TEST Engineering & Management

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

Comparative Study of Performance Analysis of Query performed on RDBMS and Solr on Digital Data of Heritage

Viratkumar K. Kothari[1] and Dr. Sanjay M. Shah[2]
[1]Ph.D. Scholar, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat
[2]Director, Narsinhbhai Institute of Computer Studies & Management, Kadi, Gujarat

Article Info Volume 83
Page Number: 4111 - 4119
Publication Issue:
July - August 2020

Article History
Article Received: *25 April 2020*
Revised: *29 May 2020*
Accepted: *20 June 2020*
Publication: *10 August 2020*

Abstract:
There is a large amount of archival content available in physical form. This includes manuscripts, printed papers, photographs, artefacts, sculptures, buildings, audios, videos and others. These content are gradually being converted into digital format now. The digital formats have significant benefits over the physical form such as availability of it on an online or offline platform, easy to share, easy to copy, easy to transport, easy to keep multiple copies at different places, searchability by adding metadata to it. These size of the data converted into the digital format of Heritage has increased exponentially. Such data usually are hosted on a large scale web platform with search functionality. The interlinking of digital data based on search functionality is becoming a challenge due to its size and multiple types of data formats. A need for improvement in the current search functionality is required for speed and time efficiency. This will improve data linking that is to find out related data efficiently. This will also help to do its further analysis and find hidden information patterns. The distributed processing of data can help here tremendously to speed up data processing exponentially. In this research paper, we have explained the methods to convert physical data into the digital format and attach metadata and evaluated the performance of the query processing on a relational database (RDBMS) and data to improve existing search functionality for utilising processors and processing time. It was envisaged to have an improvement of about 10% – 12% in each of them.

Keywords – *Heritage, Data Processing, HDFS, Hadoop, Digital Platform, Large-scale systems, Software library, Automation, Solr, Search, Enterprize Search Platform*

I. INTRODUCTION

The physical heritage data once converted into digital format is available in the various form including texts, images, videos, 3D models, architectural drawings and many others. The digital data then can be hosted on a platform like a digital library where it can be easily accessed and searched.

The digital heritage data used for this experiment is a digitised content on Mohandas Gandhi. He performed a vital role in the movement of Indian Independence. He led the various movement for the Independence of India. He has visited about 2,500 places in India. He lived in different parts of India and initiated various movements from such locations. All such locations have become now a heritage site. Each of such sites has a huge amount of physical data, and there is a special history attached to them. The content is available in physical formats are letters, books, manuscripts, photographs, audios, videos, artefacts, and buildings. The audios and videos are in analogue format

The content is being digitised, and about 2.3 million pages of content are digitised so far. The metadata is also being prepared for the content, which helps to understand the content. The metadata is data about data. Thus, it provides information and insight about the content, e.g. for a book; metadata will include the title of the book, authors, publication year, publisher, language of the book, number of pages, translator if any, compiler if any, dimension (size) of the book etc. The other example of metadata for photos maybe size of the photograph, type of photograph (colour or black & white), photographer, date of the photo, people in the photo, place of the photograph and photo itself. The text has been generated for all the textual content like books, papers, journals using a technology called Optical Character Recognition (OCR).

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

An online platform was created where the content has hosted along with its metadata and OCRed text. The platform is called the Gandhi Heritage Portal. It can be accessed using a URL www.gandhiheritageportal.org. It is currently one of the largest authentic repositories on the life and works of Mahatma Gandhi in the entire world. The content hosted on it is searchable based on the Metadata of each and OCRed text. The data is stored in an RDBMS on which the search is performed using traditional query-based functionality. The search is prolonged and takes too much time to search. Also, it becomes unresponsive if more than three words are searched. This is because it tries to search the search phrase on OCRed text which is quite huge. On the other side, content is being added continuously to the portal along with the metadata and OCRed text. This makes the search functionality gradually slower and slower. The measures like query optimisation, indexing and code optimisation have already been taken, but it does not seem useful.

The search functionality can be enhanced in numerous ways, such as an integration of enterprise search platform, processing the queries parallelly on multiple processors, and processing the queries in the distributed environment. This shall significantly enhance the speed of searchability.

This paper is organised as follows:

Section I Introduction, Section II Related work, Section III Information about the digital platform, its architecture, development specification and hosting environment IV Environmental Setup, parameter tuning and testing scenarios, Section V Results and Observations of the digital platform, Section VI Conclusion and future work.

II. RELATED WORK

Alexa T. McCray et al. [1] has explained that digital libraries are expensive and resource-intensive. An intensive thought process is required at various stages such as requirement gathering, planning, implementation, testing and deployment. The principle underlying the design, implementation, and maintenance of the digital library apply to digital platforms which host digital heritage data also. The principles defined are, know your content, expect change, align correct people, usable design system, ensure open availability, ensure digital rights, automation, adhering standards, quality and concerned about persistence. The metadata is an essential part of any digital platform or library because it makes the data searchable and helps to interlink related data and find out hidden patterns.

Elizabeth D. Liddy [2] demonstrates the feasibility of high-quality auto-generated metadata for digital libraries through Natural Language Processing (NLP). She further explains that NLP is the best technology which enables a system to learn the things by experience and accomplish understanding like a human. It can understand document content and extract its implicit and explicit meaning. Thus, based on this information, it can generate the metadata automatically. This will help make the search more relevant and quicker.

Schatz, Bruce et al. [3] explains about some Digital Library Initiative (DLI) projects which are a fair measure of the research into large scale digital libraries. This paper has nicely explained about design and development of National Information Infrastructure. Although it does not clearly explain how distributed infrastructure may be used to scale it up. Brin, Sergey et al. [4] discusses Google, a prototype of a large-scale search engine platform which can handle the heavy load of hypertext data. It has also been explained about technical challenges while using additional information present in hypertext to generate relevant search results. It also discusses how the problems of scaling traditional search techniques can be addressed. Bast et al. [5] on the other side explains using a demo of ESTER – a search engine which addresses the challenges of the ease of use, speed and scalability of the full-text search along with the powerful semantic capabilities of ontologies. It supports various types of queries including full-text queries, ontological queries and the combinations of these. The user interface is simple and user friendly. It interactively helps users providing search fields with semantic information. Haruechaiyasak et al. [6] propose a feature called category browsing which allows enhancing the full-text search function of Thai-language news article search engine. It allows users to browse and filter search results based on some predefined categories. Li, Bo, et al. [7] explains about the fast and complex growth of digital world which requires the applications to have more user-friendly features, high performing full-text search and rich functional online data analysis. This paper also explains how to exploit the power of web search and how the parallel

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

DBMSs addresses it. Lee et al. [8] discuss the recent work of an adaptive learning algorithm which then can create the content on the fly for an e-commerce search engine. This has come from the implicit knowledge extracted by the Web text mining modules. Ding et al. [9] explain that the full-text retrieval system is quite effective and famous in the information domain. The Lucene is an open-source full-text search engine toolkit. It has an important feature to detect duplicate content and highlight them in the search results generated from full-text retrieval.

Terasawa, Kengo et al. [10] presents a quick appearance-based full-text search technique for historical newspaper images. Historical newspapers are quite different from newspapers of nowadays in terms of image quality, the font type and language. So, the OCR generates erroneous results. Therefore, instead of the OCR approach, an appearance-based approach has been used where the character is matched to the character with its shapes to generate results. Imura, Hajime et al. [11] proposes a full-text search technique for image-scanned documents where recognition of individual characters is difficult or not possible. It is as fast & efficient for a full-text search of machine-readable documents. Such a system is essential when working with historical manuscripts. The proposed method works independently of the language, font style and font size because it uses a unique pseudo-coding based approach on the statistical features of character shapes. This may be useful to generate OCR based text from the images for printed and handwritten text. Palmer [12] research examines The Peter F. Drucker Manuscript and Archives Project on five different dimensions as a digital document. The aspects discussed are versatility; text, audio/video; taxonomy; contextual cues; and intellectual property.

Ikica [13] has discussed an improved profile based text detection technique which uses a set of heuristic rules that helps to eliminate non-text areas from the document. This method is evaluated on CVL OCR DB, which is an annotated image database of text in natural scenes. Nouza, Jan [14] worked in getting transcription from audio speeches. This is an important aspect to improve search in audios. The system continuously monitors selected Czech TV station and provides an automatic transcription of its audio tracks. The transcription is performed by its own specialised Speech Recognition Engine which employs a corpus of 3,50,000 most frequently used Czech words along with its forms. The accuracy of transcription is about 90 per cent for non-noisy speeches which is quite impressive. But it drops significantly when the recordings are spontaneous speeches or noisy one.

III. INFORMATION ABOUT THE DIGITAL PLATFORM, ITS ARCHITECTURE, DEVELOPMENT **SPECIFICATION AND HOSTING ENVIRONMENT**

A. Feature Ready Digital Platform for Heritage Data in Digital Form

The Heritage or an Archival content is generally available in physical and analogue form. The only way to access such physical content is to reach to the place where they are placed. There are various drawbacks of this methodology including, i) it gets deteriorated by the passing of time, ii) one has to reach to the place where physical data has placed to access it, iii) easy accessibility is not possible due to the number of reasons like permission, time constraints etc., and iv) there are high chances of accidental damage while accessing it. Due to these limitations, such important data remain underutilised, and researchers, authors, students and general people cannot access it many a time.

Therefore, these data must be digitised and should be made available to various communities, including the general public. A digital platform which is secure, scalable, and has a powerful search feature should be built to disseminate such information. A various digital platform can be envisaged where such digital data can be hosted and make them available for the public. The digital content can further be classified based on various parameters. The first level of classification may be the type of content, e.g. photos, audio, video, archival papers, manuscripts, 3d model of artefacts & buildings, buildings drawings, virtual walkthrough and others. The images having text may be converted in the machine-readable text using Optical Character Recognition (OCR). This way, huge text can be obtained, which then can be used for detailed full-text search and interlinking of the content.

The digital platform should be able to showcase a variety of content in a different section. It should be scalable enough to hold a large amount of data, able

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

to cater to a huge number of visitors on the other side. The platform should be secure enough so that it cannot be hacked easily. The huge content hosted on such digital platform should cater full-text search based on metadata and OCRed text, and it should also link related content.

A digital platform has been created to host digital content on life and works of Mahatma Gandhi. It is called the Gandhi Heritage Portal (www.gandhiheritageportal.org). It is one of the largest authentic repositories on archival collection on Mahatma Gandhi in the entire world at the moment.

B. Digitization Specification & Workflow

There are various types of physical content available on Mahatma Gandhi. Different content requires different treatment and technology to digitise them. In the case of papers, photos & manuscripts, they should be scanned using appropriate technology, e.g. book scanners, flatbed scanners etc. Most of the papers, books and manuscripts are quite old and in brittle form. So, it is difficult or in some cases not possible to digitise them using flatbed scanners as they may damage the material. Specialised scanners with cradles are made to scan books may be used here. A custom camera-based scanner may also be prepared and used. The photographs are quite important material, which should be scanned with the utmost care. One has to digitise and convert it into the dimensions which may be easily printable in the standard photo labs.

Different type of scanning technologies may be selected, e.g. flatbed scanners for loose papers and photographs, book scanners for books, camera-based scanners with a customised cradle for archival and brittle materials.

A standard process may be defined for the digitisation of the documents. The following figure depicts the workflow for digitising documents.

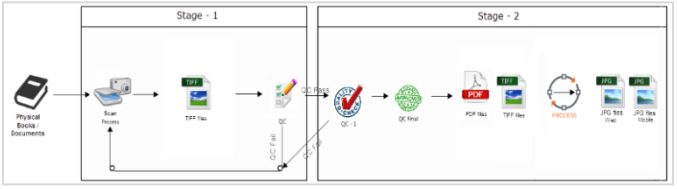

Fig. 1 Workflow for digitising documents

The documents are normally digitised using flatbed/cradle based / camera-based scanners. The images normally captured in the raw image format, e.g. .tiff. Once these images are cropped accurately, they then checked for accuracy against physical material and correction will be made if required. The images then further checked under various parameters like the number of pages, the sequence of pages, colour of pages, checking of margin on each side, dots per inch (DPI) and colour depth (bit depth) of the images. The scanning should be done at 300 DPI and minimum in RGP (8 bit each). Once the quality is found in order, the final images are further processed to generate low and high resolution .jpeg file. The low resolution .jpeg files are suitable to host on the web portal. If the physical content is the book, a .pdf file is also generated so that sequential reading is possible.

The standards for digitisation should be defined and should be updated on a regular interval to match current technology.

The quality of the digital output depends on the tuning of various parameters. The proper tune-up of various parameters of digitisation may be done to get the best image quality at best possible least size.

There are several ways to digitise objects, building, and artefacts. These may be done using textual documentation and visual documentation. Preparing drawings, capturing information, etc., are the part of textual documentation. On the other side, preparing virtual tours, preparing videos, capturing photos and 3D modelling technologies to capture objects, artefacts, building, and places are part of visual documentation. The textual documentation is an excellent way of information capturing, but visual documentation enhances it by providing minute visual details. The following figure depicts the general workflow for digitising objects in the 3D format using the scanning technique like LiDAR.

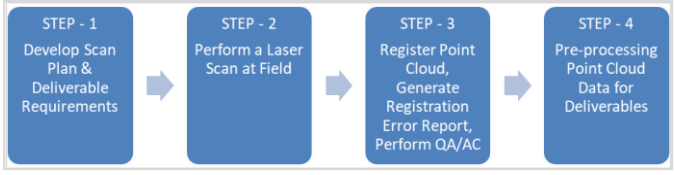

Fig. 2 Workflow for digitising objects, buildings and artefacts

Digitised images may have text in it. It is required to get the text out from the images to make the data

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

more understandable and searchable. It will generate huge text, but it provides a great benefit of searchability. Makes sense in case of digital heritage data as most of the content is in an unstructured form which is normally not searchable. Therefore, the text retrieved from the images will not only enable them to be searched but also help them to understand and interlink with other data to find-out hidden patterns out of them.

The content is once available in digital form; the greatest challenge is to make it searchable because then only the real use of the digital data will be possible. Various tasks can be done with digital data to make it effectively searchable. These include providing metadata for each content, speech to text for audio/videos and generating text from the images having text in it.

The technology to retrieve text out of the images is called OCR. The text can be retrieved out of the scanned documents having text in it using a software tool. It offers numerous benefits, including the availability of searchable data, reduction of cost & time comparing to manual tasks, reduction in errors, and eventually, it can be interlined to make it more informative. The OCR technology provides different accuracy for different languages, but it gives a quite accurate in the English language.

The output of the OCR requires a detailed quality check to correct errors, avoid publications etc. Therefore data cleansing is an important task that needs to be performed to validate the output generated using OCR. The following figure depicts the generalised workflow of the OCR activity:

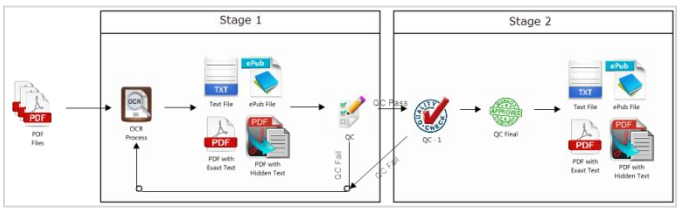

Fig. 3 Workflow for Optical Character Recognition (OCR)

C. Parameters for Search Engine

The searchability is an important factor of the data that is similar to digital heritage data also. The only difference is there are various types of data here. The data can be searched using metadata and OCRed text. The following parameters for the search feature are recommended:

- **Search Phrase:** A single field like Google using which anything can be searched. Full-text search will happen in this case with the search phrase entered in the field.
- **Serial Number:** Every item in the heritage data normally have an index/serial number. Search should be possible with it.
- **Dates:** Heritage data normally have dates attached to it; therefore, it can be easily searched with it if such parameter is provided.
- **Language:** Data may have in multi-language which may be searched using it.
- **Conversation & place:** It content is of the nature of *Letter* it may have to and from responses and the place where it is written. So, it can be searched using who has written it.

Following image shows the parameters to the search engine:

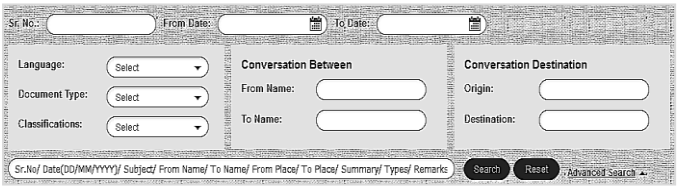

Fig. 4 Parameters to Search Engine

D. Search Matrix

The following are important factors to improve the search efficiency and find out interlined digital data::

- **Index:** All the data in the platform must be indexed. This improves the speed of the data searchability tremendously.
- Full-text Search: A full-text search should be possible based on metadata, OCRed text etc.
- **Listing order:** Once the data searched, the order of showing data is very important. By default, data must be ordered as per its relevance to the searched phrase. The listing may be further categorised. This will help the user to remain focused and find out exactly data of his need.
- **Linked content or Cross-reference:** Each content may have other related content present

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

in the data. Related content may be shown here to find connected information easily. This may be shown under various categories. The recent technologies like machine learning, natural language processing (NLP) and artificial intelligence may be useful here.

- **Efficient use of computer hardware and network:** The efficient search engine should be able to use computer hardware efficiently and should also work on the distributed hardware network environment.

IV. ENVIRONMENTAL SETUP

The digital web platform called Gandhi Heritage Portal (www.gandhiheritageportal.org) has been prepared with the use of PHP 7.0 as a scripting language and MySQL 5.5 as a database. The server has 64bit Core i5 4 processors, 16 GB RAM and 1 TB hard disk drive. This Hardware is enough in terms of data processing requirements. Following image shows hosting configuration:

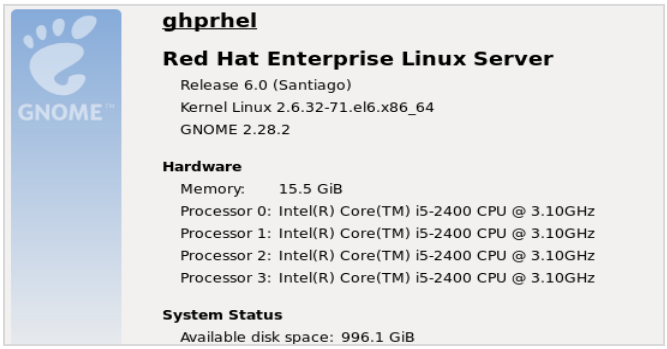

Fig. 5 System Configuration

Each document/books has been uploaded on the platform with appropriate metadata to make them searchable. The general search happens on the RDBMS using queries. The search was performed on metadata as well as on text generated using OCR, transcriptions from audio-video etc. This search is quite slow, and interlinking is not satisfactorily.

Therefore, to improve the search functionality following were measures have been taken:

- **Hardware configuration**

Hardware tune-up processes like registry cleaning, stopping unnecessary services, and tune-up of file descriptor (in our case the default 1,024 is perfect) has been carried out. The 'StartServers' and 'MinSpareServers' directives have been tuned up to free up the memory. This helped to free about 20% of memory. All unnecessary software were also removed, e.g. Openoffice. The old log has also been removed to free the space.

- **RDBMS and Query optimisation:**

Following measures were taken to tune the RDBMS and queries:

  - **Indexing:** General index was already there in the table, but the additional index has been created to the columns used in "where", "order by" and "group by" clauses. This generally considerably reduce select query execution time.
  - **Limit the Like statement:** Usage of "Like" statement is made restricted as it requires more resources. The statement "or" is used wherever possible.
  - **Table normalisation:** An adequate table normalisation has been carried out. It is also verified that correct datatype has been used.
  - **Joins in queries:** There were few joins used, so it has been made sure that those are optimised.

It is important to have data in a machine-readable format as it will give grate level of searchability. However, there will be a large amount of text available in that case. The heritage data in its original form (physical) has very limited searchability. Therefore, text generated from content like images, audio, videos, etc., will not only enable them to be searched but also help to auto evaluation and interlink with other related data. This will help find-out hidden patterns in them.

The enterprise search platform has been integrated, and the data then has also been transferred to it by performing a query to the RDBMS to improve the searchability further, These platforms offer a great level of searchability as it stores data in a very search-friendly format, i.e. key-value pair database. Following tasks has been carried out to improvise the search facility:

- **Integration of enterprise search platform**

Enterprise search platform here means a specialised tool which improves data searchability. The Solr is such a search engine platform. It's practically a Key-Value pair database specifically meant for efficient search. It stores data efficiently in the database to retrieve it quickly. It retrieves results at blazing speed. The Solr is an open-source, highly reliable, fault-tolerant, and scalable platform. It also supports replication, distributed indexing, and load balancing

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

while performing the query. It can be configured centrally and supports automated failover and recovery.

The Apache Solr is an enterprise search engine which is a fundamentally a sub-project of Apache Lucene. Apache Lucene is an indexing technology behind the Solr - most recently created search and index technology. It is a NoSQL database and has transactional support. It is a document database and also offers SQL support. It executes it in a distributed manner. The Solr is quite popular and used by Instagram, NetFlix, Disney, Internet Archive etc.

- **Parameter tune-up of Solr to improve its efficiency**

In Solr, one has to decide how much memory should be allocated to it. The Solr performs better if more memory is allocated to it. On the other hand, it will adversely affect on JVM GC in terms of overload. Also, there is a point where performance improvement does not justify the overall reduction of hardware performance. So one has to find optimal tradeoff between memory allocation and performance. Following hyperparameter has been tuned up to improve the performance:

- **Direct memory:** Generally 8 GB of memory allocation for director memory to Solr is enough. So, we allocated 8 GB to direct Solr memory.
- **Memory to JVM**: We have **allocated 8 GB** of RAM looking at the workload. If enough RAM is not allocated or JVM is not running with a large enough HEAP size, the JVM will hit the swapping and thrashing at which point everything will run quite slowly.
- **Local file system:** We have stored all the files in the **local system** to improve the **performance**. Remote filesystems are generally quite a bit slower for performing the search.
- **Linux platform:** We are **using Linux platform** which gives better **performance** in search because there is a bug in Windows in Sun's JRE.
- **Index reader:** We have set **IndexReader = True**. This gives faster search results when multiple threads are sharing the same reader, as it removes certain sources of thread conflict. Also, We are using a single IndexSearcher across queries and threads.
- **mergeFactor:** The mergeFactor generally determines the number of segments. The value inset in the mergeFactor tells Lucene index that what number of segments of equal size is to build before merging them into a single segment. This may be thought of as the base of a number system. It is not an exact number, but a type of guide on how the index should be managed on the disk. The default value of mergeFactor is 1. If we keep the higher value, it improves the performance but less frequent merges which may slow searching. If we keep the lower value of it, smaller the number of index files speeds up the search but more the segment searches slow down indexing. We **set mergeFactor value to 12,** which should give optimal result looking to our quantum of data.

V. RESULTS AND OBSERVATIONS

In this section, the observations are presented based on search performed on normal RDBMS and enterprise search engine – Solr database. We passed Following is the status comparison of CPU in the ideal position as well as different level of queries:

- **CPU in an ideal position when there is no search query running:**

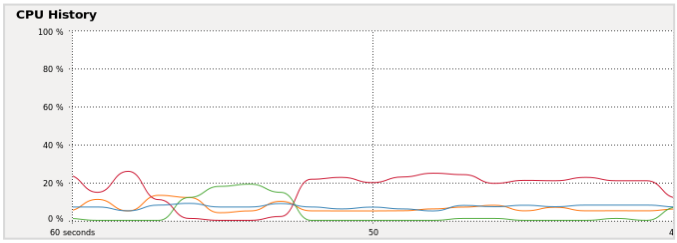

Fig. 6 CPU when there is no query in execution

- **CPU when single word searched on RDBMS and Solr respectively:**

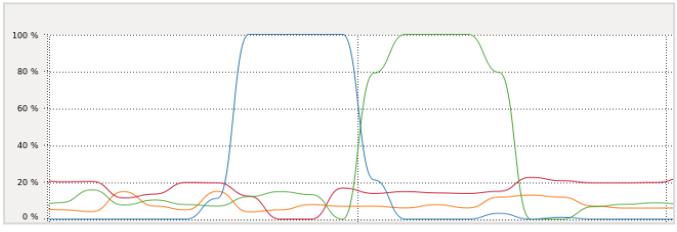

Fig. 7 CPU when single word DBMS based search

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

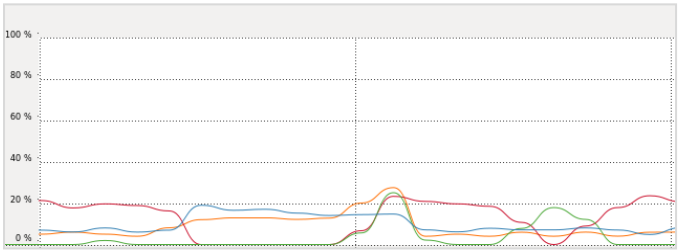

Fig. 8 CPU when single word Solr based search

- **CPU when two words (Indian National) searched on RDBMS and on Solr respectively:**

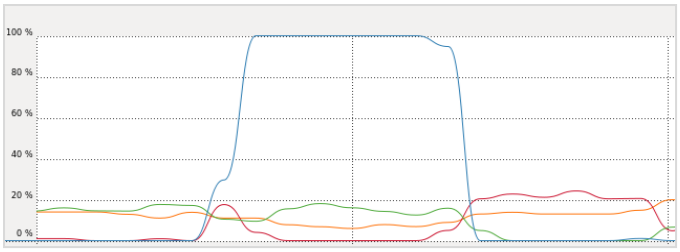

Fig. 9 CPU when two words DBMS based search

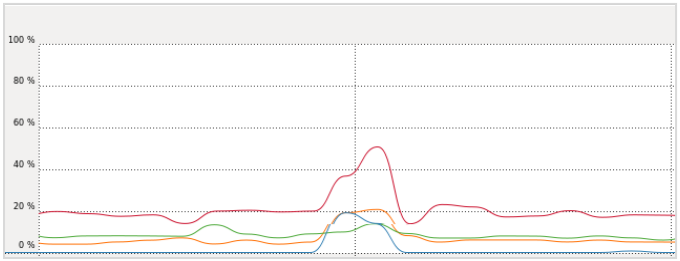

Fig. 10 CPU when two words Solr based search

- **CPU when three words (Indian National Congress) searched on RDBMS and Solr, respectively:**

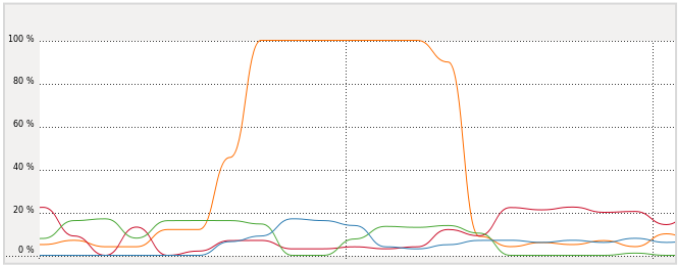

Fig. 11 CPU when three words DBMS based search

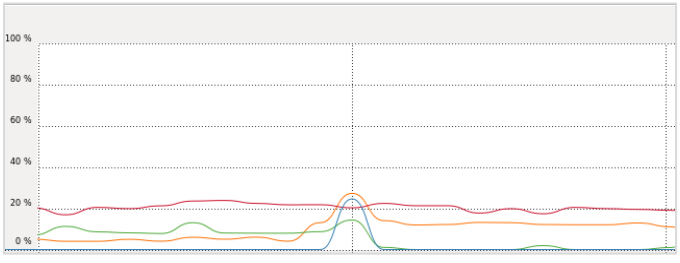

Fig. 12 CPU when three words Solr based search

Following are the findings of the tests:

- Response time has been significantly (up to 1/5) reduced in case of Solr based search in comparison of DB based search.
- Solr based search utilises multiple CPU whereas DB based search generally utilises only one CPU at a time.
- Solr based search utilises less CPU power than the DB based search and gives faster results.

VI. CONCLUSIONS AND FUTURE WORK

We found that the digital platform generally available are more of a generic nature. Most of such systems are designed to work with general metadata. On the other side, the platform designed to host heritage data has a variety of data including texts, images, audio, video, drawings, manuscripts and others which require an intelligent and quick search. Such a platform requires a user-friendly user interface, a powerful search, secure, and scalable.

A digital platform called Gandhi Heritage Portal (www.gandhiheritageportl.org) has been created which suffices such needs. It has a section-wise user interface to present the data. It had a query-based search on RDBMS but lacking powerful enterprise search platform. Therefore this search feature has been enhanced by integrating the Solr. This further can be deployed on a distributed platform. It can process huge multi-language content.

Following table represents the improvements obtained on processing queries by integrating an enterprise search engine platform – Solr over RDBMS in CPU utilisation and Processing time.

Non-complex search phrase	Query processed on RDBMS						Query processed on SOLR						Improvement in CPU (In %)	Improvement in Query Processing Time (In %)
	CPU Usage (in %)					Time Taken (Seconds)	CPU Usage (in %)					Time Taken (Seconds)		
	C1	C2	C3	C4	Total		C1	C2	C3	C4	Total			
No Query	5.1%	19.0%	1.0%	6.8%	31.9%	-	5.1%	19.0%	1.0%	6.8%	31.9%	-	0.00%	0.00%
Single word	7.0%	23.5%	8.8%	0.0%	39.3%	30	5.0%	21.0%	0.0%	8.0%	34.0%	3	13.49%	90.00%
Two words	5.0%	21.0%	8.8%	0.0%	34.8%	90	12.0%	20.0%	0.0%	0.0%	32.0%	4	8.05%	95.56%
Three words	6.9%	22.5%	0.0%	6.7%	36.1%	120	14.0%	0.9%	16.0%	0.0%	30.9%	4	14.40%	96.67%

Fig. 13 Improvement achieved in processing queries using Solr over RDBMS

The Solr has been integrated on a single node at the moment. It supports distributed processing on a distributed platform. The data will be processed on multiple nodes simultaneously and then will be collected and presented. There will be a drawback of task distribution and output collection from various nodes of distributed platform. But, it is envisaged to have tremendous improvement in processing queries

July-August 2020
ISSN: 0193-4120 Page No. 4111 - 4119

on Solr implemented on a distributed platform. So, in future, Solr will be implemented on a distributed platform.

REFERENCES

[1] McCray, Alexa T., and Marie E. Gallagher. "Principles for digital library development." *Communications of the ACM* 44.5 (2001): 48-54.

[2] Liddy, Elizabeth D., et al. "Automatic metadata generation & evaluation." *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*. ACM, 2002.

[3] Schatz, Bruce, and Hsinchun Chen. "Building large-scale digital libraries." *Computer* 29.5 (1996): 22-26.

[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30.1-7 (1998): 107-117.

[5] Bast, Holger, Fabian Suchanek, and Ingmar Weber. "Semantic full-text search with ESTER: scalable, easy, fast." *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*. IEEE, 2008.

[6] Haruechaiyasak, Choochart, et al. "Implementing news article category browsing based on text categorisation technique." *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*. Vol. 3. IEEE, 2008.

[7] Li, Bo, et al. "DIFTSAS: A distributed full-text search and analysis system for big data." *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*. IEEE, 2013.

[8] Lee, Chung-Hong, and Hsin-Chang Yang. "Developing an adaptive search engine for e-commerce using a web mining approach." *Information Technology: Coding and Computing, 2001. Proceedings. International Conference on*. IEEE, 2001.

[9] Ding, YueHua, Kui Yi, and RiHua Xiang. "Design of paper duplicate detection system based on Lucene." *Wearable Computing Systems (APWCS), 2010 Asia-Pacific Conference on*. IEEE, 2010.

[10] Terasawa, Kengo, Takahiro Shima, and Toshio Kawashima. "A Fast Appearance-Based Full-Text Search Method for Historical Newspaper Images." *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011.

[11] Imura, Hajime, and Yuzuru Tanaka. "A Full-Text Search System for Images of Hand-Written Cursive Documents." *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. IEEE, 2010.

[12] Palmer, Jonathan W. "Supporting diverse activities with digital documents: a pilot study of The Peter F. Drucker Manuscript and Archives Project." *System Sciences, 1997, Proceedings of the Thirtieth Hawaii International Conference on*. Vol. 6. IEEE, 1997.

[13] Ikica, Andrej, and Peter Peer. "An improved edge profile-based method for text detection in images of natural scenes." *EUROCON-International Conference on Computer as a Tool (EUROCON), 2011 IEEE*. IEEE, 2011.

[14] Nouza, Jan, Jindrich Zdansky, and Petr Cerva. "System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search." *MELECON 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*. IEEE, 2010.

[15] Federal Agencies Digital Guideline Initiative: http://www.digitizationguidelines.gov/guidelines/digitize-technical.html