data**	449
Database	1987	**databases**	253
Databases	1845	**database**	237
Paper	881	**big**	143
Performance	862	**performance**	106
Big	882	**based**	92
Systems	820	**cloud**	87
System	812	**system**	85

Figure 6. Tree Map

Figure 6. Shows the TreeMap of the Top 50 Author's keyword. The TreeMap represents that the keyword "nosql database", "nosql" and "big data" are the word which is arriving regularly. In the intact subject, The word "nosql database" is appearing 470 times in the dataset following, the "big data" 361 times and "nosql "312 times.

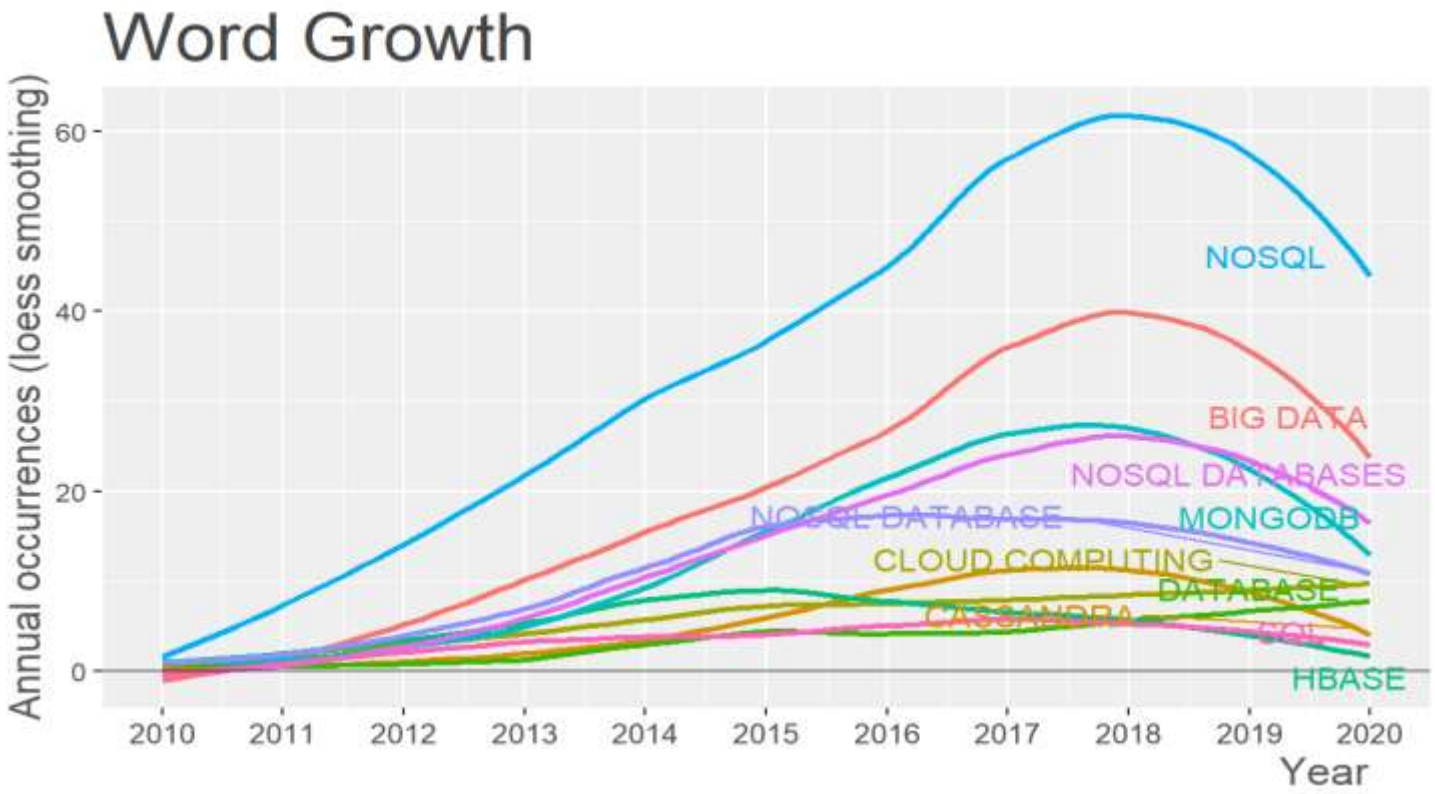

Figure 7. Word Growth

Figure 7 represent the growth of Publications by Top journals in the area of study. The Loess Smoothing is adopted. "Locally Weighted Smoothing" uses the regression analysis of the smooth line with the help of a time and scatters plot [10]. There has been growth in "NOSQL" in the period of 2010 to 2020. It also describes that there is a decreasing trend in the publication of SQL. It symbolizes that there is an escalation in the state of Big Data, NOSQL Databases and nosql. These are the most trending and embraced terms used by most vendors and databases users and researchers.

Analysis of Most Relevant Authors, Countries and Relevant Sources of Information This section presents the information concerning the Main authors and the Countries, publishing articles in the literature of nosql databases. Figure 8. represents the Top 10 Most Relevant Sources as it is derived, that "The Lectures in Computer Science Engineering" is at the top among all by having more than 100 papers and

Figure 9. Illustrate Top Most-Cited Countries; Figure shows that China is on top among all Countries and has the most Cited Articles.

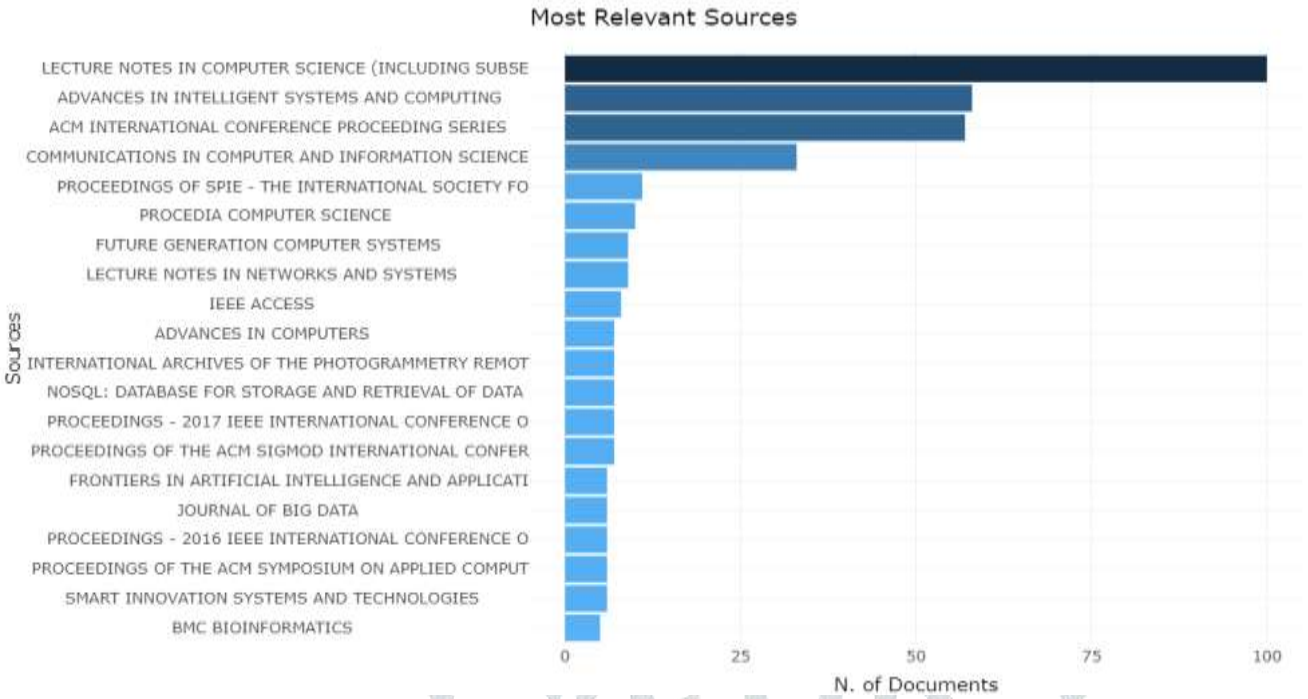

Figure 8. Most Relevant Sources

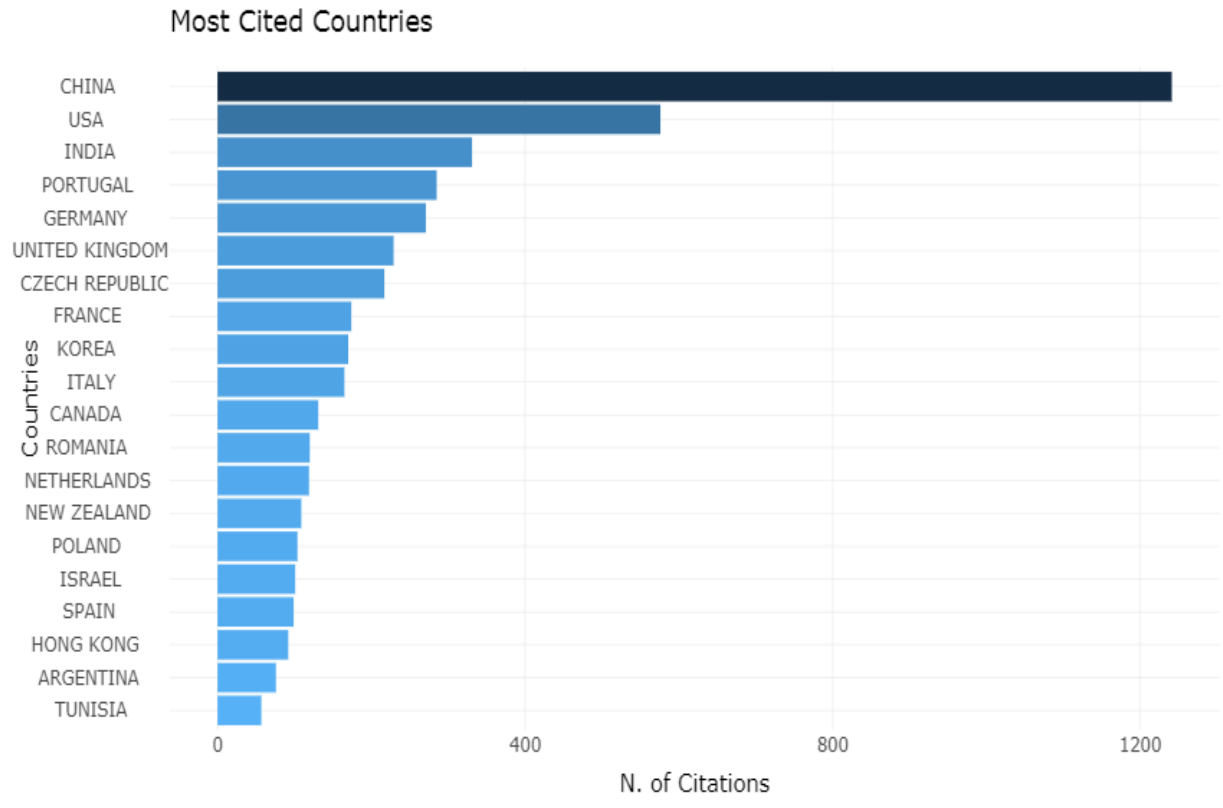

Figure 9. Most Cited Countries

TABLE 3 presents Top Authors significant in the field of nosql databases. Ranking of the top authors based on The h-Index. The author Bernardino J secured top or first rank in the list of authors with the highest impact study. In their work, they assessed the performance of nosql databases and distinguished the MongoDB and Cassandra databases [13]. The author Deters R ranked (Second) explained that the existing

data mining techniques are inadequate to mine data from the unstructured data, emphasize the need for AaaS (Analytics-as-a-Service) from the extraction of unstructured data from nosql databases [14]. Gargouri ranked third in [15], presented the transformation of a data warehouse to nosql Data warehouse transformation using the Column-oriented and Document-oriented transformations.

Table III. Top 15 Author Impact

Author	h_index	g_index	m_index	TC	NP	PY_start
BERNARDINO J	6	11	0.667	230	11	2013
DETERS R	6	8	0.667	77	10	2013
GARGOURI F	4	7	0.667	49	10	2016
HOLANDA M	3	4	0.429	24	10	2015
LOMOTEY RK	6	8	0.667	77	10	2013
POKORN J	4	5	0.444	31	9	2013
ZHANG Y	5	9	0.5	87	9	2012
ZURFLUH G	3	4	0.6	22	9	2017
AHMAD R	2	3	0.4	14	8	2017
BASRI S	2	3	0.4	14	8	2017
BELANGOUR A	6	7	2	57	8	2019
CABOT J	5	8	0.833	102	8	2016
DOS SANTOS MELLO R	3	5	0.333	30	8	2013
KLETTKE M	3	6	0.5	38	8	2016

Table IV. TOP COUNTRIES IN TERMS OF CITATIONS

Country	Total Citations	Average Article Citations
CHINA	1240	17.714
USA	564	12.818
INDIA	328	7.628
PORTUGAL	285	16.765
GERMANY	271	16.938
UNITED KINGDOM	229	20.818
CZECH REPUBLIC	211	14.067
FRANCE	174	9.667
KOREA	166	8.3
ITALY	164	9.111
CANADA	131	10.077
ROMANIA	120	40
NETHERLANDS	119	119
NEW ZEALAND	109	109
POLAND	104	6.118
ISRAEL	101	33.667
SPAIN	98	6.533
HONG KONG	92	30.667
ARGENTINA	76	76
TUNISIA	57	4.071

Analysis of Conceptual Structure The Conceptual Structure analyses the connection between the specific terms. The Conceptual Structure, of Co-Occurrence of the author's Keyword, is shown in Figure 12. Furthermore, the Thematic Evolution is represented in Figure 10. Research themes are discovered using the ThematicMap, for the interpretation of the results. The "Themes" in ThematicMap, is divided into a four-quadrant structure; The quadrants in the map represent the importance and development of the research theme. The Figure depicts, Centrality and Density are applied as a division of "Themes". In which the Centrality represented at the X-axis, the Density at the Y-axis. The Centrality helps to measure the importance of the selected theme open sources, whereas; Density helps to measure the development of the current themes in the area. The word “information system” is present at the lower-left corner of the Figure, depicting whether the theme is an emerging theme or dropped by the researchers. "NoSQL" is lying at the lower-right corner of the picture; Themes having Low Density and High Centrality means much work has been performed on these themes.

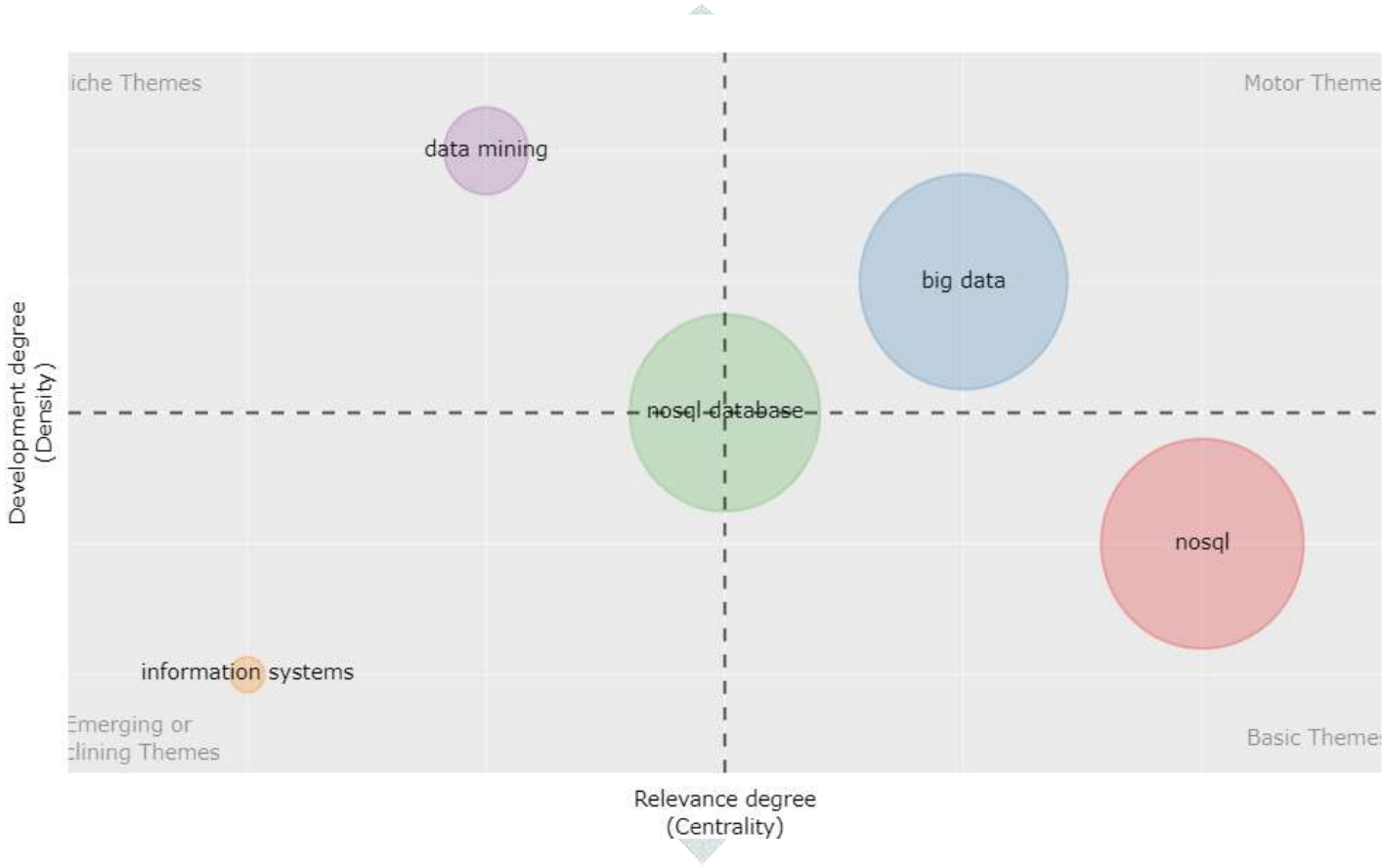

Figure 10. Thematic Map

The keyword "data mining" is present at the 2nd Quadrant or Upper-Left corner of the Figure, which depicts high Density and lower Centrality describe that the themes are well developed. The keyword “big data” is present at the 1st quadrant Upper-Right corner of the Figure are the Themes, with High Density and high Centrality the Themes are the Motor Themes, which are the well developed and vital themes of the area [10].

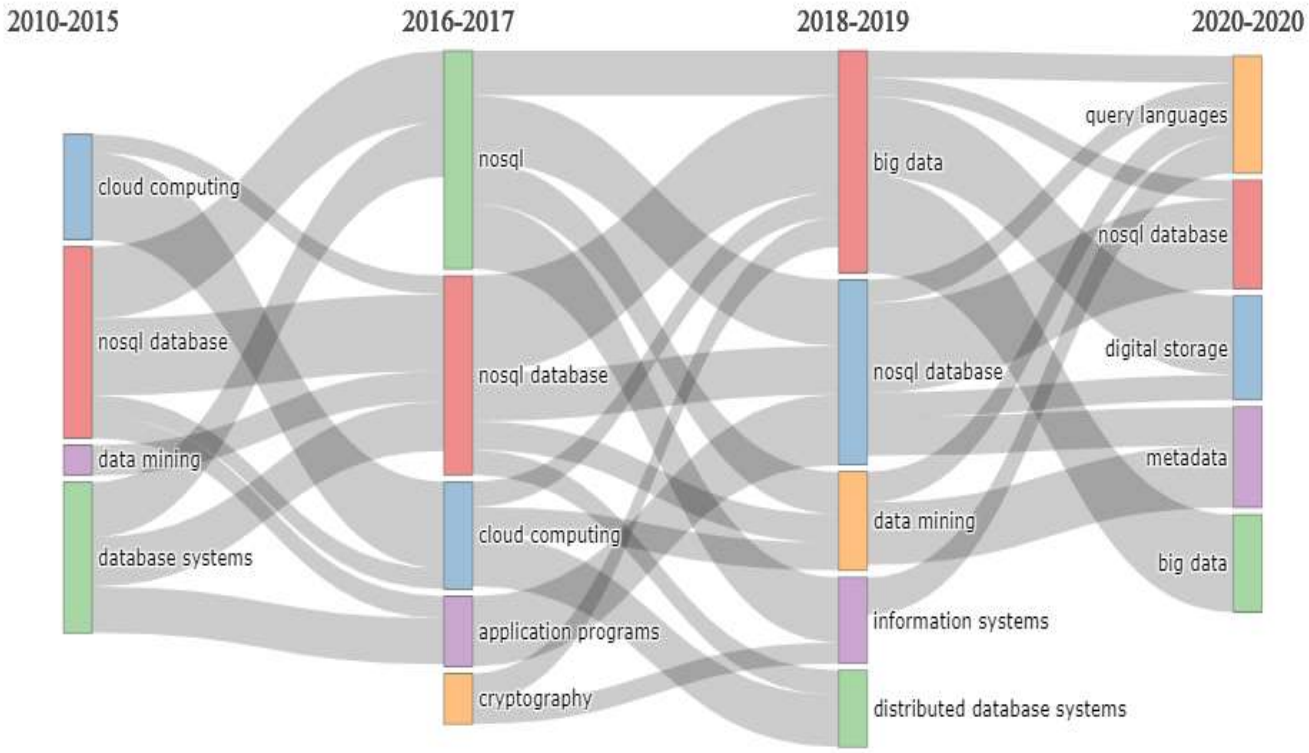

Figure 11. Thematic Evolution

Thematic Evolution also incorporated, depicts the Historical Development in the literature of "nosql databases". The keyword plus describes the history of themes and how these themes unfolded [10]. Thematic Evolution was accomplished using the Biblioshiny tool of the R-Package. The four-segment of time applied. The First segment from 2010-2015, The Second segment from 2016-2017, The Third segment from 2018-2019 and The Fourth segment is 2020. It is analyzed from the Figure that from 2010-2015, the literature of "Cloud computing and nosql databases" were the focused area of study. From 2016-2017 the research in the area of nosql, nosql databases got extended. From 2018 -2019 big data and nosql databases were the fascinating areas of study. It has also been observed that the nosql databases are the lucrative area of research in all segments of time.

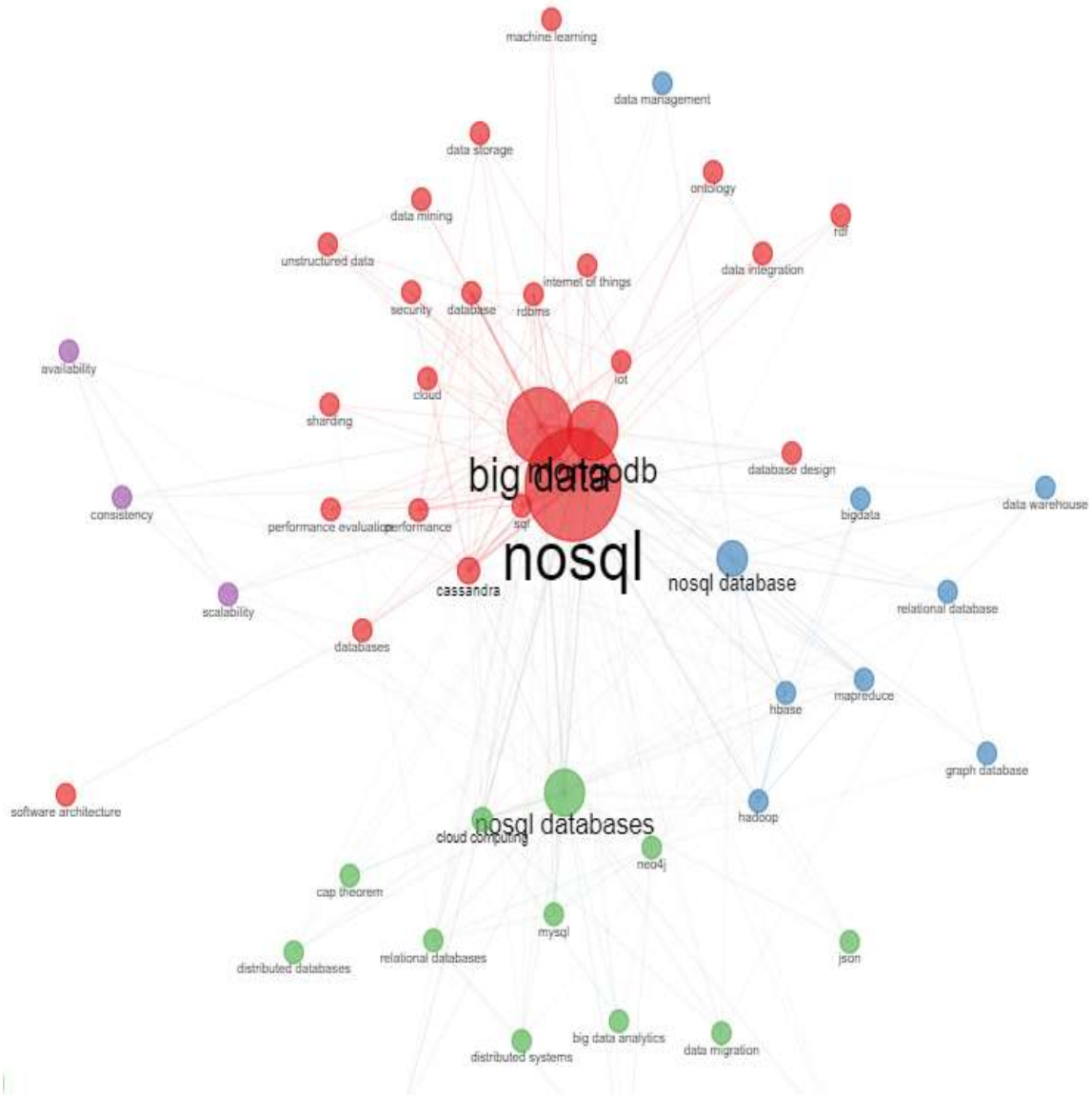

Figure 12. Co-Occurrence Network of Authors Keyword

Figure12. Shows the Co-Occurrence Network; in the network, The “Author’s keyword " is chosen as the unit of analysis. The Figure is obtained from the Biblioshiny tool of the R-Package. The Co-Occurrence Network of the Authors keyword is divided into four clusters The Red, Green, Blue and Purple Cluster. The Red cluster has the highest Centrality, Blue and Green, and Purple clusters connected in terms of themes. The” nosql databases “, "big data", and "mongodb" have the highest Centrality expresses that, currently, the authors and researchers are converging in this particular domain of study.

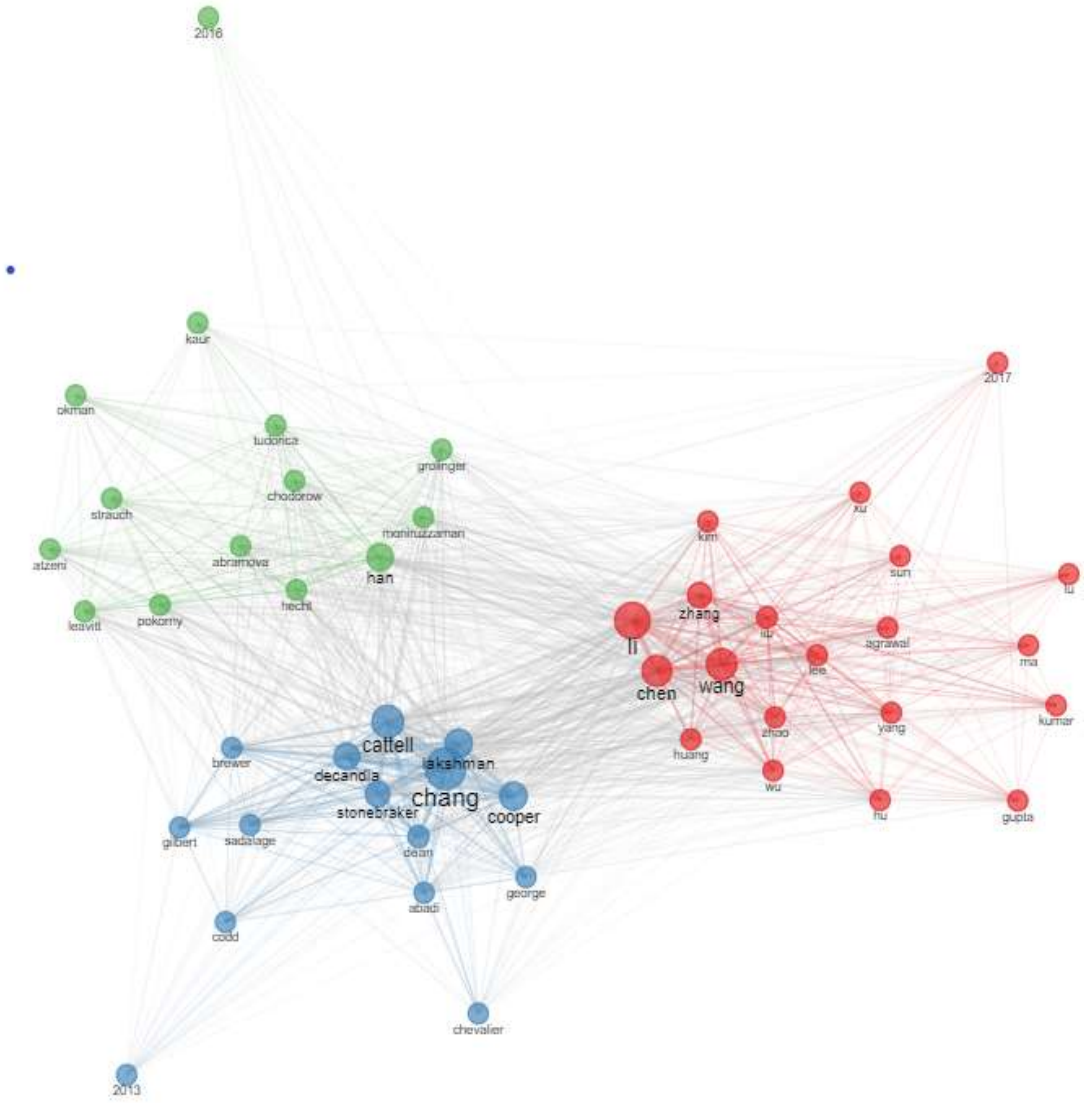

Figure 13. Co-Citation Network of Authors

Analysis of Intellectual Structure In the Intellectual structure citation analysis performed, that determines the scientific growth of a field[16]. The citation analysis represents the bibliographic coupling of Co-author, Co-word and Co-Citation. In this paper, we perform the Co-Citation analysis of the authors. The Co-Citation analysis helps us to identify and classify the literature into small subgroups according to the problem into a specific group [17].Figure13.Depicts the Co-Citation analysis of authors, which reveals the total number of Co-Occurrence of author in the network structure. The Figure is segmented into three clusters Red, Blue and Green. The authors falling in the same cluster are working collectively or in the related area of study. The authors in the Red Cluster are[18],[19], etc., involved in the area of "nosql, big data and mongodb etc.". The authors in Green Cluster are [20], [21], [22] Etc. is serving in the area of "nosql databases, cloud computing etc.

Conclusion With the continuous development of the IT Industry, the advancement of technology is producing massive amounts of data. The data in large volume, variety, veracity and velocity, is termed as Big Data. From the discussion and investigation, it has been observed that the deficiencies of the relational databases aforementioned such as no support for horizontal scalability and inadequate to process the large quantity of data efficiently, etc. These limitations of relational databases have led to the introduction of NOSQL databases. NoSQL, an approach to

handle Big Data, is a non-relational distributed database. NoSQL is not a replacement for relational databases, but this works as an alternative model. Several database models like Key-Value Store, Document-Oriented, and Column-oriented etc. are presented that can be chosen by the user as per requirements and the problem at hand. This paper gives a comprehensive overview of "nosql databases" from 2010 to 2020, using the Biblioshiny tool of R-Package bibliometric analysis conducted on research articles collected from the Scopus database. The bibliometric analysis assists us to endow the growth of "nosql databases". The outcome of the bibliometric review designates that “lectures of computer science and engineering, advances in intelligent system and computing", “Advances in intelligent system and computing” are the most prestigious journal in the field. The study acknowledges that "CHINA, the USA and INDIA" are the countries that are predominantly working in the domain of nosql databases. The authors Bernardino, Pokorn J are working in the field for the longest time. The analysis performed on keywords occurs in the title, abstract reveals that Nosql appears in all categories, which describe that nosql databases are a relevant and trending area of study. In future, the investigation in nosql databases can be grown in the realm of database systems, models of NoSQL databases and security of NoSQL databases.

References

[1] A. K. Zaki, ‘NoSQL DATABASES : NEW MILLENNIUM DATABASE FOR BIG DATA , BIG USERS , CLOUD COMPUTING AND ITS SECURITY CHALLENGES’, pp. 403–409, 2014.

[2] R. Rialti and G. Marzi, ‘Big Data and Dynamic Capabilities : A Bibliometric Analysis and Systematic Literature Review FULL TEXT (DOI)’:, 2018.

[3] A. Parlina, K. Ramli, and H. Murfi, ‘Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database’, no. Idc, pp. 1–26, 2020, doi: 10.3390/info11020069.

[4] A. Nayak, ‘Type of NOSQL Databases and its Comparison with Relational Databases’, vol. 5, no. 4, pp. 16–19, 2013.

[5] A. Gupta, S. Tyagi, N. Panwar, and S. Sachdeva, ‘NoSQL Databases : Critical Analysis and Comparison’, pp. 293–299, 2017.

[6] T. Losser and G. Harmon, ‘NOSQL VS RDBMS - WHY THERE IS ROOM FOR BOTH’, 2013.

[7] W. Nosql and N. Leavitt, ‘Will NoSQL Databases Live Up to Their Promise?’, pp. 12–14, 2010.

[8] R. Zafar, E. Yafi, M. F. Zuhairi, and H. Dao, ‘Big Data : The NoSQL and RDBMS review’, no. May, pp. 120–126, 2016.

[9] I. Zupic, ‘Bibliometric Methods in Management and Organization’, vol. 18, no. 3, pp. 429–472, 2015, doi: 10.1177/1094428114562629.

[10] A. Nasir, K. Shaukat, I. A. Hameed, S. Luo, and T. Mahboob, ‘A Bibliometric Analysis of Corona Pandemic in Social Sciences : A Review of Influential Aspects and Conceptual Structure’, vol. XX, 2020, doi: 10.1109/ACCESS.2020.3008733.

[11] C. Forliano, P. De Bernardi, and D. Yahiaoui, ‘Entrepreneurial universities : A bibliometric analysis within the business and management domains Technological Forecasting & Social Change Entrepreneurial universities : A bibliometric analysis within the business and management domains’, *Technol. Forecast. Soc. Chang.*, vol. 165, no. April, p. 120522, 2021, doi: 10.1016/j.techfore.2020.120522.

[12] M. Aria and C. Cuccurullo, ‘Science Mapping Analysis with bibliometrix R-package : an example Install and load bibliometrix R-package Section 1 : Descriptive Analysis Although bibliometrics is mainly known for quantifying the scientific production and

measuring its quality and', pp. 2007–2017, 2021.

[13] V. Abramova and J. Bernardino, 'NoSQL Databases : MongoDB vs Cassandra', pp. 14–22, 2013.

[14] R. K. Lomotey and R. Deters, 'Towards Knowledge Discovery in Big Data', 2014, doi: 10.1109/SOSE.2014.25.

[15] R. Yangui, A. Nabli, and F. Gargouri, 'Automatic Transformation of Data Warehouse Schema To NoSQL Data Base : Comparative Study', *Procedia - Procedia Comput. Sci.*, vol. 96, no. September, pp. 255–264, 2016, doi: 10.1016/j.procs.2016.08.138.

[16] Y. Li, Z. Xu, X. Wang, and X. Wang, 'A bibliometric analysis on deep learning during 2007 – 2019', *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 12, pp. 2807–2826, 2020, doi: 10.1007/s13042-020-01152-0.

[17] G. Surwase *et al.*, 'Co-citation Analysis : An Overview ISBN : 935050007-8', no. November 2015, 2011.

[18] J. Chen and W. Lee, 'An Introduction of NoSQL Databases Based on Their Categories and Application Industries †', 2019.

[19] X. Liu, R. Sun, S. Wang, and Y. J. Wu, 'The research landscape of big data : a bibliometric analysis', vol. 38, no. 2, pp. 367–384, 2020, doi: 10.1108/LHT-01-2019-0024.

[20] W. Nosql and N. Leavitt, 'to Their Promise ?', pp. 12–14, 2010.

[21] L. Okman, N. Gal-oz, Y. Gonen, E. Gudes, and A. Cassandra, 'Security Issues in NoSQL Databases', 2011, doi: 10.1109/TrustCom.2011.70.

[22] J. R. Lourenço, V. Abramova, M. Vieira, and B. Cabral, 'NoSQL databases : a software engineering perspective'.