Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 9, September 2014, pg.69 – 81

RESEARCH ARTICLE

OPTIMIZING BIG DATA

[1]Anshul Sharma, [2]Preeti Gulia

[1]M.Tech Scholar, CSA Department, Maharshi Dayanand University, Rohtak

[2]Assistant Professor, CSA Department, Maharshi Dayanand University, Rohtak

anshu.sharma215@yahoo.com; preetigulai81@gmail.com

Abstract - When sales representatives and customers negotiate, it must be confirmed that the final deals will render a high enough profit for the selling company. Large companies have different methods of doing this, one of which is to run sales simulations. Such simulation systems often need to perform complex calculations over large amounts of data, which in turn requires efficient models and algorithms. This paper intends to evaluate whether it is possible to optimize and extend an existing sales system called PCT, which is currently suffering from unacceptably high running times in its simulation process. This is done through analysis of the current implementation, followed by optimization of its models and development of efficient algorithms. The performance of these optimized and extended models is compared to the existing one in order to evaluate their improvement.

The conclusion of this paper is that the simulation process in PCT can indeed be optimized and extended. The optimized models serve as a proof of concept, which shows that results identical to the original system's can be calculated within < 1% of the original running time for the largest customers.

Keywords: - PCT, optimized, algorithms, simulations

I. INTRODUCTION

Big data is a slightly abstract phrase which describes the relation between data size and data processing speed in a system. A comprehensible definition of the concept is “data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time." [1]. This means that a scenario where innovative

optimization of both models and algorithms is required to handle large amounts of data might well be classified as a big data problem. Big data is data of the order of Petabytes, Exabytes, Zetabytes, Yottabyte , Xenottabyte, Shilentnobyte , Domegemegrottebyte (1033 bytes) and the list will go on. Handling this huge volume of information has been the most challenging task for researchers today. Since data generated is so voluminous, a lot of information can be derived from this data.

Big data is characterized by the following features.

- **Volume-** The huge amount of information generated every second has become the biggest challenge which researchers have to handle. Social networking sites are generating huge amount of information. For example Head of BI and Analytics Competency analyzed how big data can be used to deliver maximum impact for society. They have recorded that the first episode of the Indian T.V show Satyameva Jayathe generated 40000 tweets in the first 90 minutes of the show. These tweets were later used to improve the telecast according to user views and Variety. Since the information generated is from various sources, each source will have its own format of data. The data may be as images, text, video, audio etc. Again in images it may be Jpeg, bmp, Text may be .txt, .csv, .doc, .html etc. Another challenge faced today is how to handle these different varieties of information.

- **Standard**: This is another issue which comes up when dealing with big data use. Formats used by different users may be different. One user on the web may tweet that "World Wide Web is a huge repository of information "and another may tweet "WWW is a huge repository of information". In this context both WWW and World Wide Web refer to the same context. When making a prediction both these tweets should be considered as one but this is possible only if WWW and World Wide Web are treated the same. Similarly one doctor may write his prescription as a "heart attack" whereas another may write it as "cardiovascular Infraction". Both mean the same in medical terms. But understanding this is the issue. This assumption is another challenging task to researchers.

- **Value:** Whenever data is preprocessed a major challenge is how to deal with missing values and null values. When the data is as huge as big data, this becomes more complicated. If the null values or missing values are not processed then it will lead to incomplete analysis or inaccurate analysis. Identifying which is the relevant data is another tedious task. This depends on the perspective of the

user. Whenever clinicians read through a set of patient records. They found that the types of information which were relevant for each one was different. Hence the need of the hour is to develop a model that considers the preferences of users to develop a model in determining whether the data will be relevant or irrelevant.

- **Authenticity:** When we talk about huge volumes of data being generated, it is difficult or almost impossible to keep track of the generator of this information. In such cases, the big question is whether the data so generated is valid and authentic or has it been tampered with.

- **Velocity:** The speed at which the data is generated is referred to as the velocity of data. The speed at which social media generates data is unimaginable. A lot of analysis is being done based on this data generated.

II. RELATED WORK

The term big in big data was a result of information explosion, the milestones in the history of sizing data volumes and a few reasons to this information explosion are cited below [4]. Fremont Rider estimated that American University Libraries were doubling in size. With the rate at which the data is growing, Yale has predicted that in 2040 we will have approximately 200,000,000 volumes which will occupy 6,000 miles of shelves. Derek Price suggested that the number of journals were growing exponentially rather than linearly, doubling every fifteen years and increasing by a factor of 10 during every half century. B.A.Marron and P.A.D de Maine published "Automatic data compression" stating that information explosion noted in recent years makes it essential that storage requirements for all information be kept to a minimum.[4]. These are just a few to state but there are many more instances cited for the term big in big data.

The voluminous information available on the World Wide Web can be of use to many people like doctors, industrialists, economists, sociologists, common man and anyone only if information can be retrieved out of this raw data. Mc Kinsey report stated that big data is that data which our traditional RDBMS cannot handle. A major challenge is to handle this huge information.

In existing PCT, the big data challenge arises from the huge amounts of data needed in order to run simulations for large customers. In some cases more than fifty thousand historical order rows may have to be handled, with multiple possible conditions and discount rates applied to every single one of them. While the data set itself is not extremely large by today's standards, the complex operations and calculations which have to be performed on each one of them adds new dimensions to the simulation procedure. Discounts are for example inherited through a large tree structure containing tens of thousands of nodes and the results must be presented to the user within a reasonable amount of time. The reasonable time limit has been defined as ten seconds for the simulation procedure in PCT. This value is based on research [2, 3] showing that a system user who has to wait even further for results of complex calculations will lose focus - something which could prove devastating during a negotiation with a customer.

An ideal simulation procedure would always return the results within just a few seconds, since this would mean that simulations could take place during normal conversation without requiring any waiting at all.

III. PAPER OUTLINE

This paper is divided into four parts - Simulation, Method, Results and Discussion.

The Simulation part begins with a detailed description of how discount rate simulations work and the problems which the current implementation has introduced. The second part contains a specification of the scaling simulation functionality and an explanation of the technical difficulties which are introduced by this extension. The Method part describes the models and algorithms which have been developed. It also contains a theoretical analysis of these and comparisons between the current implementation in PCT and our solution. In the Results, the performance of PCT as well as of our solutions for both the optimized customer discount model and the scaling extension are presented. This is split up into a set of test cases, with motivations of their relevance for actual usage scenarios. The first goal of this paper is to optimize the existing discount simulation algorithm in order to reduce its running time. The discount simulation's purpose is to apply given discounts to articles and article categories, in order to evaluate whether they will generate an acceptable profit for the selected customer. The second goal is to create a model with associated algorithms for a scaling extension of the system's simulation functionality. The purpose of this extension is to make it possible to apply different discount rates depending on the volume of individual orders. This will encourage customers to place a few large orders every year instead of several small ones, thus decreasing shipping and warehouse charges for the company without reducing the sales volumes.

IV. SIMULATION

When a sales representative negotiates with a customer, one can think of it as a sort of balancing problem. The sales representative wishes to maximize the profit gained by keeping discounts at a minimum, while the customer wants to minimize his or her costs by maximizing the discounts. This is where the simulation process comes in handy -by simulating the effects of new discounts, it is possible to decide whether they are profitable enough or not. When both the sales representative and the customer are satisfied with the results, they can save the discounts as conditions in the system's database. Discount rates from such conditions will then be applied to the customer's future orders.

Customer Discount Simulation

Customer discount simulations are currently fully implemented in PCT. By running a simulation over the data described in section 5(i), a sales representative will find out which profit would be gained if the customer bought the same articles as in the historical period but using current pricing conditions. Even more importantly, new discount rates can be applied to the simulation meaning that the sales representative can see which effects they will give and whether they seem profitable enough or not. The details of the simulation process are described first in section 5.(ii), Understanding of the underlying concepts is a great advantage when trying to gain insight into the workings of the simulation process.

i) DATA NEEDED FOR A CUSTOMER DISCOUNT SIMULATION

A simulation is based on data from the following sources:

a) Article tree - A tree structure where branch nodes represent article categories and leaf nodes represent articles
b) Sales history - A set of aggregated order rows, containing information about previous sales history
c) Existing customer conditions - Agreed discount rates from existing contracts, which set a certain discount rate to a specific node in the article tree
d) User input - Various parameters that specify which historical data and discount rates to use in the simulation.

ii) THE SIMULATION PROCESS

The sales representative starts by entering which customer he is negotiating with and selecting a path in the article tree for which discounts will be entered. Next up, a start and stop month is specified and now the system is ready to run the first simulation. Since no discount rates have been entered at this point, all nodes in the path will use their existing discount rates if any such exist in the active conditions and 0:0% otherwise. All price level 1 nodes which are not affected by the existing conditions will also have their discounts set to 0:0%. Due to the concept of discount inheritance, all other nodes will inherit their parent's discount rate top-down if they do not have an existing condition. This means that the results of the first run will always show the economical results that will follow if the same item quantities are sold as in the historical data used for the simulation, taking only currently active conditions into account.

Conditions may have been added or removed since the historical orders were handled, so it is not enough to just aggregate the values and profits from the history database. Instead, the “base value" (which one can think of as the price for the order rows if no discounts had been applied) must be calculated for each article. By applying discount rates from existing conditions to these base values, the system finds out how much the customer would have to pay for the same orders if they had been placed using current conditions. In the next step, the sales representative sets discounts for the nodes in the selected path and runs another simulation over the same data. Any conditions affecting discount rates for the path nodes will be overrun by the discount rates set by the sales representative, while conditions affecting other nodes will still be taken into consideration. The user specified discount rates will then be inherited down through the article tree just like the ones from the conditions. The result will thereby correspond to the profit which would be achieved if these new rates were added to the conditions database and the same orders as in the historical data were then placed again by the customer. This simulation step will typically be run multiple times with different discount rates for the nodes in the path, until they are balanced in such a way that both the customer and the sales representative are satisfied with the results. Running multiple simulations with different discount rates for the same time period and historical data until one gets satisfying results is referred to as going through a simulation process.

SIMULATION OUTPUT

So far, the output of simulations has been described in terms of "profit" and "value". The actual values computed during a simulation are of course more specific than that and as such, the specification of requirements presents guidelines for the output data layout.

The specification indicates that the output should be presented as a table, where each node in the selected path is represented as a row. There is also a top row labeled "Total", which shows the total simulation values of all articles in the whole article tree. A print screen showing how this looks in the current version of PCT is shown in figure 4.1.

Customer Total	Volume (kg)	Value (EURO)	CO	CO%	Actual Discount	Agreed Discount
Total	128,167	471,233	365,257	77.5	86.4	

Price Level 1	Volume (kg)	Value	CO	CO%	Actual Discount	Agreed Discount	Avg. Agreed	Target	Avg. Target
PL1_10	114,635	446,862	355,066	79.5	87.1	44.0 * 0.0	0.0	58.5	66.2

Price Level 2	Volume (kg)	Value	CO	CO%	Actual Discount	Agreed Discount	Avg. Agreed	Target	Avg. Target
PL2_01	10,349	24,014	15,670	65.3	87.6	67.1 * 0.0	0.0	66.9	66.9

Price Level 3	Volume (kg)	Value	CO	CO%	Actual Discount	Agreed Discount	Avg. Agreed	Target	Avg. Target
PL3_5751F1	26	187	161	86.0	78.1	12.4 * 0.0	0.0	54.7	54.7

Figure 4.1: A print screen from PCT showing how simulation output is presented in the current system

Discount inheritance

Discounts can be applied to nodes on any level of the article tree - from price level 1 down to specific articles. It is intuitive that a discount which is set for a single article will only affect the price of that specific article. When it comes to discounts set on article groups or price level nodes, the system uses a concept called "discount inheritance" to let this affect underlying nodes. In order to determine which discount rate to apply to a given node, the method presented in algorithm 4.1.1 is used.

Algorithm 4.1.1: find Discount Rate (Node n)

Input: A node n from the article tree

Result: The discount rate which should be applied to n

```
1  if n is a node in the path for 1 which a discount rate d is set then
2  return d
3  else if n is not a node in the path AND n has an active condition c then
4  return the discount rate from condition c
5  else if n is a price level 1 node then
6  return 0:0%
7  else
8  parent := n's parent node in the article tree
9  return find Discount Rate (parent)
10 end
```

The concept of discount inheritance is easy to visualize due to the tree structure of the article database. An example tree with some existing discount rates is shown in figure 4.2. Existing discount rates are written directly onto the grey nodes to which they belong, while nodes without such rates are white. The final result of the discount rate inheritance in the same tree can be seen in figure 4.3, where arrows show how discount rates are passed down through the tree.

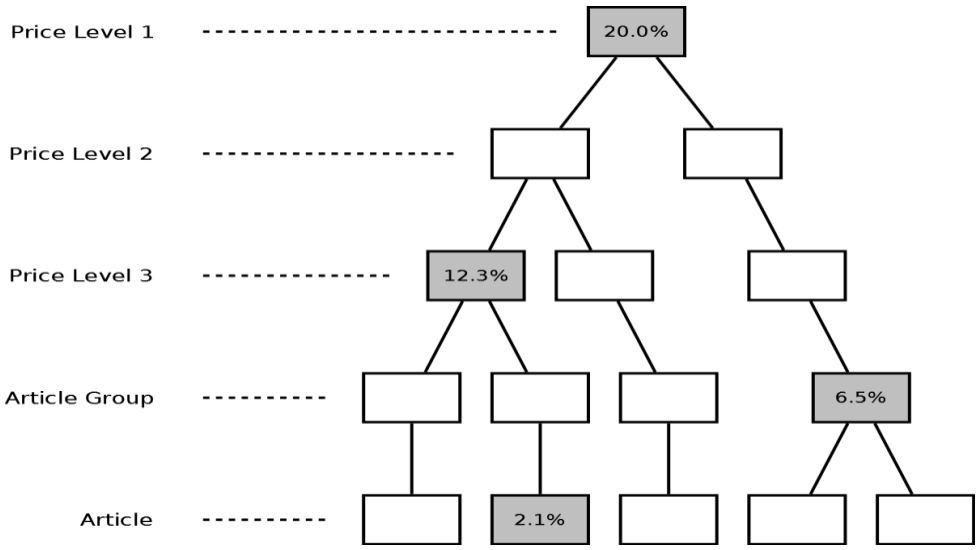

Figure 4.2: An example article tree where discount rates have been set for four nodes

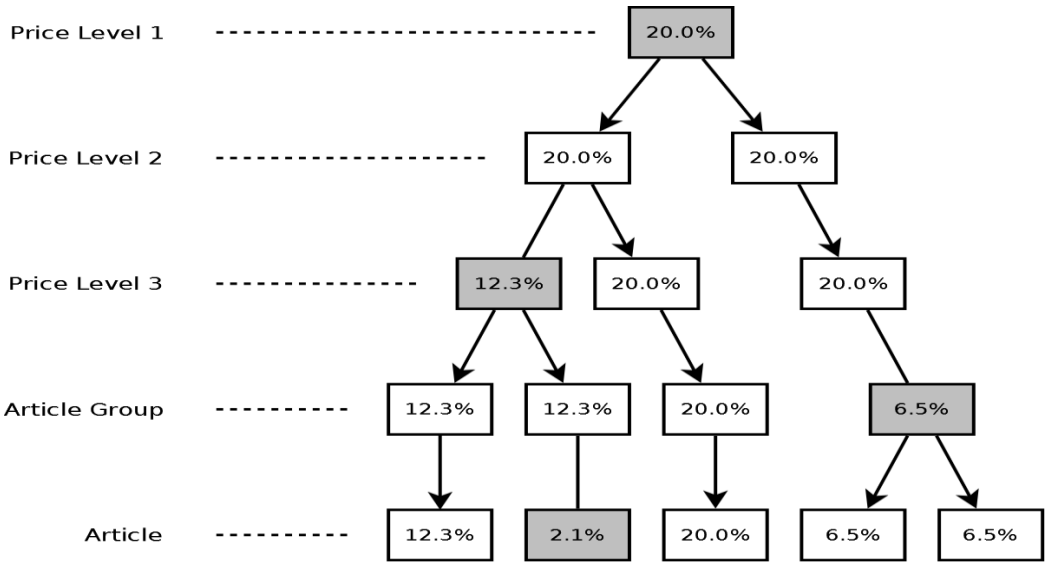

Figure 4.3: Discount inheritance in the example article tree from figure 4.2

iii) CURRENT IMPLEMENTATION

The current implementation of PCT suffers from critical performance issues. Since the source code of this system is not allowed to be included in this report, the problems of its algorithm have to be explained in terms of bad structure choices and complexity rather than examples and excerpts from the actual code.

A (very) rough outline of the algorithm structure used to perform simulations in PCT is presented in algorithm 4.1.2. While it does not motivate or explain the details of each step, it does provide enough information to analyze its complexity. To give the reader some sort of idea of the actual magnitude of the implementation of this algorithm, its Java source code takes up several hundred kilobytes (not including GUI, server connections, database handling and other parts which are not directly related to the algorithm). In other words, a line describing e.g. criteria matching means running a separate algorithm which in turn has a complexity worth mentioning.

Algorithm 4.1.2: Structure of the simulation process in PCT

1 if this is the first 1 run of the simulation process then
2 initialize connection to each input data element in the GUI [O(k)]
3 end
4 for each price level in the article tree [O(k)] do

5 match condition level [O(k)]
6 match price level [O(k)]
7 for each item in the customer's cache [O(n)] do
8 match criteria [O(k)]
9 end
10 retrieve target discount [O(k)]
11 for each article in the article tree [O(a)] do
12 for each article in the customer's cache [O(n)] do
13 match criteria [O(k)]
14 for each price level in the article tree [O(k)] do
15 retrieve data and calculate results
16 end
17 end
18 retrieve agreed discounts [O(k)]
19 compare discounts to target discounts [O(k)]
20 end
21 end
22 for each article in the customer's cache [O(n)] do
23 calculate results for articles under price level 1 nodes $\notin$ path
24 end

In the pseudo code above, the complexity has been included on each line where O notation is applicable. The meaning of each occurring variable in the O notation is presented.

The total complexity of the implementation of the current simulation algorithm is
$O(k+k(k+k+nk+k+a(n(k+k))+k+k)+n) = O(k+5k^2+nk^2+2ank^2+n) = O(ank^2)$

It should also be noted that the complexity of repeated runs of the algorithm is
$O(k(k+k+nk+k+a(n(k+k))+k+k)+n) = O(5k^2+nk^2+2ank^2+n) = O(ank^2)$

V. RESULTS

This section contains running times of customer discount simulations. Running times for our implementation are shown together with corresponding running times for PCT for the same underlying data.

Articles	Running time PCT [ms]	Running time our model [ms]
1	723	130
10	876	156
40	741	89
105	1,142	91
206	1,879	118
366	4,671	131
483	6,473	148
789	9,141	161

Table 5.1: Running time for First simulation in PCT and our model

Running time [ms]	PCT			Our model		
#articles	Run1	Run2	Run 3	Run1	Run2	Run 3
40	741	782	692	89	< 1	< 1
206	1,879	1,801	1,707	118	< 1	< 1
366	4,671	3,879	4,665	131	< 1	< 1
483	6,473	6,773	4,240	148	< 1	< 1
789	9,141	6,780	6,087	161	< 1	< 1

Table 5.2: Running time for repeated simulations in PCT and our model

Articles	Running time PCT [ms]	Running time our model [ms]
100	736	150
500	1,295	161
1,000	1,694	157
1,500	1,835	157

2,000	2,161	153
3,000	2,884	150
5,000	4,335	152
10,000	8,463	146
20,000	20,314	166
30,000	33,671	210
40,000	45,892	253

Table 5.3: Running time for first simulation in PCT and our model over generated data

VI. FUTURE WORK

Another interesting approach would be a comparison between the performances of our models using different database solutions. NoSQL database systems could prove effective in handling the big data problems introduced by the scaling extension. Particularly, an implementation using a graph database would be interesting due to this technology's great performance when dealing with tree structures. For example, the graph database Neo4j has shown promising results in multiple studies such as [5], where Neo4j is concluded to be up to ten times faster than MySQL for traversals and [6], where the results show that running times for MySQL increase much faster than for Neo4j as the data magnitude grows.

VII. CONCLUSION

This project has consisted of analysis, optimization and implementation of the existing simulation algorithms in PCT as well as modeling and implementation of its upcoming scaling extension. The results show that the implementation of the optimized customer discount model provides large enough performance improvements to guarantee reasonable running times even for the largest customers. The results for the scaling extension prove that implementation of the desired functionality in PCT is possible as well, as long as the big data issue is handled in an efficient way.

A final conclusion of this project is that optimization of existing algorithms is not always sufficient in order to improve the performance of a system. Creating new, optimized models and developing fast algorithms for these can prove far more efficient than optimization of existing algorithms based on improficient models.

REFERENCES

[1] A. Jacobs, The pathologies of big data, Commun. ACM Vol. 52 (8) (2009) pp. 36{44.

[2] R. B. Miller, Response time in man-computer conversational transactions, in: Proceedings of the December 9-11, 1968, fall joint computer conference, part I, AFIPS '68 (Fall, part I), New York, NY, USA, 1968, pp. 267{277.

[3] S. C. Seow, User and system response times, in: Designing and Engineering Time: The Psychology of Time Perception in Software, Addison-Wesley Professional, 2008, pp. 33{48.

[4] Era Li Yang, Dan Hu . : The Dissemination Model of Digital Music in Big Data

[5] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, D. Wilkins, A comparison of a graph database and a relational database: a data provenance perspective, in: Proceedings of the 48th Annual Southeast Regional Conference, ACM SE '10, ACM, New York, NY, USA, 2010, pp. 42:1{42:6.

[6] S. Batra, C. Tyagi, Comparative analysis of relational and graph databases, International Journal of Soft Computing and Engineering (IJSCE) Volume 2 (Issue 2) (2012) pp. 509{512.

