International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 1

An Examination of Factors Influenced in the Quality Checking of Data in a Data Warehoused

Ghadge Nagnath G.
Asst. Professor, Department of Computer Science, Mahatma Basweshwar Mahavidhyalaya, Latur, Maharastra, India

Panchal Vishwanath D.
Asst. Professor, Head of Department CS & IT, Rajarshi Shahu Mahavidhyalaya, (Autonomous) Latur, M.S.

Shaikh Riyaj
Asst. Professor, Department of CS & IT, Rajarshi Shahu Mahavidhyalaya, (Autonomous) Latur, M.S.

ABSTRACT

Data quality checking in a data warehouse is a key success factor for each Business Intelligence project. In fact, it has a direct impact on taken decisions. If the Data quality checking is good enough for decision makers, the decision support system is very helpful for them. It allows them to have the right inputs to take the right decisions wherever and whenever they need them. But when the data warehouse is of poor Data quality checking, it can have serious impacts on taken decisions that may be even disastrous.

Considering this importance of Data quality checking in data warehouse, we aim in this study to investigate the influence of such contingency factors as top management commitment, Data quality checking management practices, external expertise, Data quality checking at the source, Team Working and technology factor, on the one hand, and Data quality checking in data warehouse, on the other.

We developed a conceptual model where we formulated the relevant hypotheses (Zellal & Zaouia, 2015) and then we established the measurement model (Zellal & Zaouia, 2016). We conducted the survey in Morocco and we used a structural equation modeling technique to analyze the collected data.

The objective of identifying the most critical factors is to enable stakeholders to better use their scare resources while implementing a data warehouse by focusing on these key areas that are most likely to have a greater impact on the Data quality checking in data Warehouse.

Keywords: *Data quality checking Check, Business Intelligence, Data Warehouse, Influencing factors*

1. Introduction

Since data warehouse has been coined by Inmon in 1990, it has known a great expansion in the IT world. It is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process (Inmon, 1992). It allows organizations to consolidate and summarize data from different systems in order to have only one decision support system (DSS).

The data available in the DSS should be accurate enough, timely enough and consistent enough for the organization to survive and make reasonable decisions (Orr, 1998) .In other words, the data in data warehouse should be of good quality to support the decision making process. In this context, we define the Data quality checking in data warehouse as defined in the Data quality checking model standard ISO/IEC 25012: "The quality of a data product may be understood as the degree to which data satisfy the requirements defined by the product-owner organization".

In the case of poor Data quality checking in data warehouse, the managers may take the wrong decisions. Hence, decision support system may have adverse consequences and impact negatively on the performance and the benefits.

That's why poor Data quality checking has been considered, both in literature and by practitioners, as one of the factors that cause the failure of data warehouse (Briggs, 2002). Because of the huge importance of Data quality checking in data warehouse and because of the lack of academic research concerning data warehouse success, there has been a call for rigorous empirical studies to examine data warehousing success factors (Lee, Lee, & and Suh, 2001).

In this context, the purpose of this study is to examine the factors influencing the Data quality checking in data warehouse. We begin in the first section by presenting the hypotheses of our reviewed research model. In the next section, we present the adopted research methodology and the data analysis. Finally we discuss the results.

This study is interesting on two levels:

- On theoretical level, this research will highlight the influence of different contingent factors on Data quality checking in data warehouse.
- On a practical level, it will help the practitioners and the stakeholders to focus on the most significant factors influencing Data quality checking in data warehouse in order to build a decision support system of high Data quality checking.

2. Literature background and Research Model

Despite the recognition of data warehouse as a strategic information source for decision makers, academic research has been lacking concerning data warehousing practices and its critical success factors (Shin, 2003).

Hence, to find the factors influencing Data quality checking in data warehouse, we referred first to different articles concerning data warehousing, Data quality checking management and information systems implementation. Then, the research model has been reviewed and updated after discussions with researchers in the same field (while international conferences WCCS 15 and CIST 16) and after assessment of factors influencing Data quality checking in data warehouse in 3 large firms in Morocco: ONCF (Railways National Office), ANP (National ports Agency) and OCP (National Office of Phosphate).

In this paper, we present the reviewed model that we examined empirically. The research model can be divided to structural model and measurement model:

- The structural model relates the latent variables: Data quality checking in data warehouse and the influencing factors
- The measurement model relates the measured variables to latent variables

2.1 Structural research model

In this section, we present the hypothesis built in our research model as shown in the figure below.

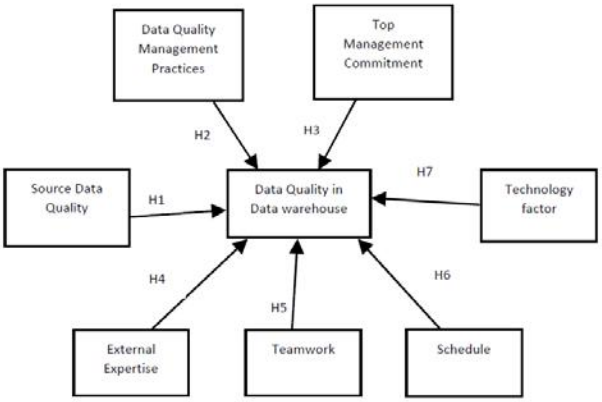

Figure 1: Structural research model

2.1.1 Data quality checking Check in source systems

The source systems are the inputs of the data warehouse system. This last one is just a logically and physically transformation of multiple operational source applications. That's why the Data quality checking in data warehouse is dependent of the Data quality checking of its input data even if this influence can be moderated by the data warehousing process. For example, in the case of multiple data sources, Data quality checking is impacted by the synergy between the different sources. The Data quality checking issues may be either on the schema level when data models and schema designs are heterogeneous, or on instance level such as semantic heterogeneity (Amit & Emilie, 1999) or varying timeliness of data sources.

Furthermore, the quality of data particularly in the source systems was considered in literature review as crucial for any Business Intelligence system implementation because of its impact on the quality data available in the data warehouse (Yeoh & Koronios, 2010). Thus, it is hypothesized that:

H1: Data quality checking in source system (s) influences the Data quality checking in data warehouse.

2.1.2 Data quality checking management practices

As defined by Weber et al. Data quality checking management is the quality-oriented management of data as an asset, that is, the planning, provisioning, organization, usage, and disposal of data that supports both decision-making and operational business processes, as well as the design of the appropriate context, with the aim to improve Data quality checking on a sustained basis (WEBER, OTTO, & OSTERLE, 2009).

Organizations that are adopting Data quality checking management practices are referring to practitioner's guides to analyze their data, to analyze the Data quality checking requirements, to identify the critical areas of data and to evaluate the cost of Data quality checking. They assign data responsibilities, assess Data quality checking, improve it and monitor it.
So in the light of this, we hypothesize that:
H2: The adoption of Data quality checking management practices improves the Data quality checking in data warehouse.

2.1.3 Top Management Commitment

Today, most companies delegate authority for managing Data quality checking to the IT department and Data warehousing Team. Although IT must be involved in the process, it doesn't have the clout to change business processes or behavior that can substantially improve Data quality checking (Orr, 1998) . It is up to top management to set up Data quality checking goals according to decision makers' needs and task decisions. It is top management duty to set up policies for Data quality checking and to allocate resources to achieve the Data quality checking goals, we propose:

H3: The higher the top management supports data warehouse implementation, the greater is Data quality checking in data warehouse.

2.1.4 External Expertise Quality

We mean by External expertise the external mediator's entities such as the BI vendors and IT consultants, who take in charge the development of the target solution, provide the training, maintenance and technical support for companies implementing data warehouse.

Surely, developing a data warehouse requires skills and deep knowledge. It requires both technical and business expertise. That's why, the external experts must give the best of their knowledge, experiences and competencies in order to build a data warehouse of high Data quality checking.

Understanding very well the business requirements, taking into account the development environment and referring to their knowledge, external experts should advise the company implementing the data warehouse to use the adequate data profiling, Data quality checking and ETL tools. They can implement validation routines, Data quality checking checks and metadata repository.

That's why the company implementing data warehouse should give the biggest importance to expertise quality while selecting the vendors and consultants, and not to refer only to the price and time of data warehouse implementation.

In addition to this, Alhyasat considers vendors and consultants as a support quality factor in his Data Warehouse Success Framework (Alhyasat, 2013) .And Thong et al. found that vendor support and consultant effectiveness are closely related to the overall information system effectiveness (Thong, Yap, & Raman, 1996) .Wang et al. also consider the system provider as an important factor in the establishment and maintenance of a quality system (Wang, Shih, Jiang, & Klein, 1996). On the light of these works, we propose:

H4: The higher the quality of external expertise for a data warehouse implementation, the greater the Data quality checking in data warehouse.

2.1.5 Team working

While implementing a Data warehouse, in addition to top management commitment, three stakeholders must work together: the external consultants, the Information System (IS) staff and the end users (decision makers).

The internal IS team plays an important role in coordinating between the different stakeholders. She is also responsible of selecting data in alignment with business requirements and giving the necessary information to external consultants. So, every role in the project is important. But the communication between all the members is the most important. It is Team Working that allows external experts to implement a system of high Data quality checking, fit for use for the decision makers, with the help of IS staff.

Hanging Xu proved that the technical factor 'Team Working' influences Information quality in the Data warehousing success model (Xu, 2008). And reviewing quality management literature, 'Team Working' is identified as one of the key success factors of Quality management (Cheng & Choy, 2007) .So the Team Working is important to produce an information product of high quality. We propose then:

H5: The greater is the quality of team working on data warehouse implementation, the greater is the Data quality checking in data warehouse.

2.1.6 Scheduling

We mean by 'schedule' the planning and the time allowed to data warehouse implementation. If it is a tight schedule in time, it pushes data warehousing team to finish quickly, and so not to give sufficient attention to Data quality checking and not to allow sufficient time to data staging.

On another side, the implementation planning should be respected in order not to heng the projects hang indefinitely in time.
In this context, Baker (Baker & Baker, 1999) and Sigal (Sigal, 1998) consider that proper planning and execution of the implementation schedule is critical to data warehouse implementation success. So, we propose:

H6: The proper schedule for data warehouse implementation, the greater Data quality checking in data warehouse.

2.1.7 Technology Factors

By the technology factor we mean ETL tools, Data quality checking tools, type of load strategy and infrastructure performance. In data warehousing project, the ETL tools are dedicated to extract, transform and load data from data source to data warehouse. In the transformation stage, cleaning and data improvement can be done depending on transformation features and Data quality checking features offered by the ETL tool used.

Data quality checking tools are also very important to get a high Data quality checking in data warehouse. When a Data quality checking tool is used in a data warehousing project, integrated with ETL tool, and depending on the different functionalities it offers, it allows Data quality checking improvement.

Loading strategy has also an influence on Data quality checking. It refers to loading type (Bulk, batch load or simple load) and loading frequency. It impacts especially on the freshness or the timeliness dimension of Data quality checking.
The performance of the platform behind the data warehousing process impacts the quality of data in data warehouse. It is the ability of platform used to execute the compiled code in an optimized and speed way.

H7: The technology factor supporting the data warehouse has an impact on data warehouse Data quality checking.

2.2. Measurement models

In this section, we present the scale items to measure each latent variable in the presented research model. The items used were taken from previously validated sources and adapted to the context. (Zellal & Zaouia, A measurement model for factors influencing Data quality checking in data warehouse, 2016).

2.2.1 Data quality checking measurements

As presented in literature, Data quality checking is a multidimensional concept (Eckerson, 2006). Which means that evaluating the quality of a dataset amounts to evaluating its completeness, correctness, accuracy, consistency and so on. In fact, there are so many dimensions of Data quality checking, and there is no general agreement on them (Fischer & Kingma, 2001). So we choose four of them which are the most frequently mentioned in literature and which constitute the focus of the majority of the authors (Emily, 1997). They are also defined as the basic set of Data quality checking dimensions by Batini et al. (Shin, 2003) after analyzing the most important classifications of Data quality checking dimensions. These Data quality checking dimensions are: Timeliness, Accuracy, Consistency and Completeness.

So we'll consider these dimensions as items of measurement of Data quality checking, but we'll give them different definitions depending on data if it is at the source or at the data warehouse.
Source Data quality checking Measurement items:

- **Timeliness** indicates if data is updated (First) according to changes known by time in the real world.
- **Accuracy** is the measure that indicates how well and how correctly is data represented in the data

base, comparing its value to the real world or to a reference data.
- **Completeness** can be defined as the measure that indicates if all the useful fields are filled.
- **Consistency** is the measure that indicates that data don't violate integrity constraints and don't conflict each other and can be considered logic referring to the business rules.

Data Warehouse Data quality checking Measurement items:
- **Timeliness** indicates if data is sufficiently updated for the decision maker and for the decision task.
- **Accuracy** is the measure that indicates the correctness and precision required to make a specific decision concerned by this information
- **Completeness** can be defined as the measure that indicates complete if the users (decision makers) can deduce any necessary information they need for their decision tasks
- **Consistency** is the measure that indicates that data is not conflicting each other and not conflicting business rules and users requirements in what concerns format and content.

2.2.2 Data quality checking management practices

The items used to measure the Data quality checking management adoption are:

- Definition of Data quality checking expectations for the Decision Support System
- Definition and use of Data quality checking dimensions accordingly
- Institution of data governance
- Agreeing to Data quality checking standards
- Monitoring Data quality checking performance

2.2.3 Top Management Commitments

The items we propose to measure top management commitment in the context of Data quality checking in data warehousing project are as follows:

- Participation and support of Top management team in the data warehousing project
- Allocation of the necessary human resources to the DW project
- Allocation of the necessary financial resources to the DW project
- Attitude to change Allocation of the necessary human resources
- End user satisfaction with the changes top management decides on Data quality checking issues
- Quality Priority: quality is treated as more important than cost and time by top management in DW project

2.2.4 External Expertise Quality

We propose the following items to measure External expertise:

1. Vending Company / Consultant adequate technical support,
2. Vending Company / Consultant credibleness and trustworthiness,
3. Vending Company / Consultant relationship and communication with organization
4. Vending Company / Consultant experience in Data Warehousing projects
5. Vending Company / Consultant quality training and services in Data Warehousing

2.2.5 Team Working

We propose these items to assess Team Working factor in our research context:
- Clear vision and elevating goals for all team members
- Competency of the team members
- Collaborative climate (sharing ideas and expertise) between the team members
- Support and recognition between the team members
- Team leadership
- Unified commitment of the team on one engagement.

2.2.6 Scheduling

We propose the following items to measure the schedule factor:
- Practical Implementation Schedule
- Stable scoping of project
- Change of the planning accordingly to any change in the project scope

2.2.7 Technology Factors

To measure the technology factor we consider:
- The transformation features of ETL tool
- The Data quality checking features of ETL and Data quality checking tools

- Loading Strategy
- Platform performance

3. Research Methodology

3.1 Data Collection

In order to test the hypotheses of our model, we conducted a survey in Morocco. The questionnaire targeted especially the BI specialists and end users of data warehouse in Moroccan medium and large accounts.

The questionnaire was sent via LinkedIn and by mails only to our professional network in order to avoid any fake response.
To ensure data validity and reliability, five knowledgeable individuals (i.e., 1 BI professor, 3 BI consultants and 1 BI managerial level user) completed the questionnaire before our mailing it, and their comments helped improve its quality.

3.2 Instrument Development

The measures used were taken from the established measurement model and anchored on a 5-point Like RT scale, ranging from *strongly disagree* (1) to *strongly agree* (5), on which participants were asked to indicate an appropriate choice.

Data quality checking in data warehouse (DQDW) was assessed with the following statements:
"In your Data Warehouse, the data is updated frequently enough to allow you to make the decisions you need at the right time", "In your Data Warehouse, the data is fairly accurate and accurate for you to make your decisions", "In your Data Warehouse, you will find all the information you need to make your decisions" and "The data in your Data Warehouse is consistent, it does not represent a conflict between them, or a conflict with business rules".

Data quality checking in source systems (DQSS) was assessed by the following:
"In your Operational Information System (OIS), the data are updated according to their variation in reality", "In your OIS, the data is precise and represent exactly their respective elements in reality" You need it for your daily operational work you find it in your OIS"," The data in your OIS are consistent and do not represent a contradiction between them "

The adoption of Data quality checking management practices (DQMP) was assessed with these statements:
"Expectations of end users in terms of Data quality checking in the decision-making system are well identified and documented", "Data quality checking measures are well defined and documented to measure the achievement of end-user expectations in terms of Data quality checking "," Your company establishes good data governance (Definition of processes, roles and responsibilities for Data quality checking) "," Standards and good practices of Data quality checking are your reference in any step relating to data: data collection, transformation, updating .
", "You set up a continuous improvement system to measure the achievement of the objectives in terms of Data quality checking".

The Top Management commitment (TM) was assessed using the following statements:
"Top Management supports the BI project by actively participating in its management", "The TM is ready to allocate all the human resources needed for the project BI", "The TM is ready to allocate all the financial resources it needs to successfully implement a DW with a good quality of data", "TM is willing to change existing work and procedures to improve Data quality checking", "End users are satisfied with the changes that Top Management decides on issues of Data quality checking in the data warehousing project", "Quality is considered by the TM as more important than the cost and time in the data warehousing project"

The External Expertise (EE) was assessed by these statements:
"Your service provider has the right technical support", "Your service provider is credible and trustworthy", "Your service provider has a good relationship with your organization (Project team and decision makers)", "Your service provider has experience in the field Decision-making", "Your provider offers high quality training and services".

The Team Working (Team) was assessed using the following statements:
"The whole team has a clear vision of the decision-making project (its objectives, deadlines, sources, users ...)", "The members of the project team are competent", "There is a collaborative climate between members Project team (sharing of ideas, experience and knowledge)", "There is help and recognition among team members project"," there is good leadership (direction) Team", "The project team has a unified commitment on which all members agree".

The Schedule (Sch) factor was assessed by these 3 statements:
"The implementation planning of the decision-making system in your organization is practical and reasonable", "The decision-making perimeter is stable throughout the project" and "the implementation planning is reviewed and modified every time perimeter of the project is modified".

The technology factor (TF) was assessed by the following statements:
"You are satisfied with the data transformations offered by the ETL used to build your Data Warehouse", "You are satisfied with the quality improvements of your tools (ETL or QD tool), For example: real-time cleaning, verification of data according to business rules", "You are satisfied with the data loading strategy (loading flow schedules), ie it does not impact Data quality checking in your Data Warehouse" and "You are satisfied with the performance of the infrastructure that supports your Data Warehouse (High Availability, Speed of Code Processing)".

3.3 The sample

The overall response rate was 33%. In total, we received 80 individual responses. The responses were received from diverse industries: Banks and Assurances (25%), Telecommunications (12,5%), Industry (11,3%), Transport (8,8%), Finance (6,3%), Consulting (6,3%) and Health sector (3,7%).

The respondents' positions in the organizations vary from junior employee to top manager. Most of them are senior employees (40%). The majority of participants work in Information Systems direction (75%) and the other minority is spread over different directions such as Business operations, marketing and control management. More than 90% of the participants have an IT background, thing that allowed them to understand and rate the questionnaire statements easily. The sample includes small (19%), medium-sized (36%) and large firms (45%). 47% of them have implemented their data warehouse more than 5 years ago.

3.4 Data Analysis

A structural equation modeling (SEM) technique was used to examine the relationships among the constructs. SEM is a powerful technique, widely used in the behavioral sciences that can combine complex path models with latent variables (Hox & Bechger).

There are two main approaches: PLS (Partial Least Squares) and covariance-based SEM. The PLS approach was chosen for its capability to accommodate small-sized samples (Chin, 1998).

Additionally, PLS recognizes two components of a casual model: the measurement model and the structural model. The measurement model consists of relationships between the latent variables and the measures underlying each construct. PLS method allows the demonstration of the construct validity of the research instrument (i.e. how well the instrument measures what it purports to measure). The two main dimensions are the convergent validity and the discriminant validity. The convergent validity (also known as the composite reliability) assesses the extent to which items on a scale are theoretically related. It reflects if the measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other. And the discriminant validity shows if the measures of constructs that theoretically should not be related to each other are, in fact, observed to not be related to each other.

On the other hand, the structural model provides information on how well the hypothesized relationships predict the theoretical model. PLS software e.g. Smart PLS 3.0 (the software we used for our PLS analysis), provides the squared multiple correlations (R2) for each endogenous construct in the model and the path coefficients. The coefficient of determination R2 indicates the proportion of variance (%) in the dependent variable that can be explained by the independent variable in the model while the path coefficients (β) indicate the strengths of relationships between constructs (Chin, 1998). Chin (1998) notes that both the β and the R2 are sufficient for analysis, and β values between 0.20 and 0.30 are adequate for meaningful interpretations.

Fornell and Larcker (1981) note that item loadings and composite reliabilities greater than 0.7 are considered adequate (Fornell & Larcker, 1981).

Assessment of the structural models

The paths coefficients (β) and the R2 were generated by Smart PLS 3.0. Values are shown in the following figure. The R2 is 0.73, which suggests that the contingency factors explained 73% of the variance in the *Data Warehouse Data quality checking* construct. This value is considered strong effect size (Moore, Notz, & Flinger, 2013).

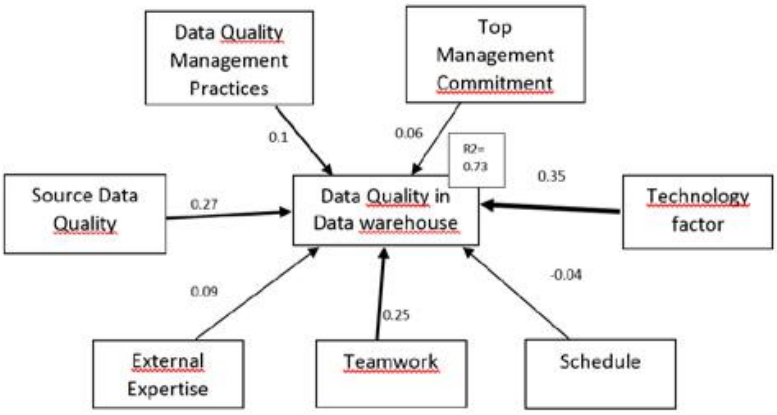

Figure 2: The Smart PLS Graph results for the research model

4. Discussion and Conclusion

The objective of this study was to find the most critical factors for Data quality checking in data warehouse while data warehouse implementation.

This empirical study reveals that the contingent factors that we included in our research model were able to explain 73% of the variance of Data quality checking in data warehouse.

The most critical factor, according to our survey conducted in Morocco, is the technology factor (β= 0.35). That means that a special attention should be given to the choice of ETL and Data quality checking tools while the data warehouse implementation. The platform performance and loading strategy are also very important.

The second most critical factor is Data quality checking in source systems (β= 0.27). This means that if an organization needs a data warehouse of good Data quality checking, it should start by improving the Data quality checking in the source systems.

Another critical factor that has been supported by the survey is the Team Working (β= 0.25). This can be considered as a key success factor for Data quality checking in data warehouse. It requires a good leadership, competent team members and a sharing and recognition spirit.

The factor "Adoption of Data quality checking management practices" has been only moderately supported by the collected data (β= 0.102), while the other factors "Schedule", "Top Management commitment" and "External Expertise" was not really supported by collected data.

REFERENCES

1. Fischer, C., & Kingma, B. (2001). Criticality of Data quality checking as exemplified in two disasters. Information & Management, 109-116.
2. Alhyasat, E. B. (2013). Data Warehouse Success and Strategic Oriented Business Intelligence: A Theoretical Framework. Journal of Management Research, 5(3).
3. Fornell, C., & Larcker, D. (1981). Evaluating structural equations models with unobservable variables and measurement error. Journal of Marketing Research, 8(1), 39-50.
4. Amit, R., & Emilie, Y. (1999). Key Issues in Achieving Data quality checking and Consistency in Data Warehousing among Large Organizations in Australia. International Conference on System Sciences. Hawaii.
5. Baker, S., & Baker, K. (1999). The best little warehouse in business. Journal of Business Strategy, 32-37.
6. Eckerson, W. (2006). Data quality checking and the bottom line: Achieving business success through the commitment to high quality data. 101 Communications LLC.
7. Briggs. (2002). A Critical Review of Literature on Data Warehouse Systems Success/Failure. Journal of Data Warehousing, 41, 1-20.
8. Cheng, E., & Choy, P. (2007). Measuring Success Factors of Quality Management in the Shipping Industry. Palgrave Journals, 234-253.
9. Chin, W. (1998). Issues and opinion on structural equation modeling. MIS Quarterly, 22(1).
10. Dale, & Duncalf. (1985). Quality-related decision making: A study in six British companies. International Journal of Operations and Production Management, 5(1), 15-25.
11. Emily, K. (1997, October). Dirty data challenges warehouses. Software Magazine.
12. Haley, Watson, H., & Barbara. (1997). Data warehousing: A framework and survey of current practices. Journal of Data Warehousing, 10-17.
13. Hox, J., & Bechger, T. (s.d.). An Introduction to Structural Equation Modeling. Family Science review, 354-373.
14. Ifinedo, P. (2008). Impacts of business vision, top management support, and external expertise on erp

success. Business Process Management Journal, 14(4), 551-568.

15. Inmon, W. H. (1992). Building the Data Warehouse. John Wiley & Sons. Lee, Y.-S., Lee, D.-M., & and Suh, C.-K. (2001). Factors Affecting Implementation of Data Warehouse. PACIS, (p. 47).
16. Moore, D. S., Notz, W. I., & Flinger, M. A. (2013). The basic practice of statistics. New York: Freeman and Company.
17. Nunnally, J. (1978). Psychometric Theory. New York: McGraw-Hill.
18. Orr, K. (1998). Data quality checking and systems theory. 41, pp. 66-71.
19. Shankaranarayanan, G., & Cai, Y. (2006). Supporting Data quality checking management in decision-making. Decision Support Systems, 302-317.
20. Shin, B. (2003). An Exploratory Investigation of System Success Factors in Data Warehousing. Journal of the Association for Information Systems, 4(1).
21. Sigal, M. (1998). A common sense of development strategy. Communications of the ACM , 42-43.

Ghadge Nagnath G. is a Asst. Professor in Department of Computer Science, at Mahatma Basweshwar Mahavidhyalaya, Latur. After One years of experience in Asst. Professor to the field of Computer Science.