REDUCE TIME PROCESS USING APRIORI ALGORITHM ON K-WAYJOIN BASED TO FIND RETAIL BUSINESS DATA RELATIONSHIP PATTERN

Joko Aryanto[1)], Yuli Asriningtyas[2)]

[1)]Department of Information Engineering, University of Technology Yogyakarta
e-mail : joko.aryanto@uty.ac.id
[2)]Department of Information Engineering, University of Technology Yogyakarta
e-mail : yuli_asriningtias@uty.ac.id

ABSTRACT

The process of running a trading business, businesses must always update information on market competition that occurs. Running a trading business is not just opening a business place and waiting for consumers to shop. Consumers will lightly come to shopping places to shop for various reasons. From cheap prices, attractive arrangements, large parking lots, to the ease of finding items to buy. From these problems, business people will compete on how to easily attract customers to their place of business.

One way to solve these problems is by structuring merchandise with the aim of customers to easily get the items they are looking for. One way to solve these problems is by structuring merchandise with the aim of customers to easily get the items they are looking for. Arrangements that are made do not originate from arranging the location of goods according to taste but are carried out on the basis of trends or trends in goods purchased by consumers when shopping.

This process is often referred to as the "Data Mining" process is one of the effective methods for finding consumers' preferences to choose the items they buy. This process can be completed using the Apriori algorithm. The principle used by the Apriori algorithm is that if an itemset appears frequently, then all subset of itemset must also appear frequently. This results in repeated checking and will require a short time. Problems that require a short time, a method is proposed, namely by developing to be able to reduce the travel time of the process. The length of time that occurs in the process of calculating the value of support and confidance and the repetition process to find the value. The method used is to manipulate the use of query languages with a k-way join research approach so that the optimal query language arrangement can be obtained. The results obtained in this study are that execution times are relatively faster, with the results of the same association rules as those produced by the Priori method without any development or modification.

Keyword : *Data Mining, Apriori improvisation, Reduce time, Apriori Development, k-way.*

1. INTRODUCTION

The retail business makes it difficult for entrepreneurs to find opportunities to get customers and make customers feel like they are a retail brand. The proliferation of department stores is the impact of changing consumer trends in shopping. The more intense competition in finding consumer markets, employers are forced to find and improve services in order to maximize sales and achieve effectiveness in winning the sales competition. One way that can be done is to process all transaction data to find trends that are common. The processing of all transaction data creates an obstacle, namely the number of transactions encountered per day is very large.

Processing data in a very large database requires a short time. But we can process very large data to get a pattern that we did not know before. The process of obtaining these patterns or rules is usually called data mining with certain techniques or methods. The pattern produced in data mining is interesting information for managers that can be developed to determine the next steps. The Market Basket Analysis method is an analysis process at the level of transaction data that can increase business productivity. The picture is in the form of an item group tendency (consisting of 2 items or more) that is purchased by a consumer in a sales transaction. This method is able to find a sales trend, grouping general products or easy to guess and can even find trends that have never been thought of before. Examples of finding common or predictable product groupings are purchasing fat-free snacks and diet soda simultaneously. Examples of finding trends that have never been thought of are purchasing toothpaste and tuna. Through these patterns entrepreneurs can capture sales opportunities that competitors miss. Market Basket Analysis is based on an analysis of consumer habits by looking for associations and correlations between different items carried out on the same transaction by consumers. [1].

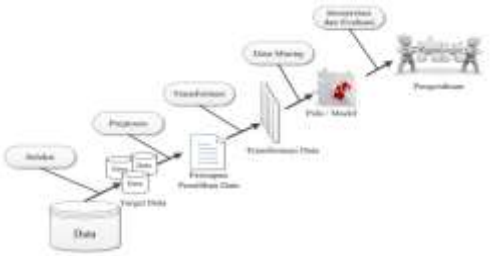
2. DATA MINING

2.1 Definition of Data Mining

Data mining is a computational process for finding patterns in large data sets involving intersection of methods from artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and convert it into a structure that can be understood for further use. Aside from the raw analysis steps, it involves database and data management aspects, pre-processing data, inference models and considerations, interestingness metrics, complexity onsiderations, post-processing structures found, visualized, and updated online [17].

Data mining is part of Knowledge Discovery In Databases (KDD). While data mining has three methods, namely clustering, classification and Rule Association. KDD is a multi-step process that converts raw data into useful knowledge (Bagga and Singh, 2011).

The stages that occur in the process of data mining or knowledge discovery according to Kenneth Collier (1998) are divided into 5 stages, namely :



a. Data Selection

The purpose of this phase is extraction from large data warehouses into data that is relevant to data mining analysis. The data extraction process helps to streamline and speed up the process.

b. Data Preprocessing

This phase is related to data cleaning and preparation of tasks needed to ensure correct results. Eliminating missing values in data, ensuring that code values have a uniform meaning and ensuring that no false data values are typical actions that occur during this phase.

c. Data Transformation

This stage converts the data into a form or format suitable for data mining needs. The normalization process is usually needed in the transformational data stage.

d. Data mining

The purpose of the data mining stage is to analyze the database according to the algorithm used so

that it finds meaningful patterns or rules and produces prediction models. Data mining is a core element of the KDD cycle.

e. Interpretation and Evaluation

While data mining algorithms have the potential to produce an unlimited number of hidden patterns in data, many of the results of these processes may not be meaningful or useful. This final stage aims to choose models that are valid and useful for making future business decisions.

2.2 Association Rules

Association mining rules is one technique used in data mining that aims to find association rules contained in a data set. The technique of association rule mining will find patterns of database. The support value of an item is obtained by the following formula:

$$Support\ (A) = \frac{Number\ of\ Transactions\ Containing\ A}{Total\ Transactions}$$

While the support value containing 2 itemset can be written in the formula as follows:

$$Support\ (A \cap B) = \frac{Number\ of\ Transactions\ Containing\ A\ \ B}{Total\ Transaction}$$

b. Establishment of association rules

After all the high frequency patterns are found, the next step is to find associative rules that meet the minimum requirements for confidence by calculating the confidence rules of associative A> B. The confidence value of rule A> B is obtained from the following formula:

$$Support\ P(B/A) = \frac{Number\ of\ Transactions\ Containing\ A\ \ B}{Total\ Transaction\ Containing\ A}$$

The association analysis is defined as a process to find all association rules that meet the minimum relationships between data that are mutually associated. This technique is also known as the market basket analysis which defines itemset in a transaction simultaneously, for example in the process of selling goods. The process begins by searching for frequent itemset by continuing the formation of association rules obtained [18].
Rules that state associations between several attributes are often called affinity analysis or market basket analysis. The basic methodology of association analysis is divided into two stages:

a. High frequency pattern analysis

This stage looks for item combinations that meet the minimum requirements of the support value in the

requirements for support (minimum support) and minimum requirements for confidence (minimum confidence).

2.3 Apriori algorithm

From the research conducted by D. M. Tank, it was explained that the Apriori algorithm is a classic algorithm for obtaining association rules from data that often appears. The basic idea of this algorithm is to use a repeat layer-by-layer approach to find the frequency of occurrence that often appears. The first process of the algorithm will get k-itemset, and then use k-itemsets to explore (k + 1) itemset. First, introducing a priori itemset knowledge that often appears, every part of the itemset frequency as well as itemset that often appears. The Apriori algorithm uses prior knowledge of itemset that often appears, first by finding a 1-itemset frequency collection, denoted L. Then use L_1 2 itemset to get L_2, and then L, and so on, until you can't find the k-itemset frequency. The Apriori algorithm mainly consists of the following three steps:

1. Step connecting: connecting k-frequency itemset to produce candisdat (k + 1), denoted by C_{k+1}. The next condition of the linking step is that the two k-itemset have the same first item (k-1) and different k-items. Shown

l_1 [j] is the first item of l_1, it can be shown as follows:
Where l1 and l2 are k-items from the set portion of Lk, l_1 [k] 2 l_2 [k] is to ensure that they can produce duplicate k-itemsets. Itemset produced by l_1 and l_2 connections are as follows:

$$\{l_1[1], l_1[2], l_1[3], \dots \dots, l_1[k], l_2[k]\}$$

2. Trimming step: To select itemset that often appears L_{k+1} comes from candidate C_{k+1}, because candidate set C_{k+1} is a superset of L itemset that often appears. According to the Apriori nature: each set of frequencies must also appear frequently, i.e. each (k-1) subset of k-items must also appear frequently. With this property we can find out whether the item k subset of C_{k+1} is in L_{k+1}, if not, then delete the candidate (k + 1) - itemset deleted from C.
3. Step counting: scanning data in the database, stacking the number of candidates k + 1 appears in the database. If the number of occurrences of candidates is less than the minimum support threshold given, the candidate itemset will be deleted[9].

Algorithms used in the search for patterns or rules of relationships between data include Apriori, FP-Growth, and CT-Pro. The use of the Apriori Algorithm is known as the association method, with the basic idea of calculating the pattern of occurrence of items that appear in transaction data with several iterations so that patterns and rules will be obtained. The Apriori Algorithm has very basic constraints in the process of finding relationships between the data that is carried out, that is, it takes a short time. Depends on the amount of data used [5].

2.4 Weaknesses of the Apriori Algorithm

The Apriori algorithm is one of the most classic algorithms for finding association rules, and has an inefficient use of time. The time needed by the Priori algorithm in consuming time lies in three main aspects as follows:

1. In the itemset relationship step, when connecting the k-itemset to produce (k + 1) itemset, it is done to compare too many items to determine whether itemsets meet the relationship conditions. When L has m itemset k-, the time complexity of the relationship step is O (k * m ^ 2).
2. In step reduction: when determining whether a subset of candidate sets C often appears in set L, the best situation is to do one scan to get results, while the worst case case is scanning as many times as k to find that part k of C_{k+1} is not in L. So the average number needs to scan and compare L.
3. In calculating steps: when calculating how much support from itemset in C_{k+1}, we need to scan data in the database for C_{k+1} times. Taking into account the three aspects of the steps above that take process time, this research presents the Apriori algorithm by performing step improvisations to reduce processing time [2].

2.5 Combination Structure

The structure of itemset here is to follow a form of combination. Definition of combinations is to combine several objects from a group without regard to the sequence. In combination, the order of objects is not considered as an example where {1,2,3} is equal to {2,3,1} and {3,1,2}. Combinations can be divided into two namely combinations with repetitions and combinations without repetition. Combination without repetition when sequence is not considered but every existing object can only be selected once, the number of combinations that are available are:

$$\frac{\boldsymbol{n}!}{\boldsymbol{r}!\,(\boldsymbol{n}-\boldsymbol{r})!} = \left(\frac{\boldsymbol{n}}{\boldsymbol{r}}\right)$$

Where n is the number of objects that can be selected and r is the number that must be chosen. For example, there are 5 different colored pencils, namely red, yellow, green, blue and purple. These

color pencils can only be selected in two colors. Many ways to combine existing colored pencils using the formula above are:

5!/(5-2)!(2)! = 10 combination.

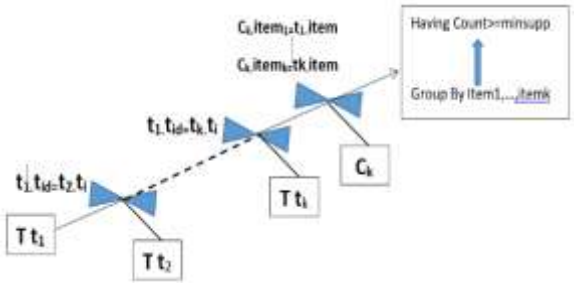
Combinations with repetitions if the order is not noticed and the object can be chosen more than once, then the number of combinations that are available is shown in the following formula:

$$\frac{(\boldsymbol{n}+\boldsymbol{r}-\mathbf{1})!}{\boldsymbol{r}!\,(\boldsymbol{n}-\mathbf{1})!} = \binom{(\boldsymbol{n}+\boldsymbol{r}-\mathbf{1})}{\boldsymbol{r}} = \binom{(\boldsymbol{n}+\boldsymbol{r}-\mathbf{1})}{\boldsymbol{n}-\mathbf{1}}$$

Where n is the number of objects that can be selected and r is the number that must be chosen. For example, there are 10 different types of donuts in a donut shop. The resulting combination if you want to buy three donuts is: (10+3-1)!/3!(10-1)! = 220 combinations.
The combination used in the priori algorithm in this study is a combination without repetition. The order of items in the combination that is formed is not considered but every item that exists can only be used once in one combination or itemset.

2.6 K-Way Concept

In determining frequent itemset is the most significant step. The approach using the K-Way method is one method to reduce workload, especially in calculating the value of support and confidence in a very large data set. The purpose of optimization and analysis using this approach is to determine the impact of optimization on datasets that have characteristics such as the number of average transactions, support, trust, and others to get heuristics that are influenced by input datasets. Although, not all optimizers always produce better timings, this study of optimization can give us better insight in formulating metadata that can be used to create dataset optimizers that have different characteristics. For support calculations, each k corresponding to the criteria will be combined in Ck,



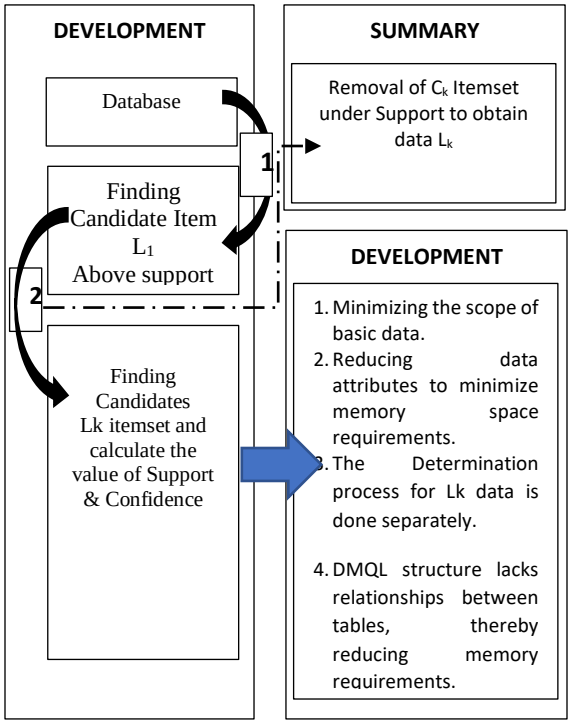
The total number of items generated from combining Ck with T equals the amount of data in Ck * the average support of the first item in Ck. By using "join" notation, the join value can be represented as a combination (Ck, R, Ck * S1). Thus, the combination of T with C will produce a table, which contains the number of support for the first item itemset in Ck. In the case of this join notation it can be represented as a join (Ck * sm-1, R, Ck * sm). The last merging value, cannot be calculated from the formula above as the value of s for itemset with length k unknown, because itemset of length k produced do not appear frequently. But the relation obtained from the last join will have a tuple record as the amount of support for each itemset in set C_k. Using this join notation can be represented as join (C_k * sk-1, R, S (C_k)). Therefore this approach can be denoted as: [21]. In SQL the process using the K-way approach can be used in the following pseudocode,

```
insert into Ck select I1.item1, ..., I1.itemk-1, I2.itemk-1
from       Fk-1 I1, Fk-1 I2
where      I1.item1 = I2.item1 and
                 ⋮
           I1.itemk-2 = I2.itemk-2 and
           I1.itemk-1 < I2.itemk-1


insert into Fk select item1, ..., itemk, count(*)
from       Ck, T t1, ... T tk
where      t1.item = Ck.item1 and
                 ⋮
           tk.item = Ck.itemk and
           t1.tid = t2.tid and
                 ⋮
           tk-1.tid = tk.tid
group by   item1, item2 ... itemk
having     count(*) ≥ minsupp
```

3. MODELING

In its development, to obtain the desired results, as a reference, the work chart as a guide for the steps to be taken is as follows,



The development step is explained as follows,

1. The first step is to reduce the scope of basic data which has a large amount to be less so that the data search process becomes easier.
2. Data with a smaller scope, because it is certain that the data will be processed, it is possible to reduce its attributes so that it can reduce the problem of data filters.
3. Determination of candidate item sets is done separately after the data is placed in a temporary table.
4. The process of DMQL (Data Mining Query Language) in one table without the relation between tables will affect the performance of the processing time.

The location of time trimming that will occur is in the preparation of the query language used. The researcher divides into several stages to reduce the processing time, namely:

1. **The process of selecting candidate items.**
   This process can reduce time because it will directly retrieve data that has a support value equal to or greater than the value we have set. The above process will get the results of items, support or the number of contributions, and the confidence value of the items that are candidates. In the next process there will be no re-selection for data pruning. In contrast to traditional priori, in the search for candidate data items will be displayed all items without regard to which items have the value of support and confirmation with the limits that we specify. The candidate search syntax for traditional algorithms is as shown below:

```
//Item SELECTION
select * from
(SELECT j.ITEM,count(J.trans) as juml,
(COUNT(J.trans)/@ttl)*100 as pross FROM D_Item
j,TRANs TJ where Tj.trans=j.trans and tj.TGL >= @tgl1
and tj.TGL<= @tgl2 GROUP BY ITEM) as DataTabel
where pross>=@suport
```

   In the syntax above there are two processes, namely displaying all candidates that appear, then proceed with the selection of data that entered the predetermined support limit criteria. From the comparison of the process of finding candidate items in the traditional a priori algorithm with the a priori algorithm that has been developed, the difference lies in the number of processing steps that will directly affect the processing time in finding candidates.

2. **In the process of determining itemset candidates**
   The process of finding candidate itemset pairs that determine the number of support and confidance is a process that requires not a little time. In the traditional algorithm process, a process is carried out to produce

Ck data, which is to find all candidate pairs of item sets which are then processed to find Lk data where data in Lk is data that has been pruned data with support and confidence values below the values we have previously specified.

In contrast to the process of developing algorithms that researchers do that is to directly find data in Lk without looking for Ck data first. So the process is done without going through the Ck scanning process first.

```
while data @c1 <> eof
begin
 set @itm1=(item1 from @c1 where nomor=@x1+1)
 set @x2=1
 while (data @c1 where nomor=@x2+1)<eof
 begin
  set @itm2=(item1 from @c1 where nomor=@x2+1)
  if (@itm1<>@itm2) and (@supp>=@suport)
  insert @c2 values(@itm1,@itm2,@supp,
(@supp/@ttl)*100,
(@supp/@conf)*100)
  set @x2=@x2+1
 end
 set @x1=@x1+1
end
```

The query syntax above is used in its development to find itemset candidates by finding Lk not to find Ck first. So that it won't pass the Ck data trimming process to get Lk data. Thus the process here will reduce the time in finding all itemset candidates, and eliminate the time of the trimming process. From the modeling step scheme above, here is the form of the pseudocode

```
L1:={Kandidat itemset}
K:=2
While (Lk-1 <>0) do
Begin
   Forall transaksi Lk-1  do
    Lk := select count(*) from kandidat pasangan
yang
    berasal dari Lk-1
     K :=K+1;
End
Answer :=Uk Lk

Insert into L k-itemset
Select * from (select p.item1,p.item2,…,p.itemk-1,
p.n_sup,p.n_conf  From L(k-1)-itemset p, L(k-1)-itemset q
 where p.item1=q.item1,…, p.itemk-2=q.itemk-2 ,
p.itemk-1 < q.itemk-1) as Tabel
where n_sup>=sup and n_conf>=conf
```

4. MEASUREMENT AND DATA TEST

4.1 Data Attributes

The data used comes from the Retail Management Information System database MAGA Supermarket. The data used is transaction data from December 2014 to December 2016 with the number of records for the transaction table reaching more than 5479519 records.

There are three important partitions that influence this research process, namely,

Attribute	Data Type	Information
Tgl_jual	Datetime	Transaction group
Kode_jual	Varchar	T_{id}
Kode_brg	Varchar	T_{item}

4.1.1 Installing Candidate Item Set k-way Join

In developing the model carried out by minimizing the scope of the search in the "JUAL" table. The step will be to take transaction data that occurs on the intended date limit so that the total amount of data will be less than the total data as a whole. The steps taken from the syntax are as follows:

1. Make a shelter table to hold itemset pair data in this case is table @ L_1.
2. Using the "while" looping command is done pairing candidate items from table C1.
3. In the process of pairing items, at the same time calculation of the value of support and confidence is carried out at the same time.
4. If the value of support and confidence is in accordance with the prescribed limits, it will be included in table L_1.

While data @c1 hasn't started yet line 1
Begin
  @item1=item1 dari @c1 line 1
  While data @c1 hasn't started yet line 2

```
 Begin
   @item2=item1 dari @c1 line 2
   @sup= count the number of transactions
       that   contain Item1 dan item2
   @conf= count the number of transactions
       that contain item1
 If (@itm1<>@itm2) and (@itm2<>
   @itm1) and (support>=limit)
  Insert
L1(@itm1,@itm2,jmltran,support,confidenc
e)
  Inc item2
 End;
 Inc item1
End;
```

4.1.2 Calculate Support and Confidence

In this stage is a process that requires the longest time, because it will be scanning the overall data from the set items that are formed. Implementing the formula to calculate the value of support and confidence is done by pseudocode,

```
@tgl1= the beginning of the transaction date
@tgl2= deadline for the transaction date
@ttl= select count(kode_tran) as jml FROM
    @C1 WHERE tanggal>= @tgl1 AND
    tanggal<=@tgl2
    AND jumlah item<=3;
@transak= select count(kode_tran) as jml
    FROM @C1 WHERE tanggal>= @tgl1
    AND tanggal<=@tgl2 A ND jumlah
    item<=3 AND (item1,item2)
@support=(@transak/@ttl)*100
```

After obtaining the value of the number of transactions containing A and B, the value of the support can be calculated,

```
// enter data into table @ L1 that matches the
    criteria
Jika (item1<>item2) and (item2<>item1)
    and (confidance>=limit)
Insert @L1 values ( @itm1,@itm2,@jml,
(@support), (@conf))
```

4.1.3 Comparison of Data Testing Process

The complete process is done by testing the method by dividing 3 groups of data based on the transaction time span.

1. **Test group data 1**

   Producing itemset candidate data amounting to 5 items is transaction data in the period 1-10 of December 2016, the number of transactions that occur as many as 77812 with data C1 = 5 items, L1 = 2 items, L2 = 0 items generate travel time for traditional syntax in seconds are 69 seconds while syntax development produces 65 seconds with candidate items,

Item	Qty Trans	%
100038	1310	11.02
100201	1940	16.32
100336	1220	10.26
100338	2030	17.08
753797	1280	10.76

The results of installing candidate items formed by the L1 association rule are as follows:

a. If you buy 100201 then 100338 total transactions are 140 transactions and the confidence value is 72.16%.
b. If you buy 100338, then 100201 transactions are 140 transactions and the confirmation value is 68.96%.

Item 1	Item 2	Qty Trans	Confidence (%)
100201	100338	140	72.16
100338	100201	140	68.96

2. **Test group data 2**

   It is transaction data in the period of 1-20 of November 2016, the number of transactions that occur as many as 144245 with data C1 =

4 items, L1 = 2 items, L2 = 0 items resulting in travel time for traditional syntax in seconds is 85 seconds while for syntax development takes 77 seconds.

Item	Qty Trans	%
100038	2650	11.21
100201	3760	15.91
100338	3900	16.50
753797	2440	10.32

The results of installing candidate items formed by the L1 association rule are as follows:

a. If you buy 100201 then 100338 total transactions are 270 transactions and the confidence value is 71.80%.
b. If you buy 100338, then 100201 transactions will be 270 transactions and the confirmation value will be 69.23%.

Item 1	Item 2	qty	Confidence (%)
100201	100338	270	71.80
100338	100201	270	69.23

3. Test group data 3

It is transaction data in the period 1-30 of November 2016, the number of transactions that occur as many as 215456 with data C1 = 4 items, L1 = 3 items, L2 = 0 items produce travel time for traditional syntax in seconds is 185 seconds while for syntax development took 181 seconds.

Item	qty	%
100038	3810	11.01
100201	5140	14.86
100336	3530	10.20
100338	5680	16.42

The results of installing candidate items formed by the L1 association rule are as follows:

a. If you buy 100201 then 100338 the number of transactions is 300 transactions and the confidence value is 58.36%.
b. If you buy 100338, then 100201 transactions will be 300 transactions and the confirmation value is 52.81%.

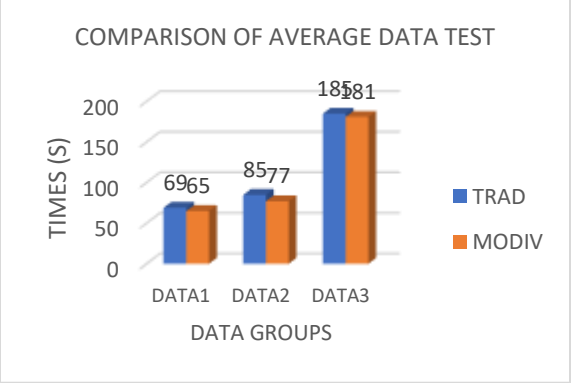
Item 1	Item 2	Qty Trans	Confidence (%)
100201	100338	300	58.36
100338	100201	300	52.81

4.1.4 Comparison of Test Data

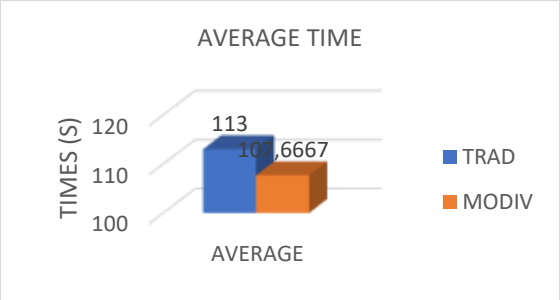
Of the three groups of data that were tested, the calculation of the average travel value was obtained as follows:

TYPE	DATA 1	DATA 2	DATA 3	AVG	DIFF (%)
TRAD	69	85	185	113	49,5
MODIV	65	77	181	107,66	

In the form of a graph can be seen as follows,



Comparison of the average values of the two methods is in the following graph,



Of the three groups of data that were tested by the algorithm, there was a time reduction value of 4.95%.

5. Conclusion

The results of the algorithm test in the 2 methods that were made were concluded to be used as a reference for the next study,

a. Removal of itemset removal processes that do not meet the criteria can reduce the scanning process time. In this section, searching for itemset pairs is done while calculating the value of support and Confidence. If the value is included in the criteria, it will be entered into L1 as the Ck data collection. In the classic a priori algorithm all candidate pairs that have already calculated the support and confidence values will be accommodated into one, then after the calculation process is complete, the process of deleting data that is not included in the specified criteria will be carried out.
b. The approach using the k-way join method in the application of a priori algorithms can improve performance in the iterations that occur. This approach can also reduce the workload, especially in calculating the value of support and confidence in a very large data set. The focus of the k-way join process is DMQL or Data Mining Query Language. By determining the operator sign formula in this study, the reduction of clause group by, between, and the relation of many partitions from the database with the data model selected using vertical data models can reduce the burden with the results of the average travel time for traditional a priori 71.66 seconds while using the k-way join approach reaches 61.11 seconds, resulting in a decrease in travel time of 10.55 seconds.
c. The k-way join method approach reduces workload, ie the process that occurs with a decrease in travel time is done directly affecting memory consumption so that the personal needs of computers can be carried out without the last-generation specification prerequisites.

As a reference for further research, there are several things that must be considered based on this research,

a. Fill in the "kode_brg" attribute that is used using the minimum number of characters. The number of characters used will affect the searching process for longer. Although it will only have an effect in a very large number of records.
b. The selection of the right operator sign in the DMQL process is very influential. The difference in the DBMS used causes different result differences even though using the same operator sign.
c. For further research it can be done using a server-based operating system with a number of records in terabytes.

Bibliography

[1] V. Mohan and D. S. Rajpoot, “Matrix-Over-Apriori : An Improvement Over Apriori Using Matrix,” vol. 5, no. 01, pp. 1–6, 2016.

[2] J. H. and M. Kamber, *Data Mining :Concept and Technique, S. Edition*, S.Edition. 500 Sansome Street, Suite 400, San Francisco, CA 94111 This: Diane Cerra, 2006.

[3] D. Edwards, “Data Mining: Concepts, Models, Methods, and Algorithms,” *J. Proteome Res.*, vol. 2, no. 3, pp. 334–334, 2003.

[4] M. B. Nichol, T. K. Knight, T. Dow, G. Wygant, G. Borok, O. Hauch, and R. O'Connor, "Fast Algorithms for Mining Association Rules," *Ann. Pharmacother.*, vol. 42, no. 1, pp. 62–70, 2008.

[5] C. D. Cleaning, D. D. Transformation, A. T. Analysis, E. D. Mining, and B. D. Selection, "Three phase iterative model of kdd," vol. 4, no. 2, pp. 695–697, 2011.

[6] S. A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation," vol. 3, no. 7, pp. 1–4, 2012.

[7] J. Singh and H. Ram, "Improving Efficiency of Apriori Algorithm Using," vol. 3, no. 1, pp. 1–4, 2013.

[8] J. Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules," vol. 2, no. 1, pp. 25–27, 2013.

[9] D. M. Tank, "Improved Apriori Algorithm for Mining Association Rules," *Int. J. Inf. Technol. Comput. Sci.*, vol. 6, no. 7, pp. 15–23, 2014.

[10] S. Chaudhari, M. Borkhatariya, A. Churi, and M. Bhonsle, "Implementation and Analysis of Improved Apriori Algorithm," pp. 70–78, 2008.

[11] J. Kaur, R. Singh, and R. K. Gurm, "Performance evaluation of Apriori algorithm using association rule mining technique," vol. 2, no. 5, 2016.

[12] A. Ansari, A. Parab, and S. Kadam, "Apriori - A Big Data Analysis - A Review," pp. 3517–3520.

[13] Q. Liu and J. Xin, "An improved Apriori algorithm based on data stream classification," *J. Comput. Inf. Syst.*, vol. 10, no. 23, pp. 10259–10266, 2014.

[14] R. S. Dm, V. Saldanha, and S. Sebastian, "Apriori Algorithm and its Applications in The Retail Industry for Analyzing Customer Interests," vol. 2, no. 3, pp. 46–51, 2015.

[15] B. S. Dhak and M. Sawarkar, "Apriori : a promising data warehouse tool for finding frequent itemset and to define association rules," vol. 4, no. 1, pp. 60–65, 2016.

[16] N. Gutierrez, "Demystifying Market Basket Analysis," *http: //www.information-management.com*, 2006. .

[17] ACM SIGKDD, *Data Mining Curriculum*. 2016.

[18] M. Kantardzic, *Data Mining : Concepts, Models, Methods, and Algorithms*, Vol. 2. No. Proteome Research, 2003.

[19] Pete Chapman., *Step-by-step data mining guide*. The CRISP-DM Consortium, 2000.

[20]. Tan, P.N., Steinbach, M., Kumar, V., *Introduction to Data Mining*, Addison-Wesley, Boston.,2006

[21]. L. Notes, C. Science, and C. Growth, "Performance Evaluation and Analysis of K-Way Join Variants for Association Rule Mining," no. January, 2003.