A Rapid Data Retrieval in Big Data Environment Using Hummingbird Algorithm

S.Selvakanmani[1], C. Kanchana [2]

[1] Assisstant Professor Dept. of Computer Science & Engg.
Velammal Institute of Technology Chennai, India

[2] PG Student Dept. of Computer Science & Engg.
Velammal Institute of Technology Chennai, India

Abstract

A Range Aggregate Query (RAQ) is a mechanism where the range servers are tenacious aggregate queries inquest information in numerous servers i.e. splitting the queries and inquest the information. Disparate data are cluster into single page. Data records are temperate and uploaded by the admin where the uploaded files are unstructured files; admin alone can recast the content and update it. Data records are stored in Hadoop Distributed File System (HDFS) which act as database. Balance Partitioning is done in numerous queries. If any queries are not found, redirection process takes place. Queries process in the format of m:n i.e. m aggregate columns and n index columns in a same record. Hummingbird algorithm is used to retrieve the queries accurately and efficiently. Irrelevant features are ejected from the database and elite inquest are appeared to the end user.

Keywords: Hadoop, HDFS, Cloudera, Hive, Hummingbird algorithm.

1. INTRODUCTION

Big data is a broad term used to describe the exponential growth and availability of data for data sets where the data are compared with capacity, storage and tools. Process of storing and analyzing peta or gigabytes of data to make sense of it for the betterment of organization. Google gets 2 million searches every minute and deals with 20 petabytes of data each day. Face book deals with 3 or 4 petabytes data and 34 thousand likes every minute. 3v's in big data volume, velocity and variety. volume consist of terabytes, records, transactions, tables, and files. velocity has batch, real-time, streams, near-time. variety like structured, unstructured, semi-structured. Types of data in big data are structured, unstructured and semi-structured. structured data are data which have pre-set format for example address book, banking, and transaction. unstructured data has no pre-set format for example movies, audio, text files, web pages, computer programs, social media. semi structured data are that can be put into a structured by available format descriptions.

Hadoop is a open source java frame work given by apache software. Hadoop keep your data in local file system and process it takes less time instead of storing it in somewhere else. The hadoop run applications on system with thousands of nodes involving thousands of terabytes. It was inspired by Google's Map Reduce (MR), a software framework in which an application is broken down into numerous small parts. any of these parts also called as fragments or blocks. The hadoop framework is used by major players including Google, yahoo and IBM. The preferred operating systems is windows and Linux but hadoop can also work with BSD and OS X. Benefits of hadoop are distribute data and computation, tasks and independent, simple programming model, flat scalability, HDFS store large amount of information.

Hadoop distributed file system (HDFS) is a technique for storing data with cluster of commodity hardware i.e. cheap hardware like pc, lap etc. HDFS has streaming access pattern we can write once read any number of times the content of files is called streaming access pattern. It can be run anywhere i.e. platform independent. By default the size of the block is 64MB. Five services in HDFS are Namenode, Secondarynamenode, Jobtracker, Datanode, and Tasktracker. Namenode is the node which stores the file system metadata i.e. which file maps to what block locations and which blocks are stored on which datanode. Namenode maintains two in memory tables, one which maps the blocks to datanodes (one block maps to 3 datanodes for a replication value of 3) and a datanode to block number mapping. Secondary namenode purpose is to have a check point in HDFS. Datanode is where the actual data resides. All datanodes send a heartbeat message to the namenode every 3 seconds to say that they are alive. If the namenode does not receive a heartbeat from a particular datanode for 10 minutes, then considers that datanode to be dead/out of service and initiates replication of blocks which were hosted on that datanode to be hosted on some other datanode. The datanode can talk to each other to rebalance data, to keep replication high. Jobtracker primary function is resource management i.e. managing the tasktrackers, tracking resource availability and task life cycle management. Tasktracker has a simple function of following the orders of the Jobtracker and updating the Jobtracker with its progress status periodically. Features of HDFS are job performance and fault tolerance. Some of the hadoop related tools are Hive, HBase, Cassandra, Pig, Sqoop, Zookeeper.

Hummingbird algorithm is a nothing but Google new search algorithm it uses to sort through all the information it has when you search and come back with answers. It's called Hummingbird. The search algorithm is a technical term Google uses to sort through billions of web pages and other information it has, in order to return what it believes are the best answers.

2. RELATED WORK

Twitter's real-time related query suggestion spelling correction[1] involves two individual systems working to recover the problem. The first one is hadoop based implementation and second one is memory processing engine which is designed for processing a task. In the future enhancement work is capable of handling big data and fast data both together. Limitations in this paper is search engine is not efficient, everything is based on query spelling and suggestions. Quantifying trading behavior in financial markets[2] using Google trends is based on human interaction with the internet in financial trade and stock markets. Data sets are collected by human behavior i.e. number of clicks on search results. Google trends are used to determine the searches like how many searches n(t-1) have been carried out for a specific search term such as in a week t-1, where Google defines weeks as ending on a Sunday. Search volume data changes slightly. Both Google trend data and stock market data are used for decision making. They find this strategy is based on search volume data for U.S. market than strategy using global search volume data. Limitation in this paper is collecting large scale.

Characterization of the structural robustness of data center networks[3] paper describes the data center network (DCN) architectures. DCell architecture degrades under all the failure types such as Fattree and Threetier architecture. Because of the connectivity pattern, layered architecture, and heterogeneous nature of the network, the result demonstrated that classical robustness metrics are insufficient. The main role of data center network (DCN) to deliver the required quality of service (QOS) and satisfy service level agreement (SLA). Limitation in this paper is higher in cost. HyperLogLog algorithmic engineering of a state of the art cardinality estimation algorithm[4] reduces the memory requirements and increases its accuracy in the range of cardinality. This algorithm is implemented in Google and evaluated. The advantage of this approach is that the hash values for any given expression only n process which exceeds the single computing resource so distributed computing resource is used by commodity hardware. Geoavailability[5] is a light weight distributed spatial indexing structure which is used to eliminate the storage resource that does not hold the relevant information. This leads to effective utilization of resources and response time. Overall this approach collects data volumes across number of distributed computing resources and allows fast and flexible retrieval of information for analysis and processing. Limitation are time consumption.

Hive a petabyte scale data warehouse[6] using hadoop. Hive is a large size dataset collected and analyzed in industry which is very expensive. Hive support query language (QL) executed using hadoop. Hadoop is popular tool which is open source which consist of map reduce and hadoop distributed file system. Metastore store the metadata and partitions into relational database and it is used for API client access. In the future enhancement work is JBC drive which is software enables java application to interact with database and ODBC is an open database connectivity programming language middleware API for access database management systems. Multiquery n way in a single map reduce jobs. Limitation are time complexity. Distributed online aggregations[7] is extended to distributed hash table where sites are maintained (Table 1). Distributed online aggregate is based on iteration and accurate results. Here local aggregate are combined into global aggregate using from each set of random samples final distributed to the processing sites. Number of processing nodes increases in distributed online aggregate as per the sample size increases. Limitation are single table query is focused.

Range queries in dynamic OLAP[8] data cubes describes range query are implemented in aggregation operation on over all selected tuples of OLAP data cubes. Size of the data cubes are exponential in number of its dimensions, rebuilding the entire data cube is costly. To tangle this problem, a new approach has been introduced which bring out constant time per range sum query. For each update cost within $0(n^{d/2})$. Then introduce the technique called Double Relative Prefix Sum Approach for range sum query problem. The proposed algorithm makes use of three data structures: relative prefix array RP, relative overlay array RO and the block prefix array BP, respectively. Assume that data cubes are stored by array A in range query it takes $o(n^{d})$ in worst case because all the array RO every overlay array is obtained by partition of array A which has $[n/r]^{d}$ of box cells it can be further divided as $(n/rr_1)^{d}$. Data structures are used in this technique we check for each update and the time for each update is $0(n^{d/3})$. Block prefix BP contains $(n/rr_1)^{d}$ and volume V. The proposed technique reduces the update time complexity to 0(n/3). In the existing system time complexity is $o(n^{d/6})$.

Table 1: Different approaches

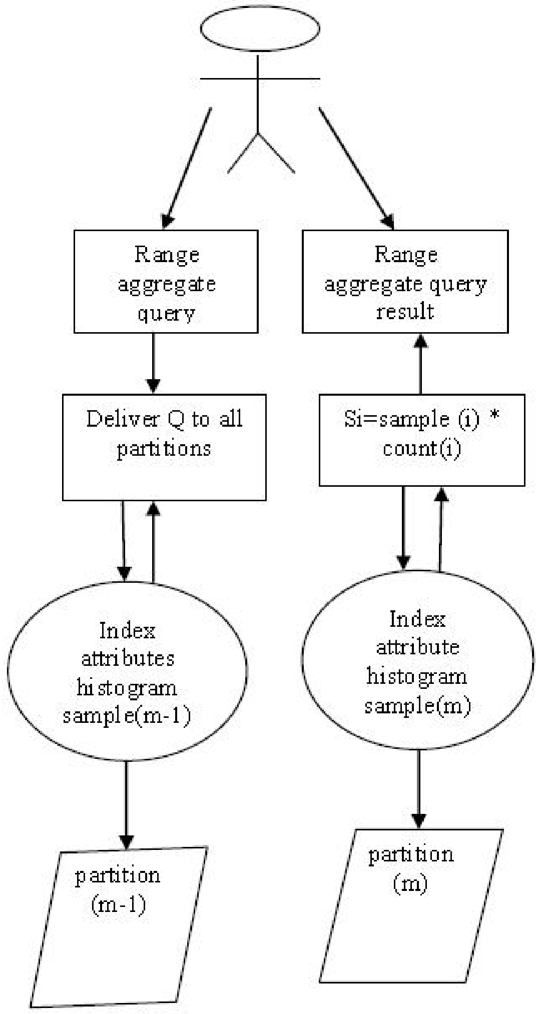
S.No	Paper titles	Approaches
1	Fast data in the era of big data: Twitter's real-time	Session-based technique

	related query suggestion architecture	
2	On the characterization of the structural robustness of data center networks	Threetier, Fattree technique
3	HyperLogLog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm	Column-stores technique
4	Polygon-Based query evaluation over geospatial data using distributed hash tables	Hierarchical distributed hash table
5	Integrity for join queries in the cloud	Probabilistic approach

3. EXISTING METHOD

In existing system, prefix-sum cube (PC) is used in OLAP (Online analytical processing) to boost the performance of range-aggregate queries. All the numeric attribute values are sorted and any range aggregate query on a data cube can be answered in constant time. However, when a new tuple is written into the cube, it has to recalculate the prefix sums for all dimensions. Online aggregation (OLA) is an important approximate answering approach to speeding range-aggregate queries, which has been widely studied in relational databases and cloud systems. The OLA systems provide early estimated returns while the background computing processes are still running. The returns are progressively redefined and the accuracy is improved in subsequent stages, shown in Fig.1.

Fig1: Fast RAQ framework


Range aggregate query
Range aggregate query result
Deliver Q to all partitions
Si=sample (i) * count(i)
Index attributes histogram sample(m-1)
Index attribute histogram sample(m)
partition (m-1)
partition (m)

Existing system have some loss of credit they are given as follows:

User cannot obtain an approximate answering with satisfied accuracy.
It cannot acquire acceptable approximations of the underlying data sets, when data frequency distributions in different dimensions vary significantly.

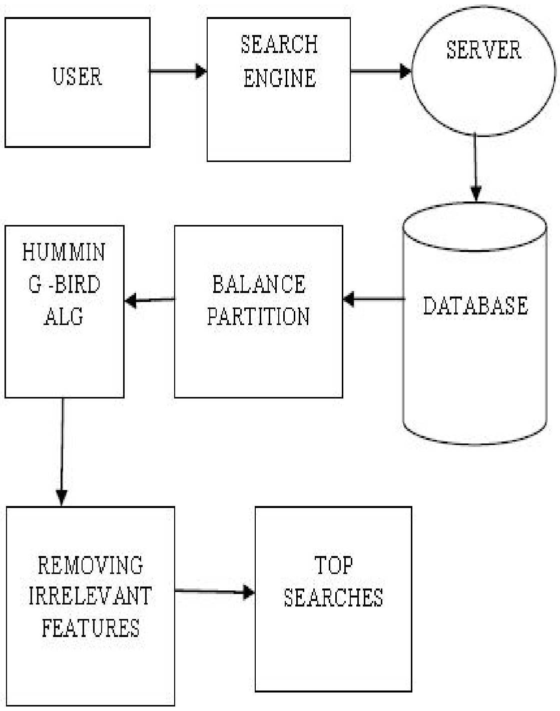
Hive is lower than that of FastRAQ. Major drawback of this method is that it can solve only 1:n format range aggregate queries i.e. there is one aggregate columns and n index columns in a record.

4. PROPOSED METHOD

In the proposed system the admin collects the data records and uploads the files. The collected files are stored in a database here we use HDFS (Hadoop Distributed File System) to store the data records. Uploaded files are modified and updated by admin alone. Balance partitioning is used to handle large keywords which partition the keywords into equal parts.

Hummingbird algorithm focuses on each individual word in a search query the whole sentence or conversation or word is taken into account, rather than particular words. The goal is that pages matching the meaning do better, rather than pages matching just a few words. Hummingbird places greater emphasis on page content making search results more relevant and pertinent and ensuring that Google delivers users to the most appropriate page of a website, rather than to a home page or top level page. Irrelevant features are removed and the top searches are appeared to the end users.

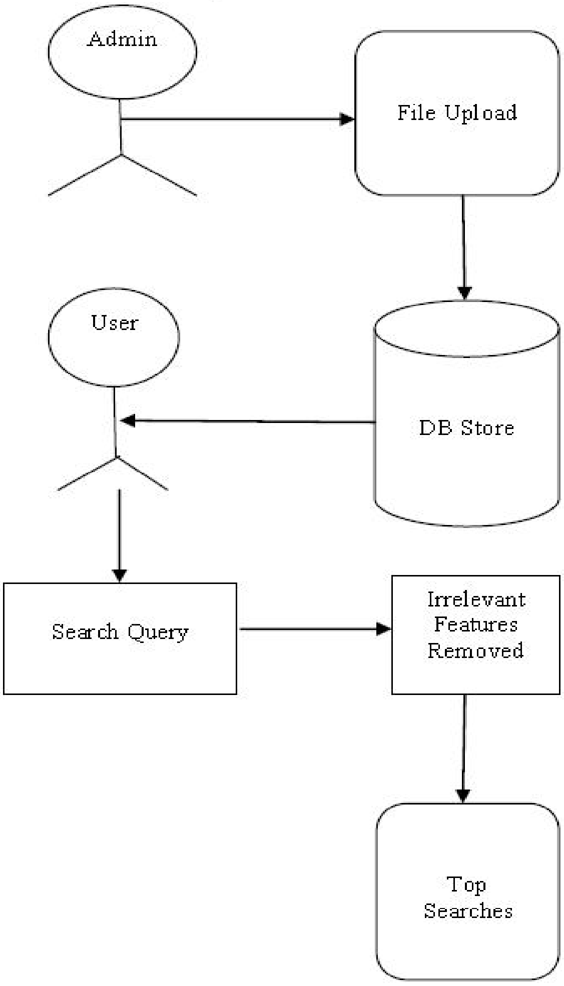
Fig 2 : Architecture diagram



An unstructured data is taken as a programming languages. The search engine obtains the data from the user and searches for appropriate result. A storage is done in Hadoop Distributed File System (HDFS) where the processed results are stored. Balance partitioning technique is implemented to partition the data into equal files. This technique is one of the efficient used. An Hummingbird algorithm is used to provide an accurate result. The work of this algorithm is to partition the data into each and every word as in if a sentence is taken then each variable is taken as a partitioned value. Therefore all the irrelevant data is been removed and top searches are obtained.

a) FLOW CHART

Fig 3: Flow chart



Explanation

The above flow chart is been explained. we have two types of login page one for admin and the another one for user. In admin login page admin alone can use it where as in user login page all the users are allowed to login or visit the page. Admin collect the data from different websites and upload it in Hadoop Distributed File System (HDFS). Admin alone can modify the content and upload the files. User searches the query in search engine relevant results are appeared to the user by eliminating irrelevant content. Hummingbird algorithm is used to find the most perfect link from the given word or sentence. This algorithm is given by Google which delivers most appropriate page of the website, rather than home page or top level pages. In a single page many links or content are hidden by clicking the hidden links user can view the needed information. In top searches ranking are provided by the admin from the views of the users.

b) IMPLEMENTATION

1. File upload

Data records are collected from different websites and uploaded. Files are uploaded in HDFS (Hadoop Distributed File System) using put and get commands. For putting files on Hadoop we use Hadoop fs - put / < local machine path > / < hdfs path >. Files are copies from local file system to Hadoop distributed file system (HDFS) before copying the files directories are kept empty. Finally files are copied.

2. HDFS Store

Multiple files are clubbed into one sequence files. Images and videos are stored in Hadoop Distributed File System (HDFS) using Map Reduce Input Format, Output Format, and Record Reader in order to split them properly. The map step inputs data and breaks it down for processing across nodes within a Hadoop instance. These "worker" nodes may in turn break the data down further for processing. In the reducer steps, the processed data is then collected back together and assembled into a format based on the original query being performed.

3. Modify already existing content

Uploaded files are modified and updated by the admin Hadoop act as the middleware user cannot modify the content because of Hadoop in normal Wikipedia user can edit or modify the content whereas here admin alone can change the content.

4. End user

Whenever the user search the query the relevant information appeared by removing the irrelevant information. Top searches are displayed first. The link is given by URL like Wikipedia here user cannot edit the content more information are hided or mined inside the page.

c) ALGORITHM

Step 1: Initializing the data from the website location.

Step 2:Collecting the data from website location for uploading.

Step 3: Partition the data into equal data sets.

Step 4:uploading of file can be done and again partitioned (i=n)times.

Step 5: Analysis is done for partitioned data set to obtain top search.

Step 6: Increment the data at website location to follow the loop.

5. TOOLS AND EXPERIMENTAL RESULT

Ubuntu is a linux based operating system which is a open source. Operation of Ubuntu is under GNU General Public License (GPL). Ubuntu tool is mainly used for security our goal is to be secure.

Windows is also called as Microsoft windows. It consist of several families of windows some of the active families are DOS-based, Windows 9X, Windows NT, Windows Embedded, Windows Mobile and Windows Phone. Features of Windows are recovery, troubleshooting.

Hadoop is Apache open source framework written in java. It allows distributed processing of datasets across clusters of computers with the help of simple programming models. Hadoop tool is mainly used for quickly analyzing the data by distributing the data across multiple machines, utilizes the parallelism of CPU. The main advantage of Hadoop is highly available and fault-tolerance. Another big advantage is apart from open source, it is compatible on all the platforms since it is java based.

HDFS (Hadoop Distributed File System) is one of the most common file system used by Hadoop. It is based on Google File System (GFS) and provides distributed file system to run on large clusters. The main usage of HDFS is files are split into several blocks and those blocks are stored and processed.

Cloudera is an ecosystem of open source platform it provides Hadoop distribution. It is one of the sponsor of Apache software foundation. The main usage of cloudera is storing, accessing, managing, analyzing, searching the data.

Hive is a data warehouse infrastructure tool to process the structured data in hadoop. The main usage of hive in this paper is to summarize the big data, and makes querying and analyzing easy. SQL type language is provided for querying called HiveQL or HQL (Hive Query Language). Some of the features of Hive are it is scalable, familiar, fast and extensible.

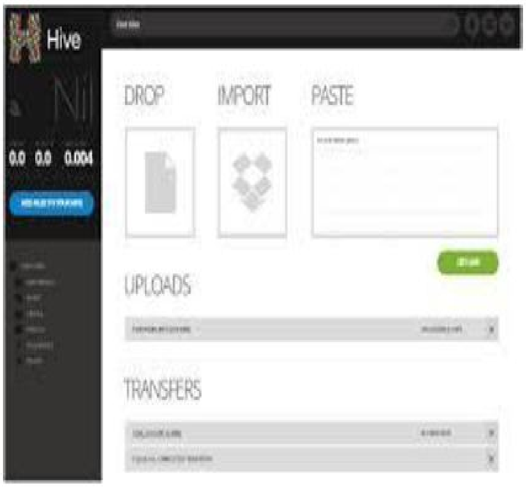
Experimental result are like we have two types of login form one for admin and the another one for user. In that login format both of them have separate user name and password. Admin alone can edit the content whereas the user can view the content. User searches query in search engine then the required content will be displayed it like normal Wikipedia but in normal Wikipedia user can modify the content here we are using Hadoop so user cannot be able to change the content he/she can only view the content. Data are collected from A to Z programming languages each alphabets have different languages. Files are stored in

HDFS (Hadoop Distributed File System). Cloudera is a ecosystem used for security purposes. Ranking is provided based on user searches. Hummingbird algorithm is focuses on each individual word in a search query rather than particular word or whole sentences.

The main goal is that page matching the meaning do better, rather than matching just few words. It removes irrelevant features and obtained the top searches accurately to the end users.

Fig 3: Admin and user login

Fig 4: Uploading files in HDFS



6. CONCLUSION AND FUTURE WORK

Hummingbird is a new Google announced algorithm that allows the search engine to process and sort its index more efficiently. With the help of this algorithm, Google is better able to understand the meaning of a phrase and return more precise results to complex search queries. These transformation produce brands with an opportunity to become more relevant and useful to consumers by expanding their content to add more informational content on their sites. Google also claims that Hummingbird does a better job of linking users to the specific page or answers they are seeking. Google delivers the most appropriate page of a website. Range servers are fixed which aggregate the query information search using multiple server like splitting the server as s1, s2,s3..etc, different data are collected and grouped into single page. It ranges from m:n format queries i.e. there are m aggregate columns and n index columns in a same record. For example in m:n format m is Java and n is Information. Finally, this paper provides a faster search result which takes less time to produce an appropriate result.

This paper although provides an efficient search technique still it could be improvised in the case of different algorithm in mere future. Also user edit does not takes place so as to provide an sincere result to the end user. An authentication could be provided such that no mal practice as in providing dummy links or spam links can be present. Although this system provides faster result it also has few demerits so those demerits can be overcome with better security as it cannot be breached.

REFERENCES

[1] G. Mishne, J.Dalton, Z.Li, A. Sharma, and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 1147-1158.

[2] T. Preis, H. S. Moat, and E. H. Stanley, "Quantifying trading behavior in financial markets using Google trends," Sci. Rep., Vol. 3, p.1684, 2013.

[3] K. Bilal, M. Manzano, S. Khan, E. Calle, K. Li, and A. Zomaya, "On the characterization of the structural robustness of data center networks," IEEE Trans. Cloud Comput., Vol. 1, no. 1, pp. 64-77, Jan-Jun. 2013.

[4] S. Heule, M. Nunkesser, and A. Hall, "Hyperloglog in practice: algorithmic engineering of a state of the art cardinality estimation algorithm," in proc. 16th Int. Conf. Extending Database Technol., 2013, pp. 683-692.

[5] M. Malensek, S. Pallickara, and S. Pallickara, "Polygon-based query evaluation over geospatial data using distributed hash tables," in Proc. IEEE/ACM 6th Int. Conf. Utility Cloud Comput., 2013, pp. 219-226.

[6] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, "Hive-a petabyte scale data warehouse using Hadoop," in Proc. IEEE 26th Int. Conf. Data Eng., 2010, pp. 996-1005.

[7]S. Wu, S. Jiang, B. C. Ooi, and K.L. Tan, "Distribute online aggregations," proc. VLDB Endowment, vol. 2, no. 1, pp. 443-454, Aug. 2009.

[8] W. Liang, H. Wang, and M. E. Orlowska, "Range queries in dynamic OLAP data cubes," Data Knowl. Eng., vol. 34, no. 1, pp. 21-38, Jul. *2000.*

[11] N.Pansare, V. Borkar, C. Jermaine, and T. Condie, "Online aggregation for large MapReduce jobs," Proc. VLDB Endowment, vol. 4, no. 11, pp. 1135-1145, 2011.

[12]T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, J. Gerth, J. Talbot, K. Elmeleegy, and R. Sears, "Online aggregation and continuous query support in MapReduce," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 1115-1118.

[13] Y. Shi, X. Meng, F. Wang, and Y. Gan, "You can stop early with cola: Online processing of aggregate

queries in the cloud," in Proc. 21st ACM Int. Conf. Inf. Know. Manage., 2012, pp. 1223-1232.

[14] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati, "Integrity for join queries in the cloud," IEEE Trans. Cloud Comput., vol. 1, no. 2, pp. 187-200, Jul-Dec. 2013.

[15] D. Mituzas. page view statistics for wikimedia projects. (2013). [Online]. Available: http://dumps.wikimedia.org/other/page-counts-raw/

[16] R. Sharathkumar and P. Gupta, "Range-aggregate proximity queries," IIT Hydrabad, Telangana 500032, India, Tech. Rep. IIT/TR/2007/80, 2007.

[17] P. Flajolet, E. Fusy, O. Gandouet, and F. Meunier, "Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm," in Proc. Int. Conf. Anal. Algorithms, 2008, pp. 127-146.

[18] E. Cohen, G. Cormode, and N. Duffield, "Structure-aware sampling: Flexible and accurate summarization," Proc. VLDC Endowment, vol. 4, no. 11, pp. 819-830, 2011.

[19] [Online]. Available: http:// research. neustar.biz/2012/12/17/hll-intersections-2/, 2012.

[20] H. Choi and H. Varian, "Predicting the present with Google trends," Econ. Rec., vol. 88, no. s1, pp. 2-9, 2012.

[21] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach M.Burrows, T. Chandra, A. Fikes, and R. Gruber. Bigtable: A distributed storage system for structed data. In OSDI, 2006.

[22] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In SIGMOD, 2004.

[23] J. Xu and W. Croft. Improving the effectiveness of information retrieval with local context analysis. ACM TOIS, 18(1):79-112, 2000.

[24] J. Rocchio. Relevance feedback information retrieval. In G. Salton, editor, The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall, 1971.

[25] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," ACM SIGCOMM Computer Comm.Rev.,vol.38, no. 4, pp. 63-74, 2008.

[26] K. Bilal, S.U. Khan, L. Zhang, H. LI, K. Hayat, S.A. Madani, N.Min Allah, L. Wang, D. Chen, M. Iqbal, C. Xu, and A.Y. Zomaya, "Quantitative Comparisons of the State-of-the-Art Data Center Architectures," Concurrency and Computation: Practice and Experience, vol. 25, pp. 1771-1783, 2012.

[27] K. Bilal, S.U.R. Malik, O.Khalid, A. Hameed, E. Alvarez, V. Wijaysekara, R. Irfan, S. Shrestha, D. Dwivedy, M. Ali, U.S. Khan, A. Abbas, N. Jalil, and S.U. Khan, "A Taxonomy and survey on Green Data Center Networks," Future Generation Computer Systems, 2013.

[28] Wikipedia Contributors. (2013) Geohash. [online]. Available: http://en.wikipedia.org/wiki/geohash.

[29] Hive Wiki at http://www.apache.org/hadoop/hive.

[30] E. Zeitler and T. Rich, "Massive scale-out of expensive continuous queries," Proc. VLDB Endowment , vol. 4, no. 11, pp. 1181-1188, 2011.