ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Big Data Analysis with Data Compression

Shantanu Walunj, Kiran Sarpale, Pravin Salve, Sataym Pawar, Dr. S.M. Chaware

B. E. Student, Dept. of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research, Pune, India

B. E. Student, Dept. of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research, Pune, India

B. E. Student, Dept. of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research, Pune, India

B. E. Student, Dept. of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research, Pune, India

Asst. Professor, Dept. of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research, Pune, India

ABSTRACT: several organizations square measure currently coping with great deal of information. Historically they used relative information. However these days they're speculated to use structured and semi structured information. To figure effectively these organizations uses virtualization, multiprocessing in compression etc., out of that the compression is simplest one. The information transmission of high volume typically causes high coordinated universal time. This compression of unstructured information is instantly done once the information is being transmitted from shopper to information Node. At first once unstructured or semi-structured information is prepared for transmission, the information is compressed victimisation some code tools or procedures. This compressed information is transmitted through bound medium that undertakes a good transmission.

KEYWORDS: Big Data, Unstructured Data, Compression, Optimization, Data Node, Data Transmission.

I. INTRODUCTION

The Corporation or organizations' success fully depends on however these companies or organizations with success manipulates or uses the Brobdingnagian quantity of unstructured knowledge. These unstructured knowledge primarily comes from web site, XML files, Social Networks, etc. a number of common such examples includes transmission, internet contents, satellite and medical contents.

- **Big Data**
  Big knowledge may be a giant set of unstructured knowledge even over tera and peta bytes of knowledge. Massive knowledge [1] will be of solely digital one. Knowledge Analysis becomes a lot of sophisticated as a result of their magnified quantity of unstructured or semi-structured knowledge set. Predictions, analysis, needs etc., ar the most things that ought to be done exploitation the unstructured massive knowledge. Massive knowledge may be a combination of 3 v's those are particularly Volume, rate and selection. Massive knowledge can primarily processed by the powerful laptop. However because of some ascendible properties of the pc, the process gets restricted.
- **Unstructured information**
  Unstructured knowledge may be a information set those are within the sort of logs. There won't be any things like rows, columns, records, etc., some unstructured information includes log details of web site, transmission contents, images, videos, satellite image contents, medical contents, etc. This unstructured information became a lot of complexes to be modified. This unstructured information doesn't have any predefined information model. These unstructured information instead of having some text, can have Brobdingnagian quantity of knowledge like date, number, fact, etc., the unstructured information will ne'er be without delay classified. This type of information can't be contained in program or electronic database like structured data.

DOI: 10.15680/IJIRCCE.2017. 0504286

- **Compression**

  The compression [2] could be a technique of reduction in size of huge quantity of structured or unstructured knowledge. By mistreatment compression one will save memory area and will be additionally ready to minimize the UTC. The compression might be created doable to be in deep trouble entire transmission unit or too sure knowledge content.

  By compression of information, the additional area characters will be removed. By introducing single repeat character for giant continual characters, subbing smaller bits will minimize the file up to five hundredth of its own contents. Algorithms can even be such as to see however the compression ought to happen Graphical knowledge like footage, videos, animations, etc. are designed in such the way that it supports the information compression with none problems. The compression on this knowledge will be of 2 types and that they are:

  1. Lossy compression 2. Lossless compression

  In lossy compression, any sure info gets loss whereas the information is being compressed. In lossless compression there won't be any info loss throughout the information compression.

- **Need for Compression**

  Now relational knowledge base in conjunction with unstructured and semi-structured data evolves like xml, video, audio and pictures. This unstructured knowledge ar known as giant objects. Organizations are expected to agitate great deal of knowledge volumes; any these reasonably given data, the standard knowledge storage strategies and tape will doesn't work any longer.

II. LITERATURE SURVEY

1) **Graphics processors Performance Study of General-Purpose Applications on Graphics Processors Using CUDA.**

   (GPUs) provide a vast number of simple, data-parallel, deeply multithreaded cores and high memory bandwidths. GPU architectures are becoming increasingly programmable, offering the potential for dramatic speedups for a variety of general-purpose applications compared to contemporary general-purpose processors (CPUs). This paper uses NVIDIA's C-like CUDA language and an engineering sample of their recently introduced GTX 260 GPU to explore the effectiveness of GPUs for a variety of application types, and describes some specific coding idioms that improve their performance on the GPU. GPU performance is compared to both single-core and multicore CPU performance, with multicore CPU implementations written using OpenMP. The paper also discusses advantages and inefficiencies of the CUDA programming model and some desirable features that might allow for greater ease of use and also more readily support a larger body of applications.

2) **CUDA: Speeding Up Parallel Computing**

   Author: Maria Andreina F. Rodriguez.

   You're faced with imperatives: Improve performance. Solve a problem more quickly. Parallel processing would be faster, but the learning curve is steep – isn't it? Not anymore. With CUDA, you can send C, C++ and FORTRAN code straight to GPU, no assembly language required. Developers at companies such as Adobe, ANSYS, Autodesk, Math Works and Wolfram Research are waking that sleeping giant – the GPU -- to do general-purpose scientific and engineering computing across a range of platforms.

3) **A Universal Algorithm for Sequential Data Compression, May 1997**

   Author: J . Ziv and A. Lempel

   A universal algorithm for sequential data compression is presented. Its performance is investigated with respect to a non-probabilistic model of constrained sources. The compression ratio achieved by the proposed universal code uniformly approaches the lower bounds on the compression ratios attainable by block-to-variable codes and variable-to-block codes designed to match a completely specified source.

4) **Compressed Data Transmission Among Nodes in BigData".**
Authors: Thirunavukarasu B, Sudhahar V M,Vasantha Kumar U, Dr Kalaikumaran T, Dr Karthik S.
Many organizations are now dealing with large amount of data. Traditionally they used relational data. But nowadays they are supposed to use structured and semi structured data. To work e_ectively these organizations uses virtualization, parallel processing in compression etc.,out of which the compression is most e_ective one. The data trans-mission of high volume usually causes high transmission time. This compression of unstructured data is immediately done when the data is being transmitted from client to DataNode. Initially once unstructured or semi-structured data is ready for transmission, the data is compressed using some software tools or procedures. This compressed data is transmitted through certain medium that undertakes an effective transmission.

5) **CUDA – Supercomputing for masses.**
Author: Peter Zalutaski
Are you interested in getting orders-of-magnitude performance increases over standard multi-core processors, while programming with a high-level language such as C? And would you like that capability to scale across many devices as well? Many people (myself included) have achieved this level of performance and scalability on non-trivial problems by using CUDA (short for "Compute Unified Device Architecture") from NVIDIA to program inexpensive multi-threaded GPUs. I purposefully stress "programming" because CUDA is an architecture designed to let you do your work, rather than forcing your work to fit within a limited set of performance libraries. With CUDA, you get to exploit your abilities to design software to achieve best performance on your multi-threaded hardware -- and have fun as well because figuring out the right mapping is captivating, plus the software development environment is both reasonable and straightforward.

III. COMPRESION AND DECOMPRESSION TECHNIQUE

In projected technique, the compression of unstructured information is created in effective manner. Essentially this affiliation is created through the communications protocol affiliation. The communications protocol affiliation are additional advantageous because it continuously replies with a respond. The consumer receives the Acknowledgement from the info Node concerning that information Node ought to be employed in the Rack. The consumer then as per the directions given by the Name Node can effectively place the Blocks on the info Node.

➢ **Compression**

The isolated information is initially created probably sorted. Then the sorted similar quite information is saving for sorted. Once the info is sorted on some form of similarities, the info Blocks area unit shaped. Then on this information Blocks the compression is created. By this sort of compression enormous quantity information might be eliminated from transmission. Since the info is significantly reduced, the UTC clearly gets reduced. The compressed information then transmitted to the info Nodes. The info is operated at the info Node. In information Node the Map cut back rule is dead. By map the info is once more sub divided. Once the Sub division is created, the Map rule can execute or can give some form of multi programmed technique to perform the desired operation on those huge set of information.

➢ **Decompression**

Once the info is compressed and reaches the info node, there for the operations, the compressed information is once more decompressed at information Node level. By Decompression, the first Dataset is extracted, so the operation is performed on these original dataset. When the operations area unit performed, the info output is recovered. This recovered information is distributed back to the consumer. The consumer receives this processed dataset. The Analysis on the info is largely created at this consumer facet finally this steps area unit processed.

IV. MERITS AND DEMERITS

There square measure essentially 2 types of compression strategies. They're lossless and lossy compression. If the lossless compression methodology is employed, information when compression is extracted or decompressed with none data miss. However encase if the lossy compression is ready, then there'll be some quantity info or information loss when the info extraction or decompression. Some blessings of compression includes the subsequent,

- Reduction of price.
- Performance will be increased by achieving improvement.
- Great data on that information the compression to require place.

V. SYSTEM ARCHITECTURE

In existing methodology, the Hadoop design consists of master node or Name Node and knowledge Node [4]. Since BigData may be a conception of distributed system. The replicas of knowledge are created and placed in varied Basic Blocks [5]. Bound set of Algorithms is used for these duplicate placements.

Initially the massive knowledge set is split into n variety of blocks. These blocks are then replicated into some variety of copies. Then, the blocks ar placed in knowledge Nodes as suggested by the formula such by the Name Node. The shopper then uses that knowledge Node. The information set is transmitted to those knowledge Nodes with none compression,

Node, the Mapping and Reducing is completed for the execution of the information set. Once the execution gets completed, this knowledge set with output is once more remanded to the shopper. This knowledge sent was additionally a non-compressed one. The desired analysis is created at the shopper aspect for business or structure functions. Here during this existing technique, it consumes longer for the transmission of the information. The blocks therefore created are organized on the information Node putting every duplicate of block in serial nodes.

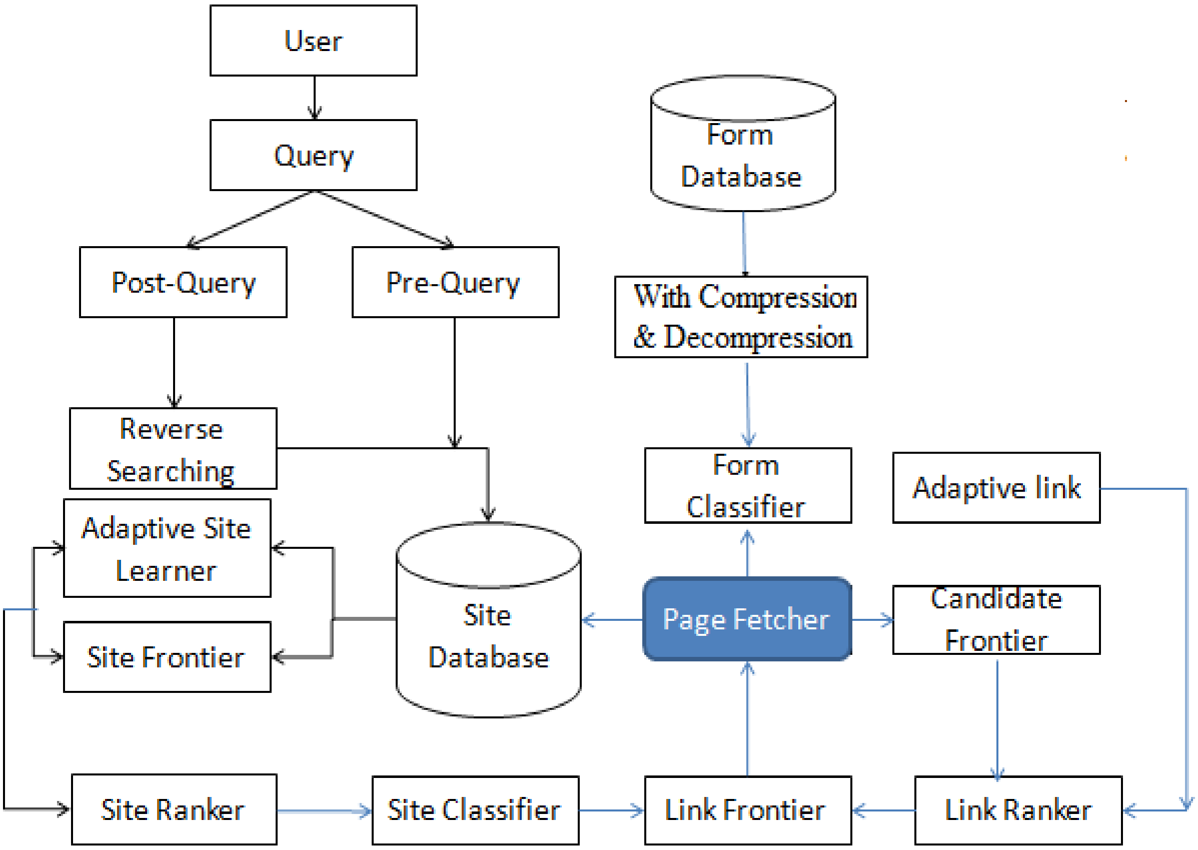


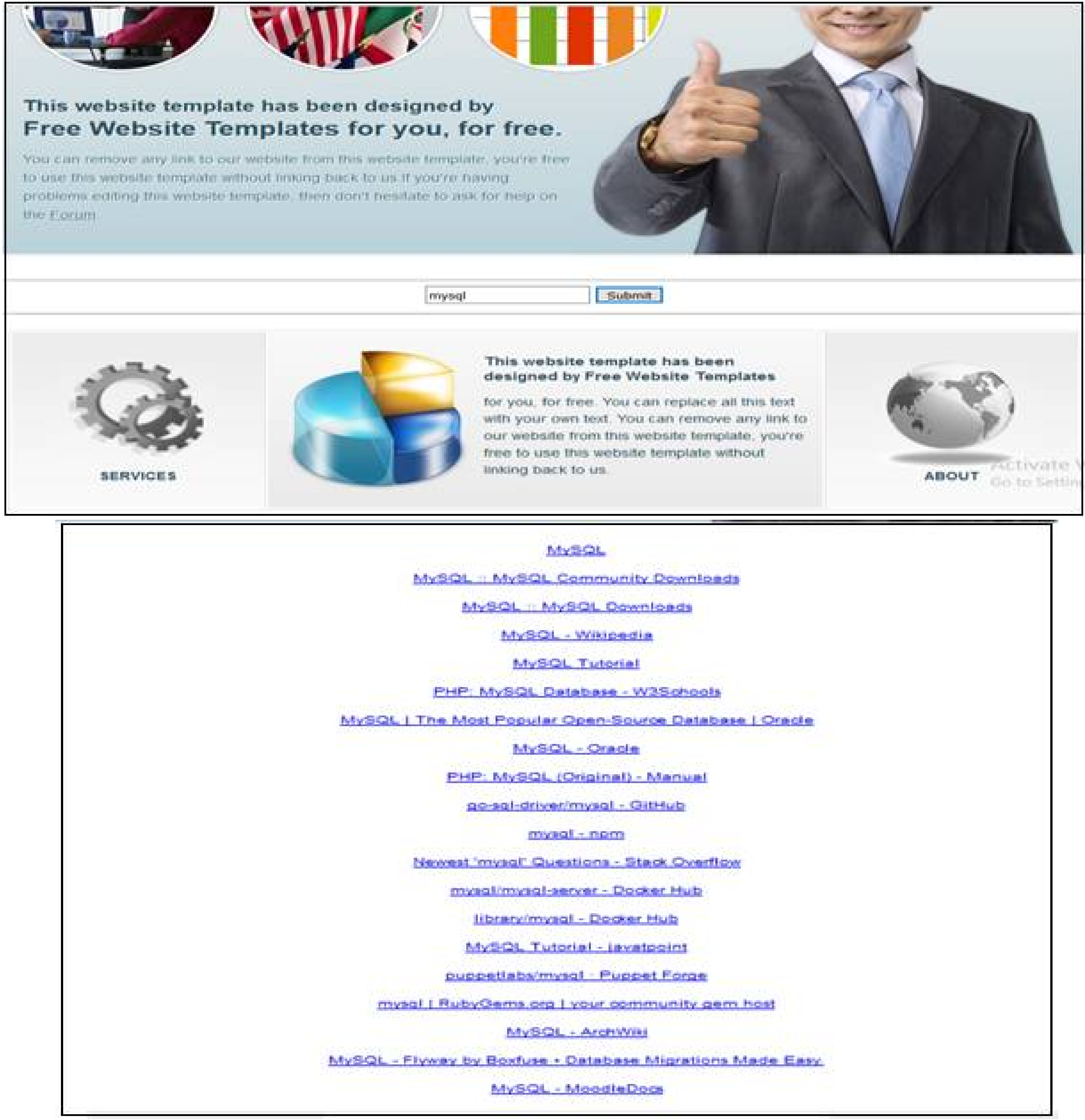
Fig.1: System Architecture

DOI: 10.15680/IJIRCCE.2017. 0504286

VI. RESULT AND CONCLUSION

RESULT:

1. **Search result through web:**



This website template has been designed by
Free Website Templates for you, for free.

You can remove any link to our website from this website template, you're free to use this website template without linking back to us if you're having problems editing this website template, then don't hesitate to ask for help on the Forum

mysql

Submit

SERVICES

This website template has been designed by Free Website Templates

for you, for free. You can replace all this text with your own text. You can remove any link to our website from this website template, you're free to use this website template without linking back to us.

ABOUT

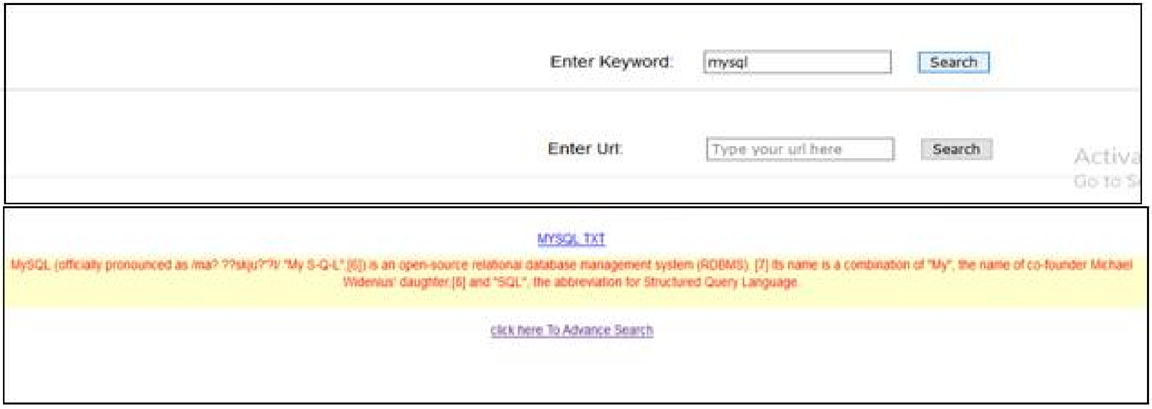
Activate W
Go to Setting

MySQL
MySQL :: MySQL Community Downloads
MySQL :: MySQL Downloads
MySQL - Wikipedia
MySQL Tutorial
PHP: MySQL Database - W3Schools
MySQL | The Most Popular Open-Source Database | Oracle
MySQL - Oracle
PHP: MySQL (Original) - Manual
go-sql-driver/mysql - GitHub
mysql - npm
Newest 'mysql' Questions - Stack Overflow
mysql/mysql-server - Docker Hub
library/mysql - Docker Hub
MySQL Tutorial - javatpoint
puppetlabs/mysql - Puppet Forge
mysql | RubyGems.org | your community gem host
MySQL - ArchWiki
MySQL - Flyway by Boxfuse • Database Migrations Made Easy
MySQL - MoodleDocs

Fig.2: Search result through web

DOI: 10.15680/IJIRCCE.2017. 0504286

2. Search result as per keyword:



Enter Keyword: mysql Search

Enter Url: Type your url here Search

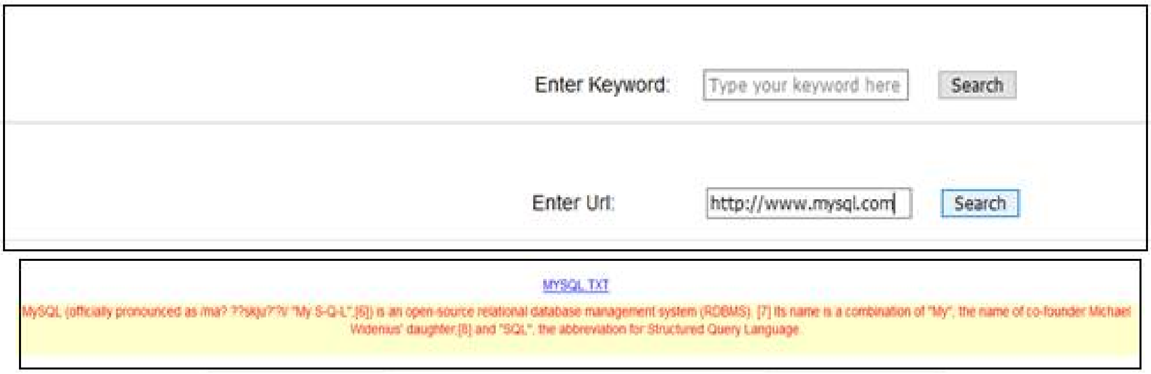
MYSQL TXT

MySQL (officially pronounced as /ma? ??skju?'?l/ "My S-Q-L",[6]) is an open-source relational database management system (RDBMS).[7] Its name is a combination of "My", the name of co-founder Michael Widenius' daughter,[8] and "SQL", the abbreviation for Structured Query Language.

click here To Advance Search

Fig.3: Search result as per keyword

3. Search result as per url:



Enter Keyword: Type your keyword here Search

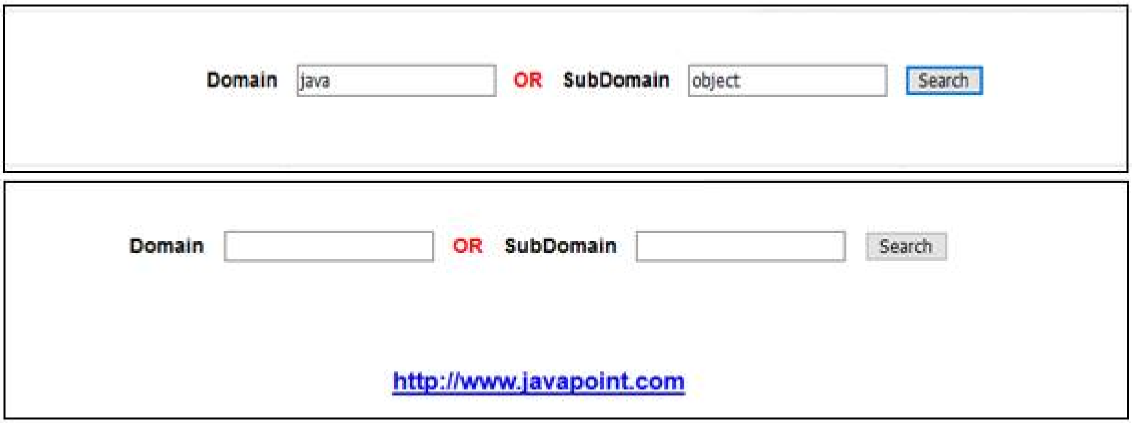
Enter Url: http://www.mysql.com Search

MYSQL TXT

MySQL (officially pronounced as /ma? ??skju?'?l/ "My S-Q-L",[6]) is an open-source relational database management system (RDBMS).[7] Its name is a combination of "My", the name of co-founder Michael Widenius' daughter,[8] and "SQL", the abbreviation for Structured Query Language.

Fig.4: Search result as per url

4. Search result as per domain or sub-domain:



Domain java OR SubDomain object Search

Domain OR SubDomain Search

http://www.javapoint.com

Fig.5: Search result as per domain or sub-domain

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

V. CONCLUSION

Thus as a result, the big knowledge set or BigData is transmitted once being playacting the compression. Since knowledge is totally compressed, the dimensions get reduced. Because the knowledge size is reduced, it's abundant enough to transmit the lesser quantity of information. This transmission of lesser quantity of information achieves solely an awfully low quantity of coordinated universal time. Therefore the performance is increased. The speed of execution additionally gets raised, because the coordinated universal time is reduced.

- Transmission time is reciprocally proportional to quantity of compression.
- Execution time is directly proportional to coordinated universal time
- Execution time is reciprocally proportional to quantity of compression.

Generally the overall performance of the Hadoop system gets raised. The improvement of the dataset is achieved by the improved knowledge Compression.

REFERENCES

[1] Shuai C., Michael B., Jiayuan M., David T., Jeremy W. S., Kevin S., Performance Study of General-Purpose Applications on Graphics Processors Using CUDA
[2] Maria Andreina F. Rodriguez, “CUDA: Speeding Up Parallel Computing”.
[3] Anthony Lippert – “NVIDIA GPU Architecture for General Purpose Computing”
[4] David Kirk/NVIDIA and Wen-mei Hwu, 2006-2008 – “CUDA Threads”
[5] Yadav K., Mittal A., Ansari M. A., Vishwarup V., “Parallel Implementation of Similarity Measures on GPU Architecture using CUDA”
[6]Direct Compute Programming Guide (http://developer.download.NVIDIA.com/compute/DevZone/docs/html/DirectCompute /doc/DirectCompute_Programming_Guide.pdf)
[7] Singh B.M., Mittal A., Ghosh D., Parallel Implementation of Niblack’s Binarization Approach on CUDA.
[8] Practical Applications for CUDA (http://supercomputingblog.com/cuda/practical-applicationsfor-cuda/)
[10] NVIDIA Corporation. NVIDIA CUDA Compute Unified Device Architecture Programming Guide, June 2008.
[11] Danilo De Donno et al., “Introduction to GPU Computing and CUDA Programming: A Case Study on FDTD,” IEEE Antennas and Propagation Magazine, June 2010.
[13] Big Data Processing with Hadoop-MapReduce in Cloud Systems, Rabi Prasad Padhy,Senior Software Engg, Oracle Corp.,Bangalore, Karnataka, India.
[14] J . Ziv and A. Lempel, “A Universal Algorithm for Sequential Data Compression,” IEEE Trans. On Information Theory, pp. 337-343, May 1997.
[15] J. Ziv and A. Lampel, “Compression of Individual sequence via variable-rate coding,” IEEE Trans. On Information Theory, pp.530-536, sept. 1978.
[16] A. Wyner and J. Ziv, “The sliding window Lemple-Zi algorithm is asymptotically optimal,” Proc. IEEE, pp. 872-877, June 1994.

BIOGRAPHY

Shantanu Walunj, Pursuing B.E. in Department of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research (BSCOER), Pune, India.

Kiran Sarpale, Pursuing B.E. in Department of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research (BSCOER), Pune, India.

Pravin Salve, Pursuing B.E. in Department of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research (BSCOER), Pune, India.

Sataym Pawar, Pursuing B.E. in Department of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research (BSCOER), Pune, India.

Dr. S. M. Chaware, Professor of Department of Computer Engineering, TSSM Bhivarabai Sawant College of Engineering Research (BSCOER), Pune, India.

DOI: 10.15680/IJIRCCE.2017. 0504286