ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

A Framework to Process Range Aggregate Query Using MongoDB

Arkaprabho Roy[1], Akash Sampathkumar[2], Anto Melvin K.F[3], Meghana Venkatesh[4]

UG Student, Department of Information Science, New Horizon College of Engineering, Bangalore, India[1]

UG Student, Department of Information Science, New Horizon College of Engineering, Bangalore, India[2]

UG Student, Department of Information Science, New Horizon College of Engineering, Bangalore, India[3]

UG Student, Department of Information Science, New Horizon College of Engineering, Bangalore, India[4]

ABSTRACT: Range searching is a fairly well-structured problem in computational geometry. Big Data deals with class of problems called Range Aggregate Query problems, the aim is to deal with some composite queries involving range searching, where one needs to do more than simple range reporting or counting. A range query applies an aggregate function over all selected cells of an OLAP data cube. The essential idea is to precompute some auxiliary information that is used to answer ad hoc queries at runtime. In order to analyze and process range aggregate query M-AQ : A framework is proposed in this paper. Existing approaches were dealt only with adhoc queries and results yielded were not satisfactory. Here M-AQ is implemented on linux platform and performance is evaluated on very large park data records .M-AQ supports range queries and also runs multiple servers. When a range aggregate query arrives it is split based on the Balanced Partitioning algorithm and distributed across multiple shards (A shard is nothing but a master with one or more slaves). Queries here return specific fields of documents and also includes user defined JavaScript functions. JavaScript is used in queries, aggregate functions (such as MapReduce) and sent directly to the MongoDB to be executed. M-AQ has O(1) time complexity for the updates of data and time complexity for range aggregate queries where N happens to be the unique tuples, P happens to be the partition number B happens to be the bucket in each of the histogram. This M-AQ framework there by reduces the cost of both network communication and local file scanning and has better performance compared to hive.

KEYWORDS: Big Data, MapReduce, MongoDB, Multiple Servers, Range Aggregate Query.

I.INTRODUCTION

We are presently living in the computerized world. With the expansion in digitization the measure of organized and unstructured information being made and put away is blasting. The information is produced from different sources – exchanges, online networking, sensors, computerized pictures, recordings, sounds and snap streams for areas including human services, retail, vitality and utilities. It is expanding getting to be basic for associations to mine this information to stay focused. The volume, assortment, and speed of Big Data[1] causes execution issues when made, oversaw and broke down utilizing customary information handling methods. Colossal data examination is done to discover examples of various social viewpoints and slants of various individual practice schedules. This structures a stage to explore essential request concerning the brain boggling world. These included related attempts to gather a profitable hypothesis strategy, and examined the huge behavioral data sets related to account and gave back an advantage of even 326 percent higher than that of a discretionary endeavor procedure. Choi and Varian [2] acquainted gage representations with figure financial markers, for instance, social unemployment, vehicles bargain, and even destinations for individual voyaging. It is currently crucial and required to give capable systems and gadgets to huge data examination .So in handling this huge amounts of information a fundamental issue emerges which is to make a rundown which manages estimated question replying. Arbitrary Sampling is yet another technique which yields adaptable synopses and backings subset-whole inquiries and certainty limits. In any case, when Classic example based summaries [3] are concerned which are fundamentally intended for discretionary subset inquiries check the structure of the keys. So it can be comprehended here that the particular structure, for example, progressive system, request or item space makes range inquiries more pertinent for Big information investigation. Range total inquiries assume a vital part in OLAP (On-line logical Processing Systems) and GIS(Geographic data Systems) in condensing data. Range total questions are likewise

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

utilized as an essential instrument as a part of choice administration, online proposal, pattern estimation etc. So it is a testing errand to gauge and give exact results for reach total inquiries in huge information situations. Prior Prefix-whole block method [5] was utilized as a part of OLAP to build the execution of reach total inquiries alongside Online collection. Be that as it may, with these methodologies clients can't acquire an agreeable replying with surmised precision as ahead of schedule as first and foremost stages.

II.RELATED WORK

The reach total question issue has been contemplated by SharathKumar and Gupta[3] and Malensek[4] in computational geometry and Geographic Information Systems(GIS).The work is essentially centered around approximating range total inquiry for continuous information investigation in OLAP. Ho et al. was the first to present Prefix-Sum Cube way to deal with illuminating the numeric information 3D square collection [4] issues in OLAP. The crucial thought of PC is to pre-process prefix totals of cells in the information 3D shape, which then can be utilized to answer range-total questions at run-time. In any case, the redesigns to the prefix wholes are relative to the measure of the information 3D square. Liang et al. [6] proposed a dynamic information 3D shape for reach total questions to enhance the overhaul cost. The prefix total methodologies are appropriate for the information which is static or once in a while overhauled. For enormous information situations, new information sets arrive constantly, and the up and coming data is the thing that the experts need. The PC and other heuristic pre-processing methodologies are not pertinent in such applications. A vital estimated noting approach called Online Aggregation was proposed to speed range-total questions on bigger information sets [7]. OLA has been broadly concentrated on in social databases [8] and the present cloud and spilling frameworks [9], [10]. A few learns about OLA have additionally been led on Hadoop and MapReduce [10], [11], [12]. The OLA is a class of strategies to give early returns assessed certainty interims persistently. As more information is prepared, the appraisal is dynamically refined and the certainty interim is limited until the fulfilled exactness is acquired. Be that as it may, OLA cannot react with worthy exactness inside wanted day and age, which is essentially vital on the examination of pattern for specially appointed inquiries.

III.EXISTING SYSTEM

The FastRAQ framework is a surmised noting approach that outlines exact estimations rapidly for extent total inquiries in situations including enormous information. It first partitions the information lumps into discrete free parcels by utilizing an adjusted dividing calculation, and by which a neighborhood estimation for every segment is produced. So on the landing of the reach total inquiry, FastRAQ gets the outcome by abridging the nearby estimation of each individual issue. The estimation representation is a multi-dimensional histogram which is assembled by means of the scholarly information conveyance. In FastRAQ, the characteristic qualities can be both numeric and alphabetic. The Key thought is so produce a nearby inquiry result utilizing the adjusted parceling calculation with the stratified testing model. So in FastRAQ the numerical space of the total – segment is separated into various gatherings and an estimation representation of the gathering is gotten. At the point when another record arrives it sister sent to the parcel contingent upon the present information conveyances and the quantity of servers accessible. . It first partitions the information lumps into isolated free segments by utilizing an adjusted dividing calculation, and by which a neighborhood estimation for every allotment is produced. So on the landing of the reach total inquiry, FastRAQ gets the outcome by outlining the nearby estimation of each individual issue. Fragment family development for FastRAQ, which fuses three sorts of area families related to degree complete inquiries. They are collection segment family, list portion family, and default section family. The gathering portion family joins an aggregation fragment, the record fragment family fuses diverse rundown areas, and the default segment family fuses distinctive sections for further expansions. A SQL-like DDL and DML can be portrayed adequately from the diagram.

A. DISADVANTAGES

- Cost is produced by data transmission and synchronization for aggregate operations.
- Scanning og local files to search selected tuples.
- The updating process includes delivering the record each time to the specified partition.
- Range Cardinality tree produces additional overhead.
- Cost of transmitting the local result of a partition cannot be negligible.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

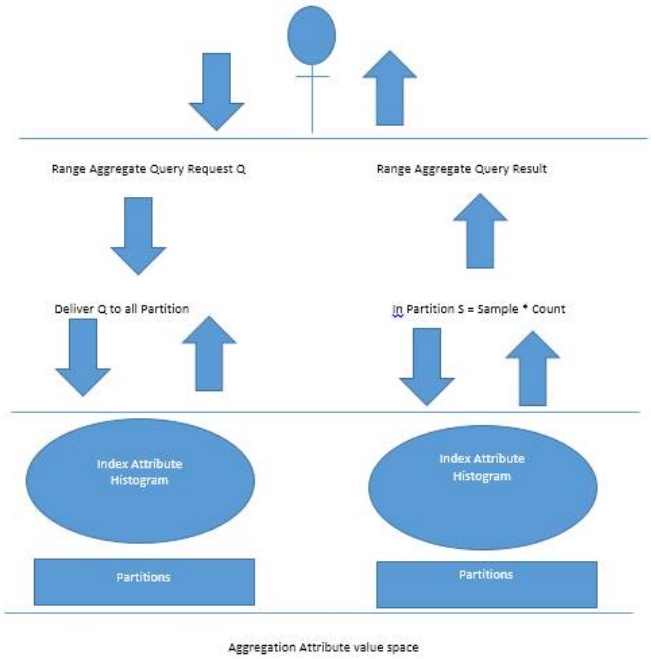

Fig 1: FastRAQ framework..

IV.PROPOSED SYSTEM

Here in the proposed part, I layout a balanced section count which works with stratified testing model. In each segment, an example for estimations of the amassing area and a multi-dimensional histogram is kept for estimations of the record portions. Exactly when an achieve complete inquiry request arrives, the adjacent result is the aftereffect of the example and a normal cardinality from the histogram. This declines the two sorts of cost at the same time. Prior FastRAQ gives a good starting stage to becoming ceaseless taking note of procedures for colossal data examination. Right when a request arrives, it is passed on into every allotment. The cardinality estimator (CE) is shaped for the addressed domain from the histogram in each package. By then we the examination regard is processed in each apportioning, which is the aftereffect of the case and the assessed cardinality from the estimator. Additionally I make utilization of the MongoDB here in the proposed segment. M-AQ joins inspecting, Histogram and information parcel ways to deal with create exact estimations including enormous information. It is intended for disseminated range total questions and it is appeared to accomplish better execution results on both inquiry and upgrade preparing in enormous information. M-AQ essentially accompanies MongoDB. It is a cross stage report arranged database. MongoDB bolsters field, range questions, standard expression looks. Inquiries can return particular fields of records furthermore incorporate client characterized Java Script capacities. MongoDB gives high accessibility imitation sets. A reproduction set comprises of two or more duplicates of the information. Every imitation set part may act in the part of essential or optional copy whenever. The essential reproduction plays out all composes and peruses as a matter of course. Optional reproductions keep up a duplicate of the information of the essential utilizing worked as a part of replication. At the point when an essential imitation falls flat, the copy set consequently leads a race procedure to figure out which optional ought to wind up the essential. Secondary can alternatively perform read operations, yet that information is in the end steady of course.

A. ADVANTAGES

- M-AQ can be used as a tool in DBaaS.
- It can be used to find solutions of m*n format problem. When there are m aggregation columns and n index columns of the same record.
- M-AQ achieves 26 times of performance improvement on count queries than Hive.
- Can apply on large data sets.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

- M-AQ achieve better performance improvement on range aggregate queries than Hive.
- It can search different index-columns of queried ranges.

The cost of merging due to union statements is negligible.

V. SYSTEM ARCHITECTURE OF PROPOSED SYSTEM

The information appropriations are measured by the bunching estimations of all record segments and they likewise make utilization of the scholarly learning to fabricate what is known as the histogram. So here the component vectors are separated from the educated information set which will create the vector set through which the last groups are shaped. The basic K-implies grouping technique is utilized to deliver the bunches. Every bunch is doled out with a novel ID.M-AQ underpins multi-dimensional extent questions which may incorporate various cans of the same histogram. It utilizes an exceptional ID for every record. The histogram is executed as a various leveled tree structure which is called as the extent cardinality tree (RC tree). A commonplace RC tree is delineated in fig 3. The RC-Tree [13] incorporates three sorts of hubs. They are root hubs, interior hubs and the leaf hubs. The root hub or the inside hub dependably indicates its kids hubs. A leaf hub relates to one basin in the histogram. The leaf hub just keeps the data and the tuples qualities are constantly put away in the pail documents. Containers are autonomous of each other, the RC-Tree structure and its development procedure is very like the B+ tree. To enhance the throughput of RC-Tree, a hash table [14] for recently approaching information is presented for incremental overhauling process. The hash table comprises of different hubs which are indistinguishable to the RC-Tree's leaves hubs. On the off chance that another record is coming, it first composes into the hash table, makes hub on the off chance that it doesn't exist, and after that adds the tuples values into a brief information document. At the point when the quantity of hubs in the hash table achieves an edge, the hash table flushes hubs into the RC-Tree, and adds the impermanent documents to the formal container information records. The incremental upgrading process [15] will enormously enhance the throughput of RC-Tree in huge information situations. The accompanying upgrading calculation clarifies the incremental redesigning process in RC-Tree.

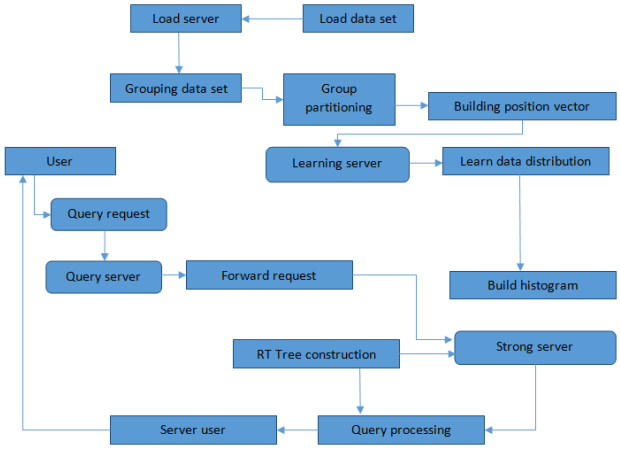

Fig.2. System Architecture of M-AQ

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

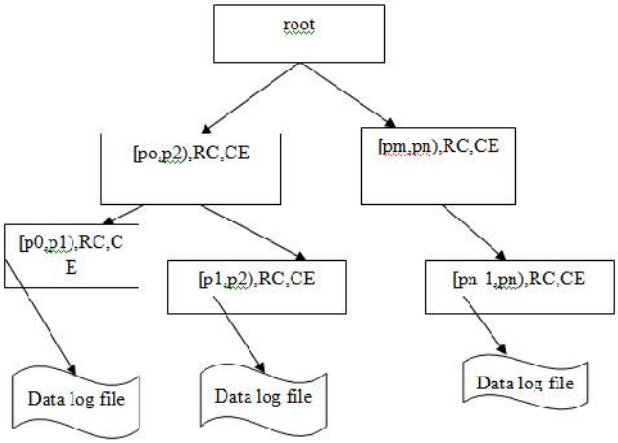

Fig.3. RC-Tree structure.

Fig.3. RC-Tree structure.

Algorithm 1:Grouping

Step 1*:* Parse value of index-columns into key-value pairs.
Step 2*:* Search in bucket spreads.
Step 3*:* Search in hash table and get the target node.
Step 4*:* End

To query cached data in hash table, the process is the same as Algorithm 2 to obtain cardinality estimator of the cached data, and then the result is merged to the estimator into CEmerge to compute the final cardinality estimation

Algorithm 2: Range Carnality Query algorithm.

Input:(Q,T,ho);
Q: Select distinct count;
T: the RC-Tree;
ho :the edge range cardinality ratio.
Output : R;
R:the range cardinality queried result.
Step 1: Locate the first node in RC-Tree by ColName;
Step 2:Scan the bucket data file;
Step 3:Merge into the cardinality estimator CEmerge;
Step 4: R<-h(CEmerge);
Step 5:return R.

VI.SYSTEM DESIGN

Apportioning is a procedure of doling out every record in an extensive table to a littler table in view of the estimation of a specific field in a record. It has been utilized as a part of server farm systems to enhance sensibility and accessibility of enormous information. The partitioning[13] step has turned into a key determinant in information examination to support the question handling performance[14]. The quantity of allotments ought to be kept under some edge in an appropriate framework. In huge information situations, an allotment is a unit for burden adjusting and nearby range-total inquiries. In every allotment, a dynamic specimen is figured from the current stacked records. Right now, M-AQ utilizes a mean estimation of accumulation segment as the example, which is Sample ¼ SUM=Counter, where SUM will be total of qualities from collection segment, and Counter is the quantity of records in the present segment. A nitty gritty adjusted allotment calculation is appeared in Algorithm 3.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Algorithm 3: Partitioning.
Input:(R,VP);
R: an input record;
VP: the partition vector set.
Output : PID;
PID: a partition identifier for partition p.
Step 1: Parse the input record R ;
Step 2: Compute the GID;
Step 3: Get the partition vector Vpi from VP with the GID,and let Vpi = < GID,Vr >;
Step 4: Set the target partition identifier;
Step 5: Build the sample in partition PID;
Step 6: return PI

To guarantee that information is adjusted on every server, the segment calculation isolates every gathering into various segments and sends to one server contingent upon the information dispersions. The info record R is sent to a parcel given by PID which is produced from its comparing collection section. M - AGE utilizes inexact noting methodologies, for example, inspecting, histogram, and cardinality estimation and so forth., to enhance the execution of extent total questions. We utilize relative mistake as a factual instrument for exactness investigation. Relative mistake is generally utilized as a part of a rough noting framework. Additionally, it is anything but difficult to figure the relative mistakes of consolidated appraisal variables in a circulated situation for M - AGE. In this segment, we examine the evaluated relative mistake and the certainty interim of conclusive reach total question result.

Theorem:
Proof: According to Algorithm 3,the range aggregate query result in each partition is expressed as follows :
$= Count * Sample$, (1)
Where *Count* is the estimated range cardinality obtained from the histogram, *Sample* is a sample of values of aggregation-column in the queried partition.

VI. RESULT AND ANALYSIS

The results that are shown are with FastRAQ in comparison with Hive. FastRAQ is better than Hive in terms of performance and it reduces the two types of cost significantly.

Here the results are shown with M-AQ and its performance with MongoDB. M-AQ acts as a tool to boost the performance in DBaaS. It can be shown that M-AQ is significantly better than FastRAQ.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

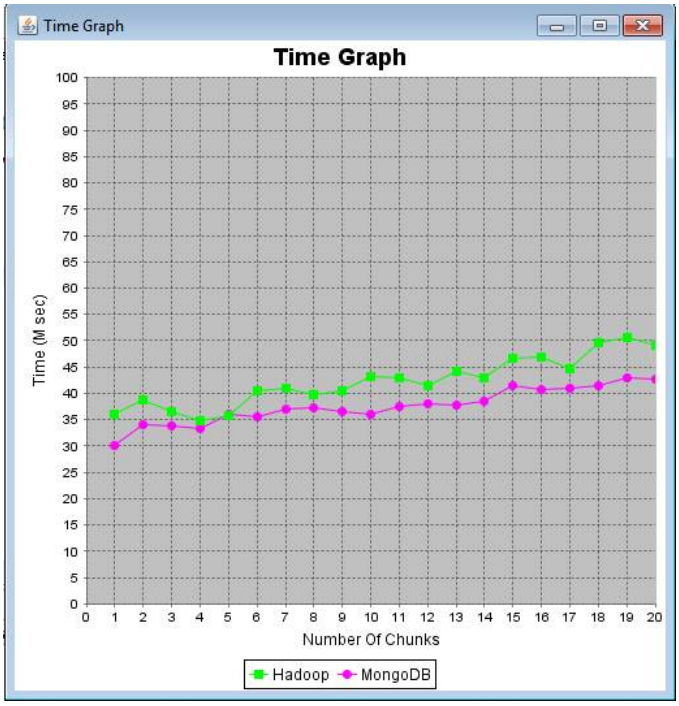

Fig 4.Time Comparison Graph

VI.CONCLUSION

The key thought was to diminish the two sorts of cost which were noted before in the current frameworks. M-AQ is another system conveyed to guarantee the same. It is for the most part to handle and give precise results to noting range total inquiries in enormous information climates. M-AQ settles the 1:n arrangement range total inquiries issue, i.e., there is one collection section and n list segments in a record.

- Real time answering method.
- Range aggregate queries in big data environments.
- Can be used as a tool in Dbaas.
- Can perform faster to access huge data.

Partitioning and Clustering mechanism improves the searching process.

REFERENCES

[1]. P. Mika and G. Tummarello, "Web semantics in the clouds," IEEE Intell. Syst., vol. 23, no. 5, pp. 82–87, Sep./Oct. 2008.
[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]. H. Choi and H. Varian, "Predicting the present with Google trends," Econ. Rec., vol. 88, no. s1, pp. 2–9, 2012.
[4]. C.-T. Ho, R. Agrawal, N. Megiddo, and R. Srikant,, "Range queries in OLAP data cubes," ACM SIGMOD Rec., vol. 26, no. 2, pp. 73–88, 1997.
[5]. 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,
[6]. 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.
[7]. J. M. Hellerstein, P. J. Haas, and H. J. Wang, "Online aggregation," ACM SIGMOD Rec., vol. 26, no. 2, 1997, pp. 171–182.
[8]. P. J. Haas and J. M. Hellerstein, "Ripple joins for online aggregation," in ACMSIGMOD Rec., vol. 28, no. 2, pp. 287–298, 1999.
[9]. E. Zeitler and T. Risch, "Massive scale-out of expensive continuous queries," Proc. VLDB Endowment, vol. 4, no. 11, pp. 1181–1188, 2011.
[10]. 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.
[11]. 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.
[12]. 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.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

[13]. 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.
[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]. 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.