Analysis of Telecommunication Database using Star Schema

Dr. Sanjay Srivastava
Professor, Department of Computer Science & Information Technology
MGM College of Engineering and Technology, Noida, Uttar Pradesh, India

Kaushal,Srivastava
B.Tech(CSe) Scholar
MGM College of Engineering and Technology, Noida, Uttar Pradesh, India

Akhil Sharma
B.Tech(CSe) Scholar
MGM College of Engineering and Technology, Noida, Uttar Pradesh, India

Avinash Pandey
B.Tech(CSe) Scholar
MGM College of Engineering and Technology, Noida, Uttar Pradesh, India

Abstract-Today, the number of customers are increasing gradually day by day in telecommunication industry. It has become a challenging as well as cumbersome and time consuming task for extracting the required information from large heterogeneous data warehouse. In the Telecom Industry there are millions of customers who are increasing day by day so managing their data and updating their need as well as providing quick requested result to them, is not a simplest task at all. In telecommunication industry, they have to come with new plans and policies to retain our existing customer by providing best facilities to them and attract new customers by providing them new catchy offers so we have association rule mining for telecommunication industry to find the most frequent items-set that a generally referred by most of the customers. In this Paper, the information of telecommunication industry about the different operators who are providing telecommunication facilities and comparing their rate plans about internet depending on streaming and services like 2G, 3G has been acquired and it is to find how many customers are finally using which services under that operator and at what cost to predict which plan and operator is preferred most frequently.

I. INTRODUCTION

Data warehousing is a technique for data warehouse. Data warehousing has become a good platform for most large companies worldwide. The data stored in the data warehouse captures many different aspects of the business process such as manufacturing, distribution, sales, and marketing. This data reflects direct and indirect customer patterns and trends, business practices, strategies, know-how and other characteristics. Therefore, this data is of vital importance to the success of the business whose state it captures, which is why companies choose to engage in the relatively Expensive undertaking of creating and maintaining the data warehouse. While some information and facts can be reap from the data warehouse directly, much more remains hidden as implicit patterns and trends. The discovery of such information often yields important insights into the business and its customers and may lead to unlocking hidden potentials by devising innovative strategies. The discoveries go beyond the standard on-line analytical processing which mostly serves reporting purposes.

Association-rule mining is one of the most important and successful methods for finding new patterns. Typically, if an organization wants to employ association-rule mining on their data warehouse data, it has to acquire a separate data mining tool. Before the analysis is to be performed, the data must be retrieved from the database repository that stores the data warehouse, which is often a complex and time-consuming process. The vendors of data management software are becoming aware of the need for integration of data mining capabilities into database engines. While this does represent an improvement, it still does not eliminate the need for additional software. In this paper association-rule data mining, has been used within data warehouse that utilizes the query processing power of the telecommunication data warehouse itself without using a separate data mining tool. In addition, this approach is capable of answering a variety of questions based on the entire set of data stored in Telecommunication data warehouse. In this paper, it has been analyzed that which plan (3G or 2G) is using by customers in different regions of India.

The remainder of this paper is organized as follows: Section 2 gives a brief overview of association-rule mining in Telecommunication data warehouse environment: Section 3 introduces the implementation of architecture for Association-rule Mining; Section 4 describes the results of an experimental study. Finally, Section 5 gives the conclusions and indicates the future work.

II. LITERATURE SURVEY

In Association Rule Mining, we find new patterns which are frequent data item set. Generally, we use data mining tool and data ware house for finding the frequent-item set. In this Our Approach is reduce the cumbersome and time consuming approach for general purpose of frequent-item set extraction. We describe the direct approach to association rule data mining within the data warehouse which utilizes the query processing power of data ware house itself using a separate data mining tool by acquiring more information with the help of Extended Association rule using other non-item dimension of data ware house which results in more detailed and ultimately actionable rule. Our Aim is also to define Association rule for aggregated, that is non-transactional data.

III. ASSOCIATION-RULE DATA MINING IN TELECOMMUNICATION DATA WAREHOUSE

With the help of association rule mining the decisions about market activities like promotional offers, discounts offered and combo items (Services) at the point of sale.

Association rule mining can find those item sets which can satisfy the minimum support and confidence from given data base. Dimensional modeling is the most extensive technique for modeling data warehouses, organizes tables into fact tables containing basic quantitative measurements of a business subject and dimension tables that provide descriptions of the facts being stored. The data model used by this method is known as star-schema. Figure 4.1 shows a star schema model of a data warehouse for telecommunication.

IV. STAR SCHEMA

One of the simplest way to represent data of data warehouse by showing their logical relationship. Star schema consists of one fact table which is used for giving references to many dimension table. Fact table consists of foreign keys and various other units for performing measurement. Dimension table defines the physical structure of various entities. Dimension table contain less records as compared to fact table but it is used to describe the fact table data. Star schema is used for representing telecommunication data warehouse. In the given figure, star schema shows the representation of sales department of telecommunication industry. This representation is used for the analysis and various queries to show that which service is used by the customers. This schema is used to show the majority of 3g or 2g Customers.

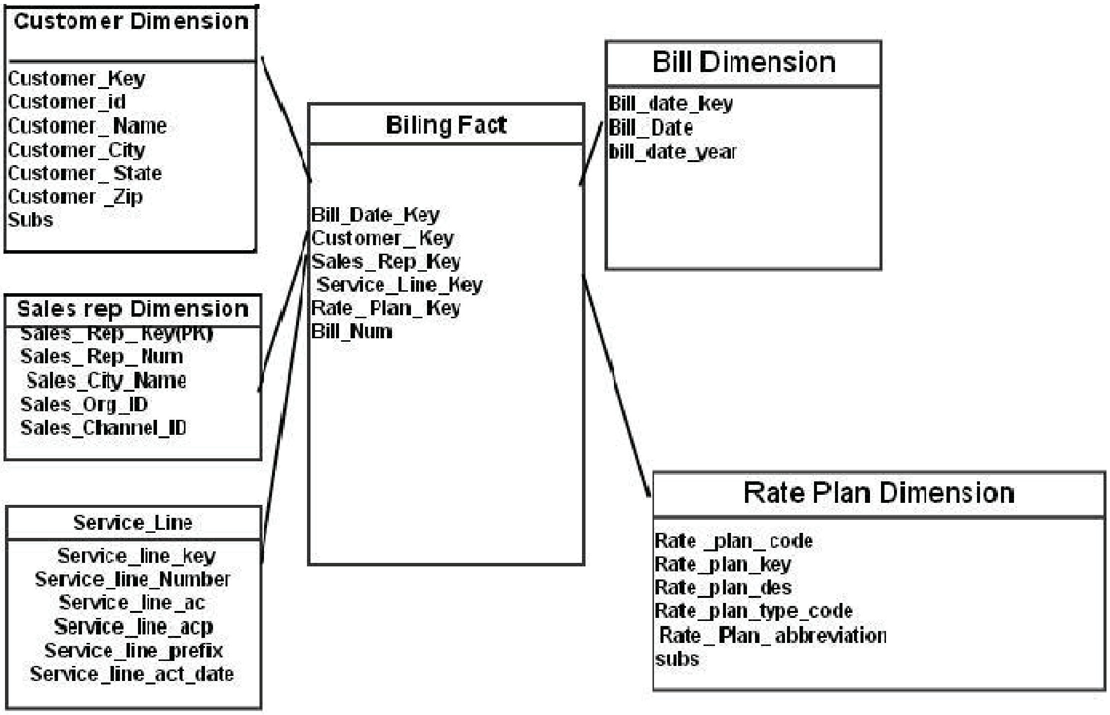


Fig 4.1 Star-Schema

The fact table contains the Billing figures for each bill transaction and the foreign keys that connect it to the five dimensions: Bill, Customer, Sales_rep, Service_line, Rate_Plan. Standard association-rule mining discovers correlations among items within transactions (the prototypical example of utilizing association-rule mining is determining what things are found together in a basket at a checkout line at the supermarket; hence the often used term: market basket analysis). The correlations are expressed in the following form:

Transactions that contain A are likely to contain B as well

Letters A and B represents set of items. There are two important quantities measured for every association rule: *support* and *confidence*. The support is the fraction of transactions that contain both A and B. The support measures the significance of the rule, so we are interested in rules with relatively high support. The confidence is the fraction of transaction containing A, which also contains B. The Confidence measures the strength of the correlation, so rules with low confidence are not meaningful. In the second phase, the association rules among the frequent item sets with high confidence are constructed. The star schema above represents the multi-dimensional model of telecommunication data warehouse.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

Following section describe the implementation of extended association rules with in Telecommunication data warehouse. The implementation is quality independent as it can accommodate both transaction and non-transaction level data. The basic architecture of our system is shown in Figure 5.1.

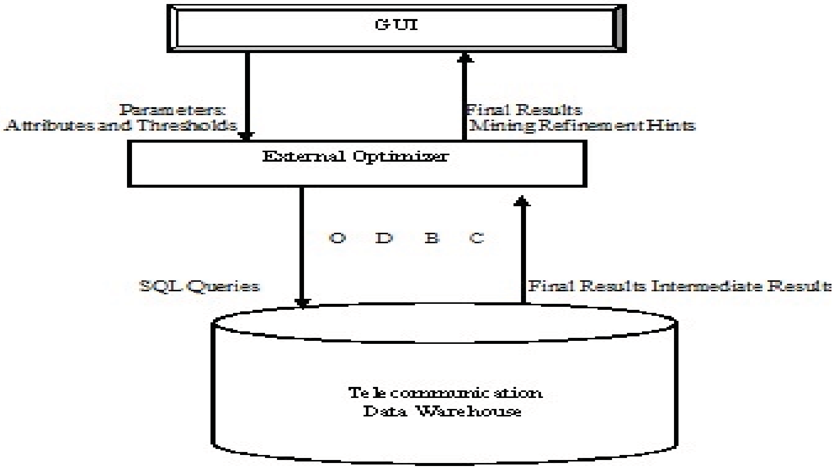


Fig 5.1. System Architecture for the Data Mining System

The most notable feature of the system architecture is the tightly coupled integration with the relational database that powers the data warehouse.

VI. OPTIMIZATION ALGORITHM

The algorithm involves several different parameters that are database specific. The following example illustrates the general optimization principles without going into system specific issues.Example 2: Consider mining extended association rules from the data warehouse shown in figure 1 based on the following question: Find the list of customers who are living in city, fill in service _line_ac having rate plan in the month of Bill_date. This question involves four tables, namely Customer, Service_line, Bill, Rate Plan, and Billing Fact. Typically, the size of a fact table (such as Billing Fact) will be several orders of magnitude bigger than the size of any of the dimension tables. In order to make the example concrete, suppose that the sizes and attribute cardinalities for these tables are as follows:

- Table Customer, Service_line, Bill, Rate_Plan has 10 thousand tuples (records)
- Table Billing Fact has 1 lake tuples

Furthermore, suppose that the support threshold is 70.

The cost of this join is likely to dominate the cost of the mining process so to optimization is to reduce the size of the portion of Billing Fact before we do the self-join. The optimization is crucial for the success of our tightly coupled approach. Mining without such optimization will be slow and may require massive additional storage space for the internal intermediate result. The completely different approach to solving the problem might be to aggregate the admin and discharge rows together in a single pass through the table and then compute the results from that. The query might be written like as:

```
 Select cust_city, S.service_line_ac, R.rate_plan, b.bill_date

  From bill_fact f INNER join customer C ON f.cust_id=c.cust_id

  INNER join Service_line S  On f.service_line_key=S.service_line_key

  INNER join rate_plan R ON f.rate_plan_code=r.rate_plan_code

INNER join bill_date b ON f.bill_date_key=b.bill_date_key

Where b.bill_date='1-jan-2010'

And R.rate_plan='2G'
And S.service_line_ac=P199

Group by

c.cust_city, S.service_line_ac, r.rate_plan, b.bill_date
```

This query will scan through the table once and compute this result in O (n) operation which makes the better performance of the execution of query.

VII. EXPERIMENT

Now for implementing which plan or which mode of service customers are using i.e. 3G or 2G, following experiment has been performed based on star schema of telecommunication data warehouse .Following query has been performed to check the most preferred network In the given query below, the query is performed to see the number of 3g customers in the telecommunication data ware house. The query is performed on oracle 10g and the result which we obtained is 1009 records.

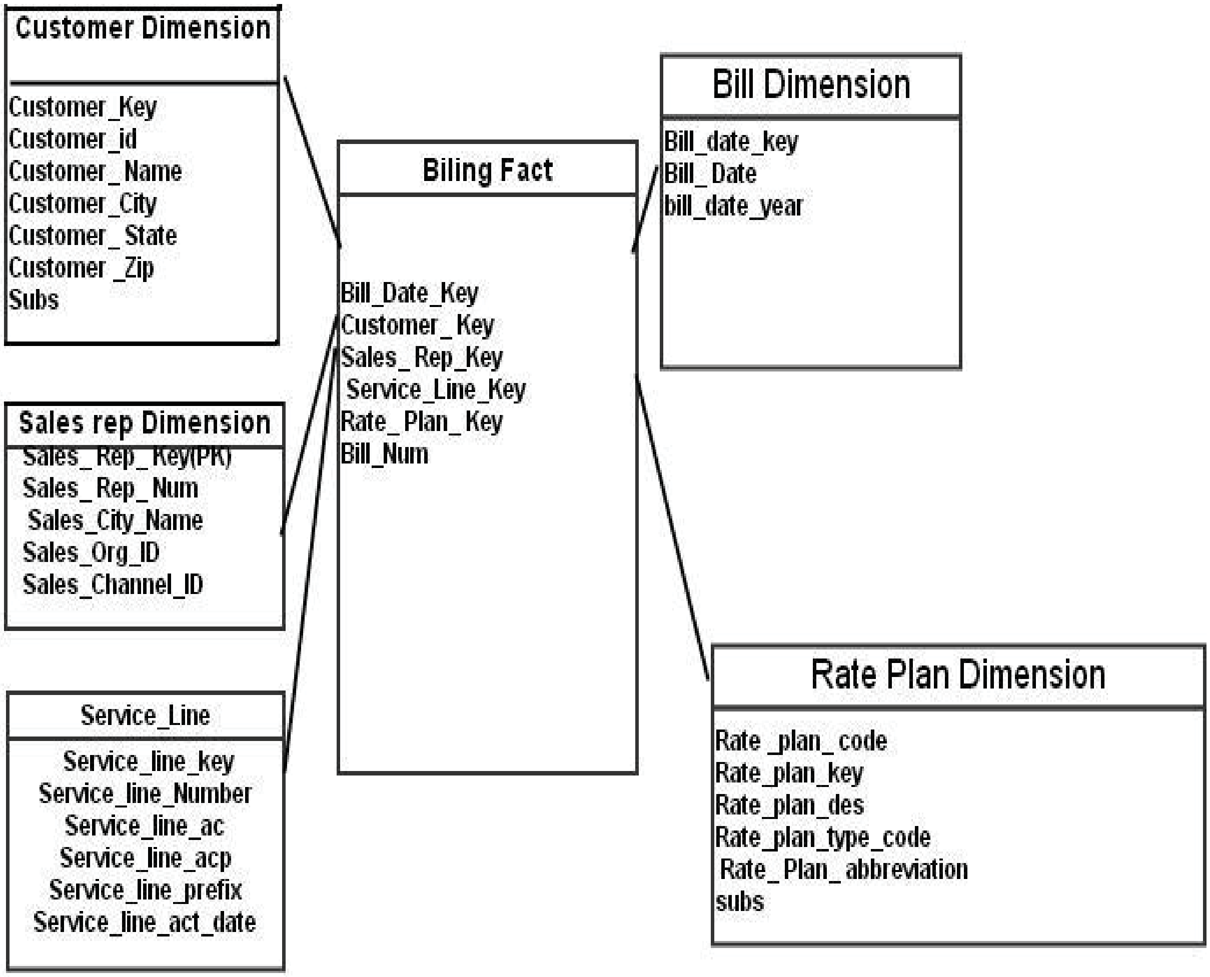


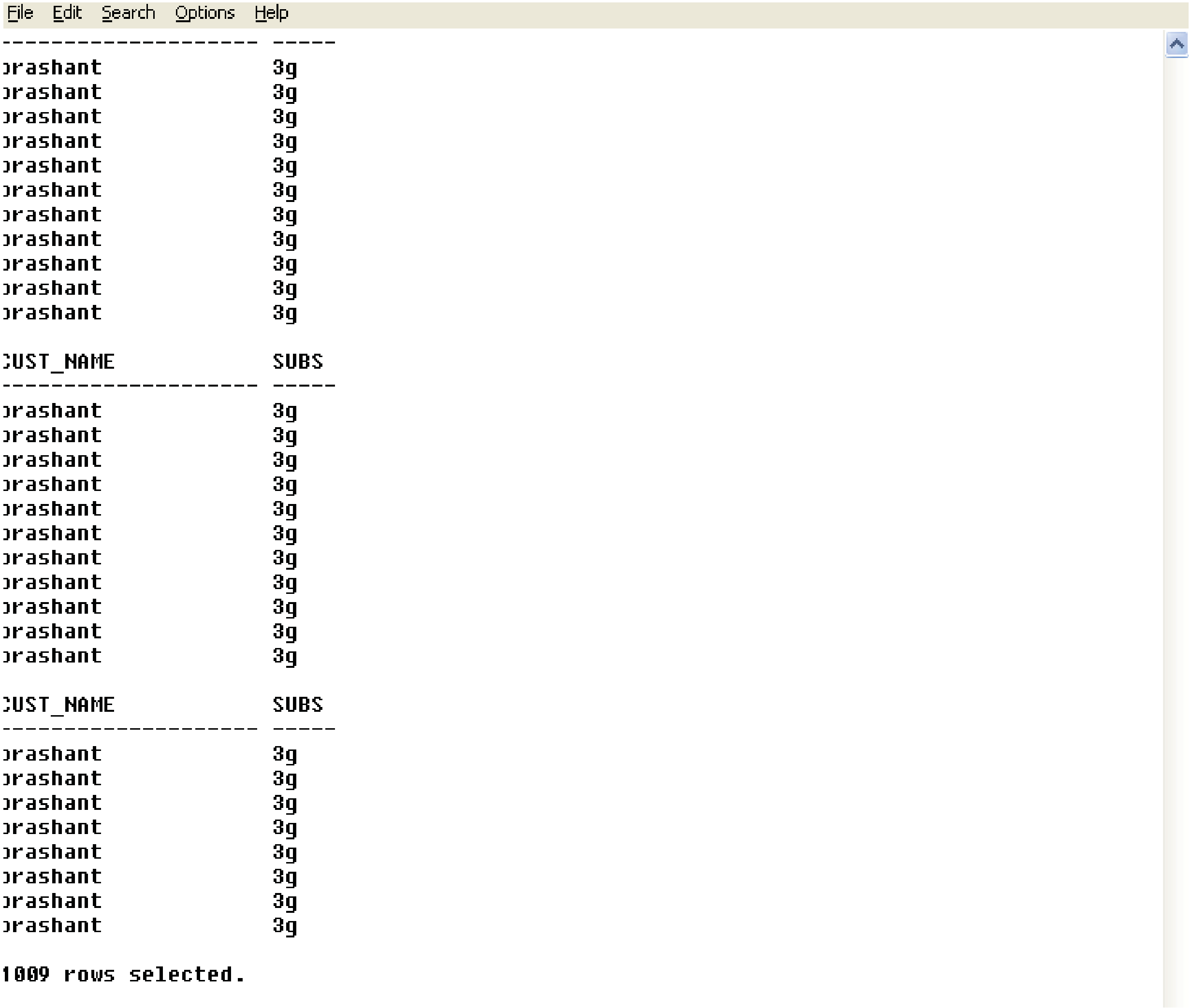
Fig.7.1 (a)Star Schema

The name of the customer is same in the given output because there may be more than one customers of same name .But customer Key is different in this case.

1. *select cust_name, subs from customer where subs='3G'*;

Similarly, query2 describe about number of 2g customer in the telecommunication data ware house. The query is performed on oracle 10g and the result which we obtained is 1009 records .The name of the customer is same in the given output because there may be more than one customers of same name .But customer Key is different in this case.

The result of following query is given below:-



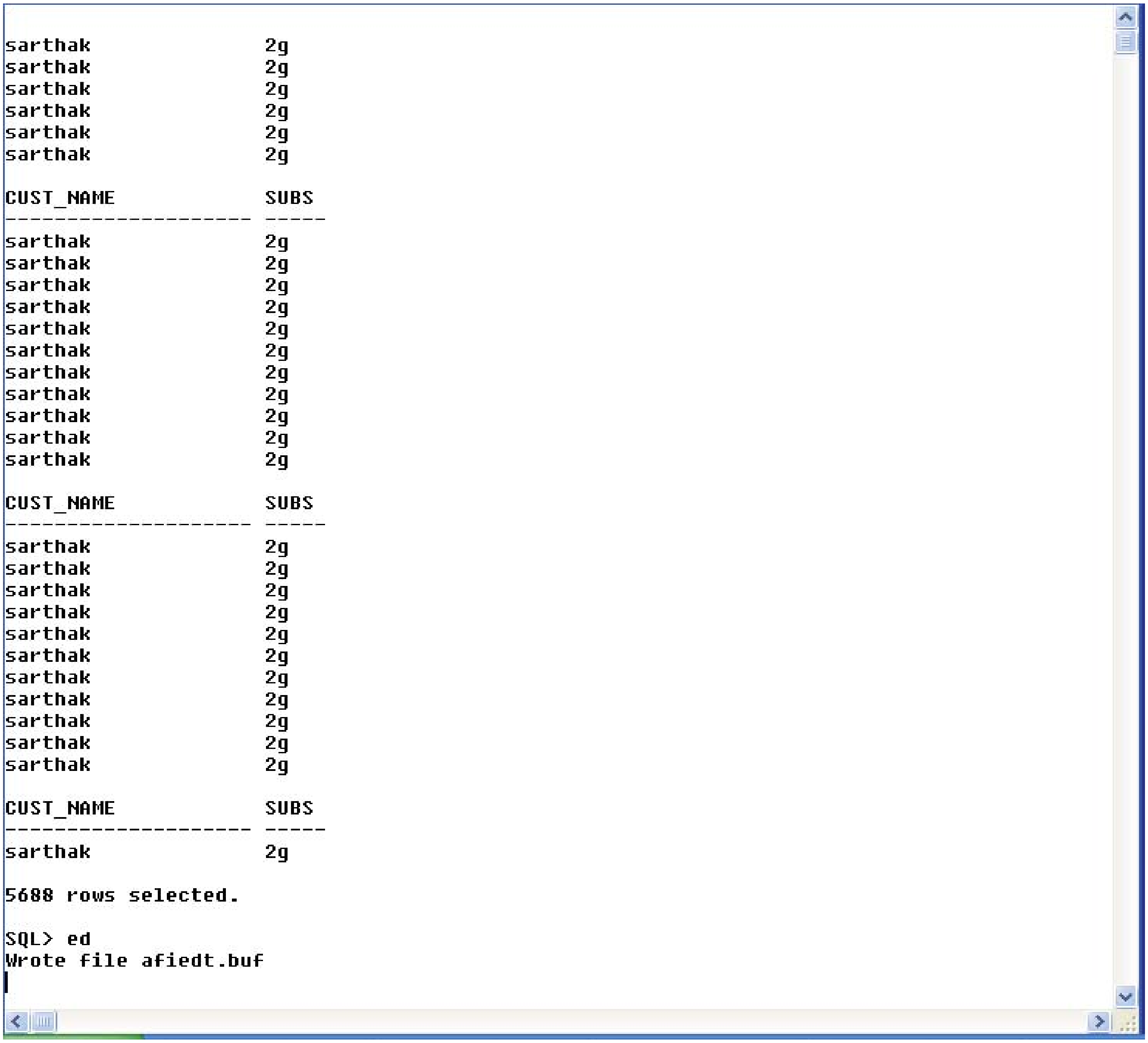
```
-------------------- -----
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g

UST_NAME             SUBS
-------------------- -----
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g

UST_NAME             SUBS
-------------------- -----
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g
rashant              3g

1009 rows selected.
```

Fig.7.1 (b)Analysis1

2. *select cust_name,subs from customer where subs='2g';*

Here from the above two queries we can clearly see that number of 2g customers are the number of 3g Customers that is 1009(3g customers) and (5688 2g customers) are there.

The result of following query is given below:-



```
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g

CUST_NAME            SUBS
-------------------- -----
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g

CUST_NAME            SUBS
-------------------- -----
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g
sarthak              2g

CUST_NAME            SUBS
-------------------- -----
sarthak              2g

5688 rows selected.

SQL> ed
Wrote file afiedt.buf
```

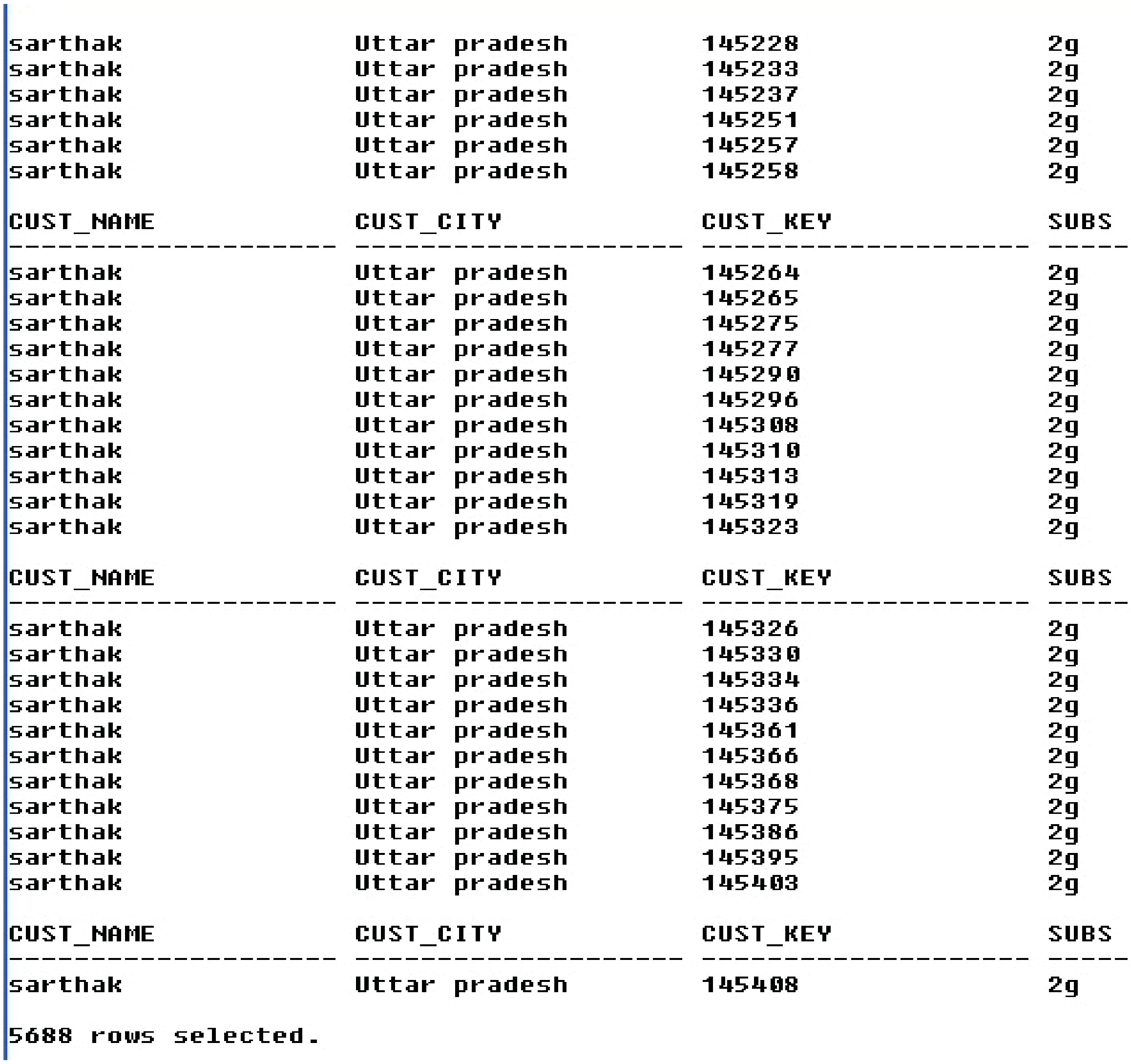
Fig.7.1(c)Analysis2

3. *select cust_name,cust_city,cust_key,subs from customer inner join rate_plan using(subs);*

Query third show the comparison between number of 2g and 3g customers according to the 2g and 3g rate plan. So after performing this we get the maximum number 2g customers as compared to 3g customers here, it can be clearly seen that number of 2g customers are more as compared to number of 3g customer according to rate plan. The output can't show full since oracle buffer show last 100 records only on the screen.

Customer key in this case is different for each customer as compared to customer name because in this case customers name might be same but customer key can't be same because it is assigned by the company itself to each customer. So after analyzing final query, the result obtained is that the 2g customers are more as compared to 3g customer in current scenario of telecommunication industry. In query it showed number of distinct people preferring the 3g and 2g network. Cost factor making the difference between the consumption of 2g.

The result of following query is given below:-



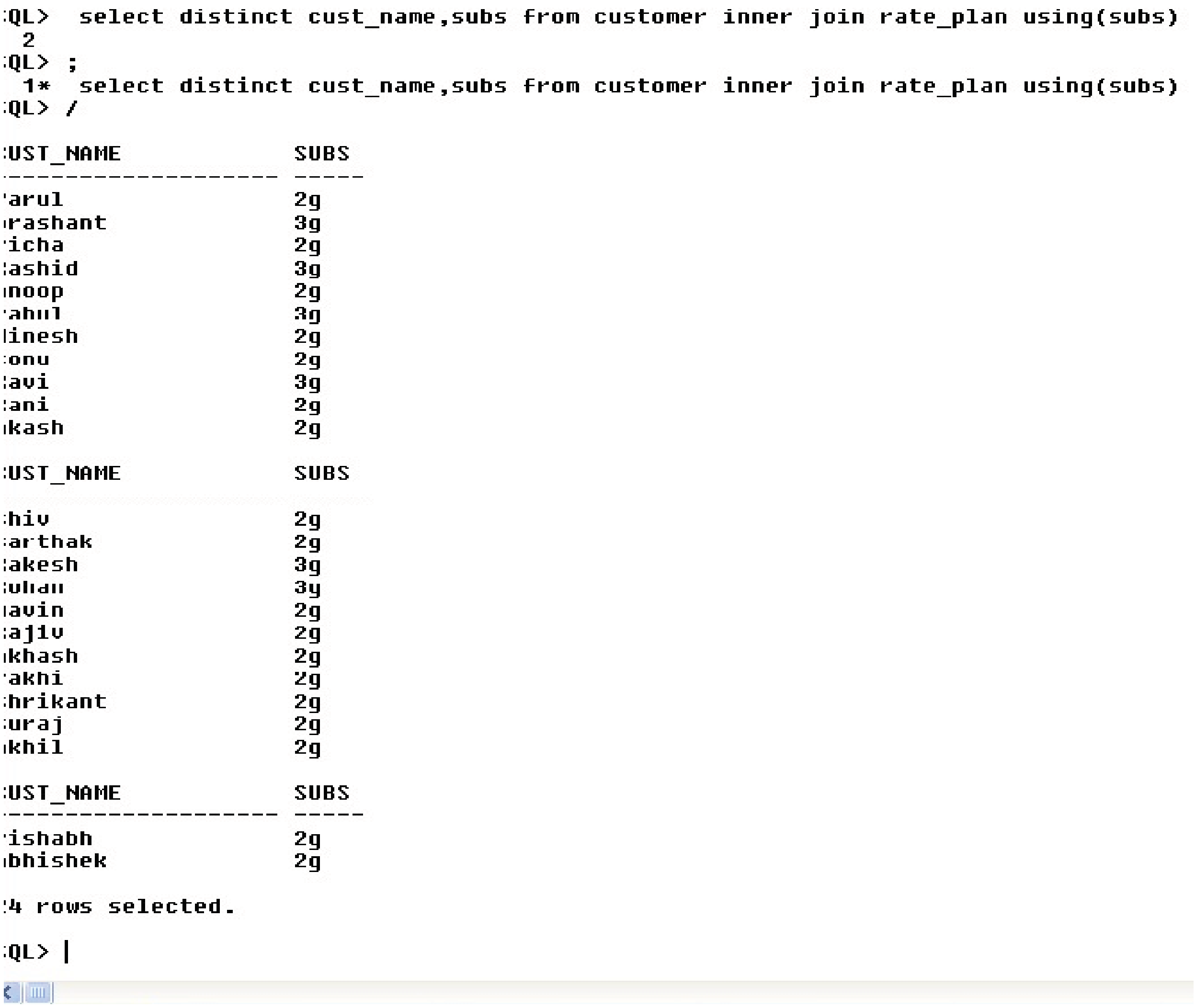
```
sarthak              Uttar pradesh        145228               2g
sarthak              Uttar pradesh        145233               2g
sarthak              Uttar pradesh        145237               2g
sarthak              Uttar pradesh        145251               2g
sarthak              Uttar pradesh        145257               2g
sarthak              Uttar pradesh        145258               2g

CUST_NAME            CUST_CITY            CUST_KEY             SUBS
-------------------- -------------------- -------------------- -----
sarthak              Uttar pradesh        145264               2g
sarthak              Uttar pradesh        145265               2g
sarthak              Uttar pradesh        145275               2g
sarthak              Uttar pradesh        145277               2g
sarthak              Uttar pradesh        145290               2g
sarthak              Uttar pradesh        145296               2g
sarthak              Uttar pradesh        145308               2g
sarthak              Uttar pradesh        145310               2g
sarthak              Uttar pradesh        145313               2g
sarthak              Uttar pradesh        145319               2g
sarthak              Uttar pradesh        145323               2g

CUST_NAME            CUST_CITY            CUST_KEY             SUBS
-------------------- -------------------- -------------------- -----
sarthak              Uttar pradesh        145326               2g
sarthak              Uttar pradesh        145330               2g
sarthak              Uttar pradesh        145334               2g
sarthak              Uttar pradesh        145336               2g
sarthak              Uttar pradesh        145361               2g
sarthak              Uttar pradesh        145366               2g
sarthak              Uttar pradesh        145368               2g
sarthak              Uttar pradesh        145375               2g
sarthak              Uttar pradesh        145386               2g
sarthak              Uttar pradesh        145395               2g
sarthak              Uttar pradesh        145403               2g

CUST_NAME            CUST_CITY            CUST_KEY             SUBS
-------------------- -------------------- -------------------- -----
sarthak              Uttar pradesh        145408               2g

5688 rows selected.
```

Fig.7.1(d) Analysis3

4. *select distinct cust_name,subs from customer inner join rate_plan using(subs);*

This query show the name of number of distinct customer which shows that there are different customers exists along with their mode of using the plan.On the basis of following query, following analysis can be made. In query 1 and query 2, result shows that the number of 2g and 3g customers along with their names. The customers might have same name but have different customer id and customer key as allocated by telecommunication companies. In the query 3 and query 4, result shows the customer name, customer city or state, customer key and subscriber detail that is whether they are 2g or 3g customers. So, result is that there are more 2g customers as compare to 3g customers.

The result of following query is given below:-



```
:QL>  select distinct cust_name,subs from customer inner join rate_plan using(subs)
  2
:QL> ;
  1*  select distinct cust_name,subs from customer inner join rate_plan using(subs)
:QL> /

:UST_NAME              SUBS
-------------------- -----
'arul                  2g
ırashant               3g
'icha                  2g
:ashid                 3g
ınoop                  2g
'ahul                  3g
|inesh                 2g
:onu                   2g
:avi                   3g
:ani                   2g
ıkash                  2g

:UST_NAME              SUBS
-------------------- -----
:hiv                   2g
:arthak                2g
:akesh                 3g
:uhan                  3g
ıavin                  2g
:ajiv                  2g
ıkhash                 2g
'akhi                  2g
:hrikant               2g
:uraj                  2g
ıkhil                  2g

:UST_NAME              SUBS
-------------------- -----
'ishabh                2g
ıbhishek               2g

:4 rows selected.

:QL> |
```

Fig.7.1(e) Analysis4

VIII. CONCLUSION

Through this paper, it can be concluded that the different operator who are providing Telecommunication services exist into the market on the basis of quality and cost. Through ARM it has been observed which operator is preferred most frequently. For different Internet Plans the subscriber prefer

a) 2G network (Urban) more frequently comparatively than 3G due to reliability and affordability.
b) Those who prefer 3G network they look for better streaming and robustness.

The result after querying telecommunication data warehouse is that 2g consumers are more as compared to 3g customers in specific area. Due to the rate plan which is cheap as compared to 3g.

Further, there is need to introduce a method which optimizes by reducing the cumbersome and time consuming approaches for queries analysis.

REFERENCES

[1] Online Diagram Making Tools (https://www.gliffy.com/).

[2] The data warehouse toolkit: the complete guide to dimensional modeling by Ralph Kimball, Margy Ross. — 2nd ed. (ISBN 0-471-20024-7).

[3] Agrawal R., Imielinski T. and A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proceeding of ACM SIGMOD International Conference. (1993), 207-216.

[4] Svetlozar Nestorov and Ad-Hoc Association-Rule Mining within the Data Warehouse Proceedings of the 36th Hawaii International Conference on System Sciences - 2003

[5] Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber, Jian Pei-2nd ediion(ISBN-9780080475585)

[6] Agrawal R. and Srikant R. Fast Algorithms for Mining Association Rules. Proceeding of International Conference On Very Large Databases VLDB. (1994), 487-499.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.Verkamo. Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996).

[8] Berry, M. and Linoff, G. Data Mining Techniques for Marketing, Sales and Customer Support. Wiley (1997).

[9] Chaudhri, S. and Dayal, U. An overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26 (1), (1997), 65-74.

[10] Hilderman, R.J., Carter, C.L., Hamilton, H.J., and Cercone, N. Mining Association Rules from Market Basket Data Using Share Measures and Characterized Itemsets. International Journal of Artificial Intelligence Tools. 7 (2), (1998), 189-220.

[11] Inmon, W. H. Building the Data Warehouse. Wiley (1996).

[12] Kimball, R., Reeves, L., Ross, M., and Thornthwhite, W.The Data Warehouse Lifecycle Toolkit. Wiley (1998).

[13] Leavitt, N. Data Mining for the Corporate Masses. IEEE Computer. 35 (5), (2002), 22-24.

[14] S. Sarawagi, S. Thomas, R. Agrawal. Integrating Mining with Relational Database Systems: Alternatives and Implications. Proceedings of ACM SIGMOD Conference (1998), 343-354

[15] S. Tsur, J. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, A. Rosenthal. Query Flocks: A Generalization of Association-Rule Mining. Proceedings of ACM SIGMOD Conference, (1998), 1-12.

[16] Wang, K., He Y., and Han J. Mining Frequent Itemsets Using Support Constraints. Proceedings of International Conference on Very Large Databases VLDB, (2000), 43-52

[17] Watson, H. J., Annino, D. A., and Wixom, B. H. Current Practices in Data Warehousing. Information Systems Management. 18 (1), (2001).