INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# FEASIBILITY OF FUZZY DATA WAREHOUSE IMPLEMENTATIONS

**PPG Dinesh Asanka[1], Dr. Amal Shehan Perera[2]**

[1]*Department of Computer Science and Engineering, University of Moratuwa, Moratuwa , Sri Lanka*
*dineshasanka@gmail.com*
[2]*Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka*
*shehan@cse.mrt.ac.lk*

**Abstract: -** Many organizations are using Data warehouse as a strategical decision-making tool. Over the years, major emphasis was placed to analytics purposes. Uncertainty is one of the major challenge faced in many areas such as data warehouse, database etc. In case of uncertainty, fuzzy technique can be utilized. This research paper is to carry out a feasibility study whether is the fuzzy data warehouse implementation would be a viable solution. Also, this research paper explorer the domains where fuzzy data warehousing can be implemented.

**Index Terms** — Fuzzy Systems, Data Warehouse, De-Fuzzification

## 1. Introduction

A multidimensional data model is typically organized around a central theme transaction. The basic components of a multidimensional model are fact and dimensions. The fact table consists of the measurement, matrices or facts of a business process. It provides the additive or semi-additive values that acts as an independent variable by which dimensional attributes are analyzed. The facts are numerical measures like quantity, amount, count, etc. [Abello A., Samos J., Saltor F., 2006]. Data warehouse is used in many domains to improve analytical capabilities in those domains to make strategical decisions for organizations to sustain in the aggressively competitive market environment.

Use of data warehouse in agriculture for providing information and recommendation on efficient utilization of fertilizer [Chai K.L et all, 1994], reducing herbicide use [Coble H.D., 1994], plant protection [Jørgensen L.N., 2007], variety-specific information [Detlefsen N.K., Jensen A.L., 2004], management of environment risks [Bazile D., 2005], Integrated Nutrient Management [Mosseddaq F., 2007], Forest Management [Riberir R.P., Borges J.G., 2005], Crop disease control [Ibragimov T.Z. et all, 2005], Agricultural practices and extension [Sodtke R., 2005], Farm mechanization [Suarez de Cepeda M., 2005], Seasonal Climate Prediction [Hansen J.W., 2002], Integrated Pollution Control [Kseniya L., 1999], labor requirements and land use planning [Matthews K., Buchan K., 2003], profit maximization and risk minimization [Meinke H., 2001] etc., have been designed and implemented successfully, mainly for improving economic returns, changing farming practices or minimizing environmental risks.

Data warehouse plays an important role in CRM, enabling the assembly of customer information and the creation of customer knowledge, which is critical for maintaining customer relationships [Cooper, B.L, 2000]. In case of banking industry, there are few research done with respect to data warehouse [Chavda M.B., Solanki A. 2014] [Gadda K.R., Dey S., 2014],

Fuzzy logic is used to mitigate uncertainty in many domains such as agriculture [Ashraf, A., Akram, M., Sarwar M., 2014], medicine [Sapna S., Pravin K., 2015] [Dong W., Huang Z., Duan H., 2014], power systems [Wang Q., Wang Y., Zhang H., Sun Y., 2016], production [Bhargava A.K., Singh S.R., 2014], sports [Zeng W., Li J, 2014], transportation [Wang Y.C., Chien C.J., Wang C.H., 2016] etc.

As mentioned before in the research paper, the fuzzy techniques can be implemented into many areas such as agriculture, sports, finance, human resources and also there are many other domains where data warehouse is implemented. The Data Warehouse Toolkit [R. Kimbal., M. Ross, 2002], [ R. Kimbal., M. Ross, 2013] is a main source for several research articles as it contains several real-world cases in different domains. Since second and third editions have focused on different business domains, these books are used as references for possible implementations of data warehouses. In these books, Retail Sales, Inventory, Procurement, Order Management, Customer Relationship Management, Accounting, Human Resources Management, Financial Services, Telecommunication, Transportation, Education, Health Care, Insurance sectors are discussed in detail. Technically, it is possible to implement fuzzy techniques for data warehouse as discussed in several research papers [Sapir L., Shmilovic A., 2008] [Delgado M., et. al. 2004]. By going through the cases of these books, feasibility of fuzzy data warehouse implementation is analyzed.

## 2. Retail Sales

Retail Sales are one of the most common business scenario for many technologies. Data warehouse are being used to enhance business value not only for Retail Sales but also for many other data warehouse implementation date dimension is used.

In Date Dimension of a Retail Sales, Selling Season column is set to the name of the retaining season. Seasons can be one of, Thanks giving, Easter, Valentine's Day, Christmas, Fourth of July, etc. Since season has a long date range, sales will be different, when it is closer to the season date. Also, if there are holidays/weekend, sales will be in a different order. Normally, seasons starts in advance. For example, though the Christmas day falls on December 25th, retail sales will start around end of November. Then sales increase gradually towards Christmas day and weightage will be different if it is a weekend. Hence, the seasonal factor will be different from time to time. The factor should be derived from using fuzzy techniques.

Also, in the promotion dimension, promotion impact is measured too. Proposed design from Kimball is to measure the impact. However, there are different ways of calculating the promotion impact as listed in Appendix E. Due to the different types of calculation, promotion impact can be a fuzzy dimension attribute.

Experience is key attribute in clerk dimension where it will be used to find sales depending on the experience of the clerk. Experience is stored as number of years. However, experience can be categorized as high, medium and low which can be treated as a fuzzy attribute. Also, linguistic fuzzy attributes can be defined to perform more analytics like very old, Not young AND old etc.

In product dimension, products are typically categorized to sub categories, categories and groups. These groups are more likely to be a crisp group. However, in reality, one product may belong to multiple categories with different weight, which needs to be handled with fuzzy concepts.

## 3. Inventory

Every inventory transaction identifies the date, product, warehouse, vendor, transaction type etc. In most cases, a single amount representing the inventory quality impacts caused by the transaction. Star schema for the inventory is shown below and the Warehouse Inventory Transaction Fact table is maintaining a granularity of one row per inventory transaction as shown in Figure 1 below.

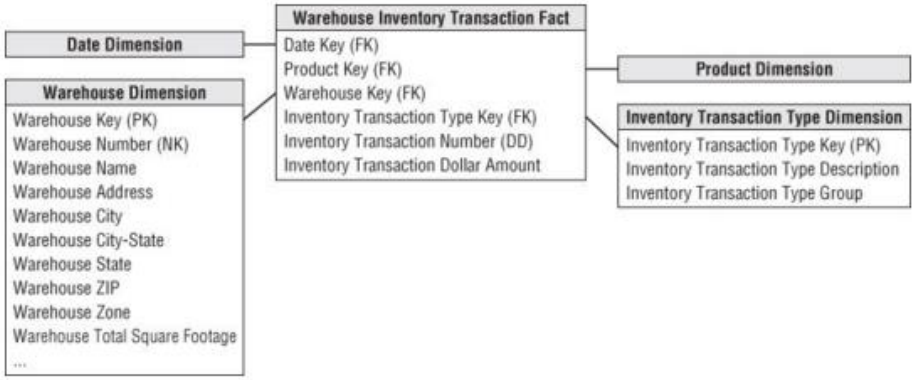


Figure 1: Table Structures for Fact Less Fact Table

In the above model, Warehouse Total Square Footage can be treated as fuzzy attributes so that the store warehouse size can be analyzed as size of the warehouse is an important attribute when deciding the storage of the goods.

In this fact table, there are measure columns such as Quantity Received, Quantity Inspected, Quantity Returned to Vendor, Quantity Placed in Bin, Quantity Shipped to Customer, Quantity Returned to Customer, Receipt to inspected Lag, Receipt to Initial Shipment Lag, Initial to Last Shipment Lag which needs to be analyzed for comparison. When these measures are analyzed, they will be set to a crisp set. For example, lag measures are important to analyze delays between stages. Typically, these lags are measured via crisp values such as high, medium and law. These can be converted to fuzzy attributes so that precious analysis can be done.

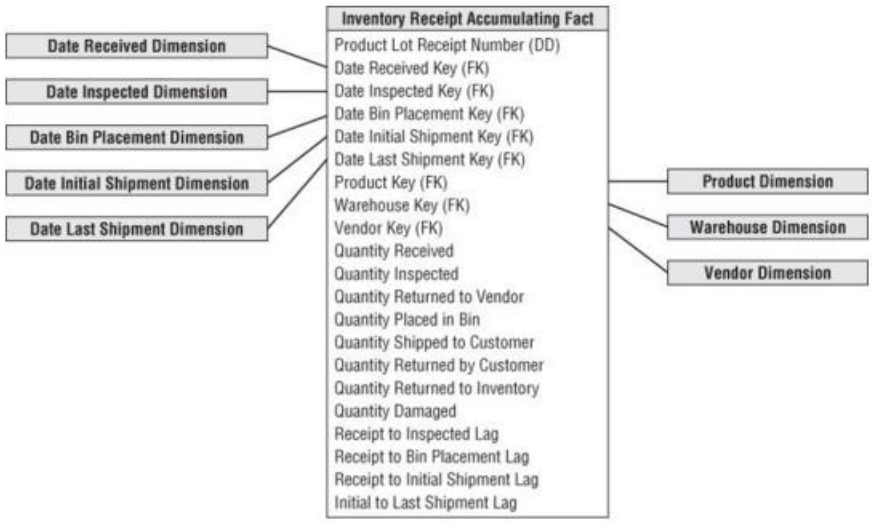


Figure 2: Data Model for Inventory Receipt Accumulating Fact Table

Apart from measure columns, in the product dimension there are some products which falls into multiple product categories. For example, some products fall under Appliance and Apparel. This means that there is an uncertainty which needs to be handled by means of fuzzy techniques. In this scenario, products are categorized into multiple product categories with different weightages. Fuzzy techniques need to be designed to calculate different weightages. These weightages may be different depending on the season and business which will make calculation of weightages are complex.

## 4. Order management

Order management is a common business scenario for data warehousing as almost every business needs order management. In case of order management, data warehouse, customer dimension contains important attribute

called Customer Credit Rating which will define how good customer's credit rating is. This is again a crisp value set which can be configured as a fuzzy attribute so that credit rating can be measured to higher accuracy. Similar to promotion dimension in the retail sales warehouse, there is a dimension named Deal in order management. Therefore, similar to promotion n dimension, deal dimension can be converted to fuzzy dimension.

Also, important matrix in the Order Management is Customer Satisfaction. Business users often are interested in customer satisfaction metrics such as whether the line item was shipped on time, shipped complete or damage free. These are added to the fact table for each line item level satisfaction metrics. These new fact columns are populated with additive one and zeros. By using these three fact measures customer satisfaction junk dimensions were created. All three of these parameters can be fuzzy attributes and by implementing rules and De-fuzzification techniques customer Satisfaction attribute can be derived.

## 5. Customer Relationship Management

In Customer Relation Management (CRM), customer segmentation plays a huge role so that better relationship with customer can be maintained. Some of the most powerful attributes in a customer dimension are segmentation classification or scores. These attributes are gentle by business context. Gender, Ethnicity, Age, Income, Status, Referring Source are the important factors among many other attributes when defining customer segmentation. However, since Age and Income are fuzzy attributes, customer segmentation will become fuzzy attributes. Like in the Customer Order Management, defining rules and de-fuzzification techniques, function of customer segmentation can be defined.

Recency is another important factor indicating whether the given customer is an important customer to the organization. One popular approach for scoring and profiling customers looks at the Recency (R), frequency (F), and Intensity (I) of the customer behavior. These are known as the RFI measures; sometimes intensity is replaced with monetary (M), so it's also known as RFM. Recency is how many days has it been since the customer last ordered or visited your site. Frequency is how many times the customer has ordered or visited, typically in the past year. And intensity is how much money the customer has spent over the same time period. When dealing with a large customer base, every customer's behavior can be modeled as a point in an RFI cube as depicted. Typically, Recency is decided by Date of last purchase, frequency (transaction count), and intensity (net purchase amount). Since frequency and intensity are fuzzy attributes, Recency will become a fuzzy attribute. Like in the Customer Order Management, defining rules and de-fuzzification techniques, function of Recency can be defined which is fuzzy attribute itself.

County Demographics Out trigger dimension can be introduced to identify different county level attributes as shown in the below figure.

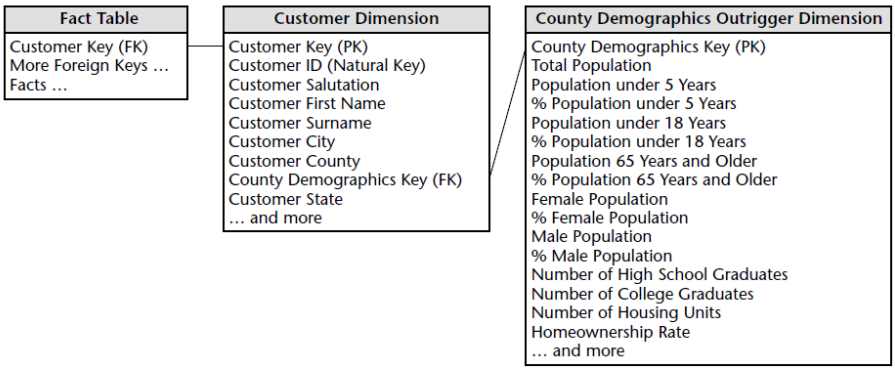


Figure 3: County Demographics Out trigger Dimension

Table 1 shows different ranges for age and income level for analysis. In classical data warehouse analysis, continuous variable into ranges are called bucketized in data warehousing so that better analysis can be done.

Table 1 : Range Values

| Demographic Key | Age | Gender | Income Level |
|---|---|---|---|
| 1 | 20 – 24 | Male | < $ 20,000 |
| 2 | 20 – 24 | Male | $20,000 - $ 24,999 |
| 3 | 20 – 24 | Male | $25,000 - $ 29,999 |
| 18 | 25 – 29 | Male | $20,000 - $ 24,999 |
| 19 | 25 – 29 | Male | $25,000 - $ 29,999 |

Both age and Income level columns are fuzzy columns as they have data in the form of ranges as shown in the above table. Customers can be categorized using Age, Gender and Income Level. Customer Categorization too can be a fuzzy variable and de-fuzzification techniques can be utilized to find the customer category. This can be represented in a following de-fuzzification model.

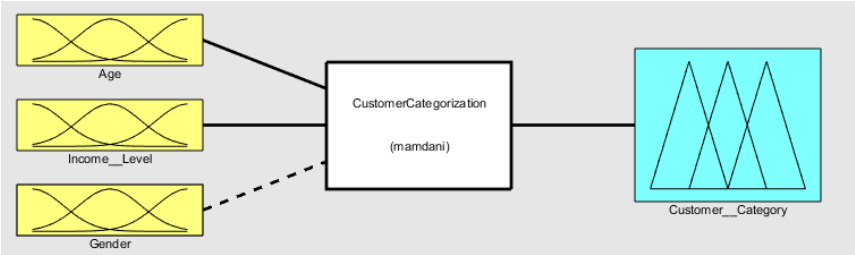


Figure 4: Extended Customer Dimension

In a CRM system, customer behavior is an important key performance indicator to measure in which design is shown in the below figure.

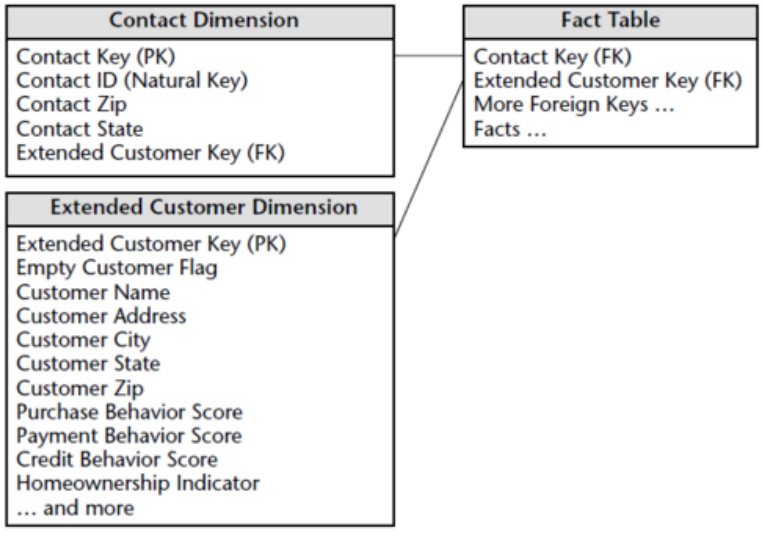


Figure 5: Extended Customer Dimension

In this data model, Purchase Behavior Score, Payment Behavior Score, Credit Behavior Score can be considered as KPIs which are candidates for fuzzy attributes. Rather than analyzing these in bucketized or as range values, by introducing fuzzy attributes, better analysis can be done.

## 6. HUMAN RESEOUCE Management

Human resource management (HRM) is a key area to compete with other organizational systems. In most of the organizations, Human Resource Management system is the heart of organization.

Skills of employees are stored and analyzed in a HRM system. Since each employee will have a variable, unpredictable number of skills, the skill dimension is a prime candidate to be a multivalued dimension keywords by their nature, usually are open ended. New keywords are created as shown in the below figure.

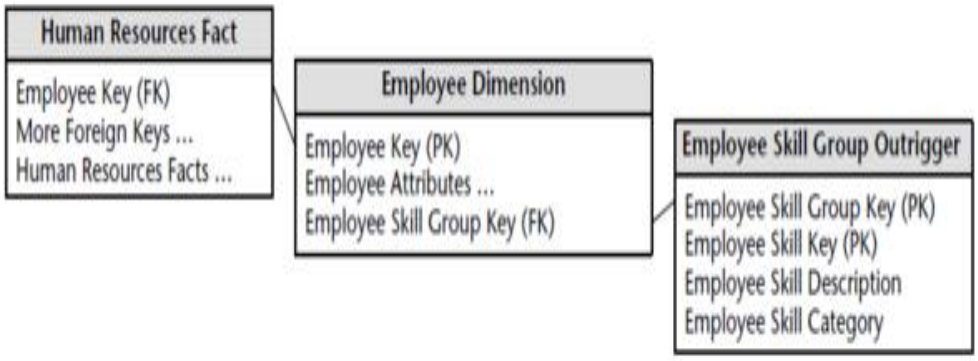


Figure 6: Employee Skill Group Dimension

Since an employee has multiple skills, there is an uncertainty about the skill levels. Therefore, skill is an excellent parameter to consider a fuzzy attribute.

Apart from skills, there is another important fact table named head count snapshot fact table in the HRM system which is shown below figure.

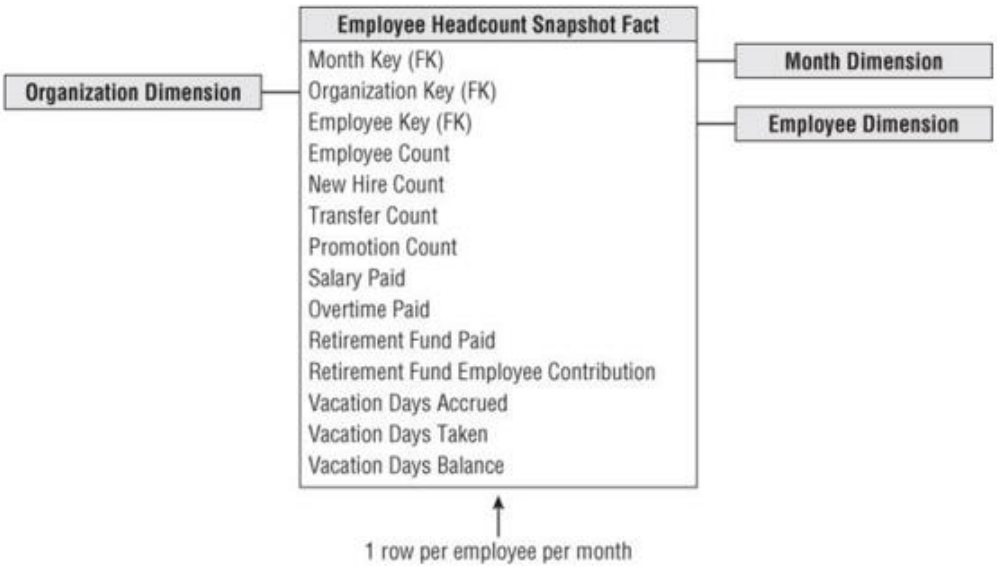


Figure 7: Data Model for Employee Headcount Snapshot Fact Table

In the above fact table, most of the measures are count for a given month for one employee. When it comes to analysis, typically range values will be used. However, the better option would be changing the ranges to the fuzzy dimensions for better analysis.

## 7. Financial Services

Financial service is one of the main sectors which have been benefitted from the data warehouse technology over the years. The following data model shows example for the finance service which maintains customer account balances in a bank.

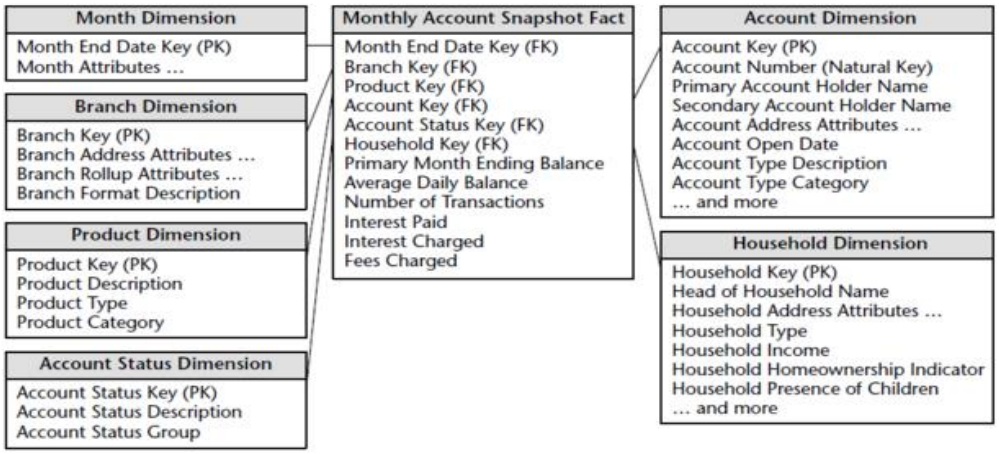


Figure 8: Data Model for Monthly Account Snapshot Fact Table

Household Income will be a candidate for fuzzy attribute as it can be analyzed as low, medium and high. Month Ending Balance, Interest Paid, Charged, and number of transaction measures can be converted to range dimensions for better analysis. Therefore, those columns are candidates for fuzzy attributes.

Apart from the bank account balance, there are other important measures in the bank such as number of deposits, ATM usages, etc. which the bank management will be interested at. Those fact tables are shown in the below data model diagram.

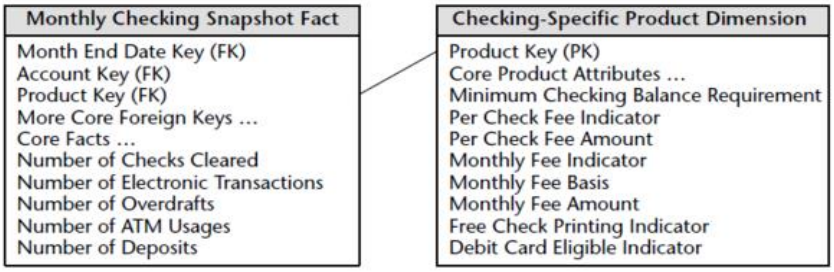


Figure 9: Data Model for Monthly Checking Snapshot Fact Table

Those measures are better off when analyzing with ranges which are better candidates for fuzzy attributes.

## 8. Health Care

In healthcare domain, there can be multiple diagnosis which are currently handled by a weighting factor according to Ralph Kimbal.

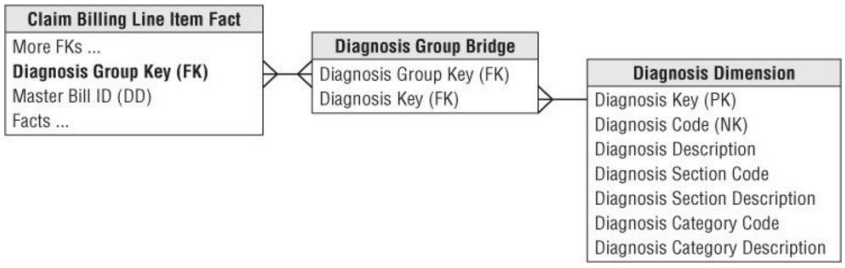


Figure 10: Data Model for Diagnosis Fact Table

If a patient has three diagnoses, he is assigned a diagnosis group with three corresponding rows in the bridge table.

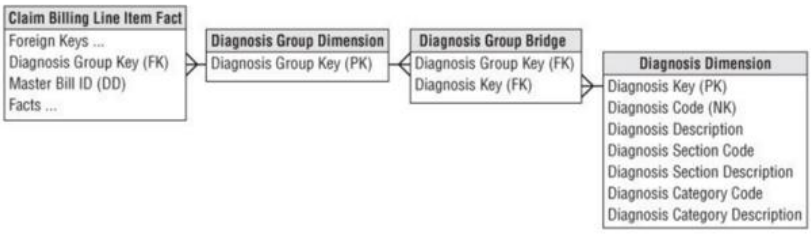


Figure 11: Data Model for Diagnosis Group Dimension Table

However, in the real world scenario, it is difficult to manage this with the crisp values. Better analysis and predictions can be done if it is converted into fuzzy data warehouse techniques.

## 9. Conclusion

This research paper looks options where fuzzy techniques can be implemented in data warehouse. Retail Sales, Inventory, Order Management, Customer Relationship management, Human Resource Management, Financial Services, Health Care cases studies from various research papers and main reference Kimball Ralph book were identified. In many cases, range dimensions can be treated as fuzzy data warehouse attributes. Also, weighted contributions such as shown in the health care example can be considered as fuzzy implementation. Also, by defining the rules, de-fuzzifications can be implemented in the data warehouse for better analysis.

## 10. REFERENCES

[1] Abello A., Samos J., Saltor F., Understanding facts in a multidimensional Object-Oriented Model. In Proc. of the 4th Int. Workshop on Data warehousing and OLAP (DOLAP), pages 32-39, ACM press, 2006.

[2] Bazile D., Le Page C., Dembélé S. & Abrami G., Perspectives of modelling the farmer's seed system for in situ conservation of sorghum varieties in Mali, 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment and 3rd World Congress on Computers in Agriculture and Natural Resources., 2005, Universidade de Trásos- Montes e Alto Douro, Vila Real, Portugal.

[3] Cooper, B. L., Watson, H.J., Wixom, B.H and Goodhue, D.L, "Data Warehousing supports corporate strategy at First American Corporation", MIS Quarterly, Minneapolis, vol.24, Issue 4, pp: 547-567, 2000.

[4] Chavda M.B., Solanki A., Innovative banking products: Win-Win situation for Customers and Banks, IRACST – International Journal of Commerce, Business and Management (IJCBM), ISSN: 2319–2828 Vol. 3, No.6, December 2014.

[5] Chai K.L., Costello T.A., Wells B.R., Norman R.J., App- eng-agric., 1994, 10, 849-855.

[6] Coble H.D., Proceedings of the annual meeting of the Northeastern Weed Science Society, 1994, 48, 155-159.

[7] Delgado M., Molina C., Sanchez D., Vila A., Rodriguez-Ariza L., A Fuzzy Multi-dimensional model for supporting imprecision in OLAP, 2004 IEEE international Conference on Fuzzy Systems, 3:1331 – 1336, 2004.

[8] Detlefsen N.K., Jensen A.L., Agricultural Systems, 2004,81, 55-72.

[9] Gadda K.R., Dey S., Business Intelligence for Public Sector Banks in India: A Case study- Design, Development and Deployment. Journal of Finance, Accounting and Management, 5(2), 37-58. ,2014.

[10] Hansen J.W., Agricultural Systems, 2002 74, 305-307. Ibragimov T.Z., Ibragimova I.T., Sanin S.S., Decision support systems for cereal crop disease control, 2005, EFITA/WCCA, Vila Real, Portugal.

[11] Jørgensen L.N., Noe E., Langvad A.M., Jensen J.E., Ørum J. E., Rydahl P., EPPO Bulletin, 2007, 37(2), 374-377.

[12] Kimbal. R., Ross M., The Data Warehouse Toolkit, The Definite Guide to Dimensional Modelling., Third Edition., John Wiley & Sons, Inc., 2002.

[13] Kimbal. R., Ross M., The Data Warehouse Toolkit, The Definite Guide to Dimensional Modelling., Second Edition., John Wiley & Sons, Inc., 2013.

[14] Kseniya L., Pollution Management Discussion Note: In focus, 1999, 3.

[15] Matthews K., Buchan K., Proceedings of the Modelling and Simulation Society of Australian and New Zealand, 2003, 4, 1534- 1539.

[16] Meinke H., Baethgen W.E., Carberry P.S. Donatelli M., Hammer G.L., Selvaraju R. & Stockle C.O. , Agricultural Systems, 2001, 70, 493-513.

[17] Mosseddaq F., Dnidane S., Lahlou M., Integrated Wheat N Nutrition Management in Morocco: A Decision Support Model, Progress of Information Technology in Agriculture Proceedings of the 4th International Symposium on Intelligent Information Technology in Agriculture (ISIITA). 2007.

[18] Riberir R.P., Borges J.G., MetaForest- a web-based decision system to support forest management involving multiple-ownership, EFITA/WCCA, 2005, Vila Real, Portugal.

[19] Sapir L., Shmilovic A., A Method for Design of a Fuzzy Data Warehouse, Department of Information Systems Engineering, Ben-Gurion University, Israel, 2008. Sodtke R., Multifunctional of Landscapes - Analysis, Evaluation and Decision Support, International Conference, Univ. Giessen, 2005, 193.

[20] Suarez de Cepeda M., Recio B., Rubio F., Decision Support System for Farm Mechanization, 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment and 3rd World Congress on Computers in Agriculture and Natural Resources, 2005, Vila Real, Portugal.