IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

Smart Web Search Engine Framework with Personalization and Privacy Preservation with Hadoop

Shri Nidhi P S[1], Bindiya M K[2], Mohan H S[3], Ravi Kumar G K[4]

IV Sem, M.Tech, CNE Department, SJBIT, Bengaluru India[1]

Research Scholar, VTU, Belgaum, India[2]

Prof& Head, Department of ISE, SJBIT, Bengaluru, India[3]

Prof& Research Guide, VTU, Belgaum, India[4]

Abstract: Now a day's data of an individual is sensitive. Everybody does not want to disclose their sensitive information. To achieve this user need to secure the sensitive data. Search engine also need to be personalized. This paper intends a design or a structure for securing the sensitive data including correct delivery of search results. The user should provide keywords for search engine before only. Based upon the content otology and location ontology the appropriate results will be displayed. User entered keyword will be encrypted. This paper also focuses the storage of data, so hadoop is used in order to store any type of data.

Keywords: Sensitive data, personalized search engine, content ontology, location ontology.

I. INTRODUCTION

Due to the rapid development of computing technology, most of the people drawn their attention towards these technologies. One of the major things here is the storage of the people private data in the servers, which is handled by the third party. Here the storage and access cost is low but the security issue arises. Because of rapid development of information digitization so many data or information is generated. Data can be categorised into three categories. Namely unstructured data, structured data and semi structured type of data. All types of data are generated, and sorting of those data, analysing them is an issue. Provider should take care of the security of the user's individual sensitive data. People data will be in huge and that need to be stored securely and processed.

There are four safety issues regarding the security of sensitive data. First, transmission of the data from the source to destination in an big data platform. Second, security issues regarding the storage of data and the computation of users sensitive data. Third, security issues in the cloud platform. Fourth, this issue is about the safe destruction of the data [1].

The technologies, which are existing, will concentrate on the aspects of data sharing and privacy of the data. They did not consider the entire lifecycle of the data security. Big data platform is a system, where any number of stakeholders can involve in that system. They also called as multi stake holders. In this system nobody will support any type of security leak in the loss of the user's sensitive data. In this paper , the major aspect is of analysing the issues regarding the security and the sharing of data in the entire lifecycle and due to the present of personalised engine users can get accurate results. With the help of proxy re-encryption technology in big data, the model will ensure security of sensitive data. Adoption of the search engine is the research problem statement. This adoption of search engine has many applications. Mainly, there are two ways of adoption took place. One is the specification of query adoption, in order to provide the accurate results for the user query from different types of data stored. Second is to adoption of the user specification that should meet the expectations of different users in the results, which they get. In this adoption of search engine the two factors need to be handled. One is identification of the preferences given by the user, second is to optimization based on the ranking. The ranking is based on the first step preferences.

Ontologies are the one, which are expensive to construct the framework, but with the ontology concept it is easy to get the results quickly [2]. In order to represent the knowledge of the domain accurately, ontologies plays the major role. In the same way re-using characteristic of ontology is the major advantage of ontology. This ontology will easily adopted to any of the new technology, which is of knowledge based. Recently, many search engines has adopted this ontology concept and the ranking process, which is used for accurate results are still in the initial stages. Swoogle rank is the technology adopted in Google's page rank system. Here the proposing architecture is client-server, where the information will stored in server and the clients will get the data through the search engine provided by server.

In many of the search engines, searching is the most commonly performing task. In the personalised search engine, the result will be provided to the user based on the

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

keyword specified by the user. If the search engine is perfect, then it should compile through all the data, which are stored in database then the results will be displayed. Now a day in many of the search engines, the results displayed for the query is non-relevant. In order to solve these problems, the personalised search engine is the solution for the accuracy and the efficiency. Rest of the paper holds the creation of personalised search engine, client server architecture, ranking based search results, content ontology, location ontology and the storage of the data or information in the efficient hadoop storage.

II. BUILDING OF AN PERSONALIZED SEARCH ENGINE

If the user enters query, "types of data" then the search engine will provide so many results will provide so many results for the entered query. User required result is the data types, but the search engine will provide the first result as structured data, unstructured and semi structured data. Here, the accuracy will not reach the maximum. So the search engine should be personalised. So that user can get accurate results.

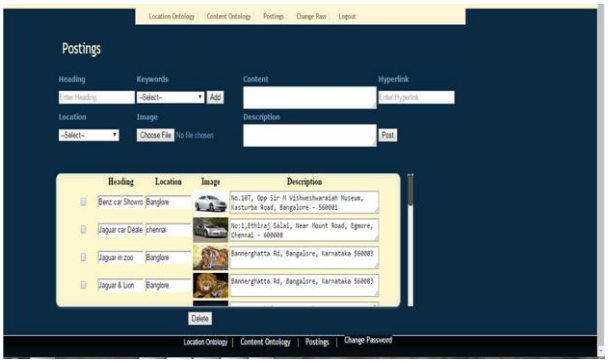

Fig 1: personalised search engine

Figure 1, tells about the keyword need to be specified and the overall requirements which the user can specify. The above the personalised search engine and it is of client-server architecture. In server this keywords need to be stored. This storing procedure will be the admin part.

Due to the providence of more relevant results, the personalised search engine gains more popularity. According to the research, existing search engine provides the percentage of 52 out of 20000 queries[3]. In personalised search engine the percentage will be more and the accuracy will be high. In order to achieve this there are mainly two ways, one is Contextual computing[4] , this refers to the improvement of the users interaction. In this, computing based on the context, application and the information of the user specified. Contextual computing is nothing but the building of a model based on the user behaviour and the preferences, which are specified. Each and every point of the user data, computation need to be takes place.

Next is the Content-based technique [5] or approach, where natural and statistical language techniques are used to get the results. In this technique, it contains a set of words or definition, but it cannot differentiate the documents, by its priority. By this there are set of problems, which are generated the names given for this is "author relevancy" technique. This technique states that the results will be provided for the user query based on the authors notes or materials, which are stored.

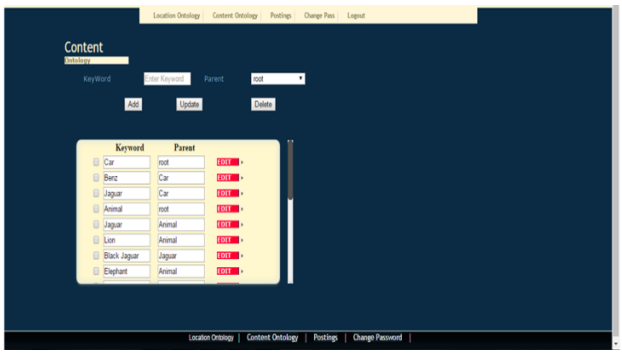

Fig 2: search using content ontology

Figure 2, tells the contents which are present in search engine are accessed through the content based. In this content based ontology the user entered word will be search through the content based and it is displayed on the basis of parent and ranking process. One among the easiest way to show the relevant results for their query can be based on the user location. This also called as personalization ranking [6]. To personalize the results, search engine has to record what you are doing and that rings privacy alarm. Others can see what you have looked in search engine. Your search history will be stored in browser or in the server. We can able to search securely by additional authentication technique. A history of any user is kept for 180 days. You can delete that history at any time, but even if you do not, it can't actually be viewed. Here security issues will come for the users.

To maintain the user privacy, user entered search word will be encrypted and that encrypted word will match the database, which is present in the server. Here we are developing an personalised search engine[6] in the client server architecture not on the web search. In order to provide the best results and the relevant information to the user the search engine need to be optimized. By this the user will attract and it motivates the service provider to load the high quality contents, very interactive applications and attractive interface for the user. In the personalized search engine, the user need to search in the limits, this is one of the disadvantages mentioned in the previous survey. In this paper the proposed system will overcome from this disadvantage. In this search engine all the related results will displayed based upon the priority, hence there is no loss of related information or results.

III. SECURING SENSITIVE DATA IN BIGDATA PLATFORM

Hash Tag Generation for Query Terms and Privacy Data. Algorithm is used to User location and query's are stored into server which is hash coded by MD5 (One way encryption).

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

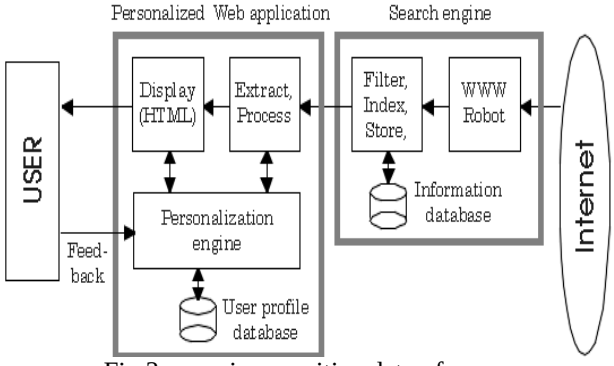

Fig 3: securing sensitive data of users.

The above figure 3 shows how the personalised search engine works and the storage of the data. In this paper we proposes a one way encryption for the user query[7]. The keywords which are given by the user are encrypted and stored in the database. The given query is also encrypted and if the encrypted query is matched with the contents in database then the user wanted appropriate result will be displayed to the user. one way encryption using md5 algorithm. This algorithm provides 5 layers of security for the data. Issuing and retrieving of sensitive data, to achieve this a mechanism is required in the semi trusted big data platform. There are four problems related to the safety. Reliable submission, storage must be in a safe place, the usage of the data must be riskless and the destruction of data in a secure way. This is also called as the lifecycle of the sensitive data. Encrypting the user query will be the additional security for the user privacy. The other security mechanisms like authorization, authentication, encryption and decryption will also be provided. In the case of cloud, the security issue will be more, when compared to hadoop storage.
Md5 algorithm accepts the user text or data as input then generate the output of fixed length. The length of the generated output will be less than the input. This output is the cryptographic hash function or the message digest. In this proposed work md5 algorithm has the properties namely, the one way encryption. By this, it is very difficult to get the data or text of user by any of the snooper. Md5 algorithm should be collision resistant, when we store the query without encryption then it is easier to find the data regarding that query.

Message digest 5 algorithm [8] is as follows,

Step 1: Add the padding bits behind the input message.

This step is to extend the initial message and make its length be congruent to 448 mod 512. First, a single bit "1" is appended to the message. Then, a series of "0" bits are appended so that
Length (the padded message) □ 448 mod 512
For example, suppose the initial message has 1000 bits. Then this step will add 1 bit "1" and 471 bits "0". As another example, consider a message with 448 bits. Since the algorithm doesn't check whether the initial length is congruent to 448 mod 512, one bit "1" and 511 bits "0" will be appended to the message. Therefore, the padding bits length is at least one and at most 512.

Step 2: Add a 64-bit binary string, which is the representation of the message's length.

Here, please pay attention to the meaning of the 64-bit binary string. You should not regard it as first 64 bits of the initial message. It is actually the binary representation of the length of the initial message. For example, suppose the message is 1000 bits length. The message's 64-bit binary representation would be 0x00000000000003E8. If the message is very long, greater than 264, only the lower 64 bits of its binary representation are used.

Step 3: Initialize four 32-bit values

These four 32-bit variables would be used to compute the message digest. We denote them by A, B, C, D. Their initial values are:
A = 0x67452301
B = 0xEFCDAB89
C = 0x98BADCFE
D = 0x10325376

Step 4. Processing Message in 512-bit Blocks.

This is the main step of MD 5 algorithm, which loops through the padded and appended message in blocks of 512 bits each.

Step 5. Output.

The MD5 message digest algorithm is simple to implement, and provides a "fingerprint" or message digest of a message of arbitrary length. It is conjectured that the difficulty of coming up with two messages having the same message digest is on the order of 2^64 operations, and that the difficulty of coming up with any message having the given message digest is on the order of 2^128 operations. The MD5 algorithm has been carefully examine for weaknesses. It is, however, a relatively new algorithm and encourage security analysis is of course justified, as is the case with any new proposal of this sort.

The operation in md5 are as follows:
- Boolean operation takes place in bitwise.
- Modular addition operation.
- Shift operation takes place in cyclic manner.

In 32 bit machine, all these operation will be fast, so the md5 is fast.

In the proposed work hadoop is used for the storage. All data and the encrypted keywords, queries are all stored in hadoop. It is open source software. It has both the storage part and the processing technique effectively and efficiently [9]. Hadoop is having a distributed environment so that it can maximize form a single to 'n' number of machines. Hadoop are designed to overcome the hardware failures and that failures should be handled automatically by the framework. Hadoop consists of clusters of computer to store the data.
Hadoop contains all information, utilities, and library needed by the modules present in hadoop. Commodity machines are used to store the data in the HDFS. Hdfs,

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

stands for hadoop distributed file system. It is having ancharacteristics of distributed file system. In clusters, because of Hdfs provides an high bandwidth, which is of aggregated. In hadoop the resource management will be taken place by the YARN. This also present in the clusters only. Scheduling of the resources will be taken care by the yarn only. For the large scale data processing, map reduce is the technique used in hadoop. Cloud storage can also be used for storage but in cloud there is no security guarantee and processing of storage of data like hadoop. Hdfs is an low cost and fault tolerance hardware. Hadoop hdfs holds the large amount of data and it is very easy to access the data. In order to store the huge data in hadoop it requires multiple machines for storage purpose. The data which are stored in hadoop are in redundant fashion to rescue from the data loss or failure. Hdfs is an parallel processing system.

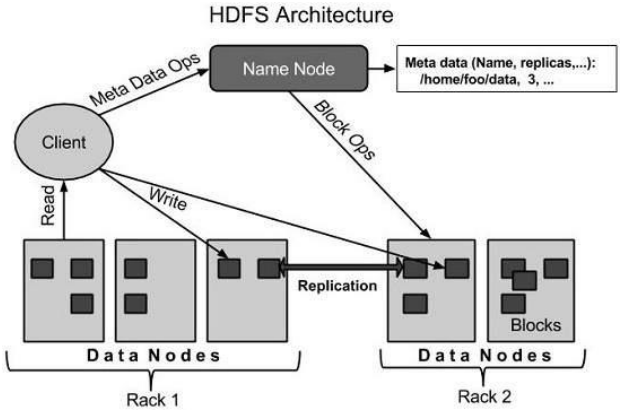

Fig 4: hdfs architecture.

Figure 4 explains the architecture of the hadoop distributed file system. Client, name node, data node are all explained in below section.

In hadoop the storage machine are attached in horizontal from. Hence any number machine can attach to it. Hadoop will provide the easy interface and it provides command interface also. Name node and data node are two built in servers helps the user to check this states of the cluster. The security like authentication, authorization will be provided by hadoop.

Name node, it contains the operating systems like Linux, Ubuntu etc. It has the commodity hardware. The name node is present in Hdfs.

Data node, in Hdfs each and every node present in the cluster, there will be a data node. It will maximize the data in system. It will perform operations based on the client request.

Blocks, the user file will be divided into one or more segments. These segments will store in data nodes. This file system are called as blocks. Fault detection and recovery, since hadoop has distributed file system there will be a huge data and the detection of fault and the recovery of data recovery from fault will be in an quick time. This hadoop has the automatic mechanism of detection and recovery. In order to maintain the huge data sets the hadoop, Hdfs should have so many nodes in a single cluster. Then only it can manage the huge data. In this the user query or task will be solved effectively and very fast. The traffic present in the network will be reduced and the overall throughput of the process will increase.

Based on java, map reduce[10] processing technique and a programming model is developed in an distributed manner. There are two main tasks in map reduce technique, one is mapping and another is reduce. Individual elements break into tuples i.e. sub elements(key/value pairs)in a group of data or set of data. This is the task of mapping. Secondly, reducing the task, the output from the mapped task will taken and then it will process the task. After all the processing data, then it combines those data tuples and produces the output. After mapping only reduce technique will taken place, over multiple computing nodes it is easy to scale the over all big data. This will be the major advantage of the map reducing. There are two primitives present in map reducing. One is the mappers and other is the reducers. All users will attract to this model only because of scalability. Scaling an application can run over hundreds, thousands and many more.

IV. HOW USER GETS ACCURATE RESULTS

User will get results through the search engine and user will get the accurate results through personalized search engine. The search result will based on the user keyword. The results will validate using ontology management systems, Personal Behavior Collections using SPY NB Algorithm [11], Personalized Searching using Adaptive Ranking Technique. The ontology management will classified into two categories one is content ontology, second is location ontology.

Content ontology [12], Based on the content present in the server. The ontology process will take place; consider an example if the user enters the query tiger, then the root of the entered query will be animal. This makes the incremental of the domain knowledge for representation. This technique is the major advantage of the ontology concept. The root will be saved and it used for other word also, which is related to the root animal. Like this any of the word can be considered. Content ontology has an character of easy deployment into any of the platforms. The corps of the content will be analyzed to identify the related terms in that domain. With the help of the frequency measures the representation of the corps will be highly effective. Based upon the ranking methods the contents will be ranked as positive and negative. Positive rank is for the active search content and the negative rank is for the passive search content.

Location ontology [12], Based on the user location, the entered query's result will be displayed. If the keyword in the data base, which are predefined by the user it will be shown or displayed depending on the location of the user situated. Location concepts are extracted from the documents, and it is difficult to extract similarity and parent-child relationship from full documents because a limited amount of location concepts are present in the document. As all the locations are almost identified, it is possible to create ontology by organizing all cities under

DOI 10.17148/IARJSET.2016.3502

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

their province or state, all provinces under their regions, and all regions under their country, like an hierarchy level.

Personal Behavior Collections using SPY NB Algorithm: Spy NB algorithm works based on the user behavior model and the preferences mentioned. When the user click on certain document then that document is treated as an positive sample and rest of the relevant results appeared will consider as negative sample or un labeled. In order to predict the result, the navies bayes classifier will help in prediction and the "spy" technique is used.

Spy NB algorithm is used to differentiate between the negative and positive data. Here clicked result will be treated as positive and the unclicked result is treated as the negative data set. Whenever the user search any query then he will get the results related to the query. Primarily user will provide the keyword for the search engine. Based on that keyword the authorized user will get the results. Then based on the priority user will get all results. For next search the search engine provide the rank according to the click through. This Spy NB algorithm is very useful to calculate the results. The below is the Spy NB algorithm.

Spy NB algorithm:
Step 1: Get the query keyword from the user Let it be K.
Step 2: Compare K with Post Link Content Keywords and Extract the Result Set RS.
Step 3: Re-rank the RS with Location Ontology concept
Step 4: List the Re-ranked RS to User Display.
Step 5: Record the Links which are selected by the users.
Step 6: Let clicked Result Subset is Positive and Unclicked
Result sub sets are Negative Set for that search.
Step 7: Record Positive Set and Negative Set in that User account.
Step 8: Repeat the process for all the search queries.
Step 9: Stop.

The above algorithm explains that first step is to enter the query and that query is received by the server. The entered query will be K. Then that query will be compared with the predefined result set R. This result set is the one, where the users keywords are stored in the set and that called as results set. After this based upon the re ranking technology, the entered query will be compared with the content and location ontology. By mentioning the positive and negative set results again it will be re-ranked. Next step is to mention the positive and negative rank for the results. Whenever the user clicked the result, that result will be positive and the other results will be the negative one. After this all the records are saved in the user account. Then repeat the process for all search queries and last step is the stop the process.

Adaptive ranking technique [13], this ranking is employed to learn, how the personalized ranking function works and the adoption of the search results based on the user location and the content mentioned in the preference table or list. For a given query, a set of contents, preferences, and location concepts need to be extracted from the previous search results. While giving the preferences the pairs of preferences will also possible with the help of the linear ranking function.

An adaptive ranking method is used to evaluate the personalized search result of the user. For the large data base, this adaptive ranking method is used to rank the user results. In this method, it will provide the rank for each and every level of the result. The method calculates the distance between adjacent words and the complex terms and then compensates it. By this the user will get appropriate results. This method is very much useful for the large database and the ranking achieves the results on these datasets.

V. EXPERIMENTAL RESULTS AND FUTURE ENHANCEMENTS

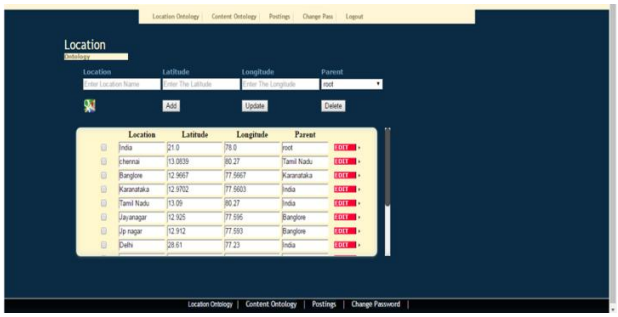

Fig 5: Location based search result.

The above figure is the location based result, based upon the user location the search engine will provide the result in terms of the location. The content based search result is shown in figure 2.

This paper discusses about the client server architecture. This is for the client request who are using that any particular server. This is very useful for any organization to store and retrieve any type of data. The enhancement of this paper is for web search engine. When this is implemented in web search engine the obtain results will be more accurate and the efficiency will be in large extent. Only disadvantage in implementing this technique for web search is individual user need to specify their keyword to the search engine.

VI. CONCLUSION

This paper proposed a work for the securing the sensitive data of the user. Through the personalized search engine the results for the user query will be accurate and the efficiency will be more. The user can get the results for their query based on content and the location of the user. The query entered by the user will be encrypted and the user specified keyword will be in the encrypted text. The link should be provided in between the encrypted keywords and the data stored in the database. The storage will be in hadoop. Hadoop can store all type of data and hadoop can process any type of data. Hadoop is efficient because of volume, efficiency, security and open source.

DOI 10.17148/IARJSET.2016.3502

ACKNOWLEDGMENT

We thank all the staff members of Department of Information Science and Engineering for their help during the course of this paper. Last but not the least we thank our parents, family members & friends, for their Continuous and great support and encouragement throughout this paper.

REFERENCES

[1] Xinhua Dong, Ruixuan Li_, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu, "Secure Sensitive Data Sharing on a Big Data Platform", IEEE and T SINGHUA SCIENCE AND TECHNOLOGY, Vol :20, Number1 February2015.
[2] Matthew Jones, Harith Alain, "Content-Based Ontology Ranking".
[3] "Personalised Search", September 2002/Vol. 45, No. 9 communications of the acm.
[4] "Context-Aware Computing" by Constantin Schmidt Computer Engineering Bachelor of Science Berlin Institute of Technology, Germany.
[5] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng "fundamentals of content-based image retrieval".
[6] Wenhai Sun, Wenjing Lou, Y. Thomas Hou, and Hui Li, "Privacy-Preserving Keyword Search Over Encrypted Data in Cloud Computing" , S. Jajodia et al. (eds.), Secure Cloud Computing, DOI 10.1007/978-1-4614-9278-8__9, Springer Science+Business Media New York 2014.
[7] T.Sathiyabama, Dr. K. Vivekanandan, "Personalized Web Search Techniques -A Review", Global Journal of Computer Science and Technology,Volume 11, Issue 12, Version 1.0 July 2011.
[8] R. Roshdy, M. Fouad , M. Aboul-Dahab "DESIGN AND IMPLEMENTATION A NEW SECURITY HASH ALGORITHM BASED ON MD5" International Journal of Engineering Sciences & Emerging Technologies, Volume 6, Issue 1, August 2013.
[9] Serge Blazhievsky, Nice Systems " introduction to hadoop, mapreduce, and HDFS for big data applications", SNIA education.
[10] http://www.tutorialspoint.com/hadoop/mapreduce.
[11] Lin Deng, Wilfred Ng, Xiaoyong Chai, and Dik-Lun Lee, "Spying Out Accurate User Preferences for Search Engine Adaptation", Department of Computer Science Hong Kong University of Science and Technology {ldeng, wilfred, carnamel, dlee}@cs.ust.hk.
[12] Vemula, Divya, "Ontology Based Personalized Search Engine" (2014). Technical Library.Paper 182.
[13] Ihab f. ilyas walid g. aref and ahmed k. elmagarmid and hicham g. elmongui and rahul shah and jeffrey scott Vitter, "Adaptive Rank-aware Query Optimization in Relational Databases" Purdue University.