البيانات الضخمة: البرامج مفتوحة المصدر فتحت الأبواب للابتكار في المكتبات

Big Data: How the open-source software will open the doors for innovation in libraries

أ.م.د. زينب عبد الواحد سلمان

الجامعة المستنصرية، كلية الآداب، قسم المعلومات وتقنيات المعرفة

Assist. prof. D. Zainab Abdullahid Salman

Al-Mustansiriya University - College of Arts - Department of information and Knowledge technical

Email: **drzselman@gmail.com**.

المستخلص :

في هذا البحث تم استعراض كيفية التعامل مع البيانات الضخمة، إذ إن جمع وتخزين المعلومات والتحليل للبيانات اصبح مسألة لا يمكن السيطرة عليها من خلال برامج معالجة البيانات بأحجامها التقليدية، سابقا كانت واحدة من أكبر العقبات التي تواجه شركات التكنولوجيا والمؤسسات الناشئة ومحللي البيانات هي كيفية القدرة على معالجة مثل هذه الأحمال الكبيرة من البيانات والتي كانت عائقًا كبيرًا للعديد من الشركات الناشئة أو المؤسسات البحثية غير الربحية ، لكن البرامج الحديثة مفتوحة المصدر مثل Hadoop وغيره أزالت هذه الحواجز، Hadoop هو منصة أو إطار عمل برمجي يسمح بتخزين ومعالجة البيانات على نطاق واسع، وهو مجاني ومتاح لجميع المبرمجين والمطورين. يهدف البحث الى تحديد كيفية التعامل مع مجموعة من البيانات التي يتجاوز حجمها قدرة برامج قواعد البيانات المعروفة لالتقاطها وتخزينها وإدارتها وتحليلها والذي يتطلب أشكالا مبتكرة وفعالة لمعالجتها تختلف عن معالجة البيانات العادية بحيث تمكن مستخدميها من تحسين الرؤيا واتخاذ القرارات. عينة البحث هي الرسائل والاطاريح الجامعية المتاحة بشكل رقمي بصيغة PDF وبصيغة Word والمتوفرة في المكتبة المركزية للجامعة المستنصرية، وقد بلغت (١٠٧٣٤٥) رسالة وأطروحة جامعية تمثل ٢،٤٩ تيرا بايت مقابل ٢٥٦٦١ كتابًا إلكترونيًا مخزنة في هذه المكتبة، وتمثل ٥٨٥٢ ميغا بنصها الكامل، وبهذا بلغ العدد الإجمالي للبيانات المؤرشفة ٣،٠٨

تيرابايت. وعلى الرغم من تنوع قواعد البيانات بين مكتبات الجامعة المختلفة، لكن السمة الغالبة في البحث هي حسب الموضوع أو المؤلف أو العنوان. ويتم استخدام طريقة البحث هذه في معظم أنواع قواعد البيانات المكتبية، ومن خلال عدة معايير منها الوقت والدقة وحجم المصادر التي يتم استدعاؤها في وقت واحد توصلت الباحثة في نتائج بحثها بان الوضع الحالي غير مرضي وربما يستمر كذلك في المستقبل بسبب التزايد المستمر في أعداد وأحجام الرسائل والاطاريح الجامعية وما يقابله من تنافس قوي من قبل البحوث العلمية إذ اصبح الباحثين يتوجهون لها في الوقت الحاضر ومع تعقيدات الوصول الى المعلومات الكاملة لمحتوى تلك الرسائل والاطاريح وعدم إتاحتها بالنص الكامل في اغلب قواعد البيانات وذلك بسبب عدم استخدام التقنيات الملائمة للتعامل مع البيانات الضخمة واستيعاب هذا الكم من البيانات فهذا يعني تردي الطلب على الاطاريح الجامعية قياسا بتزايد قوة الطلب على البحوث العلمية بسبب تعقيدات الوصول لمحتوياتها بالنص الكامل وعدم ملاءمة استراتيجيات البحث التقليدية لمواكبة احتياجات المستفيدين خاصة مع تزايد إتاحة الكتب بشكل رقمي رغم وجود بعض المحددات للوصول الى المحتوى الرقمي الكامل للكتب الرقمية. أوصت الباحثة انه من الضروري استخدام تقنيات تستجيب لاستراتيجيات البحث خاصة في البيانات الضخمة والبحث المتقدم عبر استخدام برنامج Hadoop لتغطية المخرجات الفكرية في المستقبل، وإمكانية استثمار Hadoop في مجال البيانات الضخمة واختيار المكتبة المركزية في الجامعة المستنصرية نموذجاً للتعامل مع البيانات الضخمة وكيف يمكن إن تساهم في تنظيمها.

الكلمات الدالة (البيانات الضخمة، المكتبات الجامعية، منصة هادوب، معالجة البيانات، الأنظمة مفتوحة المصدر)

Abstract

Big Data, defined in this paper as the gathering and storage of information and analysis on a scale typically untenable for traditional, mass-market data-processing software, has previously been one of the biggest obstacles facing tech companies, startups, and analytic researchers. The ability to process such large data loads has been a significant barrier of entry to the market for many young companies or not for profit research organization, but recent open-source software, such as Hadoop, have removed those barriers. Hadoop, a programming framework that allows for large-scale data storage and

processing, is free and available to all developers. This software allows independent developers,

the exceeds size ewhos data of set a with deal to how determine to aims research The which analyze, and manage store, capture, to programs database known-well of ability data ordinary from differ that processing of forms effective and innovative requires research The .making-decision and vision improve can users its that so processing in digitally available theses and theses university is samplePDF and Word and format (١٠٧٣٤٥) to amounted , University Mustansiriya-Al of library central the in available this in stored ooksb-e ٢٥٦٦١ to compared terabytes, ٢,٤٩ representing thesis, and theses of number total the thus and , text full its in megabytes ٥٨٥٢ representing and library, the between databases of diversity the Despite . terabytes ٣,٠٨ reached data archived author subject, by is esearchr the in feature dominant the libraries, university different through and , databases library of types most in used is method research This title. or at called are that sources the of size the and accuracy, time, including criteria, several from competition strong corresponding the and theses and theses University . time one the with and time present the at them to turning are researchers as research, scientific and theses and theses these of content the for information full accessing of complexities techniques appropriate of lack the to due databases stmo in text full in available being not in deterioration a means this data, of amount The this absorb and data large with deal to scientific for demand increasing the to compared theses, university for demand the the and text full the in contents its accessing of tiescomplexi the to due research the of needs the with pace keep to strategies research traditional of inadequacy the despite digital, in books of availability increasing the with especially beneficiaries, The digital. books of content digital full the access to slimitation some of presence search to respond that techniques use to necessary is it that recommended researcher using by research, advanced and data big in especially strategies,Hadoop program to

the in investing Hadoop of possibility the and future, the in outputs intellectual cover a is University Mustansiriya-Al at library central The choosing and data big of field it organizing to contribute can it how and data big with dealing for model

Keywords (big data, university libraries, Hadoop platform, data processing, open source systems)

Introduction

With the rise of the Internet and recent technological advances, information has taken on a whole new role in our modern world. Whereas information was simply equated with knowledge, it has now become something of a commodity that can be gathered, stored, analyzed, processed, and used to create new methods of understanding. The information discussed in this context is data—data driven from clicks, browsing, purchases, and information given voluntarily by users of the Internet. Over the past two to three decades, the research and management of this information has become an entire field of study and commodity market that of which has never been seen before.

Data is growing at an astronomical rate, with more being produced in the last few years than in the entirety of prior human history. Our volume of information grows exponentially every day, and it will never cease in its expansion. This growth will only continue, with some reports claiming that “by 2022 containing nearly as many digital bits as there are stars in the universe. It is doubling in size every two years, and by 2022 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes.(IDC 2014)” With this huge amount of new information at their disposal, researchers faced a problem in how to manageable collect, store, and analyze this constantly expanding network of data. This issue and the processes designed to deal with it are commonly referred to as “Big Data.”

In his 2001 report, “Application Delivery Strategies,” Doug Laney put words to the formal idea of “Big Data” and introduced the idea of the three V’s in data management: volume,

velocity, and variety.(Laney 2001). The volume vector deals with the considerable size of the quantities of data being processed, while the velocity measures how fast that data is coming in, with the variety vector considers the unstructured and incredibly varied types of data that will be sorted. Through this assertion, Laney clearly explained the double-edged sword of data management. On one hand, volume, velocity, and variety of information is wonderful for researchers, but it also presents a significant challenge when it comes to managing these factors. Volume is the main and most pressing issue for small researchers as each piece of data represents memory that must be processed and occupies space. The volume of this information is constantly increasing and expanding which benefits the variety of knowledge but requires a complex system to handle its storage. Up until recently, this data management was limited to costly software or frameworks that existed largely in the private sector or within large technology firms.

Laney compares the three V's in e-commerce situations. He notes how e-commerce channels increase the depth of data available about a transaction, allowing an enterprise to collect up to 10x the quantity of data about an individual transaction, increasing the volume of data to be managed. As this data appears to be an asset to the company, they are reluctant to discard the data so they turn to purchasing online storage. This becomes an issue as the volume of large data collected in every transaction begins to pile up and the company runs the risk of poor financial justification for protecting the data. As the company begins to grow and sell more online, there becomes an increase in data velocity – now not only is storage of data an issue, the company is acquiring more data more quickly. Finally, Laney points out that as we grow online, so too does the style of data we create. The variety of data that is created, shared, and stored is abundant and so too must be our plan for dealing with Big Data. (Laney 2001)

Since Laney's first introduction to the study of data collection, numerous organizations and researchers have suggested other qualities that contribute to the quality and properties of Big Data. For example, SAS, software management firm based in the United States writes on their website that they "consider two additional dimensions when it comes to big data:

challenges, many faces researchers from data big with Dealing :**problem Research** limit a reach can it and second yever produced is data world, s'today in that including and produced, is it as soon as data this to access Quick imagination. beyond goes that delivery its up speeding and data this analyzing of feature this achieve must platform the .it from benefit desired the achieve to, i be to which in stored is data the cases many n platforms cloud in analyzed, b cannot data their and complex very are resources data ig data big in data the Therefore, columns. and rows of set simple a into organized be annual If data. embedded the describes atth way a in explained be must resources exponential an and year every collected data the of doubling a to lead developments .it in growth

relevant is query the from extracted data the whether determine to impossible almost is It when and query the of purpose intended the to correspond that iespropert of set a has i.e. with query the to relevant resource data big the in data the all represents it combined .values data described-well

bservations,o collecting for resource important an is data Big :**research of importance** the provides also It uses. various its test to used be can method experimental the and with us provides it as information, of use the forecasting studying of possibility and trends chresear their of prediction future and users by handling data of knowledge comparison. by reached have we data the of validity the verify to opportunity an gives .studies data small With

objective Research: t data of set a with deal to how determine to aims research he store, capture, to programs tabaseda known-well of ability the exceeds size whose that processing of forms effective and innovative requires which analyze, and manage and vision improve to users its enable to as so processing data normal from differ .making-decision

theses university of study case A :**sample hresearc** the And **Methodology Research** in digitally available theses andPDF and Word central the in available and format thesis, and theses (١٠٧٣٤٥) to amounted It , University Mustansiriya-Al of library and library, this in stored books-e ٢٥٦٦١ to compared terabytes, ٢,٤٩ representing data archived of number total the thus and , text full its in megabytes ٥٨٥٢ representing . terabytes ٣,٠٨ reached.

Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage, even more so with unstructured data.

Complexity. Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.(SAS 2015)"

Other organizations, such as BBVA, suggest that qualities such as veracity and value must be taken into account when dealing with adequately processing large amounts of data(BBVA 2017). All of these factors point to one conclusion: that Big Data is an ever changing and growing field with new challenges and systems emerging every day.

Challenges of Large-Scale Data Collection

Data management varies significantly based on the application and organization processing the analytics. Private companies, like Google or Facebook, utilize data research and management programs in order to create better products for the consumer market and generate profits. Because the majority of practical data generation has been used for and funded by private interests, the majority of conversation about Big Data has been focused on how to provide **corporations** with the software tools they need for **profit generation**. The conversation, therefore, has inherently left out a major segment of software users: academia

and research libraries. These, often public, institutions are an ever-growing segment of the research and data management landscape and better software for the processing and storing of data has opened the doors on analytic possibilities for such organizations.

In his 2018 workshop on “Effective Management of Big Data and Research Data within Academic Libraries,” Marshall Breeding made the following observation:

“Libraries associated with research-oriented universities are increasingly involved with providing services in support of managing the data produced through research projects. The data repositories of each academic department have traditionally been managed internally and often informally. As these data repositories increase in scale, the provision of adequate storage and management infrastructure presents challenges. Many organizations that provide funding or oversight of scientific research now make specific stipulations regarding the treatment of data. The National Science Foundation in the US, for example, requires all grant proposals to include a Data Management Plan, which states how the data produced through the research will be managed, preserved, and made available. Many libraries have initiated services to work with university faculty members and departments in support of the development of research data plans.(Breeding 2018)”

New open-source programs like Apache Hadoop provide libraries and academic researchers with a new solution for managing the substantial quantities of data often required for analytic study. By allowing anyone the power to run significant processing and storage of information, Apache Hadoop's software takes away the previous barrier to entry for data research and management. This has the ability to revolutionize academia and learning as a whole and gives an entirely new set of free and easy to use tool to those who study the field of quantitative data research.

In Switzerland, the Large Hadron Collider is one of the most powerful machines in the world that is equipped with over 150 million sensors that create a petabyte of data every second. Since this data has been growing since the beginning of CERN, in both size and complexity, the researchers run a Hadoop cluster to help deal with the massive volumes of data acquired. While a company as large as CERN could benefit from a large distributed database or a

RDBMS, Hadoop offers a lot more flexibility and scalability with big data that isn't offered in RDBMS.

The software framework of Hadoop works extremely well with all kinds of data, whether it is structured, semi-structured, or unstructured, in large sizes from Tbs to Pbs. This variability also allows Hadoop to support a wide variety of data formats in real time, while RDBMS works very efficiently with only structured, average sized data. An RMBDS requires that the entity-relationship flow is defined perfectly, while Hadoop is the better choice when the need for big data processing does not have dependable relationships. The consistency and rigidity of RMBDS allows it to excel in dealing with online transaction processing and is often chosen by large companies dealing with financial and structured data. However, Hadoop's ability to process all forms of structured and unstructured data, as well as the capabilities to alter the program to users' needs at a low cost make it more accessible to companies. The analysis and storage of Big Data are made easier only with the help of Hadoop and its scalable, data-intensive program, rather than the traditional RDBMS. (Bista 2018)

In comparison:

RDBMS	HADOOP
Is mainly used for structured data	Used for structured, semi, and unstructured data
Can handle average data (GBS)	Can handle large data (TBS and PBS)
High costs for the license	Free
Used mostly for online transaction processing.	Used mostly for analytics of data and data discovery.

Since Hadoop is an open-source framework that anyone can learn or work within, it represents a large-scale democratization of Big Data management. By this, we mean that programing frameworks like Apache Hadoop allow anyone with coding knowledge to being their own process of gathering, analyzing and collecting data for research purposes. This means that startups, independent developers, entrepreneurs, libraries, and academic researchers all

now have the same tools at their disposal as major technology corporations like Facebook, Google, or Yahoo! Previously, to handle the enormous load of information in their system a research center or library would have had to build their own framework and processes reliant on expert programming knowledge and extensive hardware, or purchase costly data processing software. Now, these features are at the disposal of anyone with access to a computer and a basic knowledge of programing. All of this is possible because of frameworks like Apache Hadoop, which will be explored further in the following section.

Apache Hadoop

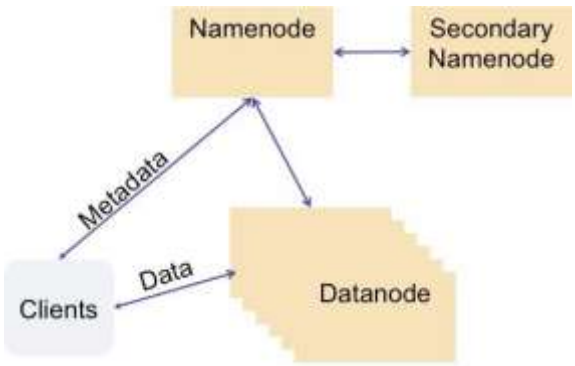
Apache Hadoop is a new open-source software, which had its stable release in December, 2017 (though it existed in numerous private use or beta forms before said release) and has been making waves in the programing world for its exciting new application in the field of Big Data management. The software is largely written in Java programming language making it accessible to even the most basic of developers. Anyone who can learn basic coding is able to access and use the Apache Hadoop framework integrated into his or her data processing.

Because it is open source, there is no single way to use Hadoop and many ways to customize it to the particular needs of the user. For example, Facebook programmer Andrew Ryan explained the way the company uses Hadoop in their data management by stating:

"HDFS clients perform file system metadata operations through a single server known as the Name node, and send and retrieve file system data by communicating with a pool of Data nodes. Data is replicated on multiple data nodes, so the loss of a single Data node should never be fatal to the cluster or cause data loss.

But the loss of the Name node cannot be tolerated. All metadata operations go through the Name node, so if the Name node is unavailable, no clients can read from or write to HDFS. Clients can still read individual data blocks from Data nodes if the Name node is down, but for all intents and purposes, if the Name node is unavailable, HDFS is down, and users and applications that depend on HDFS won't be able to function properly."

The following graph explains the way this meta data is collected and held within their systems:



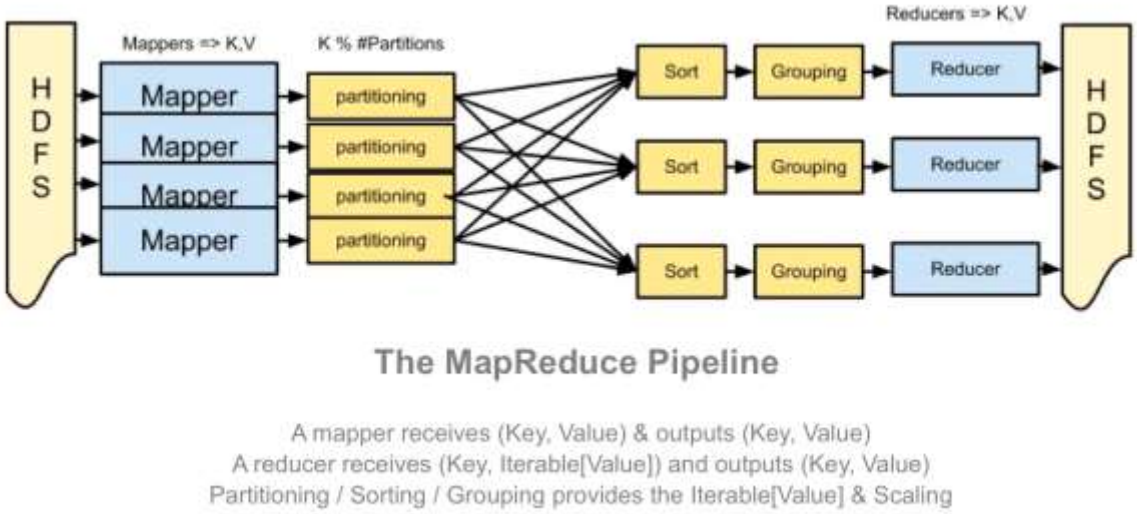
Simplified HDFS Architecture: Namenode as SPOF

Furthermore, Apache Hadoop is able to process information on such a significant scale because it uses a Map/Reduce method of organizing the data it does collect. That method is detailed below:

"Map/Reduce is a programming paradigm that expresses a large distributed computation as a sequence of distributed operations on data sets of key/value pairs. The Hadoop Map/Reduce framework harnesses a cluster of machines and executes user defined Map/Reduce jobs across the nodes in the cluster. A Map/Reduce computation has two phases, a *map* phase and a *reduce* phase. The input to the computation is a data set of key/value pairs.

In the map phase, the framework splits the input data set into a large number of fragments and assigns each fragment to a *map task*. The framework also distributes the many map tasks across the cluster of nodes on which it operates. Each map task consumes key/value pairs from its assigned fragment and produces a set of intermediate key/value pairs. For each input key/value pair *(K, V)*, the map task invokes a user defined *map function* that transmutes the

input into a different key/value pair *(K', V')*. Following the map phase, the framework sorts the intermediate data set by key and produces a set of *(K', V'*)* tuples so that all the values associated with a particular key appear together. It also partitions the set of tuples into a number of fragments equal to the number of reduce tasks.(Ryan 2012)"



(Mothilal 2016)

In 2010, Facebook became the single largest user of Apache Hadoop mapping software. The company published a note on their platform in 2012 entitled "Under the Hood: Hadoop Distributed File system reliability with Name node and Avatar node," explaining the way they used the framework for the data processing systems within the company. They state, "The Hadoop Distributed File system (HDFS) forms the basis of many large-scale storage systems at Facebook and throughout the world. Our Hadoop clusters include the largest single HDFS cluster that we know of, with more than 100 PB physical disk space in a single HDFS file system. Optimizing HDFS is crucial to ensuring that our systems stay efficient and reliable for users and applications on Facebook.(Ryan 2012)" The company helped pioneer this application on a scale previously unthinkable as they began collecting, storing, and analyzing the data of the now billions of users. Every like, comment, message or share began to be housed in a massive data retention operation on the company's servers. This information was used to help

determine future visitor habits, but it ended up serving a much larger goal as a test pilot for Big Data retention in any contest. In this way, organizations like Facebook and Google have served as beta testers and modern stand-ins for the traditional research library.

Because of their capital and financial interests, they were able to work with creators and perfect a software framework for use in the Big Data field. These advances can now be used in an open-source framework by researchers, libraries, and academic information catalogues. In many ways, the framework and tools which will continue to propel academic and scientific research forward would not have been possible without technology corporations like Facebook and Google.

Possibilities for the Future

The possibilities for application of software frameworks like Hadoop within the academic and library field are endless. Nearly all fields of study will be able to use programs like Hadoop to grow their understandings and depth of study, but some will benefit even more than others. Medicine, for instance, is one of the fields to benefit most from this new data management framework. Using programs like Hadoop, doctors and medical researchers will soon be able to track patient analytics on every scale.

Imagine a world in which every piece of information gathered about illness or a patient's medical history can be collected, tracked, and studied against the data of others. Big Data Management has made that a very real possibility. Doctors and medical researchers are now able to look at every step of the illness process, analyze the benefits and side effects of medication, track surgical procedures, and map genomes to predict future disease. All of this information at the hands of researchers and libraries will open the door for technological advancement previously unimaginable, all because data is now able to be stored and mapped on every level.

In his workshop on "Effective Management of Big Data and Research Data within Academic Libraries," Marshall Breeding makes the point that academic and library institutions are going to benefit from the Big Data management advancements the most. He writes:

"Organizations engage with big data to gain specific results. Scientists create data through research processes to make discoveries, test hypothesis, or other related goals. The commercial realm collects and analyzes big data to identify spending trends, power recommendation engines, to enable highly targeted advertising, to refine design of products, or a host of other possibilities to maximize their business outcomes.

Libraries and educational institutions will have their own goals in mind as they begin working with big data. One possible application involves using big data to inform the development of services offered by the library. By collecting and analyzing usage data, some of the aspects of its services and operations can be assessed and refined:

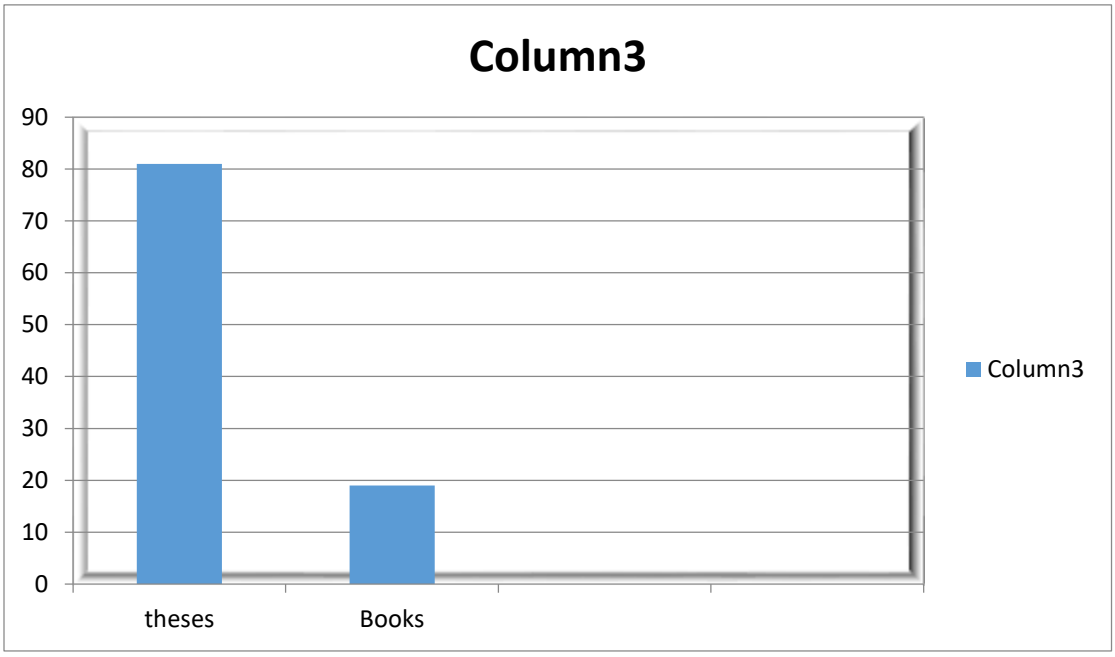
- Design of digital services: understand user behaviors, resource usage, navigational pathways, and other aspects of user interfaces and content offered. Google Analytics can be considered an example, creating a data warehouse of use data and analytic tools able to inform optimal web design to maximize desired user actions. While designed to inform the use of advertising, organizations can also use Google Analytics to optimize other desired actions such as digital downloads, successful search strategies, or other goals.
- Refinement of information systems. An information discovery environment requires a sophisticated understanding of the content involved and user expectations to function optimally. Fundamental features such as relevancy rankings rarely work well using the default algorithms. Additional context data can improve the quality of search results, such as the discipline of interest and associated specialized vocabularies.(Breeding 2018)"

When the barriers of managing data and technology are removed, the potential advancements are endless.

Practical aspect

For the purpose of using the various sources of information stored in the databases of Iraqi universities libraries by beneficiaries. These sources have to be managed in such a way as to ensure that the data is communicated to the adherent without a lack of information through the use of best practices by university libraries and information institutions. Here it is necessary to sweat the information available in the library databases of Iraqi universities.

Despite the diversity of information sources, the digital format available from those sources has been formed by two important authorities, which are theses and university theses, with a rate of 80% of the total information sources available in (doc. Pdf) libraries (the subject of study), which are in two forms (Pdf)and 19% represented in digital books, it is in the form of:



The total number of archived databases reached 3,08 terabytes which represents (107345) theses or university theses represent 2,49 Tera bytes, compared to 25661 electronic books stored in the databases of these libraries, which represented 5852 megabytes within its full text there are only hyperlinks associated with its biographical information. So, there is a preference for university theses that have been better explored through several entrances and allow searching within the text, when examining the databases, the researcher found, and despite the diversity of databases among the various university libraries, the dominant feature is in the search operations by subject, author or title. It is used in most other types of databases. It has a large body of research in this field, and through several criteria, including time, accuracy, and the size of the sources that are called at one time, the researcher finds these results unsatisfactory in the near future.

1- Enlarging of intellectual production, especially in theses and university theses, and the steady increase in them annually, as indicated by many local types of research and studies, which are steadily growing annually in their sizes.

2- Entering scientific research as a strong competitor to a university thesis, especially those published globally to be within the local scope of local databases, after the great assurances made by the Iraqi ministry of higher education and arguing in this field.

3- Increasing the quantities of books authored by teachers and experts in several fields and their direct availability in digital form, it was made into one of the cornerstones of the aspirations of the beneficiaries.

4- The inadequacy of traditional search strategies to keep pace with the needs of the beneficiaries, especially as they depend on the main features in database tables without taking Connections between these fields.

Therefore, it is necessary to use techniques that are responding to the search strategies, especially in reversed exaggerations and double search, which is the use of the Hadoop program or other programs to cover the intellectual outputs in the future.

Advantages and Disadvantages of Big data:

Disadvantages	Advantages
the quality of data	enhanced productivity (improvement Productivity)
rapid change	Butter costume services (beneficiaries for)

Challenges of Big Data in libraries:

lack of professionals (of shortage The field this in specialists)	improve decision making
Cyber security risks	minimization the cost (reduction cost)
flexibility and malleability)	Revenue increased (revenue Increase)
hardware needs (needs Equipment)	fraud detection
	increased agility and stay in competition (Increase competitive stay and activities)

1. Needs for management across dissimilar data sources**.**
2. Lack of professionals how to know big data analysis**.**
3. Achieving correct business insights out of big data**.**
4. Management of voluminous data**.**
5. Storage issue**.**
6. security and privacy of data

results Search

1. Despite the diversity of databases between the different university libraries, the dominant feature in searching for information sources is by subject, author or title. This search method is used in most types of office databases , and with the complexities of accessing the full information of the content of these dissertations and theses, and not being available in full text in most databases, due to the lack of appropriate techniques to deal with large data and absorb this amount of data And the inadequacy of traditional search strategies to keep pace with the needs of the users

2. Big data statistics suffer from many fundamental flaws when the data is large enough that we can find any data lurking within it somewhere, and the observations that can be found may be statistically significant but without actually making any sense. Whenever we choose a subset of big data, we may have no way of knowing the .significance of the data that has been left out
3. Possible applications in the use of big data to inform the development of services .provided by the library
4. By collecting and analyzing usage data of users, we can evaluate and improve some aspects of the library's services and develop its procedures.
5. Designing digital services: understanding user behaviors, resource usage and users' browsing paths, dealing with user interfaces, and handling advanced research . ,Google Analytics can be taken as an example
6. A repository can be created to use data and analytical tools capable of directing optimal web design to achieve maximum of required user actions.
7. Organizations can also use Google Analytics to improve other desired actions such as .digital downloads, successful search strategies, or other goals
8. Refinement of information systems. An information discovery environment requires a sophisticated understanding of the content involved and user expectations to function optimally. Basic features such as relevance ratings rarely work well with default algorithms. Additional context data can improve the quality of search results, such as subject of interest and related specialized vocabulary.
9. The application of open-source software such as Hadoop will provide endless possibilities in academic research and library studies. Research in the field of using programs such as Hadoop will contribute to expanding the awareness and understanding of researchers and open up broad scientific horizons for them in developing their research. Still, some libraries have not benefited from these programs effectively for

themselves and their users so far, unlike other institutions that have benefited extensively from these programs.

10. Big data needs a technological infrastructure that includes tools for capturing, analyzing and storing information to visualize the huge amounts of unstructured big data that can be modified, updated, processed, and transformed into valuable information to promote the potential growth of using that data to improve retrieval .efficiency. This cannot be accomplished by relying on old technologies

11. researcher also found in the results of her research that the current situation is not satisfactory and may continue to be so in the future due to the continuous increase in the number and sizes of university dissertations and theses and the corresponding strong competition from scientific research, as researchers are turning to it at the present time, which led to the deterioration of the demand for University theses in comparison with the increasing demand for scientific research due to the complexities of accessing its contents in full text and the inadequacy of traditional research strategies to keep pace with the needs of users, especially with the increasing availability of books in digital, despite the presence of some limitations to access the full digital content of digital books. The researcher recommended that it is necessary to use techniques that respond to search strategies, especially in big data and

advanced research, by using Hadoop program to obsorb intellectual outputs in the future.

Conclusion

Big Data is a growing and evolving field that will greatly benefit in the long run our libraries that contains literature and knowledge. Although it is effectively useful, it presents its challenges and obstacles, especially in implementing big data management programs and platforms. There's a need for libraries to Provide a new frameworks and methods of work that are compatible with the changing needs of the users, increase of their numbers, the large amount of data produced by researchers and specialists, and working on investing in modern technologies through open source software in this field, such as Apache Hadoop software, which allows a continuos tracking and collecting big data that is continuously growing and expanding. This new program will allow institutions such as libraries and academic research centers to use the same advanced big data handling techniques that were the monopoly of private companies in the field of Technology such as Facebook or Google, as these new technologies have the potential to open doors for research and allow significant progress in academic communit,. and creating innovation in research areas.

Works Cited

1. BBVA. *BBVA.* 05 08, 2017. https://www.bbva.com/en/five-vs-big-data/ (accessed 04 04, 2018).
2. Bista, Narayan. *EDUCBA.* 02 16, 2018. https://www.educba.com/hadoop-vs-rdbms/ (accessed 04 04, 2018).
3. Breeding, Marshall. "Effective Management of Big Data and Research Data within Academic Libraries." *24th Annual Conference and Exhibition of the Special Libraries Association.* Oman: Special Libraries Association, 2018.
4. IDC. *EMC2.* 04 01, 2014. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm (accessed 04 04, 2018).

5. Laney, Doug. "3D Data Management." *Gartner Blog Network.* 02 06, 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 04 04, 2018).
6. Mothilal, Deepika. *SQL Fanatic.* 02 05, 2016. https://sqlfanatic.wordpress.com/2016/02/05/what-is-hadoop/ (accessed 04 04, 2018).
7. Ryan, Andrew. *Under the Hood: Hadoop Distributed Filesystem reliability with Namenode and Avatarnode.* 06 13, 2012.
8. SAS. *SAS.* 05 04, 2015. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html (accessed 04 04, 2018).