ISSN:0975-9646
Prasadkumar Kale et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2871-2875

Big Data Analysis using Partition Technique

Prasadkumar Kale[1] and Arti Mohanpurkar[2]

[1]*Department of Computer Engineering, Pune University, Maharashtra*
Dr. D Y School of Engineering and Technology, Lohegaon, Pune

[2]*Assistant Professor, Department of Computer Engineering, Pune University, Maharashtra*
Dr. D Y School of Engineering and Technology, Lohegaon, Pune

Abstract:-Analyzing a Big data is a challenging task because of its characteristics and presence of data in large amount. For large scale data analysis hadoop technology provides a key role. Aggregate queries executes on number of columns simultaneously and difficult for large amount of data .In this paper we are proposing FRAQ and Balanced Partition technique which gives better performance with help of PIG and generate a histogram for the respective partition. This Histogram is generating according to the specified column by user in interface. Histogram gives an effective result according to requested query and data set coming online.

I. INTRODUCTION:

Analysis of Big data is nothing but the breaking a large amount of data into smaller parts for better understanding. Big data is generated as every person in world is connecting with internet and access social, commercial, educational and Business sites for understanding. Each person is wants connected to his friends, colleagues and family with help of internet. New updates and learning the technologies those are over the internet for making life more beautiful and easy. Big data is generated through all these things. And another thing is business that is only way of earning money and here comes logic. The world is moving quickly and the expression gets to be genuine World turns into a Town. Each individual human needs to get to system for staying joined with the world. These clients may get into a great deal of information identified with Topographical regions, political issues, neural system, wellbeing data and numerous more.

There is an alternate thing identified with Big Data is social locales and the media. Social locales like Google for Gmail and most ideally for the web index, Facebook, whatsApp are hit consistently by billions of individuals as far and wide as possible. These locales enhance learning of human long range interpersonal communication, mathematicians, doctors and numerous more science field by trade of data in a little measure of time [2]. All these individuals seek important data in only a single click. Enormous information preparing is the fundamental errand.

In this transforming a few structures are Hive, pig, Jaql like innovations assume an imperative part depicted in [4][5] [6]. On the sixth Oct. 2014 Flip-kart declares an offer which is extremely shabby. Bringing about high disjoin preparing is an exceptionally low little measure of time. As per Flip-kart there are billions of appeals hit inside 30 min. For transforming huge measure of information and dissect that information different advances are being used as specified previously. The most basic test for Enormous Information applications is to investigate the expansive volumes of information and concentrate helpful data or learning for future activities. In numerous circumstances, the knowing extraction process which must be exceptionally efficient and near to continuous as putting away all watched information is almost unachievable. The exceptional information amounts require successful information investigation and forecast stage to attain to quick reaction and constant classification of such Huge Information.

The primary center is on how information is examined, recovered by and an efficient way. [3] Gave HACE hypothesis to classifying the information into individual trademark also examined the information mining difficulties. Presently a day's Guide Diminish outline work is utilized for master accessing on OLAP and OLTP frameworks, which are upgraded occasionally. Guide decrease method [13] has one greatest trademark, i.e. parallel execution. For the handling vast measure of information HADOOP [14] uses parallel handling procedures in which Guide Diminish method is generally utilized. This system is straightforward from whatever remains of the others. Bunch and Allotment calculations are utilized for transforming on the enormous information. These things are viably giving yields, however, not in fulfillment and their comprehension level gets to be more mind boggling than oth-ers. Question mapping gets to be more confused with scientific databases. Mapping of questions of Big Data web sources [12], presents an explanatory meta - dialect for comprehending the importance of questions and guide them into individual. The greater part of query improvement calculations [7] [8] are utilized charts to examine and work effectively. The pattern matching calculation is a piece of chart investigation. Appropriated and live information can deal with this calculation. The primary significance of pattern matching calculation is finding the examples that are identified with the friend or approaching information. Most time the DAG is utilized for inquiry improvement. DAG is steered non-cyclic chart which does not have any cycle means better way a tree, so finding information won't end in Stop manner. The example matching calculation is for the most part known to distinguish the assaults and keep the assault, yet here we are utilizing it for finding the related inquiry. Big Data is a massive technology where a still lot of research has to be taken and most of developers are interested in this technology.

II. LITERACHURE SERVEY

Xiaochun Yun [1] proposed FastRAQ- Big Data query execution in a reach total questions approach. An adjusted segment calculation is utilized first to gap enormous information into free parcels, at that point nearby estimation portrayal created for each one part. FastRAQ gave come about by condensing nearby estimation from all parts. The Linux platform is useful for actualizing FastRAQ and execution assessed on billions of information records. As per the creators, FastRAQ can give great beginning stages of ongoing enormous information. It explains the 1: n group range aggregate question issue, however m:n designed issue still outside there.

For Big Data investigation, i.e. abnormal state dataflow framework an extensible and dialect autonomous system m2r2 is depicted in Vasiliki Kalavri [9]. This model execution is done on the Pig dataflow framework and the results took care of consequently in discovering, normal sub query matching revamping as well as junk gathering. Assessment is carried out utilizing the TPC-H benchmark for pig and report lessening in query execution time by 65%.

Characterization of point pattern matching is carried out by the nearby descriptor called Line Diagram unearthly setting. This work is carried out by Jun Tang [10] and his partners by doing an examination of unearthly systems and intending to present a strong for positional jitter and anomaly. Multiview unearthly implanted method is utilized for discovering the like-nesses between descriptor by contrasting their low dimensional implanting.

Feng Li[11] proposed a Map-Reduce System for supporting ongoing OLAP framework. According to need and different query requirements, processing needs, access patterns and other terms. Exaction of data is done an OLAP system instantly which necessary for data analysis in that system. The open source appropriated key/value framework, they called it as Hbase and streamed Mapreduce as Hstreaming for incremen-tal redesigning. They proposed an R-store for Map-Reduce backing on Ongoing OLAP. They assess their execution comes about on the premise of TPC-H information.

III. BIG DATA AND PARTITION TECHNIQUE

A. Big Data

Big Data contains heterogeneous data, self-sufficient source and complex and evolving connections. Heterogeneous mean information that is not in the same organization in light of the fact that every enterprise, institutional and seller has an alternate arrangement as the copyright and different issues. Autonomous sources may produce the information according to the occasions are happening in the framework, for illustration, numbering the occupation and finishing different undertakings in businesses. The errand may contain examination of projects, testing of the projects. People are meeting up in view of their likenesses with one another. These likenesses may contain distractions, natural connections, what's more shared understanding of one another. For extensive development of information, the information examination uses progressed logical systems like prescient examination, information mining, statically investigation, complex SQL, information virtualization and artificial intelligence.

The Big data additionally may characterize according to following:-

1) Volume: Information created in huge scale by machine and human collaboration than a traditional information. Case in point, Information created in call focuses, which is regarding call recording, labeling of queries, request, complaints and so forth.
2) Velocity: Online networking information streams create an extensive linux of conclusions and connections significant to client relationship administration. That is similar to messages, photographs on twitter alternately Facebook and so on.
3) Variety: Conventional database utilization organized in-formation, i.e. information composition and change gradually. In inverse of that non conventional databases arrangement displays confounding rate of progress.
4) Complexity: Information administration in enormous information is exceptionally intricate assigned, when a lot of information which is unstructured originating from different sources. This must be connected, associated also correlated to handle the data.

Information investigation obliges a great deal of processing and complex time for results and comprehension. Huge information contains numerous systems for examination comprise of a considerable measure of processing which is carried out utilizing huge information calculations. For getting quick output Big data must be processed as quickly as possible. One approach to the intention Big data issue is group information into more elevated amount view where smaller gatherings ended up obvious. In Information mining likewise have numerous difficulties for the enormous information like Stage for calculations, Information semantics and application learning by offering, protection and space of information. HDFS and Map-Reduce is nearly related by conveying parallel methodology and joining the yield. HADOOP utilizes a Map-Reduce method for the examination of information. The benefits of utilizing Map-Reduce system are :

1) It will run a little measure of methods while information, breaking down,
2) All the while sorts out the relocate.

B. Pattern Matching Algorithm

In dissecting the information examples assumes a vital part in that each approaching information is analyzed. For a set of examples of a set of articles with a specific end goal to focus all conceivable matches system utilized is Rete Match Algorithm. It keeps up state data of articles which are matched and somewhat match until the article is displayed in the memory. There is an alternate example matching calculation likewise like definite example matching which utilization seeking of related examples in giving content. Knuth-Morris-Pratt is an alternate calculation which is likewise on scanning for examples utilizing Java procedures. RE example matching and graph calculations are on customary declarations and they give more than one result for a related example. The Brute

energy accurate example matching calculation uses hunt systems down to find the ex-demonstration information.
Applications of this calculation are for web crawlers, parsers, advanced libraries, screen scraper. Different calculations use DFA, punctuation and consistent statement for assessment of designs into the straight time ensures no reinforcement stream. The RE example matching calculation gives different events of examples in the content file. One of the most actively researched areas of computer science is Patten matches with numerous papers still being published. This type of category of matching problems includes following:-

1) Exact string matching:- for example, in a saying handling issue.
2) Approximate string matching:- for example, in penmanship identification and optical chaste reorganization.
3) Largest common substring
4) Information retrieval and querying
5) Spam filtering and plagiarism discovery
6) Signal processing

Execution of any matching algorithm is judged, as various string similarity measures have been produced that can come close to compare strings and focus a quantitative measure of the level of likeness generally communicated in extends. The Exact pattern matching algorithm is implemented on the basis of searching and matching contain with other data that is present at that time in the system. Patterns are also helpful for finding the impurity or intruders in the system.

IV. PROPOSED SYSTEM

A. Problem Statement

Efficient query handling on Big data using Balance partition algorithm and Exact pattern matching algorithm. This paper is performs partition on the data sets that are coming online or in real-time. The partition data is stored into a database according to cluster and applying index mechanism on respected data. The pattern matching technique used for matching query requirements with the fetched data. Related data is fetched and serve to the user.

B. System Architecture

For distribution of big data and processing of it in a parallel computing environment is done with the help of the most popular framework called Hadoop. Hadoop is open source software which uses Google's Map/Reduce framework and use Hadoop Distributed File System; this is an open source implementation of Google's File system for storing data. It allows decomposition of massive, distributed, data intensive and parallel applications into smaller jobs and executes them independently. But still parallel processing has issues like:

- Tedious task: equal chunk creation.
- Integration of intermediate results for final output

Big data's random partitioning leads to place correlated information (data) into different chunks. For batch-oriented processing jobs are optimized with Hadoop MapReduce programming rather than real-time request and response type of jobs.

Even Hadoop is providing the best performance for big data on respective data set, it has uncontrolled chunks which are balanced in this project with the help of a balanced partition algorithm. Means control the uncontrolled chunks of Hadoop Framework using a balance partition algorithm explained later on. Also the Hadoop system is designed to perform only Fig. 1: System Architectures searching and sorting of particular data from a massive amount of data.
The client will fire query as indicated by his necessity of data on the database. This query is sent as a request message to sever. The data set is the main input to the system which is the number of records present in one file. This file is first processed with the help of partitioning algorithm which is output as chunks of data. This data are then processed for generating of histogram. This histogram is scattered per partition, so it has been easily available for processing the query.

C. Partition Algorithm

Even though Hadoop is most popular for performing best for distributed computing, its simple partitioning method does not preserve correlation between data chunks. So needed partitioning Framework for FARQ in which partitioning is help for balancing data chunks into respective partitions. These partitions hold data for increasing processing speed. According to large data record field partitioning algorithm is separating and analyzing that particular record. Also, it is assigned a record from large data tables to small data tables.
For busting query performance partitioning algorithm play an important role. The sampling of data is necessary for the analysis on the big data as the data is present in huge amount. Sampling has various methods one most famous method is stratified sampling in which sampling independent groups and select only one sample for improvement and reducing errors. The project is implementing the partition algorithm on the idea of stratified sampling for maximum error value relatively. Firstly, divide space values into different groups and subdivide groups into different portions according to server space available for particular partition.
Partition algorithm is expressed for data set D s as:-

Partition (Ds) = (G, p) = (V_{id} ,, random [1, V range])
Where p- number of a partition in group G, random function is a random number in [1; V range] , and V_{id} is a Group Identifier (GID) for the group G.
Stratified sampling method is subdividing the space into independent intervals with batch of logarithmic function and each interval stand for a group. After fixing an arbitrary number N can map into unique group G. For calculation for length of group model we can take input as numerical space. For initial condition GID is equal to < 0; 0; 0 > then length of group is [0; 1]. For GID is equal to < x; 0; 0 > then length of group is [2x ; 2x + 1]. For GID is equal to < x; y ; 0 > then length of group is [2x +y ; 2x +y + 1]. For GID is equal to < x; y ; z > then length of group is [2x +y +z ; 2x +y +z + 1]. \

Algorithm 1: Balanced Partition Algorithm

Input: Record(R), Vector Set Vs
Output: Partition identifier PID
1 Record has to parse into different column families.
2 Compute Group Identifier (G_{id}) with value ranges as stated above. Get partition vector Vpi from Vs with Gid and set
3 Vpi =< Gid ; V range > Set target for Partition identifier,
4 P I D =< Gid; random [1; V pi * Vrange] > ; Build sample in partitioning P I D ;
5 counter P I D counter P I D + 1;
6 sum PID sum PID + N;
7 sample PID sum x;y ;z ;range =counter PID ;
8 RID Hash (PID; counter PID);
9 Send Record to partition PID; return PID;

Partitioning is the unit of Big data utilizes load balancing and local aggregate queries. We use the vectors set V s = (V pi :< V id ; V r ang e > j1 i M) to build partitions of incoming all records, where M - group number. From current loaded record, dynamic sample is calculated for each partition. Currently, we uses mean value of aggregation that generates samples, given as Sampl e = S U M =counter , where S U M - sum of value from aggregation-column, and counter - number of records in current partition. P I D sent to partition is generated by input record R . Value of aggregate-column is used to generate P I D s.

D. Aggregation

The aggregation query is nothing but the aggregate functions used in the query like SQL, oracle, MySql and Sybase. There is Online Aggregate (OLA) that is used for improving the interactive behavior of database. For effective operations on database, batch mode is performing a key role. The traditional way is that user query and waits till database come to an end of processing entire query. on contradict to it is OLA, the user gets estimated results side by side as query is fired. In 1997, Hellerstein proposed the OLA for group-by aggregation queries for just one table. Later on many research are taken place and more and more results are come from various types of experiments on this. Most common approach of OLA is random sampling.
OLA has a potential to process very large-scale, data-oriented computing. For processing a large amount of information is a challenging task for OLA so the map-reduce technique used with this. The stored data is in the blocks and OLA has not control of all blocks. In such situation, associated aggregate values with that block placed in random fashion. So that its easy to maintain and process a large amount of data in limited time.

E. **Querying Data (Qd)**

The user enters a query according to requirements. The query is come into users query processing block where the Querying Data algorithm is invoked. This is help for mapping queried data with histogram and go to respective partition for fetching respective data.

Algorithm 2: FARQ

Input: Q d ;
Q d : Select sum (AggCol) other ColNam where
r i 1 < C ol N am < r i 2 opr r j 1 < C ol N am < r j 2
Output: R A ;
R A : result set of aggregate query.
1 Request Q d must be delivered to all partitions.
2 for each par t i in partitions do
3 The estimate cardinality of range
r i 1 < C ol N am i < r i 2 from the histogram and let compute C E i be i th dimension estimator.
4 estimate cardinality of range r j 1 < $ColNam_j$ < r j 2 from histogram and let C E j be j th dimension estimator.
5 combine estimators C E i and C E j by the logical operator opr , and estimate combined cardinality C E combine .
6 C ount i h (C E combine) h - function of cardinality estimator.
7 Compute samples for Ag g C ol and let sampl e i be sample,
8 S U M i count i ——————————————sampl e i : //S U M i - result of local range-aggregate query.
9 end
10 set approx answer of aggregate Query R A .
11 Let R A P M
i =1 S U M i ,
12 where M- number of partitions. Return R A

V. MATHEMATICAL MODEL

Let 0 S 0 be the system implementating FARQ(Fast Result for Aggregate Query).
Thus FARQ Framework can be represented as
S = f::::: g
Let 0 I 0 be set of inputs I = fI 1 ; I 2 g
I 1 - Data set(D s)
i:e:I 1 = fD s 1 ; D s 2 ; D s 3 ::: g
I 2 - User Query(U q)
i:e:I 2 = fU q 1 ; U q 2 ; U q 3 ::: g
Thus the system can be represented as
S = fI ::: g
Let 0 F 0 be the set of functions,
F = fF par tition ; F histog r am ; F Qpr ocess g
F par tition : Function for data set partitioning.
F histog r am : Function for histogram generation.
F Qpr ocess : Function for user query processing.
Thus the system can be represented as
S = fI ; F ::: g
i F par tition
Input: Data set (I 1)
Output: Chunk generation (I 3)
Let 'O 1 ' be the output of F par tition ,
O 1 = fP 1 ; P 2 ; P 3 ::: g
Criteria:
Chunk generation is depending upon the total number of slave nodes and records within a range value for aggregate column. Also set of partition provides input to F histog r am .
Let 'I 3 ' be set of chunk generated,
I 3 = fP 1 ; P 2 ; P 3 ::: g

I = fI 1 ; I 2 g[I 3 = fI 1 ; I 2 ; I 3 g
ii F histog r am
Input: I 3
Output: Histogram per partition.
Let 'O 2 ' be the output of F histog r am ,
O 2 = fH P 1 ; H P 2 ; H P 3 ::: g
Criteria:
Value of Data entity in each column.
Let 'I 4 ' be the set of histogram,
I 4 = fH P 1 ; H P 2 ; H P 3 ::: g
I = fI 1 ; I 2 ; I 3 g[I 4 = fI 1 ; I 2 ; I 3 ; I 4 g
iii F Qpr ocess
Input: I 2 and I 4 .
Output: result of queried data(R).
On arrival of query 0 Q 0 each node returns a local result.
Let 'O 3 ' be the output of queries data,
O 3 = fR 1 ; R 2 ; R 3 ::: g
Let 0 O 0 be the set of output sets,
O = fO 1 ; O 2 ; O 3 g
Thus FARQ framework can be represented as,
S = fI ; F ; O g

VI. DATA SET

For our project implementation, we are using Wikipedia data set which is freely available from them. This data set contains information about users that are searching for the data and also bog information. The page view static file contains data such as log of actions generated on the fly, or articles from Wikipedia or from project that were sent via up to filter the tosses request from an internal host. The filter is generating information like project name, page size of the request, and title of the page requested. The project is using Wikipedia data sets that are of pagecount Dec. 2014. There are various files containing data up to 110MB and each contains at least 1 Lac of records. From these we are using some files in which are recorded above 1.5 Lac and processing on that we are installing Linux on that system and then HADOOP. The link for the data set contains in [15].

VII. EXPECTED RESULT

This project is implementing the partition algorithm on PIG technology with help of HADOOP platform and all programming is done with the help of the Java language. After that project need implementation of algorithms for processing of aggregate queries and applying an algorithm on them. The project has named node i.e. master node and data nodes (slave nodes).

In this project first of all the partition of the data set is carried out which is uncontrolled as the Hadoop system does not control data. So by use of the balanced partition algorithm, data is controlled and chunks are created in the first stage of output. Then for mapping the data and utilizing the time histogram is generated. This is the second output of the project, which gives a benefit while matching contain with the user query. Main purpose of the project is handling data efficiently for the aggregate functions which are fired on one or more column on the big data.

VIII. CONCLUSION

Big data is nothing but unstructured, uncertain, real-time data that is present in a massive amount. Querying on such data is a bit difficult task, though there are varying technologies present in today's world. In this paper, we propose Balance partition technique and exact pattern matching technique that is useful for handling queries. Balance partition technique first divides big data into partition and store in respective partition. This partition contains indexing, which are used with an exact pattern matching technique for effective handling of queries. Also, the project is implemented this on the top of PIG and HADOOP technologies which support the java language.