2021

Application of Hadoop as Big Data Infrastructure in Libraries

Gopal Ji
University of Calcutta, gopalji@drtc.isibang.ac.in

Ranjeet Kumar Singh
University of Calcutta, ranjeet@drtc.isibang.ac.in

Application of Hadoop as Big Data Infrastructure in Libraries

Gopalji
(Research Scholar, DRTC)

Ranjeet Kumar Singh
(Research Scholar, DRTC)

Abstract

In the era of data explosion and big data, it is a challenging task for the data managers or data analysts to store and process big data for analysis and visualize it. Big data imposes many challenges and libraries to have big data which can be analyzed to bring new insights which will help in managing the library and its services better. The present study shows the traces of big data in libraries and how library big data can be used as an opportunity to serve the patrons better. The present study advocates the use of the Apache Hadoop framework and ecosystem in the libraries to solve the challenges of big data in libraries. This paper talks about the features and problems of big data and the basic concepts behind Apache Hadoop and its components for solving such problems. Libraries/Librarians can assist researchers in handling big data by providing them with data curation and cleaning services.

Keywords- Big Data, Library Big Data, Apache Hadoop, Apache Hadoop Ecosystem.

1. Introduction

Why do we hear the term Big data so frequently these days? That is because the data is growing at a very high speed, and it's an assumption that around 90%-95% of web data generated over the last two years, which tends to continue. So, to address the problem, we often use the term big data. Now the question arises what is big data and what is significant in big data? The term big data refers to the datasets that are so large and complex that it is beyond the control of traditional data processing applications, and it requires some additional powerful tools and applications to process. Any datasets that are unprocessable to us or by conventional database systems are significant and can be called big data. Hence, 'big' refers to our limitations of controlling or processing a dataset.

Factors behind big data are; a) *Evolution of Technology*- Aforementioned is a significant factor behind the emergence of big data. For example, earlier telephone used to do the task of receiving and making calls. Presently smart mobile phones used for multiple works other than primary calling features like internet browsing, messaging, etc. Increased use of smartphones increases the data generation rate due to their multifunctionality. Such an evolution in technology from car to the smart car, city to smart city, desktop to cloud storage are generating big data.; b) *Internet of Things (IoT)*- The era of IoT which has a wide range of devices such as smartphones, wearables, vehicles that contains sensors and these sensors generate vast amounts of data rapidly.; c) *Social Media*- With the emergence of social media, each individual of the world produces data every second. Until 2015, there were approx 1.7 million Instagram pics, 3.4 lakhs tweets, 300 hours of video uploads on Youtube, 204 million emails, 4.2 million Facebook posts & likes, etc. in every 60 seconds. The amount and speed of data generated through social media lead to the emergence of big data.; d) *Other factors*- We can have data from various factors alike Insurance, Media & Entertainment, Education, Health Care, Government, Banking & Finance, Transportation, Weather, Retail.

Generally, big data have 3V's, but this paper considers 5 V's as big data features; i.e., volume, variety, velocity, value and veracity. (i) *Volume*[1]- The size is a primary and fundamental factor of big data. The bulk of the data is prominent in the sense that it is difficult to process by the existing systems, which makes it "big" in big data. It expected that the volume of data would increase by ten times what it is in the present. Consequently, the size of big data is a big challenge to handle. (ii) *Variety*- Not only the volume of the data but a different range of data is also going to be challenging. However, structured data in the form of tables, semi-structured data in the form of .csv, .xml, .tsv, .json, e-mails etc. and unstructured data in the form of a log, audio, video, image, etc. Structured data is straightforward to handle and process, but handling unstructured and semi-structured data is very difficult. (iii) *Velocity*- Data generated at an alarming speed. The world witnessed immense growth in data generation speed over the last few years. As the new technologies emerge, this pace will increase exponentially. Studies found that more than lakhs of new tweets, instant messages, e-mails, etc. are created in 60 seconds. Matching the processing speed of such data with its generation speed is a challenge. (iv) *Value*- Big data features the mechanism to bring the correct meaning out of the data. Not all the data that is generated nowadays is useful or valuable. A mechanism needs to develop that can bring value out of it and then only drawing correct conclusions from the data is possible. (v) *Veracity*[12]- Real-world data is full of inconsistencies and uncertainty. Dealing with such real-world data full of irregularities and outliers are other problems that are challenging.

2. Traditional Database Management Systems and their problems in ILS

Currently, traditional relational database management systems (RDBMS) are in use in almost all the Integrated Library Systems(ILS). Several times for a library that has a vast collection-physical or digital, and has an eternal user base, it becomes difficult for the traditional RDBMS to handle the enormous data generated. Traditional RDBMS such as MySQL, Oracle are involved in handling datasets with less complexity. But with the advent of data complexity, these standalone traditional RDBMS are unfit to handle the work of storing the exponentially growing datasets, processing the data faster and having the complex structures of data.

3. Problems or Issues with Big data

Big data is a problem right now across the different domains, including library and information science. The following can be said to be four main problems or issues with big data.

1. *Exponentially growing data storing*- Earlier data was in kilobytes, megabytes and gigabytes that can be stored easily for analysis. But presently, data are in terabytes, petabytes, exabytes, zettabytes and yottabytes. Also, the data generation speed is very high and alarming, and a considerable amount of data generates every second. So, storing the exponentially growing data, which is abundant in volume and can be called big data, is a big problem.
2. *Complex structures of data storing*- Data can possess various structures like structured data, semi-structured data or unstructured data. Consequently, saving the massive volume of data is a problem but also storing the complex structures of big data for analysis is a problem.
3. *Complex structures of data processing*- Data is of no use if only stores. It needs to be processed also. But, processing complex structures of data is not so easy rather a tricky task to perform. Traditional systems lack in performing any operations on complex structures of big data. Thus, processing the complex structures of data is also a big problem associated with big data.
4. *Quick treatment of massive datasets* - Along with processing the complex structured data, how to handle it faster is a big question hereabouts. Every piece of data has a specific and limited life span in which it needs processing to furnish insights and aid in decision making. After a particular period, data becomes obsolete, and if it remains

unprocessed during the specific time, then it becomes outdated. It is a tedious job for the data analysts while processing the data. Henceforth, imagine the case where data will be in terabytes or zettabytes, how much time it will take to process. Accordingly, this is another critical problem associated with big data.

4. Apache Hadoop as a solution to big data

In the era of the information economy, there is a saying that "Information/Knowledge is power". The sense of the statement is so real that if proper and sufficient information about something or desirable things to do, then only success can be achieved. With additional information, people/organization/country will dominate in every field. But the question is where to get information? For extracting information, it requires digging or diving deep into the big unused datasets generated at an alarming rate. It also demands to solve the complexity of such large datasets for processing and analyzing, which is a problem in handling big data. Apache Hadoop is the solution to these problems[4].

Apache Hadoop is a multitasking, open-source platform framework, written in Java, which is capable of taking care of all the challenges and problems possessed by big data. Apache Hadoop allows for big data analytics[7]. It stores and processes big data in a distributed manner across clusters of computers[9]. It uses simple programming models to process the data. It allows scaling up from single servers to hundreds of machines, where each device offers local computation and storage. Parallel and Distributed processing is the central core of Hadoop.

5. Hadoop Architecture

Apache Hadoop architecture consists of mainly four components, that are-Hadoop Common, Hadoop Distributed File System (HDFS), MapReduce, and Yet Another Resource Negotiator (YARN)

5.1. *Hadoop Common*- Hadoop commons, contains utilities that will help the efficient working of other Hadoop modules, regarded as the base core of the Apache Hadoop framework. This framework consists of shared Java libraries needed by all other Hadoop modules that provide essential services and underlying processes.

5.2. *HDFS*- HDFS stands for Hadoop Distributed File System, is the storage unit of Hadoop and solves the storage problem of big data discussed above. Unlike the local file system which stores the data directly into the systems, it allows dividing any kind of input data or files into small chunks called blocks and stores it in a distributed fashion across the cluster. It also solves the problem of storing the complex structures of data as it saves any kind of data, be it structured, semi-structured or unstructured. HDFS possesses a master-slave arrangement and has two main components-NameNode and DataNode. NameNode is the primary node, that is the master daemon, that maintains and manages the DataNodes and records the metadata (e.g. location of the blocks, the size of the files, permissions, etc.) about the data stored in DataNode and tracks the working status and reports of data nodes. It also keeps the records of every change that has taken place to the file system metadata. DataNodes, the slave daemons, are nothing but the commodity hardware arranged in different racks and these DataNodes store the data and execute the read and write requests from the clients. Within HDFS architecture, there exists a NameNode and can have multiple slave machines arranged in multiple racks connected to the master via core switches. Whenever a file (big data) needs to be stored, under such architecture, it splits into several blocks depending upon the size of the blocks (default block size is 128 MB that can be user adjustable). After splitting, blocks get stored in the slave machines by following the rack awareness algorithm with a default replication factor of three (adjustable) — each block to protect any loss of data due to system or node failure. Thus, Hadoop has a fault tolerance system so that if any mishap happens with any node, data will remain unaffected. Though it will use some extra hardware, nothing can price more than the data.

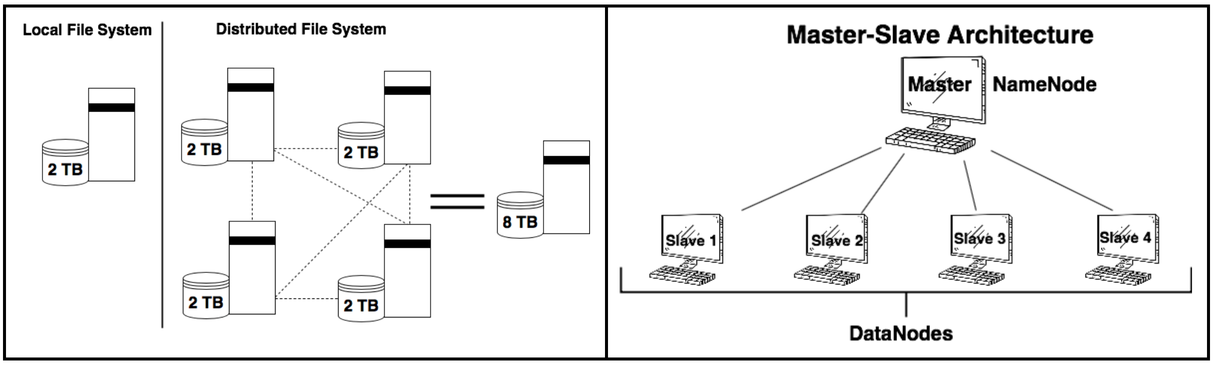

Figure 1: Commodity hardware setup showing file system and master-slave architecture.

5.3. *MapReduce*- MapReduce is the processing unit of the Apache Hadoop framework, which is capable of processing data parallelly in a distributed environment that enables fast processing and saves lots of time. It replaces the batch processing or sequential processing of the traditional systems which use to take a longer time for processing. It is a program model having features of parallel programming and performs two functions Map and Reduce which works on a key-value pair. Here the Map function (assigning key-value pairs) refers to the processing of data in HDFS, and provide the opportunity to process the data inside every block parallelly. Afterwards, each block sends the result to reduce function where aggregation of results happens to support a decision. Typically, it maps all the processing in different neighbourhoods and then reduces the time and effort to aggregate results. MapReduce saw as a remedy to solve the problem of big data as it reduces the processing time by almost ten times and can be frequently used for indexing, searching, classification, recommendation, analytics and many more. Google well adopts MapReduce and Apache Hadoop for HDFS, Pig, Hive, etc. to work efficiently.

5.4. *Yet Another Resource Negotiator (YARN)*- Aforementioned is again a critical component introduced in Hadoop 2.X. It gives the freedom and ability to run non-MapReduce jobs within the Apache Hadoop framework. It provides a generic resource management framework and an abstraction over MapReduce for implementing Hadoop applications. By this, it makes the Hadoop framework more powerful. The main idea behind the development of YARN is to divide the works of job trackers which manage the resources; i.e. job scheduling or job monitoring.

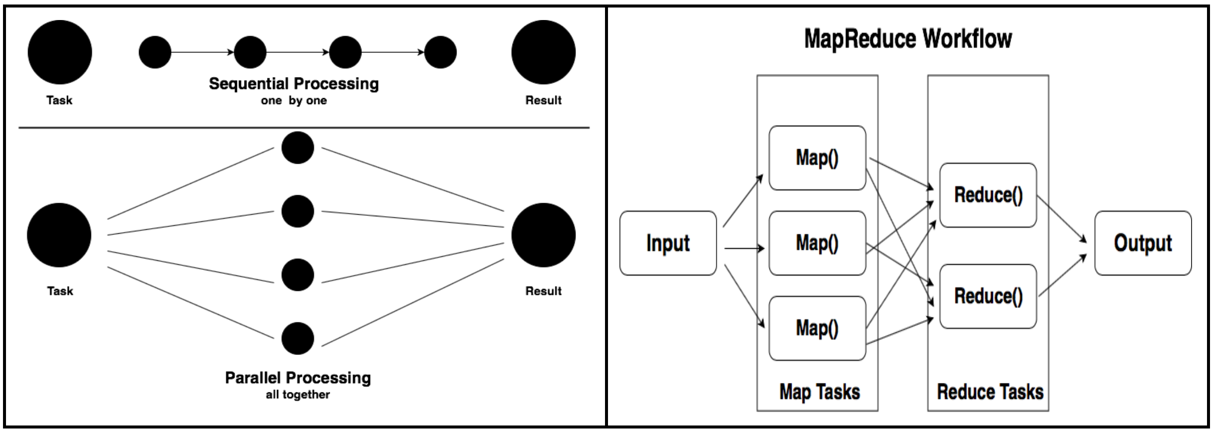

Figure 2: MapReduce working mechanism using parallel processing.

5.4.1. ***YARN Components***[3]**-** YARN has four major components which are Resource Manager, Node Manager, Application Master, and Container.

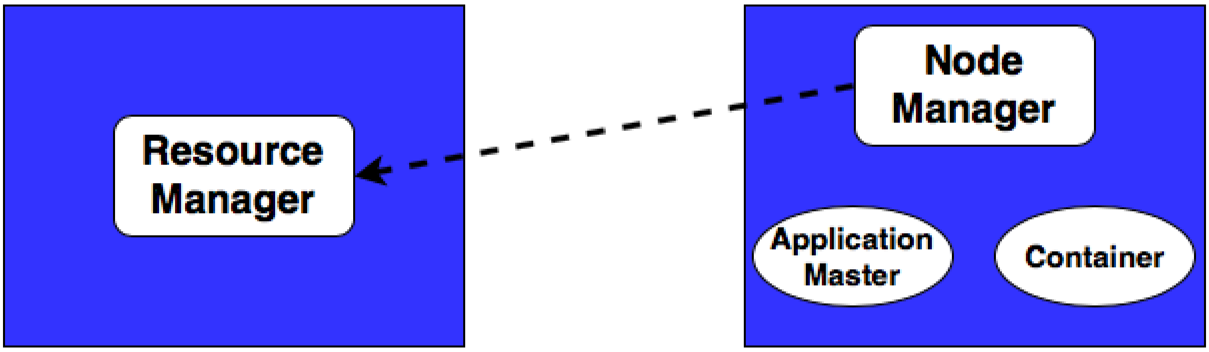

Figure 3: YARN components.

5.4.1.1. *Resource Manager* [10,11]**-** It is the central authority that settles resources among all the applications. The resource manager, the master daemon, accepts the requests for job submission and manages all other daemons. Resource Manager constitute three components-

5.4.1.1.1. *Scheduler***-** It is regarded as a central scheduler that allocates resources to the various running applications.

5.4.1.1.2. *ApplicationsManager***-** It accepts job-submissions, negotiates the first container for executing ApplicationMaster and restarts the ApplicationMaster container if it fails. It arranges the right resource containers from the scheduler, tracks and monitors their status.

5.4.1.1.3. *Node Manager***-** It is as per the slave machine framework agent. It has responsibilities towards containers, monitors their resource (CPU, network, etc.) usage and reports it to the ResourceManager.

5.4.1.2. *ApplicationMaster***-** It is one per application which coordinates and manages MapReduce jobs.

5.4.1.3. *Container***-** It allocates a certain amount of resources and includes elements such as CPU, network, etc. on a slave node; i.e Node Manager.

5.4.2. ***YARN Architecture***[6]**-** As discussed before, YARN has four components that work together to form a YARN Architecture. The client submits a job to the Resource Manager which after acceptance of the job request, it receives the node status from the Node Manager. Then the ApplicationMaster sends a resource request to the ResourceManager; furthermore, the NodeManager specifies the containers in which the ResourceManager itself will determine the first container. The container, which allocates a certain amount of resources, will send the status of MapReduce jobs to the ApplicationMaster which manages the MapReduce jobs and finally, the client will get an acknowledgement for either completion or failure of the situation.

6. Hadoop Ecosystem-

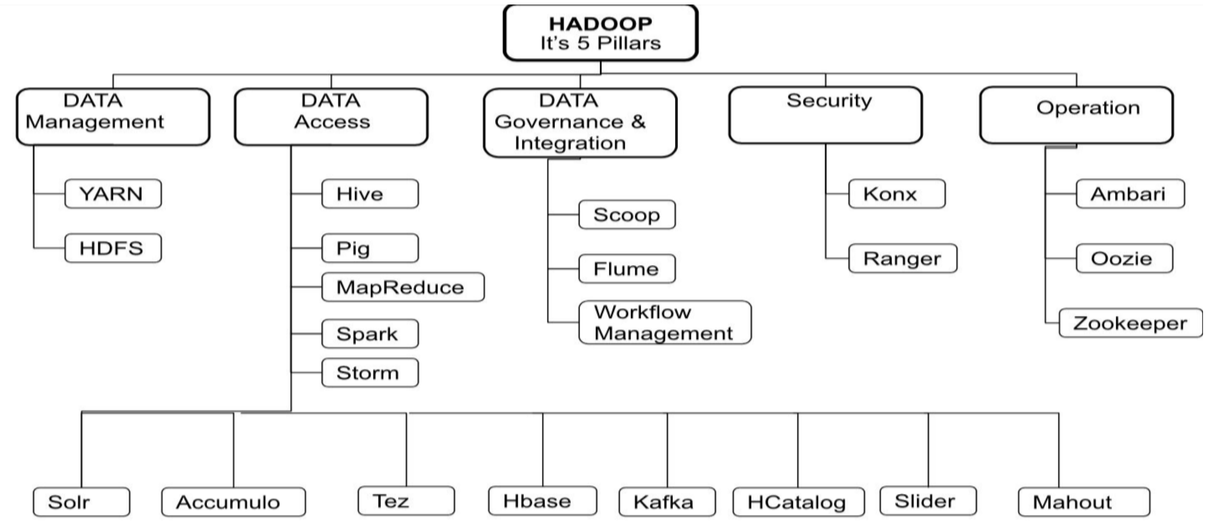

Figure 4: Complete Hadoop Ecosystem categorized in five pillars.

i. *Data Management*

Data management is the core part where only two-component falls, i.e. HDFS and YARN.

a. HDFS stores and processes vast quantities of data in a storage layer. This storage layer can be efficiently scaled-out as per the requirement.
b. YARN provides resource management and pluggable architecture. YARN prolongs the capability of MapReduce by extending its support to non-MapReduce workload models. It also allows a different option for data access method, which Hadoop provides.

ii. *Data Access*

Hadoop provides several engines that interact with data stored in HDFS in a wide variety of ways this can be a batch process, real-time process, or intermediary process. The engine used to access is shown in the figure above.

a. Apache Hive built on the MapReduce framework. It is a data warehouse where all managed and external tables are summarized. Such tables are queried via a SQL-like interface.
b. Apache Pig provides scripting capabilities that use pig-Latin, a high-level language for expressing data analysis programs paired with Tez or MapReduce. It also provides a platform for processing and analyzing big data sets[8].
c. MapReduce is a framework for parallel processing. It allows writing an application that processes a large amount of structured, semi-structured and unstructured data in parallel across a cluster of machines[5], which makes the action more reliable and fault-tolerant.
d. Apache Spark used for streaming data. The streaming application consists of Spark core which receives chunks of streaming data as Dstream at a different time of fixed division size. The receiver is a process running on an executor. Spark allows implementing fast and interactive algorithms for analytics such as clustering and classification of datasets.
e. Apache HBase for a NoSQL based database management which offers columnar No SQL storage, This provides random real-time write /read operation to big data.

f. Apache Tez is a more robust and extensible framework for executing a complex directed acyclic graph(DAG) for big data in real-time. This DAG is accomplished by generating the MapReduce model. Tez is built on top of YARN and improves the speed of a process.

g. Apache Kafka is a fast and scalable publish-subscribe messaging system, which has higher throughput, replication and tolerance to failure. Kafka is often used in place of traditional message brokers.

h. Apache Storm works on storm topology where data processing occurs in a topology. A topology consists of spout and bolt, where former bring data into topology later persist data including to HDFS.

i. Apache Solr is a searching and programming framework, which provides a platform for searches in HSFS. It also enables near real-time indexing to the website and powerful full-text search.

j. Apache Accumulo engine provides a cell-level access control which enhances the performance of data storage and retrieval system.

iii. Data Governance and Integration

The pillar allows quick and easy data load and management according to the policy defined in the Hadoop ecosystem.

a. Workflow management provides easy creation and scheduling of workflows and also monitors workflow jobs. Foundations of workflow management are on the Apache Oozie workflow engine, which allows connecting and automatizing the task execution on big data.

b. Apache Sqoop is a tool used for the import and export of data from the Hadoop cluster. The Map tasks can use any of the plugins, exist in Sqoop, to provide connectivity to various databases, which could be a relational database, enterprise data warehouse, or a document-based system.

iv. Security

In this pillar, all the tools address the requirements of accounting, authorization, authentication and data protection. Apache Knox and Apache Ranger secure every component and services.

a. Apache Knox is a Knox Gateway that provides a Single Sign-On (SSO) for its user. With a single sign-in into a Hadoop cluster, an operator and a user can control access and execute jobs respectively.

b. Apache Ranger provides central security policy administration across the Hadoop cluster's security requirement.

v. Operations

All the administrative work, like management, scheduling and coordination take place by the engines that fall in that operation pillar.

a. Apache Ambari is a lifecycle management, monitoring and administration system which allows modification in the configuration of services according to the requirement of the user.

b. Apache Zookeeper provides coordination among distributed processes to store and mediate updates to configuration information which is directed to the operations.

c. Apache Oozie is a java web application used for scheduling jobs on Hadoop. It also maintains the logical sequence of multiple tasks.

7. Traces of Big Data in Library

Big data is not a new thing nowadays, and it can be found everywhere irrespective of domains. Libraries to have big data which were unutilized or unprocessed earlier because of various reasons like librarian's reluctant attitude towards doing something new, lack of infrastructure and support, limitations of traditional databases and tools, software and hardware constraints, etc. But due to an information explosion, LIS professionals need to find out those areas where libraries have big data. They also need to find out ways to utilize such big data to come up with better solutions and decision making regarding the collection development or tracking uses of library materials and providing patrons with better services. Earlier, libraries were supposed to be storage houses of books. However, presently, it is considered an information centre which is capable of helping its patrons. The various ways for assisting patrons in understanding the issues and opportunities possessed by big data, helping them in the curation of large datasets, providing them with a personalized recommendation or reference services, making searching for materials easy and fast and many more.

These are some big data or areas of big data where libraries should look-

- **Users footfall data:** These are patrons check-in, check-out data or register entry data when a patron enters a library. If a library wants to analyze what type of patrons it has or who (by age group, by profession, etc.) are regular or frequent visitors of the library or simply understand its patrons and do some other analysis on users, then this data can be utilized. Earlier digitization and the internet was not popular, footfall in libraries was higher. But now due to digitization and introduction of Integrated Library System (like Koha, Libsys, etc.) and easy availability of internet, a patron has migrated to these new facilities and reduced the footfall in libraries, however, implicitly, there is an increase in the data usage. Now, patrons prefer to access the (digital) library through the web by logging in rather than going to the library physically. That gives rise to user behaviour study, in which the task of analyzing the users' footfall data. In turn, it makes the job easier than earlier as we do not need to feed the data from the register to any software manually. Here the log data of users' become the users' footfall data, and this can be analyzed, in real-time, easily using a big data handling environment like Hadoop. Though there is an experience in a reduction in users' physical footfall in libraries, there is an increase in library logins. Hence, analyzing these data will help in making better future policies and strategies to provide better services to the patrons.
- **Circulation data:** Circulation data refers to the data generated by the circulation of library materials (books, articles, documents, CDs, etc.). It is one of the core services a library provides, which grants its users to issue library materials for a specific period according to their need and interests. This particular service generates a substantial amount of valuable data which can bring new insights to the librarians. By analyzing these data, an analysis will be able to say which book or topic in which subject is more demanding, which book out of circulation for a long time. By answering all these questions, libraries can build our collection better, and it will help us in the acquisition of library materials. Analysis of circulation data will help any library to fulfil three basic laws of library science among five laws given by S.R. Ranganathan which are " Books are for use ", " Every reader his/her book" and " Every book its reader ". By analyzing circulation data, we can see usages of library materials and identify which uncirculated materials to make them visible and find their users. We can decide about the number of multiple copies of materials needed according to the usage data.

- **Acquisition data:** Acquisition data refers to the data generated during the library's whole life span of buying or subscribing to any library materials such as books, journals, articles, CDs, etc. There is a separate acquisition section in any library that maintains an acquisition register which contains the metadata about the whole collection of a library like date of purchase, price, source, title, publisher, author, etc. Subscription of journals or newspapers recorded daily, weekly, biweekly, monthly, bimonthly, quarterly, half-yearly, and yearly basis. This data is again a significant one that can be analyzed to bring new insights and to assist librarians in the act of acquisition. They can see the collection size of a library as a whole and collection size by particular domain or subject. They can also have an idea of year-wise spent budgets on different library materials in real-time that will further help us in the allocation of budget for different library materials for the next year.
- **Clicking behaviour of users:** Frequently libraries are using Web 2.0, where most of the communication has become bidirectional. Hence whenever a patron starts using the services of a digital library; i.e. starts clicking on the web, it generates a considerable amount of log data which is raw data for the behavioural analysis of the user. Libraries can monitor the clicking behaviour of the users using different tools and build a better recommendation system to recommend only relevant materials according to the clicking history of a user[2]. From where a library can have the idea of most searched or read books in a particular domain or subject that will help us in the recommendation of books or other materials to a new user of that domain interest.
- **Users data:** Libraries use to keep records of their users or data about the users. Why do libraries only keep those data if they can use it or analyze it to provide personalized services to its users in the form of services like Selective Dissemination of Information (SDI), Current Awareness Services (CAS) and many more innovative services? Generally, whenever a user registers himself with a library, the minimum data a library takes from him is about their educational qualification (current course or passed courses) and areas of (research) interests. So, libraries can use these data for SDI, which is nothing but recommending materials to their users. According to the user's selected areas of interests and CAS, which makes the users aware of the new addition of any further document in their areas of affairs through an email. So, this will be of great help to the researchers, scholars, teachers, professors, and scientists while doing any research as they will get updates by the current works done in their research interests areas.
- **Bibliographic data:** Bibliographic data refers to the citation data or references data that a researcher cites in his scholarly articles or theses. There is a separate branch of study in Library and Information Science called Bibliometrics which analyzes the different types of bibliographic data. Because the size of data is sometimes unstructured, a bibliometric is tedious for analyzing data. Most researchers prefer manual analysis, for that a time-saving solution for efficient big data management platforms such as Apache Hadoop for bibliographic data analysis. Generally, bibliometrics studies measure the impact of an author, institution and different sources of information by analyzing the references given by researchers. This problem is vital in the subject, and Big Data principles will make the work easier and reduce the time and energy needed to perform the job manually.
- **Preparing Indexes:** Libraries can use the concepts of big data for making indexes in large databases like JSTOR, PubMed, Scopus, etc. to facilitate searching for information more meaningful, relevant and accessible.
- **Data Curation:** Librarians can also use the concepts of big data for the curation of large datasets. Libraries can provide the services of data curation and assist the researchers in understanding their data and bringing some meaning out of the data.

8. Big Data Opportunities in Libraries

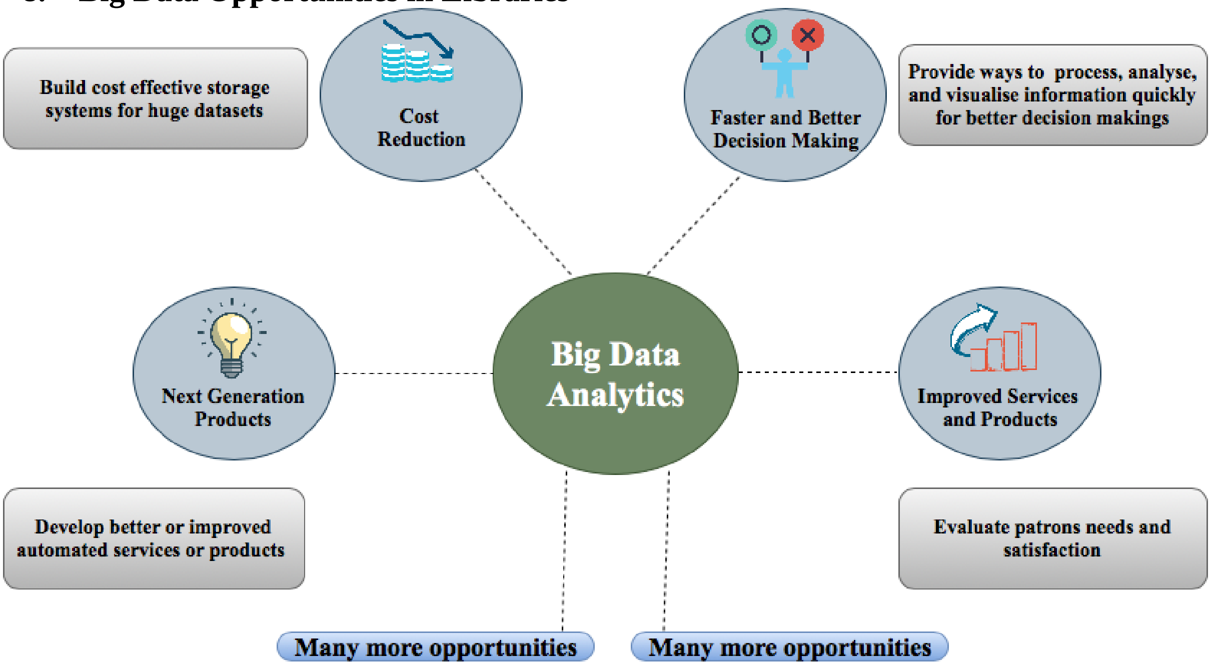
Build cost effective storage systems for huge datasets
Cost Reduction
Faster and Better Decision Making
Provide ways to process, analyse, and visualise information quickly for better decision makings
Next Generation Products
Big Data Analytics
Improved Services and Products
Develop better or improved automated services or products
Evaluate patrons needs and satisfaction
Many more opportunities
Many more opportunities

Figure 5: Opportunities in libraries when dealing with big data.

Libraries have big data and libraries can play a vital role in handling big data. LIS professionals need to understand the importance of processing these big unutilized data in libraries that can help them in providing better services to the patrons. Libraries can create opportunities out of big data in several ways that we will discuss in the later section, which will be fruitful for both the patrons and the professionals.

Libraries can-

- build cost-effective storage systems for massive datasets.
- develop better or improved automated services or products.
- provide ways to process, analyze and visualize information quickly for better decision making.
- evaluate patrons needs and satisfaction.

9. Conclusion

Modern time is the age of big data, and people can find big data everywhere irrespective of domains. Libraries also generate a considerable amount of data with their various service activities which were unutilized or unprocessed earlier because of multiple reasons. Libraries have big data in numerous forms such as circulation data, users footfall data, acquisition data, bibliographic data, and users data. The library professionals ignored earlier these data but now with the advances in technologies, the time has come to use those data, analyze them and bring new insights that can help in better decision making and improved library services.

The fifth law of library science given by S.R. Ranganathan states that "The library is a growing organism". Library and its services have changed with time. Nowadays, libraries should not only be considered as storage houses for books along with that it should also be recognized as an information centre. Libraries are capable of helping their patrons in various ways such as-

- helping them in understanding the issues and opportunities possessed by big data;
- assisting them in the curation of large datasets;
- providing them with a personalized recommendation or reference services; making searching for materials easy and fast.

Big Data Analytics is the key to fulfilling these services and Apache Hadoop with its ecosystem can help us for the same.

References

1. *ADVANCED INFORMATICS FOR COMPUTING RESEARCH: Second international*. (2019). SPRINGER.
2. Adekunjo, O. A., Adepoju, S. O., & Adeola, A. O. (2015). Assessment of users information needs and satisfaction in selected seminary libraries IN Oyo State, Nigeria. *Educational Research and Reviews, 10*(15), 2130-2140. doi:10.5897/err2015.2321
3. Bayrak, E. A., & Kirci, P. (2020). A Brief Survey on Big Data in Healthcare. *International Journal of Big Data and Analytics in Healthcare, 5*(1), 1-18. doi:10.4018/ijbdah.2020010101
4. Casado, R., & Younas, M. (2014). Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience, 27*(8), 2078-2091. doi:10.1002/cpe.3398
5. Jung, J., Kim, M., & Lee, H. (2015). A Study on Efficient Design of A Multimedia Conversion Module in PESMS for Social Media Services. *International Journal of Electrical and Computer Engineering (IJECE), 5*(4), 821. doi:10.11591/ijece.v5i4.pp821-831
6. Lahmer, I., & Zhang, N. (2016). Towards a Virtual Domain Based Authentication on MapReduce. *IEEE Access, 4*, 1658-1675. doi:10.1109/access.2016.2558456
7. Mittal, S., & Sangwan, O. P. (2018). Big Data Analytics Using Data Mining Techniques: A Survey. *Communications in Computer and Information Science Advanced Informatics for Computing Research*, 264-273. doi:10.1007/978-981-13-3140-4_24
8. Moussa, R. (2012). TPC-H benchmarking of Pig Latin on a Hadoop cluster. *2012 International Conference on Communications and Information Technology (ICCIT)*. doi:10.1109/iccitechnol.2012.6285848
9. S, S., & Tajunisha, N. (2015). A study on evolution of data analytics to big data analytics and its research scope. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. doi:10.1109/iciiecs.2015.7193065
10. Wadkar, S., & Siddalingaiah, M. (2014). *Pro Apache Hadoop*. Apress.
11. Wadkar, S., & Siddalingaiah, M. (2014). Hadoop Concepts. *Pro Apache Hadoop,* 11-30. doi:10.1007/978-1-4302-4864-4_2
12. Lang, X., Zhang, Z., Xie, L., Horch, A., & Su, H. (2018). Time-Frequency Analysis of Plant-Wide Oscillations Using Multivariate Intrinsic Time-Scale Decomposition. *Industrial & Engineering Chemistry Research, 57*(3), 954-966. doi:10.1021/acs.iecr.7b03042