Data in Data Warehouse and its Qualities Issues

Arif Ali Wani, Bansi Lal Raina

Abstract*: Data quality (DQ) is as old as the data is. In last few years it is found that DQ can't be ignored during the process of data warehouse (DW) construction and utilization as it is the major and critical issue for knowledge experts, workers and decision makers who test and query the data for organizational trust and customer satisfaction. Low data quality will lead to high costs, loss in the supply chain and degrade customer relationship management. Hence to ensure the quality before using the data in DW, CRM (Customer Relationship Management), ERP (Enterprise Resource Planning)or business analytics application, it needs to be analyzed and cleansed. In this, we are going to find out the problem regarding dirty data and try to solve them.

***Index Terms*: Data warehouse, Data Quality, Customer Relationship management (CRM), Enterprise Resource Planning (ERP).**

I. INTRODUCTION

In recent years organizations become more dependent on the data they have collected in past which is increasing in massive amount as days go on [1]. Quality of data they have collected is one of the critical issues as you need to transfer the data from one system to another or you want to retrieve the data or you want to save the data. When you are making a DW project you need to take a bit more care about the quality of data, its completeness, consistency etc [2]. One mistake anyone can easily make is selecting the data, it can be noisy data, in complete, data duplicity, missing values, inconsistence data, validity, understandable and many more. If one need to make the data effective and efficient in quality then firstly it needs to go through a quality check criteria and then further processing is done on it [3]. Satisfaction of data quality is important for high quality data. Number of attempt is made to measure the quality of data and to identify its dimensions. Extend to data quality basically includes accuracy and precision of data, is data reliable, importance of data, data consistency, understandability, responsiveness, accurateness, utility [4]. Some factors are there for data quality check given below for our research work

1. Definition
2. Redundancy (duplicate data)
3. Completeness of data
4. Integration done correctly
5. Correctness
6. Precision
7. Accessibility Punctual
8. Punctual

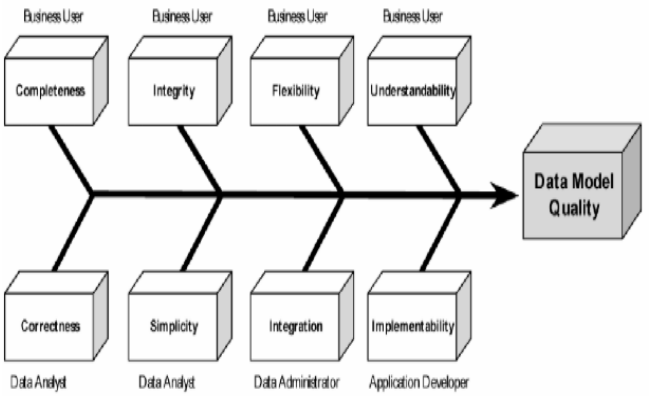

Fig 1: These are described below with brief knowledge about them.

II. PROCEDURE FOR PAPER SUBMISSION

A. Submit Definition

The chosen data should be defined in detail and its meaning should be defined properly and correctly.

B. Redundancy (duplicate data)

There should be no duplicate data in our data warehouse which increases the memory space and it is difficult for us retrieve[8].

C. Completeness of data

All the values present should be complete in all aspects. No missing values are accepted.

D. Integration done correctly

Here all the data is coming from different source so we should integrate it in the manner that there is no compromise in the quality of data[9].

E. Correctness

The data should be in a correct manner .there should be no ambiguities in it i.e. errors or bugs should not be there[9].

F. Precision

The domains values should be correct in precisions as per the requirements in the specifications in original source[10].

G. Accessibility

The data should be easily in accessible; whenever data is retrieved it should be not a difficult task but an easy one.

H. Punctual

Data is timeliness means that data is kept for years but whenever data is retrieved it should be quick.

Revised Manuscript Received on July 05, 2019.

Arif Ali Wani, Computer Science and Engineering Department, Glocal University, Saharanpur, India.

Bansi Lal Raina, Computer Science and Engineering Department, Glocal University, Saharanpur, India.

III. DATA WAREHOUSE SYSTEM

DWS is considered a core component of business intelligence; it is used for reporting and data analysis. It is a central repository of all the data generated by all the companies and unit of large organizations [5]. It has the data integrated from one or more disparate data source. It is subject oriented, time-variant, nonvolatile, integrated and summarized. Staring from 1060's Bill Inmon introduced the term data and later in early 1990's he bring us his book about DW. Ralph Kimball says "DW is a homogeneous and heterogeneous collection of different data sources organized under a unified schema" [3].

ARCHITECHTURE OF DATA WAREHOUSE

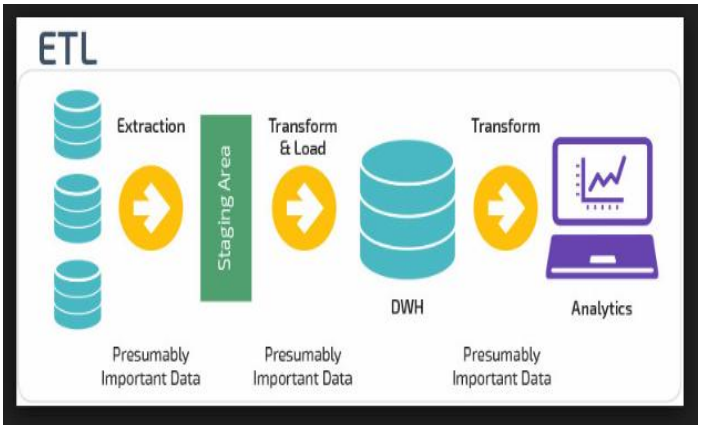

Fig 2: This contains basically 3 layers or areas our data quality will depend upon and checked

A. Data sources

These are the different places from where data comes from or data is taken, for example an operating system, ERO, CRM, flat files, databases etc[8].

B. Staging Area

This is the storage area of DW, it is required during the ETL process because data may be moss during processing .this area is also used due to the time availability is less during extraction from source system, so this increases the time period for loading [6].

C. Target Stage

This is the final stage where we need to keep the data so here it is important for us to keep the data quality high so that our DW is best for the customers and organizations[11].

IV. DATA QUALITY STEPS

Firstly we need to collect the data from different sources [7]. Here collecting and assembling the data together bring many conflicts and issues related to data quality that we have discussed above we need to remove them.

1. Analyzing the data
2. Purifying the data
3. Organizing the data
4. Integration of data
5. Improving the understanding of data
6. Final check of data

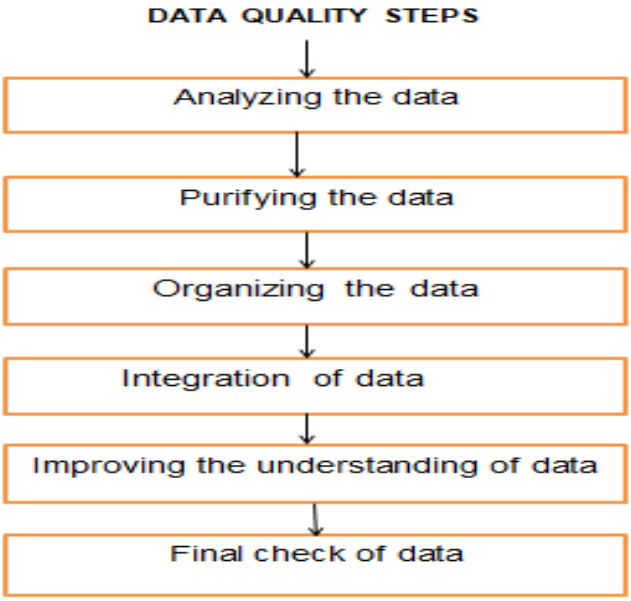

Fig 3: Steps involved for the data quality.

A. Analyzing the data

Firstly the data is analyzed with respect to the standard of the work we are doing. Analyzing the data at very first stage helps us to examine the basic problem and issues in the data and save time at the time of implementing the work [1], [2], [4]. Here we start with the unwanted things by working on them and removing them. There are several different tasks while analyzing the data some of them are column statistics, distribution of values and patterns[12].

B. Purifying the data

As soon as the data analysis is done now we need to purify all its impurities and see that we are able to connect our customers need in all ways. There is several data purification techniques used to removes the issues in data quality [3].

C. Organizing the data

Now when we have remove impurities we need to put the data in a format and need to have a certificate of standard maintained. This may include special data sets row, column, tables, etc [5].

D. Integration of data

Here we will do the integration of data into sets and forms and check if some more issues may come, if so remove them instantly and work on them. Similar and non-similar records are put together respectively and make groups accordingly [6].

E. Improving the understanding of data

As the data is now integrated we make it easier to understand, so that it gets in demand by costumer [7]. This can be done by some additional information, related data set, validation and verification of data set and so on.

F. Final check of data

After all the steps are performed a final check is made to see any ambiguity left, and the data is put in a way to be easily recognized in specific manner [4].

V. ANALYSIS OF DATA QUALITY FACTORS

A. Data Sources as a Data Factor

1. All the heterogeneous and homogeneous data came together and combines can cause issue.
2. As time may increase quality can decrease.
3. Less knowledge can also cause some quality issues.
4. Old data management can lead to issues.
5. Modification is not done is also a major problem.
6. Measurement bugs.
7. Different format can create problem.
8. Failure in updating data done can cause problem.
9. Duplicate data in different sources.
10. Presence of errors.
11. Encoding decoding issues.
12. Ambiguity problem.
13. Mismatch data issues.
14. Wrong data mapping.
15. Entities, attributes, and relationships can be an issue.

B. Staging Area as a data quality factor

1. Architecture of DW.
2. Business rules should be there to see DQ.
3. Database is important for DQ.
4. Time response is important.
5. Proper storing of data in staging area is a factor.
6. ETL tools play an important role in DQ.
7. Data cleansing and quality checking is a factor.
8. Unwanted data is seen as a big issue.
9. Incomplete data and code symbols format.
10. Unknown logics used can cause a problem in ETL.

C. Target Stage as a Data Quality Factor

1. Misunderstanding of requirements of users.
2. Choosing a model between star, snowflake and fact constellation.
3. Poor design of data model.
4. Taking more time is identifying data model.
5. Hierarchical structures can cause issue.
6. Database design can be a factor.

VI. CONCLUSION

Here in this paper we have talked about the importance of data quality, how the data is taken in account on behalf of its standard and level. We learn how data quality is checked using different ways and steps to be taken to do so. Data warehouse architecture is being discussed, how we get through the stages of data ware house and the issues they may have during seeking the data quality. We talk about overall needs to see the factors regarding data quality and its issues. In future we will be working on tools and architecture of data to work on with best quality data, within the data warehouse.

REFRENCES

1. C. Noce and S. Sartore, "The new Enel Distribuzione power quality data warehouse and its applications for smart grids," ICHQP 2010 - 14th Int. Conf. Harmon. Qual. Power, 2010.
2. P. Tiwari, A. C. Mishra, S. Kumar, V. Kumar, and B. Terfa, "Improved performance of data warehouse," Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017, no. Icicct, pp. 94–99, 2017.
3. T. N. Manjunath, R. S. Hegadi, and G. K. Ravikumar, "Analysis of Data Quality Aspects in DataWarehouse Systems," Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 1, pp. 477–485, 2010.
4. N. Zellal and A. Zaouia, "An exploratory investigation of Factors Influencing Data Quality in Data Warehouse," Proc. 2015 IEEE World Conf. Complex Syst. WCCS 2015, 2016.
5. Munawar, N. Salim, and R. Ibrahim, "Towards data quality into the data warehouse development," Proc. - IEEE 9th Int. Conf. Dependable, Auton. Secur. Comput. DASC 2011, pp. 1199–1206, 2011.
6. [6] M. L. Lee and W. Hsu, "Improving data quality Eliminating dupes & I-D-ing those spurious links," IEEE Potentials, vol. 24, no. 2, pp. 35–38, 2005.
7. W. Gongora de Almeida, R. T. de Sousa, F. E. de Deus, G. D. Amvame Nze, and F. L. Lopes de Mendonca, "Taxonomy of Data Quality Problems in Multidimensional Data Warehouse Models," Inf. Syst. Technol. (CISTI), 2013 8th Iber. Conf., pp. 1–7, 2013.
8. Arif Ali Wani, Bansi Lal Raina, "Issues and handy Solutions addressed at every stage in real time data warehousing, i.e. ETL (extraction, transformation& loading) - Literature Review." .
9. Arif Ali Wani, Bansi Lal Raina, "Discovery of knowledge by using Data warehousing as well as ETL processing," Int. J. Recent Technol. Eng., p. 10, 2019.
10. S. Praveen, U. Chandra, and A. A. Wani, "A Literature Review on Evolving Database," vol. 162, no. 9, pp. 35–41, 2017.
11. A. A. Wani, U. Chandra, and P. Jain, "International Journal of Research in Engineering and Innovation Performance analysis of FME based servers and cloud for data loading in big data and machine learning models for future data mining process , knowledge discovery in geo-spatial data," vol. 1, no. 1, pp. 68–71, 2019.
12. A. A. Wani, U. Chandra, P. Bansi, and L. Raina, "Security Challenge in Big Data for Behaviour Analytics," vol. 5, no. 7, pp. 578–581, 2018.

AUTHORS PROFILE

Arif Ali Wani received his Bachelor's degree in Information and Technology from Model Institute of Engineering and Technology (MIET) affiliated to Jammu University, Jammu India. During the 2008 and M.Tech in Computer Science and Engineering from Gurgaon College of Engineering affiliated to Maharshi Dayanand University Rohtak, during the year 2013. Pursuing Ph.D. degree in Glocal University, Saharanpur Uttar Pradesh. He is having 9 years of teaching experience, his area of business is Data Warehouse and Data mining, Computer Network. He has published and presented Research papers in journals, international and national level conferences.

Bansi Lal Raina Backed by an exceptionally brilliant academic record, Prof. Raina has been engaged in administration, teaching & research for nearly 35 years now. He was awarded prestigious national fellowship of "TATA INSTITUTE OF FUNDAMENTAL RESEARCH" (T.I.F.R), Bombay, INDIA wherein he spent four years of research work and then proceeded to USA on an International fellowship to obtain his M.Tech (Computer Science Engineering & PhD. From 'USC', USA. Did he not only write an exemplary research paper at an early age of his career of 10+2 standard published by reputed 'American Mathematical Society' (January 1969 page 48-51), but his paper (part of which is noted below just for reference) was also widely acclaimed and often cited (e.g., See A. Del Cintel, 2008-SPRINGER) which in a dramatic development helped various eminent Scientists like Prof. ANDRE WILE then at PRINCETON UNIVERSITY, to

draw a vital connection between the ELLIPTIC CURVES And MODULAR FORMS (See Ribet: Tanahama-Shimura Conjecture, 1986) leading him eventually to the famous solution in 1995 of even more famous CONJECTURE (See Annals of Mathematics, 142 (1995), which was unsolved for the last 350 years, earned Prof. Wiles a well-deserved 'KNIGHT HOOD' & the most prestigious award of 'FIELDS MEDAL'. Dr. Raina's above cited results have also immensely helped in the development of many subjects and more recently in 'CALABI-YAU' spaces & 'STRING Theory' in ASTRONOMY, thereby unifying the theories of 'NEWTON'S Gravitation, QUANTUM Physics & EINSTEIN'S Relativity.