Acta Technica 62 (2017), No. 4A, 601–610

Distributed MOLAP technology of big data analysis

Shiyu Zhang[1], Dan Luo[1], Weinstein Ben[2]

Abstract. objective: To validate the application effect of distributed MOLAP with large data analysis. Methods: The influence of large data scale effect is inevitable during the process of data storage and analysis. In order to solve these problems, a multi-distributed file system is adopted to deal with, and a distributed MOLAP technique for large data analysis is analyzed by establishing a data model. Conclusion: Distributed MOLAP techniques for large data analysis are more efficient.

Key words. Big data,distributed file system, MOLAP technology.

1. Introduction

With the development of modern information technology, big data times has quietly come, big data makes fields such as internet technology, electronic computer technology, science technology obtain wide development space, but under the background of universal development in every field, which may cause information data bursting and affect processing ability of information data in every field. Therefore, MOLAP technology has been universally applied; it can realize quick storage and processing on data information on the basis of data structure with multi-dimension.

Multi-dimension data base(MDD) stores data into one array of n-dimension, it has plenty of sparse data, in the part of event occurrence, data is integrated together and has big dense, which is regarded as dense data. Therefore, it needs to solve problems of sparse data and data integration.MDD separates sparse data from the dense data; it makes compression and storage on dense data so as to reduce occupancy of storage space. The summary data stored in MDD derives from the original detail data, so it needs one kin of mechanism to extract data from data source. The main advantages of MOLAP structure is that it can quickly respond to analysis request of decision analysis personnel and quickly feeds back analysis result

[1] Workshop 1 - 1.Tianfu College of Southwestern University of Finance and Economics; e-mail: 30691825@qq.com

[2] Workshop 2 - Computer Science, Kennesaw State University, Kennesaw; GA, United States

to users, this is profits from its unique multi-dimension data base structure and data(generally pre-processing degree is above 85% with high pre-processing degree.

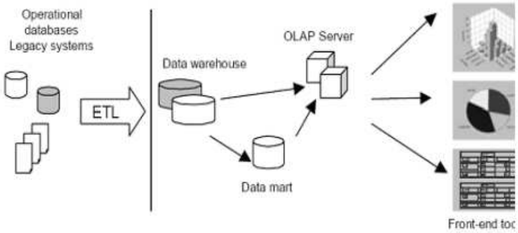

Fig. 1. MOLAP system structure

In MOLAP, it usually regards multi-dimension image as one super-cube in concept, and organizes it as one multi-dimension database in physics, the nature value of dimension is reflected into index bound of multi-dimension data, and stores the data of multi-dimension array unit, which is the core of organizing data by using multi-dimension array to store data, while multi-dimension data will form into cube structure in the multi-dimension data, drilling, rotating, section and switch are the most basic technologies.

2. Distributed MOLAP technology of big data analysis

The involved OLAP of distributed MOLAP technology of big data analysis is divided into 5 operations: scroller drill-down, cube, section and rotation. Of which, search of data range includes cube and section, data range and integration can be divided into scroller and drill-down, different images of creaming cube includes rotation. While one LLAF in distributed MOLF technology can be divided into 4 abstract tuple, of which it respectively includes Target representing data cube waiting for analysis, Flang C of data range waiting for analysis data in the cube, aggregation function Aggregation represents data analysis and Result represents data cube result. In the distributed MOLAF of big data analysis, in the process of involving with OLAF in technology will use 4 kinds of most basic estimation method, of which includes data search, block option, change of dimension grade and aggregation of information data. In the distributed MOLAF of big data analysis, when technology is making OLAF operation, it can data range Range represents waiting for analysis data in the cube set as the block coordinate of setting can meet and represent data range Range waiting for analysis data is (c1 c2 . . . cn)block scale is $(\lambda 1\lambda 2\ldots\lambda \mathrm{n})$, so the formula is as follows: $(\alpha i/\lambda i)$ $c1{*}(\beta i/\lambda i)$. The block can meet the above formula has no necessary to input all the data information block, while it just needs to make input in the process of technology making OLSF operation in the distributed MOLAF of big data analysis, which reduces search space of OLAF to great extent, it can complete it through simple coding. The detailed option algorithm is indicated by the following figure:

Make OLAP on distributed MOLAP of big data analysis, the data range waiting

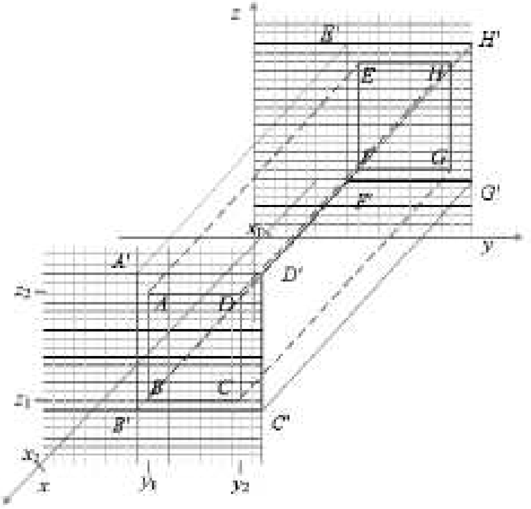

Fig. 2. Block option algorithm

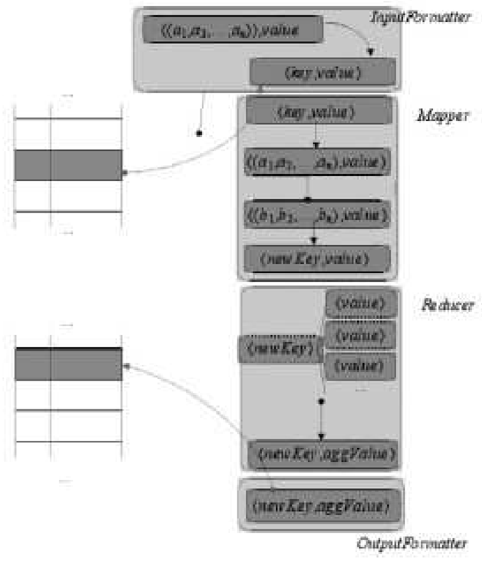

Fig. 3. Scroller operation procedure of OLAP

for analysis in the cube is represented in the operation process, which is set as ({x1?? y1?? z1); {x2?? y 2?? z2)), the coordinate in the corresponding diagram is F and D, form this it can induce that data information range needs search is between ABCD-EFGH. Make the whole cube as reference and it can find select most unrelated information to OLAP. Take the OKAP involved in the distributed MOLAP technology of big data analysis for example, its algorithm can be divided into 4 parts, which is input fonnatter, mapper, reduae and output fonnatter, its detailed scroller operation procedure of OLAP is indicated by the following diagram.

In MapReduc, before program procedure model is working, the scroller operation involved in OLAP in the distributed MOLAP of big data analysis, its 4 parts will be submitted, while system terminal will make effective verification on data information of these 4 parts, this is to prevent program invalid due to o result in search task. After this operation is completed, it obtains the required input list by block algorithm. MOLAP is to use Map Reduce as basis, it will make processing on electronic computer by many ways in the actual operation process, for example, it makes mapping on dimension and measurement of information data, this is implemented by multi-dimension model, such as operation on dimension level of information data, this operation undergoes traversal algorithm and operation of dimension coding. In short, MOLAP plays important effect in the whole information technology processing; it is also the key point of present analysis.

3. Data model

MOLAP technology adopts multi-dimension data model, its content has dimension and fact. The key point in the operation process is to seek mapping relations between dimension and fact. MOLAP needs one kind of integrated way to maintain mapping of dimension and fact, because one dimension can include many layers, every layer can include several grades; dimension and fact are relation of one to more, so dimension model has complication. In order to avoid additional storage and maintenance cost, DOLAP firstly makes simplification on dimension so as to adapt to distributed environment, meanwhile it reduces complication of OLAP algorithm.

(1)Dimension: in the multi-dimension model, after classification on all the data, dimension will place all the even data in the data structure without overlapping; meanwhile, it provides item selection method and organization method among every data. In this research, we firstly makes simplification on connotation of dimension and data model with multi dimensions, the model after simplification should obey restrictions of the following conditions, suppose a is dimension then a has dimension layer, but the number of dimension layer is 1. a is the assemble composed of n dimension grades, it is marked as { }suppose (i [1n]) is any dimension grade, then there is one dimension nature having one value. Now it regards a as the structure composed of dimension nature value of every grade, the node of the same grade includes the same sub-node number.

Figure 4 is the cube composed of 3 dimensions (x??y??z)??the smaller grid represents cell, the bigger grid represents block. In the actual operation, block may include some empty cells, which means this cell has no measurement. In the actual application, in order to reduce physical space of cube, if cell has any no measurement, this file dose not store record of this cell.

(2)Measurement: suppose measurement is one independent variable, and is regarded s MOLAP technology analysis object according to dimension value, of which the measurement with the finest particle refers to the dimension value with low grade in the dimension.

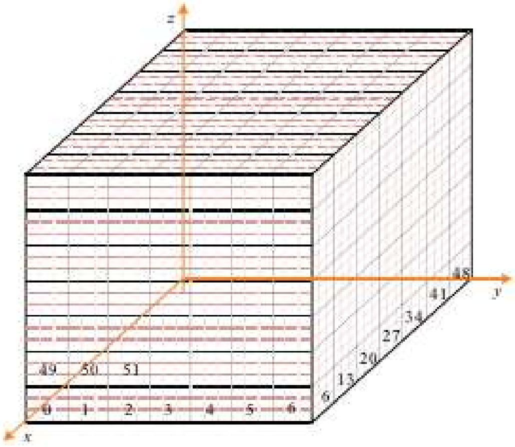

Fig. 4. Example of data cube

(3)Cell: in the whole logic view, cell is composed of different measurement, these measurements use the same dimension value as reference object, so cell can be regarded as the combination{k |k-a} of measurement.

(4)Data cube: according to the above-mentioned definitions, data cube is the multi-dimension structure of MOLAP, it is composed of many cells.

(5)Definition of block: block is the logic diagram of data cube, every data cube can take different values according to different dimensions.

4. Layered dimension integration and data storage of MOLAP

A.Layered dimension integration of AMOLAP

It adopts multi-dimension array to realize MOLAP based on multi-dimension database. In the design of category, it establishes base class of Dimension, it derives subclass of dimension. Under the standard structure dimension class, it is distinguished as the usual Star Dimension, Snow Dimension, because dimension includes layer relations, so it should establish one base class Dimension Status which can represent dimension information under the state of different layer and structure, when making processing on dimension with designated layer, it uses derived class of this base class, such as Standard Dimension. As for the detailed structure pattern, it can derive Star Hierachy, Star Level etc. In the aspect of database storage, it considers storage of layer data. This paper uses Hashtable to store hierarchical relationship. The members are respectively member and parentlndex.

```
**
* The corresponding relations of father node in level element
*
Private H ashtab le parents= new H ashtable( )
.......
```

```
**
Set number of father node of level element
@ param m em ber level element
@ param parentlndex number of father node1 indicates the highest level
*
```

This, it can set father node number of level element through parents, put0, when it needs to use information of level relations of dimension, it can obtain father node sequence of designated level element through parents, put0. Here it discusses realization of dimension integration with level. As for aggregation operation, it can establish corresponding base class of aggregation calculation; different aggregation calculation operation is derived from it. Dimension aggregation is divided into aggregation and level aggregation of dimension. As for the aggregation of the whole dimension, it can be regarded as calculation of old cube, stores the obtained result into new cube. As for algorithm of aggregation, it can consider referring to array of 1 dimension to simulate ergodic algorithm of multi-dimension array. Under the condition of unknowing multi-dimension array dimension, if it uses the usual way to indicate multi-dimension array. Generally speaking, we think that if it is 3 dimensions, it makes 3 circulations, every dimension one circulation, but it may be 4 dimensions and 18 dimensions, the multi-dimension array of one dimension change can not make traversal through designated corresponding dimension. Since this, we can also use 1 circulation to solve problem. The principle is from the lowest level, starts from 0 and adds 1 every time; it adopts way of carry to indicate changes of dimension. 0,1,10,11, the detailed code section is as follows:

As for the new cube, C ell New Array [x1][x21[xN] = Sum (I From lto aggrC ount) O ldA rray[x1][x2][.]-..[xN].The position of I is the sequence of dimension waiting for cluster.

```
w hile (idxN ew A rray[0] < new D im ensionLength[0]) {
it generates idxO ldA rray and the same part of idxN ew A rray
for(inti= 0 i< new D im ensionCount i++) {
if(i< dim ensionlndex)
idxO ldA rray[ i] = idxN ew A rray[ i]
else
idxO ldA rray[i+ 1] = idxN ew A rray[i]
) for (inti= 0?? i< aggrC ount i++) { cluster 1-dimension
idxO ldA rray[dim ensionlndex] = i
Sum = op eration A dd(sum cub e getM easures0 getE lem ent(idxO ldA rray))
) new Measures SetElem ent(idxN ew A rray sum )
Update id x New Array so as to make traversal
idxN ew Array[new D im ensionC ount-1]++
f o r (intI_ new D im ensionC ount 1 i> 0 i–) {
if(idxN ew A rray[ i]>= new D i m ensionLength[ i]) {
idxN ew A rray[ i] = 0
idxN ew A rray[ i- 1]++
}else {
break
```

}

If it considers the cluster problem of level, calculation of dimension aggregation becomes further complicated. We not only need the present level information, but also need the information of the last level. Traversal level aggregation needs the following steps :it traversal from element of the bottom level to element of the highest level.

```
while (colum n < table getC olum nC ount0 ) {
level sequence(the highest level is 0)
O bjectcell= table getC ell(row colum n)
if(1evels[ 1evelIndex]indexO f(cel1) 1) {
/ /element is not existed grade object of this level, so it is added to grade object of this level.
levels[1evelIndex] addMember(cel1)
)if(1evellndex > 0){// is not the highest level element, it needs to construct relations of father node
Objectparent= table getCell(row colum n + 1) father node element
intparentlndex = levels[ 1evelIndex. 1]. index Of( parent)// the sequence of father node element in this level grade
if ( parentlndex != 1) { father node element is existed in the corresponding level grade
levels[1evelIndex]setP arent(cell parentlndex)
break
)else {father node element is not existed in the corresponding level grade
levels[ 1evelIndex - 1]addM em ber( parent)
```

statusregister). MSCI is asynchronous static memory. It makes proper initialization for the memory access, and sets physical parameter such as access time-delay. PCR is PowerControlRegister which controls power switch of every component. Here program turns off all the powers of all the external equipment and plate set equipment such as CF card, USB, LCD etc. BCR is BoardControlRegister there are many settings in BCR??it mainly sets flash memory of system as readable. FLASHB0WP and FLASHBlWP respectively controls Flashbank0 and bank if here is 0??the corresponding bank can be used, if it is 1, the corresponding bank can be locked. BIPR is BoardInterruptPendingRegiste it will forbid using all the interrupt trigger source after zero clearing. It firstly makes zero clearing on the entire RAM by writing, and then makes initialization of some auxiliary parts, for example, start data Cache, command Cache, star MMU etc.

B. Data storage of MOLAP with level

Multi-dimension of array of DOLAP is obtained by calculation without memory, so storage cost of data cube is very small, DOLAP makes simplification on dimension, it can guarantee the code of the same level is the continuous decimal, meanwhile, every brother node has the same sub-node, so dimension information only needs to store scale of every dimension, which greatly reduces storage cost of dimension. Suppose dimension d is composed of m dimension levels which is marked as {li|i[1m]}then physical storage of d can be indicated as assemble composed of dimension level and ordered pair of dimension level scale. {li|li||i[1m]}of which li indicates the name of this dimension level. Realization system of DOLAP can use XML file to store information and stored in the main node of cluster.

In logic structure of cube and its cell or cube and its block can be analoged as data t\structure of multi-dimension array and its element. In physics, block is the memory cell of cube, makes linearization on cell in the block, block can be regarded as the independent file to make memory. In order to find address, block and cell need support linearization and reverse-linearization calculation??and this calculation is the same with linearization and reverse-linearization of this calculation. Suppose there is one array of n-dimension, its scale is marked as A1 A2. . . Ancoordinate of element X in multi-dimension is marked as (X1 X2. . . Xn) in multi-dimension space the coordinate after linearization is marked as index(X)??its linearization method is indicated by formula (1)reverse-linearization method is indicated by formula (2):

index(X)=(. . . ((Xn An1+Xn1)×An2+. . . +X3)×A2+X2)×A1+X1 (1)

temp1= index

X1=temp1%temp2=[temp1/A1]

X2=temp2%temp3=[temp2/A2] (2)

Xn=tempn% An

As for cell, the dimension of constructing cube is marked as d1,d2,...,dn. Suppose x is one cell in the cube and the corresponding dimension value of x in dimension di is vi,code(vi)=xi, then coordinate of x is (x1x2, . . . , xn), formula (1) repalces (X1, X2, . . . , Xn) as (x1x2. . . xn)A1 A2. . . An is replaced as |d1|, |d2|, . . . , |dn|, formula (2)is also the same linearization and reverse-linearization algorithm of cell??cell is stored as one piece of record in MapFile of Hadoop HDFS??this record includes the coordinate and the included measurement of this cell after linearization.

As for block suppsoe y is one block and it is the division of cube

$$|y| = \lambda 1 \lambda 2 \ldots \lambda n$$

coordinate of y is (y1y2. . . yn)coordinate of any one cell in y is (x1 x2 . . . xn)then yi=[xi /λi]in formula (1)repalce (X1X2 . . . Xn)as (y1y2. . . yn) A A2. . . An is replaced as

$$\lambda 1 \lambda 2 \ldots n$$

formula (2)is the same it can linearization and reverse-linearization algorithm of cell. In storage reality, one block is stored as one MapFile file in Hadoop HDFS the coordinate after linearization is regarded as the file name of block file, so that it can seek address of block file. Every record in the block file stores one cell in this block and block file is stored in the distributed file system Hadoop HDFS. In the file, it stores all the cell data by way of Key-Values , of which, Values is corresponds to reality data Key is the search value after linearization of cell coordinate according to formula (1). This search value is one decimal positive integer, in the big data environment, cube scale of data is very large, take Java language for example, and search value will exceed the maximum range of 264 of long. In addition, we adopt character string to indicate the number data. In addition, if we adopt Key-Values to store, storage expenditure of Key is larger than the corresponding storage expenditure of Values which wastes plenty of storage, therefore, as for one block file, it only record the minimum cell search value while storage of Key corresponds to the deviation amount of this value.

5. Algorithm of dimension coding

Algorithm of dimension coding includes 2 ways, which are binary coding algorithm and decimal coding algorithm. Binary coding algorithm has certain omission, decimal coding is relatively detailed, which can realize coding calculation of grade dimension value, but it can not directly realize mapping between coding and dimension value. In order to avoid the omission in the actual work, MOLAP technology adopts decimal coding algorithm, suppose t is one dimension grade in dimension a, then calculation is as follows:

```
Input: Dimension a : A target dimension
Function: Dimension Coding
FOR $f=1$ TO $|t(a)|$
FOR $g=0$ TO $|t(())|-1$
Dimension value of ( )
Code= g
END FOR
```

In the actual application, most data values are of value form of dimension, such as height and price etc. The dimension with data form can make different division according to difference of its co-domain. Steps with different division can confirm different dimension grade.

References

[1] Song. Jie,Guo. Chaopeng,Wang. Zhi,Zhang. Yichan: *Distributed MOLAP Technology of Big Data Analysis.* Software Journal *04* (2014),731–752.

[2] Jiang. Waiwen,Xiong. Dongping,Zhang. Xiaoxia: *MOLAP Storage, Search and Technology Research Based on Multi-database.*Computer Engineering and Application *24* (2005), 166–168.

[3] Wu. B, Shen. H: *Exploiting Efficient Densest Subgraph Discovering Methods for Big-Data.* IEEE Transactions on Big Data(2017).

[4] B. Wu,H. Shen,K. Chen: *Exploiting Active Sub-areas for Multi-copy Routing in VDTNs.*roc. of the 24th International Conference on Computer Communications and Networks (ICCCN), August 3-August 6(2015).

[5] B. Wu,H. Shen: *Shen, Discovering the Densest Subgraph in MapReduce for Assortative Big Natural Graphs.*Proc. of the 24th International Conference on Computer Communications and Networks (ICCCN) Workshop on Big Data and e-Health (BDeHS), August 3-August 6(2015).

[6] Jiang. Waiwen,Xiong. Dongping,Zhang. Xiaoxia: *MOLAP Storage and Search Technology Research Based on Multi-database.* Computer and Digital Engineering *02* (2005) 56–59.

[7] Jiang. Bo: *MOLAP Technology of Bog Data Analysis.*The Communication World *24* (2015),331–332.

[8] Zhang. Yu,Zhang. Yansong,Chen. Hong,Wang. Shan: *One Kind of Mixed OLAP Search and Processing Model of Adapting to GPU* . Software Journal*05* (2016),1246–1265

Received November 16, 2016