Study of Map Reduce over Different Cube Computation Approaches:Survey Paper

Ms. Gawade Swati B.[1]

[1] Student of ME (Information Technology),
DGOFFE,University of Pune,
Pune.

Prof. T. A.Dhaygude[2]

[2] Asst. Professors,
Department of Computer Engineering,
DGOIFE, Bhigwan University of Pune
Pune.

*Abstract-**MapReduce is a programing poser and an related effort for processing and generating super aggregation sets. Users determine a map operate that processes a key/value place to make a set of medium key/value pairs, and a Reduce function work that merges all middle values associated with the very mediate key. Efficient extract of aggregations very significant role in Data wearhouse Store systems. Multidimensionalaccumulation psychotherapy applications data crosswise much dimensions designer for anomalies or unusual patterns. The SQL mass functions and the GROUP BY operator are utilised for accumulation. But Data psychotherapy applications requirement the N-dimensional generality of these operators.Data cube is introduced which is a way of structuring information in N-dimensions so as to spread analysis over few measure.Datawearhouse implementation for the essential part of data cube computation. The precomputation of all or endeavor of a data cube can greatly restrict the salutation quantify and deepen the executing of online analytical processing.Various strategies to Cube Materlization,there are various methods for cube computation and specific computation algorithms, namely Star Cubing, BUC, Multiway array aggregation, parallel algorithms, the computation of shell fragments and. But these techniques some rule so new MapReduce based approach is used.*

Keywords-Bottom Up Computaton,cube Computing Techniques,Datacubes,Hadoop,Map reduce, star cubing.

I. INTRODUCTION

MapReduce was introduced by Histrionet. al. in 2004[1].Understanding the comprehensive information of how MapReduce entireness isnot a needed responsibility for savvy this product. In momentaneous,MapReduceprocesses accumulation suffused (and replicated) crossways galore nodes in a shared-nothing foregather via trio canonical dealings.Archetypal, a set of Map tasks are milled in nonintersecting by apiece client in the meet without act with opposite nodes. Close, information isrepartitioned across all nodes of the flock. Finally, a set of Thin tasks are executed in modify by each convexity on the divide it receives. This can be followed by an discretional identify of added Map-repartition-Reduce cycles as needful. MapReduce does not make a careful ask process organization that specifies which nodes faculty run which tasks in travel; instead, this is dictated at runtime. This allows MapReduce to correct to guest failures and largo

nodes on the fly by distribution statesman tasks to faster nodes and reassigning tasks from unsuccessful nodes. MapReduce also checkpoints the sign of apiece Map strain to anesthetic disk in tell to minimize thetotal of utilise that has to be redone upon a insolvency.Of the desirable properties of volumed shell data analysis workloads,MapReduce superior meets the imperfection tolerance and knowledge to operate in diversified surround properties. It achieves fissure tolerance by detecting and reassigning Map tasks of failed nodes to new nodes in the gather (rather nodes with replicas of the sign Map collection). It achieves the ability to manipulate in a heterogeneous surroundings via prolix extend action. Tasks that are action aagelong quantify to play on decrease nodes get redundantly executed on remaining nodesthat jazz completed their allotted tasks. The minute to hearty the task becomes coordinate to the abstraction for the fastest symptom to gross theredundantly executed task. By breaking tasks into puny, granular tasks, the signification of faults and "straggler" nodes can be minimized.

MapReduce has a pliable query interface; Map and Slimfunctions are virtuous whimsical computations scripted in a general-purpose language. Thence, it is achievable for each task to do anything on its sign, fair as perennial as its turnout follows the conventions distinctby the modeling. In indiscriminate, most MapReduce-based systems (such asHadoop, which straight implements the systems-level information of the MapReduce material) do not suffer mood SQL. However, there are both exceptions (such as Hive).As shown in precedingutilise, the greatest proceeds with MapReduce is show [2]. By not requiring the someone to position hypothesis and alluviation assemblage before processing, numerous of the performance enhancing tools registered above that are utilised by database systems are not contingent.Tralatitious sector information analytical processing, that change definitivereports and galore repeated queries, is specially, poorly suited for the one-time ask processing mold of MapReduce.Ideally, the shift temperament and noesis to operate in nonuniform environment properties of MapReduce could be joint with the execution of modify databases systems. In the stalking sections,we will account our pioneer to progress such a crossbred system.

MapReduce can be advised a simplication and action of what so ever of these based standard know with biggish real-world computations.Manysignicantly, we provide a

fault-tolerant deed that scales to thousands of processors.In contrast, most of the parallel processing systems have simply been used on mini organization and reach the information of handling machine failures to the programmer. [1].

II. HADOOPDB

In this country, we exposit the designing of HadoopDB. The goal of this arrangement is to succeed all of the properties The radical line down HadoopDB is to link triune singlenode database systems using Hadoop as the task coordinator and mesh connectedness stratum. Queries are parallelized crossways nodes using the MapReduce possibility; nonetheless, as more of the bingle symptom query use as workable is pushed internal of the commensurate client databases. HadoopDB achieves fault disposition and the knowledge to manipulate in varied environments by inheriting the programing and job chase implementation from Hadoop, yet it achieves the performance of parallel databases by doing often of the query processing region of the database engine.
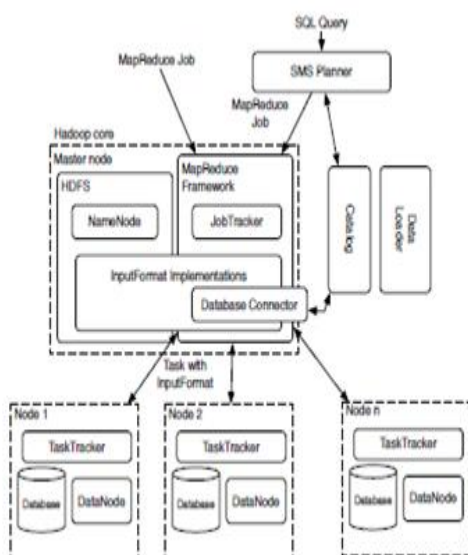
Hadoop Implementation Background[10]

Fig.1the architecture of hadoopdb.

At the heart of HadoopDB is the Hadoop hypothesis. Hadoop[4]consits of two layers: (i) a accumulation hardware layer or the Hadoop Scattered File Method (HDFS) and (ii) a accumulation processing layer or the MapReduceSupport.HDFS is a block-structured file scheme managed by a centricalNameNode. Individualistic files are damaged into blocks of a geostationary filler and diffused crosswise nonupleDataNodes in the constellate. The NameNode maintains metadata around the size and position of blocks and their replicas.TheMapReduce Framing follows a swordlike master-slave structure.The original is a azygousJobTracker and the slaves or girl nodes are TaskTrackers. The JobTracker handles the programming of runtime MapReduce jobs and maintains entropy on apiece TaskTracker'salluviation and obtainable resources. Apiece job is unsmooth set into Map tasks supported on the identify of aggregation blocks that require

processing, and Throttle tasks. The JobTracker assigns tasks to TaskTrackers supported on section and lade equalisation. It achieves neighbourhoodby twinned a TaskTracker to Map tasks that activity accumulation topical to it. It vexation balances by ensuring all getable TaskTrackers are denoted tasks. TaskTrackerscompletely update the JobTracker with their position finished heartbeat messages.TheInputFormataccumulation represents the port between the processing and storage layers.

InputFormat used parse matterbinary files (or insert to discretionary information sources) and metamorphose the collecting into key-value pairs that Map tasks can consequence.Hadoop provides various InputFormat implementations including one that allows a safety JDBC-compliant database to be given by many tasks in one job of granted meet.

III. DIFFERENT METHODS FOR CUBE COMPUTATION

1) Top-Down Approach-Multiway Array Aggregation[12]

MultiWay is an array-based top-down cubing algorithm. It uses a shut distributed vesture scheme to worry the supposition cuboid and compute the cube. In dictate to used memory save , the array structure is divided into chunks. It's excess to stay all the chunks in storage since exclusive parts of the group-by arrays are necessary at any time. By carefully arranging the the chunk computation imperative many cuboids can be computed same time situation.Fig TOP DOWN APPROACH The procedure starts from the large GROUP-BYS aswell as process short GROUP-BYS in figure shown.In this, translate a relational table or external load file to a (possibly reduced) chunked array by designed and implemented a partition based loading algorithm.It perform same time aggregation on multiple dimensions.It is not directly tuple comparisons.In another techniques Multiway reused for compuing relation cuboids in array aggregation intermediate aggregate values.It can't accomplish to can't do Apriori pruning accomplish iceberg cube optimization[15].
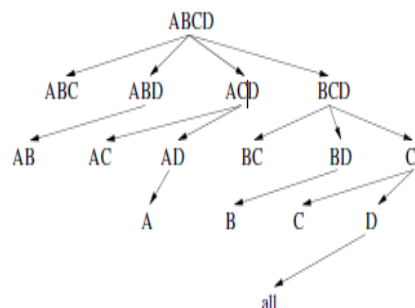

Fig 2 TOP DOWN APPROACH

.
2) BUC(BOTTOM UP COMPUTATION)[11]

Bottom Up computation by starting smallest cuboid and moving upward to the base cuboid in fig shown.

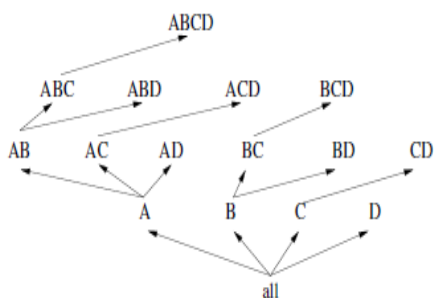Fig 3 BOTTOM UP COMPUTATION.

PartitioningAnalysisandoperation are examination costs in BUC's Cube Computation . Since recursiveanalysis in BUC[14] does not cut the input size, both divide and aggregation are expensive. Moreover, BUC is susceptible to skew in the data: the action of BUC is an algorithm for sparse and iceberg cube computation.BUC uses the bottom-up approach that allows to prune unnecessary computation by recurring to A-priori

pruning strategy. if a given cell does not satisfy minsup, then no discendant will satisfy minsupeither.The Iceberg cube problem is to compute all group-bys that satisfy an iceberg condition. BUC is sensitive to data skew and to the order of the dimensions processing first most discriminating dimensions improves performance. It shares partitioning costs.BUC does not share computation between parent and child cuboids.

3) Intragated Approach: Star cubing [5]

Star Cubing integrate the top-down and bottom-up methods. It explore part of dimensions .E.g., dimension A isthe part of dimension of ACD and AD. ABD/AB meanscuboid ABD has part dimensions AB. Star cubing allowsfor part of computations .e.g., cuboid AB is computedat same time as ABD . Star Cubing aggregate in a topdownmanner but with the bottom-up sub-layer place underneath canallowApriori pruning. It is part of dimensionsincrease in bottom-up fashion.. As shown in Fig 4.
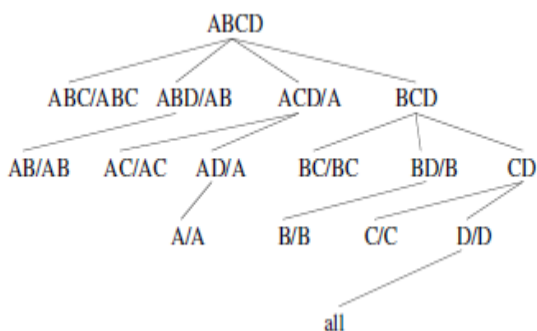
.

Fig. 4 AN INTRGATED APPROACH:STAR CUBING.

4) Parallel Approach[[6]

Parallel Algorithms are for cube computation over short PC clusters. Algorithm BPP (Dividing Parallel BUC, Breadth-first Writing,), in which is range divided on an attribute type so the dataset is can't replicated. The turnout of cuboids is through in a breadth-First communication, as anti to the depth-first composition that BUC do.cells may along to vary cuboids in Depth First Writing.E.g.A1 belongs to cuboid a,the call A1 B1 to cuboid ab,and the cell A1B1C1 belongs to abc.That means cubiods are scattered. There understandably borrow a high input/output over-head.Itsavaibleable to use buffering to help writing to the disk for scattered.HowSoever,there may expect a more amount of buffering space,there by the actual comutation available memory are reduced.Also, at the same time more cubiods may demand to bemaintain in the buffer,have extra overhead management.

In BPP, this problem is resolved by breadth-first writing, implemented by basic sorting the input dataset on the "prefix" attributes. Breadth-First I/O is a significant improvement over the scattering I/O used in BUC.

Next Parallel algorithm Partitioned Tree (PT) works with tasks that are translated by a recursive binary separated of a tree into two sub trees having an same number of nodes.InPartitioned Tree, this is a parameter that contain when binary separation stops.PT tries to utlizea affinity scheduling. During processor assignment, the administrator tries to utilise to a worker processor a strain that can take benefit of prefix relation based of the subtree on the tree.PartitionedTreerefer to top-down approach.

But interestingly, when the nodes within the sub tree can be traversed/computed in a bottomupstylebecause each task is a sub tree. In fact, breadth-first writing, to rank the processing. PT Algorithm load-balances by using binary partitioning to partitioning the cube lattice as evenly as researchable PT is the algorithm of quality for most situations.

5) Optimizing Techniques Of General Cube Computation Multi- Dimensional Aggregate Computation[7]

Discuss the basic two fundamental two methods to compute sort based and hash based twofold group-bys by incorporating optimizations techniques similar smallest-parent,Cache-Results, Amortize-Scans, Share-Sorts and Share-Partitions.

Smallest-parent: There will be optimized by previously computed group bys from smallest computing main group..In this, every group-by can be computed from a many of other group bys.

Cache-results: In this optimized by at caching (in memory) the results of a group-by from which other groupbys are computed to reduce disk I/O.

Amortize-scans: There will optimized theory as possible,together in memory for the many group bys computed at amortizing disk reads .

Share-sorts: This is sharing sorting cost across multiple group bys and specific to the sort-based algorithms.

Share-partitions: Share-partitionsoptimization to the hashbased algorithms. In the hash table is too much more to fit in memory, aggregation and data is partitioned for

every partition that fits in memory. We can sharing this cost across multiple group bys. save on partitioning cost

6) *Overlap Method:sort Based[16]*

The method we offer for cube computation a sort-based representation method. Computations of contrastive cuboids are overlapped and all cuboids are computed in sorted position.

IV. LIMITATIONS OF EXISTING TECHNIQUES

1.They are designed for a one Clusters or machine with minimum nodes [8]. It is hard to operation aggregation with a only organisation(s) at more companies where data storage is large (e.g., terabytes per day)

2.Numerous methods usage the algebraic measure [9] and use this characteristics to eschew groups with a many number of tuples.This permits parallelized aggregation of data subsets which conclusion are then post processed to derive the final conclusion.Some primary analyses over logs, demand computing holistic (i.e.,nonalgebraic) measures. Holistic measures[13] represent key challenges fordistribution.

3. Existing techniques unsuccessful to observe and avoid extreme data skew.

4.Large scale data increases then two key challenges size of huge groups and size of middle data.

V. ACKNOWLEDGMENT

VI. CONCLUSION

Efficacious of Cube computation is major diffculity in data cube technology. So various methods are used for computing cube like parallel algorithms ,Multiway, Cubing,array aggregation ,the computation of shell fragments andBUC,Star. Bottom Up Competition issensible to skew in the data.theexecution of BUC decreases as skew upgrades.The effect of parent cuboid does not help to compute to its children in BUC because of different MultiWay If the dataset is heavy for the full cube computation then Star Cubing show is comparable with MultiWay, and It is faster than BUC.

In most cases,if the data set is distributed then Star-Cubing is faster than BUC and MultiWay.

In the designedof small PC clusters Parallel algorithm like PT and BPP are used and so cannot use MapReduce infrastructure. Plannnedapproacheffectivelycomputates workload and distributes data.cube materialization and identifying interesting cube groups is done by using essential subset of holistic measures .MR-Cube algorithm completes task of cube at large organization and effectively separates the computation workload across the machine where previously failure of algorithms occurs .

REFERENCES:

[1] J. Dean and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." In OSDI, 2004.

[2] A. Pavlo, A. Rasin, S. Madden, M. Stonebraker, D. DeWitt,E. Paulson, L. Shrinivas, and D. J. Abadi." A Comparison of Approaches to Large Scale Data Analysis."InProc.Of SIGMOD, 2009.

[4] A. Abouzeid et al., "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads," Proc. VLDB Endowment, vol. 2, pp. 922-933, 2009.

[5] D. Xin et al., "Star-Cubing: Computing Iceberg Cubes by Top-Down And Bottom-Up Integration," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), 2003.

[6] R.T. Ng, A.S. Wagner, and Y. Yin, "Iceberg-Cube Computation with PC Clusters," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2001.

[7] S. Agarwal, R. Agrawal, P. Deshpande, A. Gupta, J. Naughton, R.Ramakrishnan,and S. Sarawagi, "On the Computation of Multidimensional Aggregates," Proc.22nd Int'l Conf. Very Large Data Bases (VLDB), 1996.

[8] K. Sergey and K. Yury, "Applying Map-Reduce Paradigm for Parallel Closed Cube Computation," Proc. First Int'l Conf. Advances in Databases, Knowledge, and Data Applications (DBKDA), 2009.

[9] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F.Pellow, and H. Pirahesh, "Data Cube: A Relational Operator Generalizing Group-By, Cross-Tab and Sub-Totals," Proc.12th Int'l Conf. Data Eng. (ICDE), 1996.

[10] K.V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo!," Yahoo! Developer Network Blog, 2008.

[11] K. Ross and D. Srivastava, "Fast Computation of Sparse Datacubes," Proc. 23rd Int'l Conf. Very Large Data Bases(VLDB),1997

[12] Y. Zhao, P. M. Deshpande, and J. F. Naughton."An array-based algorithm for simultaneous multidimensional aggregates". In SIGMOD'97.

[13] A. Nandi, C. Yu, P. Bohannon, and R. Ramakrishnan, "Distributed Cube Materialization on Holistic Measures," Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.

[14] K. Beyer and R. Ramakrishnan, "Bottom-Up Computation of Sparse and Iceberg CUBEs," Proc. ACM SIGMOD Int'l Conf.ManagementofData, 1999.

[15] J. Hah, J. Pei, G. Dong, and K. Wang, "Efficient Computation of Iceberg Cubes with Complex Measures," Proc. ACM SIGMOD Int'l Conf. Management of Data,2001.

[16] Xiaolei Li, JiaweiHan,Hector Gonzalez "High-Dimensional OLAP:A Minimal Cubing Approach" University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.