Implementation in Data Cube Mining for Map Reduce Paradigm

Ms.Gawade Swati B[#1]
[1]PG Student ME (Information Technology), Dattakala Group of Institutions Faculty of Engineering, Swami-chincholi(Bhigwan),Tal-Daund,District-Pune,India.
nazirkar33piyou@gmail.com

Prof.T.A.Dhaigude[#2]
[2]Ass.Prof.Dattakala Group of Institutions Faculty of Engineering
Swami-Chincholi(Bhigwan)
Tal-Daund, District-Pune,India
tanajidhaigude@gmail.com

Abstract-Computing measures for tweeter data cubes mining of cube group over data sets are impossible for many analyses in the tweeter.We have to compute the data set taken from tweeter user. You have to create a cube creation and then measure dimension setting using the roll up function.In the real world various challenges in the cube materlization and mining on web data sets. Map shuffle Reduce can be efficient extract cube and aggregate function on attribtes of tweeter.MR-Cube can be extract from efficient and effective PC cubes of holistic measures over large-tuple aggregation sets.In the existing techniques can not measure the holistic scale to the large tuples.

Keywords-*CubeMaterlization,MapReduce,Cube Mining,Holistic Measure ,Data Cube.*

****

I. INTRODUCTION

In the multidimensional data analyzing having the Data cube analysis is Powerful Tool. Parallel
Hadoop is an open-source version of the MapReduce framework,implemented by directly following the ideas described in the original MapReduce paper, and is used today by dozens of businesses to perform data analysis [1]. We deployed the system with several changes to the default configuration settings. We also allowed more buffer space for file read/write operations (132MB) and increased the sort buffer to 200MB with 100 concurrent streams for merging.Additionally, we modified the number of parallel transfers run by Reduce during the shuffle phase and the number of worker threads for each TaskTracker's http server to be 50.Moreover, we enabled task JVMs to be reused[1].For each benchmark trial, we stored all input and output data in HDFS with no replication add. After benchmarking a particular cluster size, we deleted the data directories on each node, reformatted and reloaded HDFS to ensure uniform data distribution across all nodes[1].
We present results of both hand-coded Hadoop and Hive-coded Hadoop (i.e. Hadoop plans generated automatically via Hive's SQL[1] interface). These separate results for Hadoop are displayed as split bars in the graphs. The bottom, colored segment of the bars represent the time taken by Hadoop when hand-coded and the rest of the bar indicates the additional overhead as a result of the automatic plan-generation by Hive, and operator function-call and dynamic data type resolution through Java's Reflection API for each tuple processed in Hive-coded jobs[1]. These adjustments follow the guidelines on high-performance Hadoop clusters [2].
Attributes refers to the set of attributes that the someone wants to analyze. Based on those attributes, a number Cube all possible grouping(s) of the attributes.We have to representing that attribute Where the dimension setting and roll up.
Given the hierarchical cube,the task of cube computation is to compute given measures for all valid cube groups, where a measure is computed by an aggregation function based on all the tuples within the group. MapReduce. MapReduce is a shared-nothing parallel data processing paradigm that is designed for analyzing large amounts of data on commodity hardware. Hadoop is an open-source implementation of this framework During the Map phase, the input data are distributed across the mapper machines, where each machine then processes a subset of the data in parallel and produces one or more key, value pairs for each data record. Next, during the Shuffle phase, those key, value pairs are repartitioned (and sorted within each partition) so that values corresponding to the same key are grouped together into values v1,v2, and other. Finally, during the Reduce phase, each reducer machine processes a subset of the key v1, v2 pairs in parallel and writes the final results to the distributed file system[3]. The map and reduce tasks are defined by the user while the shuffle is accomplished by the system. Fault toleranceis inherent to a MapReduce system, , which reschedule and reduce task or detects failed map the tasks to other nodes in the cluster. [3]

II. LITERATURE SURVEY

HadoopDB: an architectural hybrid of mapreduce and dbms technologies for analytical workloads.[1]
This Paper states following method to extract the trending topics from the system level hybrid over Map reduce and among Map reduce better than all technic. MapReduce outmatch the fracture tolerance and power to operate in diversified surroundings properties.Map reduce based system are very well suited due to scability, fault tolerance, and able to handle not defined data.In this system if any node will be failure then the fault tolerance will detect and reassigning Map to that node and also he ability to treat in a nonuniform surround via prolix extend process.Map Reduce performance is better than other system.Map Reduce combines all these properties for the parallel database systems[1]
The cm sketch and its applications.[4]

The CM Sketch to solve below problem improve using the time and space bound.Introduction of the sublinear space data structure from the Count –Min Sketch for summarize data streams. CM Sketch defines query in summarization of data stream defines range,point and inner product queries can be find the approximately answer quickly.In that it can be varius important problems can be solve in data stream such as finding frequent items and quantifies. Exact from $1/€^2$ to 1/€ in factor. Our CM sketch is not effective when one wants to compute the norms of data stream inputs. These have applications to computing correlations between data streams and tracking the number of distinct elements in streams, both of which are of great interest.It is an open problem to design extremely simple, practical sketches such as our CM Sketch for estimating such correlations and more complex data stream applications.

Mapreduce: simplified data processing on large clusters.[5]

Mapreduce is implementation for process and generate high data sets and it is also programming model.user define a Map function have to process a value/key twice to extract a set of middle value/key tice and a reduce function that join all middle values related with the same middle key. In this model more real world tasks are expressible.In this paper programs are extracted on a high cluster of commodity PCs and this functional style are automatically written parallelized. The run-time system takes care of managing the required inter-machine communication, handling machine failures, a set of machines extracted from schedule of the program and the details of partitioning the input data.In this the allow to does not any experience with distributed and parallel to easily usage the resources of a high distributed system[11]

MapReduce runs on a large cluster of commodity machines and is highly scalable:a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on twitter clusters every day[5].

Data Cube: A Relational Operator Generalizing Group-By, Cross-Tab and Sub-Totals [6]

Aggregation analysis applications aggregate information crosswise many dimensions search for unusual patterns.The SQL mass functions and the Grouping BY opportunist create zero-dimensional or one-dimensional answers.Applications requisite the N-dimensional thought of these operators. This production defines that cause, titled the collection cube or but solid. The cut opportunist generalizes the histogram, cross-tabulation, roll-up, drill-down,and sub-total constructs plant in most estimation writers. The cut treats each of the N accumulation attributes as a dimension of N-space. The commix of a peculiar set of judge values is a quantity in this place. The set of points forms an N-dimensional solid. Super-aggregates are computed by aggregating the N-cube to displace dimensional spaces. Aggregation points are represented by an "innumerable consider", ALL, so the tangency (ALL, ALL,...,ALL, sum(*)) represents the globose sum of all items.. Each ALL value actually represents the set of values contributing to that aggregation[6].

In this paper to solve the problem arise for group by query for SQL aggregating data various problem created by the SQL Group by operator can't solve the Histogram, Roll-up total and subtotal for drill down & cross tabulations.The data cube by using Group by query solve the problem for N-dimentional aggregates. It supports Histogram, Roll-up total and subtotal for drill down & cross tabulations[6].

III. DETAILS OF DISSERTATION WORK

Platform

Java Hadoop is the platform of my project

Installation to be done

1. Install the JDK 1.7
2. Install the ubuntu
3. Install VMVear
4. Windows XP, Windows 7

DESIGN PROCESS MODULES

1.Cube creation,Dimension Seting and roll up.

2. Cube execution: Map –Shuffle –Reduce

3. Search Module:- Aggregate functions on above attributes of twitter.

4.Graph modules for results.

Module Description

Cube creation,Dimension Seting and roll up phase

In this module, Data set is taken from 'Twitter" User, tag, retweet and time are considered. the for above attributed buckets are created.You have to Add or Remove the cube by using two function add and remove.after using the dataset above values entered the cube is created then you have to save that created cube.

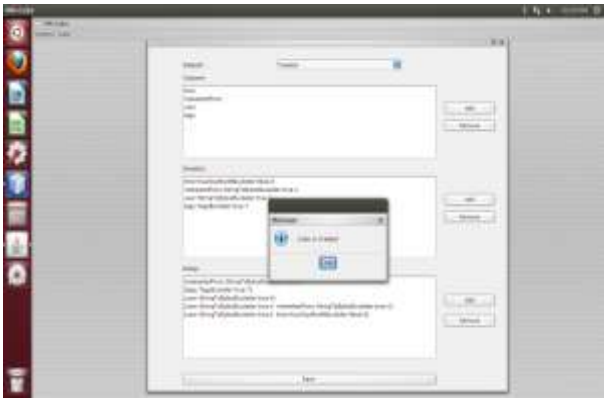

Figure.1 Cube is created.

Cube execution: Map –Shuffle –Reduce phase

 ISSN: 2321-8169
4194 – 4197

In this module,using the data seta no. of record will be created then the dimension and roll up will be executed to the cube mining.In the cube mining the mapreduce It can be implement Materlization from cubing task such as mapreduce algorithm[3] as input. By using the parent group label as the primary key and the group label as the secondary key, measures are clustered based on the parent group level,while ensuring sortedness on the group label[3]. This allows aone-pass discovery of the most interesting group for each parent group-dimension combination.using the Mining MapReduce MINING-MAP Algorithm[3].

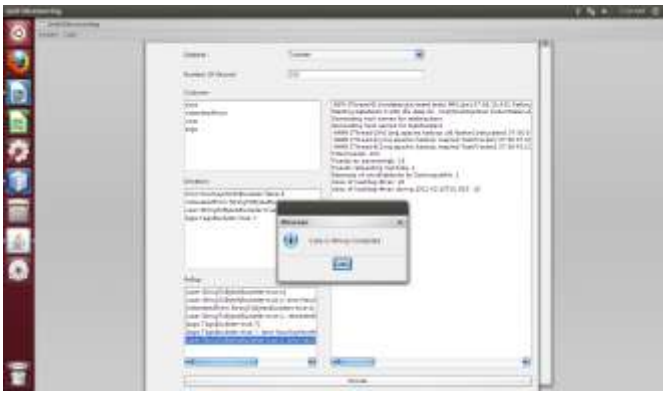

Figure.2.Cube is mining is completed.

Search Module:- Aggregate functions on above attributes of twitter.

In this module Cube analysis provides the users easy way to more from the data by computing aggregate measures[3].In that the twitter dataset are taken from the user can be search using the volume data and time.find out the duration from the twitter data set.you have to and ,or remove to the retweetfrom and user or the tag.you can also the clear dataset values using tha aggregate function. The twitter syatem can be used to compute aggregate from hierarchical measure .it can not directly used to holistic measure.in the mature database uses the Mapreduce and MapReduce Model.

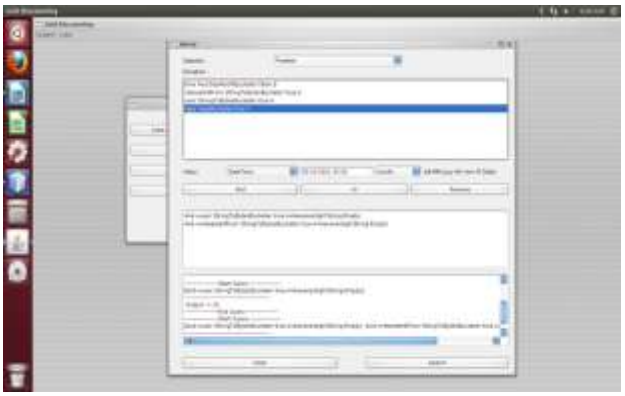

Figure.3 Search to the datasets.

Graph modules for results

Map Reduce is implemented four types of graph completed We think on four principal parameters that touch the show of the algorithm Cube mining time with vary data set size defined by twitter,Twitter data set parallelism, Hierarchies (sign and depth of the dimension hierarchies, which relate the solid lattice situation). And Mapreduce Depth hierarchies.During the map state, the grouping automatically partitions the input data into about 1M tuples per mapper. As a prove, the product of mappers is the equal for all algorithms with the self input information and we do not examine its upshot. We emphasize that, for MR-Cube, we instrument the enumerate instant including the Sampling process[3].

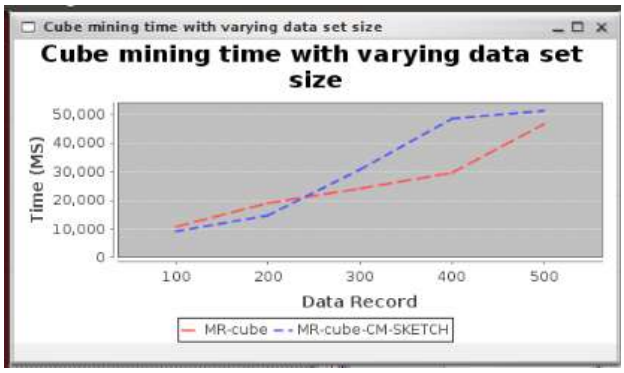

Figure.4 Cube mining time with vary in dataset size.

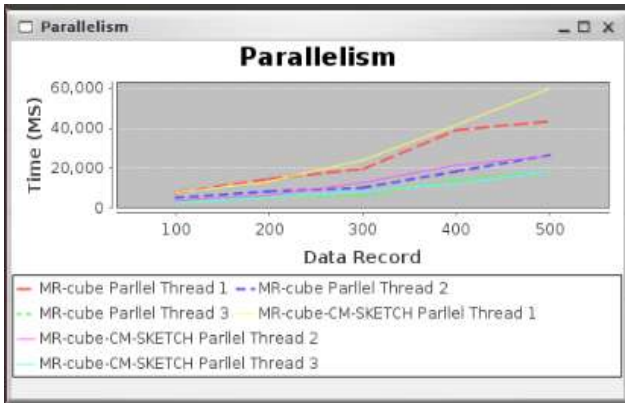

Figure.5 Parallelism in twitter dataset.

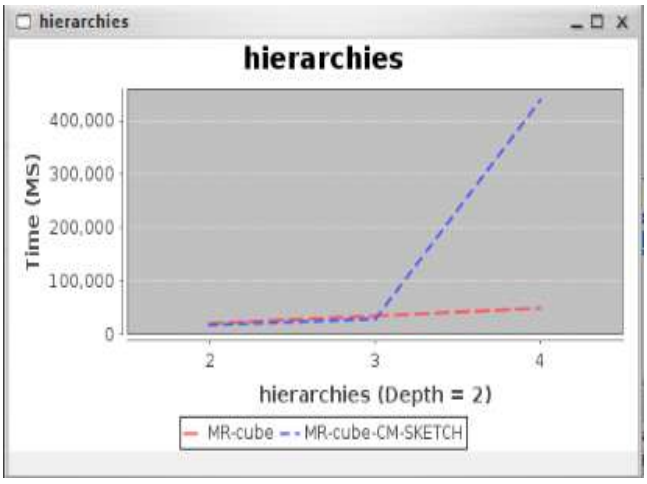

Figure 6.Hierarchies.

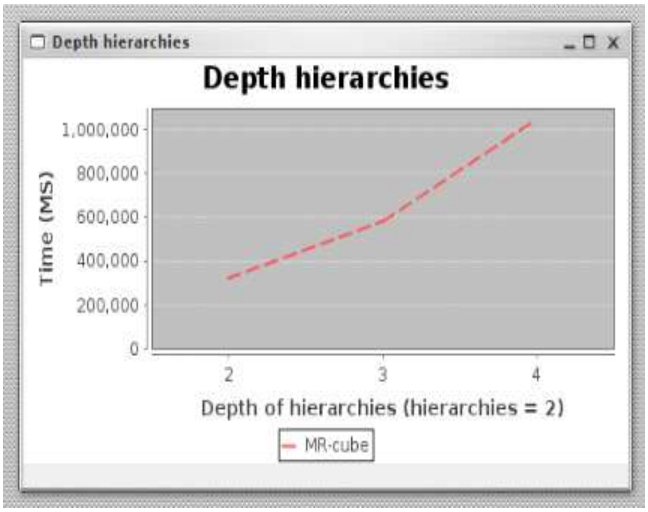

Figure.7 Map Reduce Depth hierarchies.

IV. CONCLUSION

In this paper we study the map reduce various algorithm from the twitter dataset.cube can created then cube can be executed by using the Map reduce shuffles the cube mining is completed.we also study the how to aggregate function to serch the result from dataset. Cube materlization algorithm can be used and the Map – reduce aggregate Algorithm used.finally it occur the Expeimatal resuls Impact data sets ,Parallism,Hierarchies and Depth Hierarchies of the data show that our MR-Cube algorithm distributes the computation load with the pcs and to complete cubing tasks at a scale is able then algorithms fail.

ACKNOWLEDGEMENT

I want to thank all people who help me in different way. Especially I am thankful to my guide “Prof.T A Dhaigude” for him continuous support and guidance in my work. Also, I would like thank our HOD “Prof. S S Bere”, and PG – Coordinator “Amrit Priyadarshi”.Lastly, I thank to “iPGCON-15” who have given opportunity to present my paper.

REFERENCES

[1] 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.

[2] Hadoop Project. Hadoop Cluster Setup. Web age.hadoop.apache.org/core/docs/current/cluster setup.html .

[3] Arnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan “Data cube materialization and mining over mapreduce “ IEEE transaction on Knowledge and Data Engineering, vol. 24, no. 10, Oct 2012.

[4] G. Cormode and S. Muthukrishnan, “The cm sketch and its applications,” J. Algorithms, vol. 55, pp. 58-75, 2005.

[5] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Proc. Sixth Conf. Symp. Operating Systems Design and Implementation (OSDI), 2004.

[6] 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.

[7] 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.

[8] E. Friedman, P. Pawlowski, and J. Cieslewicz, “SQL/mapreduce:a practical approach to self-describing and parallelizable user-defined functions,” Proc. VLDB Endowment, vol. 2, pp. 1402-1413,2009.

[9] V. Harinarayan, A. Rajaraman, and J.D. Ullman, “Implementing data cubes efficiently,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 1996.

[10] Y. Xie and D.O. Hallaron, “Locality in search engine queries and its implications for caching,” Proc. IEEE INFOCOM, 2002.

[11] JefferyDean,SanjayGhemawat”Mapreduce:simplified data Processing clustering on large cluster”,OSDI,2004.